



UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

**DESARROLLO DE UN SOFTWARE
PARA EL ANÁLISIS ESTADÍSTICO
BAYESIANO**

TESIS:

PARA OBTENER EL TÍTULO DE
LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA:

HERNÁNDEZ SOLANO ALAN MARTÍN

DIRECTORA DE TESIS:

M.C. NORMA EDITH ALAMILLA LÓPEZ

HUAJUAPAN DE LEÓN, OAXACA.
JULIO DE 2008

Índice general

Agradecimientos	v
1. Introducción	1
1.1. Antecedentes	1
1.1.1. La estadística bayesiana	3
1.2. Objetivos	4
1.3. Justificación	5
2. Conceptos de estadística bayesiana	7
2.1. Probabilidad subjetiva	7
2.2. Función de distribución a priori	8
2.2.1. Distribución a priori no informativa	8
2.2.2. Distribución a priori conjugada	10
2.3. Distribución a posteriori	12
3. Inferencia bayesiana	15
3.1. Estimación puntual	15
3.1.1. Error de estimación	16
3.1.2. Estimación multivariada	17
3.2. Conjuntos creíbles	17
3.2.1. Diferencias entre los conjuntos creíbles y los intervalos de confianza .	19
3.3. Contraste de hipótesis	19
3.3.1. Contraste de hipótesis unilateral	22
3.3.2. Contraste de hipótesis nula puntual	23
4. Simulación e integración Monte Carlo	27
4.1. Integración Monte Carlo	27
4.1.1. Métodos directos	27
4.2. Métodos de simulación Monte Carlo	30
4.2.1. Antecedentes	30
4.2.2. Motivación	35
4.2.3. Algoritmo de Metropolis-Hastings	35
5. Software	41
5.1. Diseño e Implementación	41
5.1.1. Funciones de distribución	44

5.1.2.	Funciones de distribución a priori	44
5.1.3.	Función de distribución a posteriori	44
5.1.4.	Estimación puntual	53
5.1.5.	Intervalos creíbles de máxima densidad a posteriori (HPD)	57
5.1.6.	Contraste de hipótesis	58
5.1.7.	Estadística descriptiva	62
5.2.	Requerimientos	65
6.	Resultados y trabajo a futuro	67
6.1.	Resultados	67
6.2.	Trabajo a futuro	69
A.	Apéndice de figuras	71
B.	Apéndice	77
B.1.	Algoritmo de Romberg	77
B.2.	Funciones de verosimilitud	77
C.	Notación	79

Agradecimientos

Quiero aprovechar la ocasión y este espacio para brindar un sincero agradecimiento y gratitud, a las siguientes personas e instituciones quienes contribuyeron de manera directa o indirecta en la realización de esta tesis.

A mis padres Cecilia Solano Olivera y Everardo Hernández García, quienes siempre me han apoyado y apoyarán en cada uno de los proyectos de mi vida.

A mi directora de tesis, la maestra Norma Edith Alamilla López, por su tolerancia, consejos, amistad y su acertada dirección en el desarrollo de esta tesis.

A mi coasesora de tesis, la maestra María Esperanza Pérez Córdoba Sánchez, por su disposición para solucionar todos los aspectos de carácter computacional del software desarrollado.

A la profesora Verónica Borja Macías, por sus grandes atenciones y su amistad brindada a lo largo de estos últimos 2 años.

A mis sinodales el profesor Vulfrano Tochihuitl Bueno y el profesor José del Carmen Jiménez Hernández, quienes me apoyaron en la revisión y corrección de este trabajo.

A la Universidad Tecnológica de la Mixteca, gracias a la cual pude realizar mis estudios.

A las instituciones Telmex y Pronabes, quienes me apoyaron con becas económicas para solventar los gastos generados por mis estudios.

Y en general a todos aquellos que de alguna manera intervinieron para hacer posible y llevar a feliz término esta tesis, gracias.

Capítulo 1

Introducción

1.1. Antecedentes

Con la finalidad de motivar el uso de la estadística bayesiana, a continuación se presentarán algunas inconsistencias encontradas en la estadística clásica y después se dará una breve descripción de lo que es la estadística bayesiana.

La estadística es una disciplina que en los últimos años ha tenido una gran injerencia tanto en los diversos campos de la ciencia como en la vida real. En particular, la estadística clásica ha sido una de las herramientas más utilizadas en el estudio de los fenómenos aleatorios, sin embargo, a través del tiempo se han encontrado algunas inconsistencias en esta teoría.

Con la intención de ilustrar esto y al mismo tiempo de motivar el interés hacia una alternativa estadística, se presentan algunos ejemplos de problemas en donde el uso indiscriminado de la estadística clásica desencadena en soluciones contradictorias o poco satisfactorias.

Ejemplo 1.1 Sea X una variable aleatoria con distribución $P_n(\lambda)$, se desea encontrar un estimador insesgado de $P(X = 0)$.

Recuérdese que

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!} \quad \lambda > 0.$$

Así, el problema se reduce a estimar $\exp(-\lambda)$, puesto que

$$P(X = 0) = \frac{\lambda^0 \exp(-\lambda)}{0!} = \exp(-\lambda),$$

de aquí que se requiera buscar $h(X)$ tal que $E[h(X)] = \exp(-\lambda)$. Puesto que el estimador es insesgado se tiene que

$$E[h(X)] = \sum_{x=0}^{\infty} h(x) \frac{\lambda^x \exp(-\lambda)}{x!} = \exp(-\lambda),$$

si y sólo si

$$\sum_{x=0}^{\infty} h(x) \frac{\lambda^x}{x!} = 1.$$

De la última igualdad se sigue que

$$\sum_{x=0}^{\infty} h(x) \frac{\lambda^x}{x!},$$

es el Polinomio de Taylor de la constante 1, y por la unicidad de los coeficientes se concluye que

$$h(x) = \begin{cases} 1 & \text{si } x=0 \\ 0 & \text{d.o.f.} \end{cases}$$

lo cual muestra la unicidad del estimador insesgado. Es evidente que estimar $P(X = 0)$ con 1 ó 0 no resulta congruente, ya que $\exp(-\lambda)$ no puede tomar estos valores. Esto debido a que

$$\exp(-\lambda) = 1 \quad \text{si y sólo si} \quad \lambda = 0.$$

pero por la definición de la función de distribución de Poisson se tiene que $\lambda > 0$. Así, $0 < \exp(-\lambda) < 1$, para cualquier $\lambda > 0$.

Ejemplo 1.2 Sean X y Y variables aleatorias independientes con función de probabilidad Poisson(λ). Sea Z otra variable aleatoria con la misma distribución de X y Y . Se desea estimar insesgradamente $P(X = 0, Y = 0) = \exp(-2\lambda)$ con base a Z .

Sea $h(Z)$ un estimador, que cumpla

$$E[h(Z)] = \exp(-2\lambda),$$

donde $h(z) > 0$ para toda z . Por otro lado se tiene que

$$E[h(Z)] = \sum_{z=0}^{\infty} h(z) \frac{\exp(-\lambda)\lambda^z}{z!},$$

por lo que

$$E[h(Z)] = \exp(-2\lambda) \quad \text{si y sólo si} \quad \sum_{z=0}^{\infty} h(z) \frac{\lambda^z}{z!} = \exp(-\lambda).$$

Resulta evidente que la suma corresponde al polinomio de Taylor de $\exp(-\lambda)$ y por la unicidad en los coeficientes se tiene que:

$$h(z) = \begin{cases} 1 & \text{si } z \text{ es par} \\ -1 & \text{si } z \text{ es impar} \end{cases}$$

Este resultado que arroja la estadística clásica presenta gran incongruencia con la definición de probabilidad, ya que $h(z)$ estima una probabilidad y por lo tanto no puede tomar valores negativos.

Ante tales inconsistencias muchos investigadores han buscado algunas alternativas para la solución de problemas de carácter estadístico.

La estadística Bayesiana se presenta como un enfoque alternativo a la estadística clásica en la solución de problemas típicos estadísticos como son:

- La estimación
 - Puntual
 - Por intervalos
- El contraste de hipótesis
- Predicción.

Esta teoría ha generado un enorme interés en los últimos años y ha tenido una gran aceptación en muchas áreas de la investigación científica.

1.1.1. La estadística bayesiana

La estadística se ocupa principalmente del análisis de los datos que presentan variabilidad, con el fin de comprender el mecanismo que los genera, pero en este análisis existe un componente de incertidumbre, un vector de parámetros θ , por lo cual el estadístico se ocupa en reducir a lo mínimo tal componente, así como también en describirla de forma apropiada.

Pero si se supone que toda forma de incertidumbre debe describirse por medio de modelos de probabilidad, en el caso de la estadística bayesiana se considera al parámetro θ , sobre el cual se desea inferir, como un evento incierto.

Entonces como nuestro conocimiento no es preciso y está sujeto a incertidumbre, se puede describir mediante una distribución de probabilidad (llamada a priori), lo que hace que el parámetro tenga el carácter de aleatorio.

La teoría bayesiana plantea la solución a un problema estadístico desde el punto de vista subjetivo de la probabilidad, según el cual, la probabilidad que un estadístico asigna a uno de los posibles resultados de un proceso, representa su propio juicio sobre la verosimilitud de que se tenga el resultado. Este juicio estará basado en opiniones e información acerca del proceso. Esta última también es una desventaja, pues algunos investigadores rechazan que la información inicial se incluya en un proceso de inferencia científica. Pero esta situación se puede evitar estableciendo una distribución a priori no informativa o de referencia, la cual se introduce cuando no se posee mucha información previa acerca del problema.

A un problema específico se le puede asignar cualquier tipo de distribución a priori, ya que finalmente al actualizar la información a priori que se tenga acerca del parámetro, mediante el teorema de Bayes y obtener la distribución a posteriori del parámetro, es con esta con las que se hacen las inferencias del mismo. Cuando un investigador tiene conocimiento previo a un problema, éste conocimiento previo puede cuantificarse en un modelo de probabilidad. Si los juicios de una persona sobre la verosimilitud relativa a ciertas combinaciones de resultados satisfacen ciertas condiciones de consistencia, se puede decir que sus probabilidades subjetivas se determinan de manera única. La estadística

bayesiana tiene una fundamentación axiomática en los llamados axiomas de coherencia¹, con el fin de tener un comportamiento coherente en el momento de tomar decisiones en un ambiente de incertidumbre. Estos axiomas son útiles en la construcción formal de las definiciones de probabilidad.

1.2. Objetivos

En esta tesis se analizan algunos aspectos importantes de la teoría bayesiana, esto con la finalidad de crear un software que permita realizar algunos de los métodos principales de inferencia bayesiana. Cabe mencionar que también se le agregará al programa algunas herramientas de la estadística descriptiva para tener un software mas completo a la hora de realizar un análisis estadístico. Las principales funciones que tendrá este software son las siguientes:

- Inferencia bayesiana
 - Estimación
 - Puntual
 - Por intervalos
 - Contraste de hipótesis
- Estadística descriptiva
 - Diagramas de caja
 - Histogramas
 - Cálculo de cuantiles
 - Cálculo de la media y la varianza muestral
- Análisis de sensibilidad

El nombre de este software será **ANESBA** (**Análisis estadístico bayesiano**) un programa destinado principalmente a la inferencia bayesiana.

La estructura del desarrollo de la tesis es la siguiente:

En el capítulo 2 se presentan todos los conceptos necesarios para el desarrollo y el entendimiento de la teoría bayesiana.

En el capítulo 3 se desarrolla la teoría correspondiente a la inferencia bayesiana, la cual es la parte principal del software que se pretende realizar.

En el capítulo 4, dada la complejidad de los cálculos necesarios para integrar la inferencia bayesiana al software desarrollado, se muestra el método de integración Monte Carlo, el cual es una poderosa herramienta para el cálculo de las funciones de distribución acumulada, necesarias en los contrastes de hipótesis y la verificación de los conjuntos creíbles. También se presenta el algoritmo de Metrópolis-Hastings la cual es una herramienta necesaria para la utilización de la integración Monte Carlo.

En el capítulo 5 se muestra la arquitectura implementada al sistema, a partir de la descripción de funciones, algoritmos y fórmulas utilizadas para la creación del software.

¹Para el lector que desee ahondar en el tema ver[1].

En el capítulo 6 se muestran como resultados, mediante un ejemplo, el fácil manejo del software y el buen funcionamiento del mismo. También se proponen algunas mejoras a ANESBA como un trabajo a futuro.

1.3. Justificación

Es basta la literatura acerca de la inferencia bayesiana, por el contrario sólo se ha encontrado un software, de nombre First Bayes², que en principio realiza este tipo de inferencia, pero desafortunadamente no se ha logrado su instalación con éxito, por lo tanto ante la poca existencia de este tipo de software se ha alentado el desarrollo de un programa computacional que integre la teoría que se desarrollará en la tesis, el cual tendrá como objetivo fundamental realizar inferencia bayesiana y estadística descriptiva.

Algunas diferencias³ que se pretende existan entre estos software son las siguientes:

Tabla 1: Comparativa de características entre First Bayes, R y ANESBA.

Características	First Bayes	R	ANESBA
Visual	Si	No	Si
Idioma	Inglés	Inglés	Español
Funciones de densidad a priori usadas	Conjugadas	No	Conjugadas y no informativas
Contraste de hipótesis unilateral	No	No	Si
Contraste de hipótesis unilateral dual	No	No	Si
Contraste de hipótesis puntual	No	No	Si
Análisis descriptivo	No	Si	Si

Con respecto a la estadística descriptiva el programa R⁴ contiene de manera similar los elementos que se están implementando en ANESBA, tal como se muestra en la tabla anterior, aunque hay que recordar que el objetivo principal de ANESBA es cubrir algunos métodos de la inferencia Bayesiana.

Por lo anterior se puede resumir que el software que se pretende desarrollar, aparte de complementar los elementos de la estadística bayesiana con la estadística descriptiva se encontrará en el idioma español.

²Este software se encuentra en la siguiente dirección electrónica <http://www.tonyohagan.co.uk/1b/>

³La información que se usó para comparar a First Bayes, fue obtenida de su manual, el cuál se encuentra en la dirección electrónica mencionada anteriormente.

⁴R presenta grandes facilidades para la implementación, por parte del usuario, de todas las funciones que realiza Abesba, pero se cae en la necesidad de programarlas. Dicho programa se puede descargar de la siguiente dirección electrónica [/http://www.r-project.org/](http://www.r-project.org/)

Capítulo 2

Conceptos de estadística bayesiana

2.1. Probabilidad subjetiva

El primer punto que se discutirá es el significado de la probabilidad concerniente a lo eventos (subconjuntos) en Θ , donde Θ es el conjunto de todos los posibles eventos (valores). El concepto clásico de probabilidad envuelve una larga secuencia de repeticiones dada una situación. Por ejemplo, la probabilidad de que al lanzar una moneda, esta caiga cara es de $\frac{1}{2}$, esto significa que en una larga secuencia de repeticiones independientes, cerca de la mitad de las veces la moneda habrá de caer cara. Desafortunadamente, este concepto frecuentista no es suficiente cuando se trata con probabilidades acerca de θ . Por ejemplo, considere el problema de tratar de determinar θ , la proporción de fumadores en México.

Dada la siguiente sentencia:

$$P(.3 < \theta < .35) = .35$$

¿Qué significado tiene? Aquí θ es según la estadística clásica, simplemente algún número que se desconoce. Claramente el parámetro está o no, en el intervalo $(.3, .35)$ pero esto no tiene nada de aleatorio.

Un segundo ejemplo, sea θ la tasa de desempleados en la ciudad de Oaxaca para el siguiente año. Es fácil imaginar a θ como algo aleatorio, desde que el futuro es incierto. Pero ¿cómo se puede interpretar

$$P(.03 < \theta < .04)$$

en términos de una secuencia de situaciones idénticas? La tasa de desempleo del siguiente año será única.

La teoría de la probabilidad subjetiva permite hablar acerca de probabilidades cuando la estadística frecuentista no lo pueda hacer.

La idea central de la probabilidad subjetiva es que la probabilidad de que un evento suceda, está determinada por la creencia del investigador. Por ejemplo, se puede tener el presentimiento de que θ (en el caso del desempleo) está entre el 3% y 4%, aunque no se le pueda asignar una probabilidad frecuentista a este evento. No hay nada de malo en este hecho, pues es común pensar en términos de probabilidad personal todo el tiempo, por ejemplo, cuando se apuesta en el resultado de un partido de fútbol o cuando se evalúa la

posibilidad de lluvia del día siguiente. La forma mas fácil de determinar la probabilidad subjetiva es comparando eventos y determinando las probabilidades relativas. Supóngase por ejemplo que se desea encontrar la $P(E)$. Simplemente se compara E con respecto E^c . Si sentimos que E tiene dos veces mas probabilidad de ocurrir que E^c , entonces claramente se asignaría $P(E) = \frac{2}{3}$ y $P(E^c) = \frac{1}{3}$.

Pero no siempre es fácil asignar probabilidades subjetivas, algunas dificultades que se presentan son por ejemplo: la asignación o manejo de las probabilidades pequeñas y las inconsistencias en la asignación de probabilidades.

2.2. Función de distribución a priori

Toda la información previa que se tenga acerca del vector de parámetros desconocido θ , se cuantifica asignándole una distribución de densidad a priori $\pi(\theta)$.

Las distribuciones a priori se clasifican en conjugadas y no informativas.

2.2.1. Distribución a priori no informativa

En ocasiones se intentará utilizar la estadística bayesiana aún cuando no se posea información previa o se posea poca. En estos casos es necesario utilizar una distribución a priori no informativa, la cual es una función que no contiene información, o posee poca, acerca de θ .

Una forma de obtener tal información a priori es mediante **el principio de la razón insuficiente**, el cual enuncia lo siguiente: “Si no se sabe nada sobre el espacio parametral Θ , entonces no hay razón alguna para asignarle una probabilidad mas alta a algunos de estos sucesos inciertos”.

Supóngase que el parámetro de interés es la media θ de una normal, así el espacio paramétrico es $(-\infty, \infty)$. Si se desea asignar una distribución a priori no informativa, el principio de la razón insuficiente llevaría asignar igual peso a todos los posibles valores de θ . Desafortunadamente, si se elige $\pi(\theta) = c > 0$, entonces π tiene infinita masa (i.e $\int \pi(\theta)d\theta = \infty$). El valor de c no es relevante así que normalmente se utiliza $c=1$, así $\pi(\theta) = 1$. Frecuentemente ésta es la a priori usada, y es llamada la densidad no informativa natural, la cual es una a priori impropia¹. Un problema más grave que el de tener masa infinita es el de la falta de invarianza ante reparametrizaciones. Esto se ilustra en el siguiente ejemplo.

Ejemplo 2.1 Sea θ la proporción de individuos con cierta característica, entonces la distribución a priori inicial no informativa por el principio de la razón insuficiente será:

$$\phi(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{d.o.f.} \end{cases}$$

Supóngase ahora que el interés está puesto sobre $\varphi = -\log(\theta)$. Resulta obvio que si no se tiene información sobre θ , tampoco se tiene sobre φ , en este caso, bajo el principio de la razón insuficiente se debería asignar la distribución

$$\phi(\varphi) = \begin{cases} c & 0 < \varphi < \infty \\ 0 & \text{d.o.f.} \end{cases}$$

¹Se conoce como a priori impropia a aquella que tiene masa infinita.

Por otro lado, la distribución para φ inducida por $P(\theta)$, usando el teorema de la transformación[11] es

$$P_*(\varphi) = \begin{cases} \exp(-\varphi) & 0 < \varphi < \infty \\ 0 & \text{d.o.f.} \end{cases}$$

La cual es claramente informativa y por lo tanto diferente a $P(\varphi)$, a este tipo de problema se le conoce como falta de invarianza.

Principio de invarianza (Regla de Jeffreys)

La carencia de invarianza de la a priori derivada del principio de la razón insuficiente, ha conducido a la búsqueda de a priori iniciales no informativas, las cuales sean invariantes ante transformaciones del parámetro, ya que por ejemplo, Jeffreys dice[6]: “Cualquier parametrización arbitraria del modelo debe de llevar a los mismos resultados”. Se han dado muchas sugerencias para resolver este problema. El método más usado es el de Jeffreys (1961), pero antes de mencionarlo se dará la siguiente definición.

Definición 2.1 Dadas dos funciones $h(x)$ y $f(x)$, decimos que $h(x)$ es proporcional a $f(x)$, y se escribe

$$h(x) \propto f(x)$$

si y solo si existe una constante $k \in \mathfrak{R}$, la cual no depende de x , tal que

$$h(x) = kf(x).$$

Sea X una variable aleatoria con f.d.p. $P(x|\theta)$ y $\theta \in \mathfrak{R}$. Basándose en la noción de invarianza, Jeffreys propuso tomar la distribución inicial no informativa para el parámetro θ como

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}},$$

donde $I(\theta)$ es la información esperada de Fisher dada por:

$$I(\theta) = -E_{X|\theta} \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right]. \quad (2.1)$$

Si $\theta = (\theta_1, \dots, \theta_p)^t$ es un vector, Jeffreys propone usar

$$\pi(\theta) \propto \{\det[I(\theta)]\}^{\frac{1}{2}}, \quad (2.2)$$

donde $I(\theta)$ es la matriz (de $p \times p$) de información esperada de Fisher bajo condiciones de regularidad, con elemento (i, j)

$$I(\theta)_{i,j} = -E_{\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta) \right].$$

La idea intuitiva de basarse en la información de Fisher es que $I(\theta)$ se interpreta generalmente como un indicador de información provista por el modelo (o por las observaciones) acerca del parámetro θ . Además parece razonable suponer que los valores de θ para los cuales $\det[I(\theta)]$ es grande deben ser más probables bajo una distribución inicial no informativa. Dicho de otra manera, favorecer los valores de θ para los cuales $\det[I(\theta)]$ es grande, es equivalente a minimizar la influencia de la distribución inicial y por lo tanto esta será lo más no informativa posible.

Ejemplo 2.2 Sea $X \sim Pn(\lambda)$ es decir

$$f(x|\lambda) = \frac{\lambda^x}{x!} \exp(-\lambda), \quad \lambda > 0, x = 0, 1, 2, \dots$$

Calcule la distribución a priori de λ usando el criterio de Jeffreys.

Primero se obtiene

$$\begin{aligned} I(\lambda) &= -E_{x|\lambda} \left[\frac{\partial^2 \log f(x|\lambda)}{\partial \lambda^2} \right] \\ &= - \sum_{x=0}^{\infty} f(x|\lambda) \frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) \\ &= - \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \exp(-\lambda) \frac{\partial^2}{\partial \lambda^2} \left[\log\left(\frac{1}{x!}\right) + x \log \lambda - \lambda \right] \\ &= \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \exp(-\lambda) \frac{x}{\lambda^2} \\ &= \lambda^{-1} \exp(-\lambda) \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda^{-1} \exp(-\lambda) \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= \lambda^{-1} \exp(-\lambda) \exp(\lambda) = \lambda^{-1}. \end{aligned}$$

Luego la distribución a priori de Jeffreys es:

$$\pi(\lambda) \propto [I(\lambda)]^{\frac{1}{2}} = (\lambda^{-1})^{\frac{1}{2}} = \lambda^{-\frac{1}{2}},$$

$$\therefore \pi(\lambda) \propto \lambda^{-\frac{1}{2}}. \tag{2.3}$$

2.2.2. Distribución a priori conjugada

En general, dada una función de densidad $f(x|\theta)$ y una distribución a priori $\pi(\theta)$, será necesario calcular la distribución a posteriori

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\theta} f(x|\theta)\pi(\theta)d\theta},$$

lo cuál no es fácil por la integral involucrada, por lo que se necesita escoger a $\pi(\theta)$ de tal manera que facilite la tratabilidad de la integral. Además, el teorema de Bayes (2.5) se puede expresar como $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ y así se puede garantizar que tanto $\pi(\theta|x)$ como $\pi(\theta)$ pertenecen a la misma familia general de funciones matemáticas, siempre y cuando se escoja a $\pi(\theta)$ con la misma estructura que $f(x|\theta)$. Es por ello, que se habla de una familia a priori que asegura la tratabilidad y fácil interpretación de esa integral, lo cual motiva a la siguiente definición.

Definición 2.2 Sea la clase F de funciones de densidad $f(x|\theta)$, una clase ρ de distribuciones a priori se dice que es una familia conjugada para F si $\pi(\theta|x)$ está en la clase de ρ , para toda $f \in F$ y $\pi \in \rho$.

Frecuentemente para una clase de densidades F , una familia conjugada puede ser determinada examinando la función de verosimilitud $f(x|\theta)$. Entonces la familia conjugada puede ser escogida como la clase de distribuciones con la misma estructura funcional. A esta clase de familia de conjugada se le conoce con el nombre de distribución a priori conjugada natural.

Ejemplo 2.3 Sea $\vec{X} = (X_1, \dots, X_n)$ una muestra aleatoria con distribución Poisson. Entonces

$$X_i \sim Pn(x_i|\theta),$$

es decir,

$$f(x_i|\theta) = \frac{\theta^{x_i} \exp(-\theta)}{x_i!}, \theta > 0$$

y

$$f(\vec{x}|\theta) = \prod_{i=1}^n \left[\frac{\theta^{x_i} \exp(-\theta)}{x_i!} \right] = \frac{\theta^{n\bar{x}} \exp(-n\theta)}{\prod_{i=1}^n x_i!}.$$

$$\therefore f(\vec{x}|\theta) \propto \theta^{(n\bar{x}-1)+1} \exp(-n\theta)$$

con $\bar{X} = \sum_{i=1}^n X_i$.

Aquí F es la clase de todas las densidades. Obsérvese que la función de verosimilitud de tales densidades se parece a una densidad gamma.

$$Ga(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$

Dado que el dominio de la densidad gamma es $\theta > 0$, una posible familia de distribuciones a priori es la clase de distribuciones gamma. Se asume que $\theta \sim Ga(\alpha, \beta)$ y se observa que

$$\begin{aligned} \pi(\theta|\vec{x}) &\propto f(\vec{x}|\theta)\pi(\theta) = \frac{\theta^{n\bar{x}} \exp(-n\theta)}{\prod_{i=1}^n x_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \\ &= \frac{\beta^\alpha \theta^{n\bar{x}+\alpha-1} \exp[-(n+\beta)\theta]}{\Gamma(\alpha) \prod_{i=1}^n x_i!} \\ &\propto \theta^{n\bar{x}+\alpha-1} \exp[-(n+\beta)\theta]. \end{aligned}$$

El factor que envuelve a θ en esta última expresión es claramente reconocida como una distribución $Ga(n\bar{x} + \alpha, n + \beta)$, entonces, la distribución a posteriori es una distribución gamma. Por lo tanto, se tiene que la clase de distribuciones gamma es verdaderamente una familia conjugada (natural) de F .

2.3. Distribución a posteriori

La distribución a posteriori de θ dado x , donde $\theta \in \Theta$, será denotada por $\pi(\theta|x)$.
Nótese que por la ley multiplicativa, θ y x tienen distribución conjunta dada por:

$$h(x, \theta) = \pi(\theta)f(x|\theta)$$

y x , tiene densidad

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta, \quad (2.4)$$

si $m(x) \neq 0$

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)},$$

es decir

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}. \quad (2.5)$$

A esta fórmula se le conoce como el teorema de Bayes. Y de esta fórmula obtenemos la siguiente observación:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta} \iff \pi(\theta|x) \propto \pi(\theta)f(x|\theta). \quad (2.6)$$

Recuérdese que la distribución a priori refleja la creencia antes de la experimentación, de forma similar la distribución a posteriori $\pi(\theta|x)$ refleja la creencia actual de θ después de observar la muestra x .

Para determinar la distribución a posteriori, es de gran ayuda utilizar el término de suficiencia. De hecho si $S(X)$ es un estadístico suficiente para θ con densidad $g(s(x)|\theta)$ el siguiente resultado puede ser establecido:

Lema 2.1 Dada una m.a. $\vec{X} = (X_1, \dots, X_n)$ con $X_i \sim f(x_i|\theta)$, $\theta \sim \pi(\theta)$. Además $S = S(\vec{X})$ un estadístico suficiente para θ (esto es, $f(\vec{x}|\theta) = h(\vec{x})g(s(\vec{x})|\theta)$, i.e. $S(\vec{X}) \sim g(\vec{x}|\theta)$), entonces se tiene que:

$$\begin{aligned} \pi(\theta|\vec{x}) &= \frac{g(s|\theta)\pi(\theta)}{\int_{\theta} g(s|\theta)\pi(\theta)d\theta} \\ &= \pi(\theta|s). \end{aligned} \quad (2.7)$$

Demostración

$$\begin{aligned} \pi(\theta|\vec{x}) &= \frac{f(\vec{x}|\theta)\pi(\theta)}{\int f(\vec{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{h(\vec{x})g(s(\vec{x})|\theta)\pi(\theta)}{\int h(\vec{x})g(s(\vec{x})|\theta)\pi(\theta)d\theta} \\ &= \frac{g(s|\theta)\pi(\theta)}{\int_{\theta} g(s|\theta)\pi(\theta)d\theta} \\ &= \pi(\theta|s). \end{aligned}$$

□

Ejemplo 2.4 Supóngase que $X \sim N(\theta, \sigma^2)$ donde θ es desconocido y σ es conocido. También que $\pi(\theta)$ es una $N(\mu, \tau^2)$, con parámetros conocidos, entonces usando la ecuación (2.4) tenemos que

$$h(x, \theta) = \pi(\theta)f(x|\theta) = (2\pi\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2} \right] \right\},$$

para encontrar $m(x)$, note que definiendo

$$\rho = \tau^{-2} + \sigma^{-2} = \frac{\tau^2 + \sigma^2}{\tau^2\sigma^2}$$

y completando cuadrados se tiene que

$$\begin{aligned} \frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2} \right] &= \frac{1}{2} \left[\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right) \theta^2 - 2 \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \theta + \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \right] \\ &= \frac{1}{2} \rho \left[\theta^2 - \frac{2}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \theta \right] + \frac{1}{2} \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ &= \frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 - \frac{1}{2\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right)^2 + \frac{1}{2} \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ &= \frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 + \frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)}. \end{aligned}$$

Por tanto

$$h(x, \theta) = (2\pi\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 \right\} \exp \left[-\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)} \right]$$

y

$$m(x) = \int_{-\infty}^{\infty} h(x, \theta) d\theta = (2\pi\rho)^{-1/2} (\sigma\tau)^{-1} \exp \left[-\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)} \right].$$

Y de esto se sigue que

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \left(\frac{\rho}{2\pi} \right)^{1/2} \exp \left\{ -\frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 \right\}.$$

De la ecuación de arriba se obtiene que la distribución marginal de X es $N(\mu, (\tau^2 + \sigma^2))$ y la distribución a posteriori de θ dado x es $N(\mu(x), \rho^{-1})$ donde

$$\mu(x) = \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) = \frac{\sigma^2}{\sigma^2 + \tau^2} \mu + \frac{\tau^2}{\sigma^2 + \tau^2} x = x - \frac{\sigma^2}{\sigma^2 + \tau^2} (x - \mu).$$

Como un ejemplo concreto, considerese la situación en la que a un niño se le aplica una prueba de inteligencia. Supóngase que el resultado de esta prueba, X , es una $N(\theta, 100)$, donde θ es el verdadero IQ del niño (en otras palabras, si al niño se le aplicaran un número considerable de similares e independientes pruebas, la calificación promedio será θ). Asíumase que $\theta \sim N(100, 225)$.

Usando los resultados anteriores, se sigue que la distribución marginal de X es $N(100, 325)$, mientras que la distribución a posteriori de θ dado x es normal con media

$$\mu(x) = \frac{(100)(100) + x(225)}{100 + 225} = \frac{400 + 9x}{13}$$

y varianza

$$\rho^{-1} = \frac{100 * 225}{100 + 225} = \frac{900}{13} = 69.23,$$

por tanto, si la puntuación del niño es de 115 en la prueba, su verdadero IQ tiene una distribución a posteriori $N(110.39, 69.23)$.

Ejemplo 2.5 Continuación del ejemplo 2.4. Considérese nuevamente el ejemplo 2.4, con función de densidad y a priori normal. Pero ahora la muestra aleatoria es de tamaño n , es decir $\vec{X} = (X_1, \dots, X_n)$, dado que $S(\vec{X}) = S = \bar{X}$ es suficiente para θ , entonces se tiene que $\pi(\theta|\vec{x}) = \pi(\theta|\bar{x})$. Pero como $f(\bar{x}|\theta) = N(\theta, \sigma^2/n) = g(s|\theta)$, y del lema (2.7) se tiene que,

$$\begin{aligned} \pi(\theta|\vec{x}) &= \pi(\theta|\bar{x}) \\ &= N\left(\theta \left| \frac{(\sigma^2/n)\mu + \tau^2\bar{x}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2} \right.\right) \\ &= N\left(\theta \left| \frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \right.\right). \end{aligned}$$

Así la distribución a posteriori para θ es :

$$\pi(\theta|\vec{x}) = N\left(\theta \left| \frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \right.\right). \quad (2.8)$$

Capítulo 3

Inferencia bayesiana

Los problemas de inferencia bayesiana pueden ser tratados fácilmente usando la siguiente idea: Dado que la distribución a posteriori contiene toda la información acerca de θ , entonces cualquier inferencia del mismo debería de ser hecha en base a ésta distribución.

3.1. Estimación puntual

La técnica mas común para estimar θ , es tomar el valor $\hat{\theta}$ como aquel que maximiza la función de densidad a posteriori.

Definición 3.1 La probabilidad máxima generalizada estimada de θ es la moda más grande, $\hat{\theta}$, de $\pi(\theta|x)$ (i.e., el valor $\hat{\theta}$ que maximiza a $\pi(\theta|x)$, considerada como una función de θ). Obviamente $\hat{\theta}$ tiene la interpretación de ser el valor más parecido a θ , dada la función de densidad a priori y la muestra X .

Ejemplo 3.1 Continuación del ejemplo 2.4. Cuando f y π son densidades normales, la densidad a posteriori es una $N(\mu(x), \rho^{-1})$. Una densidad normal alcanza su máximo valor en su media, así la probabilidad máxima generalizada estimada de θ es:

$$\hat{\theta} = \mu(x) = \frac{\sigma^2 \mu}{\sigma^2 + \tau^2} + \frac{\tau^2 x}{\sigma^2 + \tau^2}.$$

Ejemplo 3.2 Supóngase que

$$f(x|\theta) = \exp\{-(x - \theta)\} I_{(\theta, \infty)}(x)$$

y

$$\pi(\theta) = \{\pi(1 + \theta^2)\}^{-1},$$

entonces

$$\pi(\theta|x) = \frac{\exp\{-(x - \theta)\} I_{(\theta, \infty)}(x)}{m(x)(1 + \theta^2)\pi}.$$

Si $\theta > x$, entonces $I_{(\theta, \infty)}(x) = 0$ y $\pi(\theta|x) = 0$, para este caso no vale la pena encontrar un estimador. Por otro lado si $\theta \leq x$, se necesita considerar

$$\begin{aligned}
 \frac{d\pi(\theta|x)}{d\theta} &= \frac{\exp(-x)}{m(x)\pi} \left[\frac{\exp(\theta)}{1+\theta^2} - \frac{2\theta \exp(\theta)}{(1+\theta^2)^2} \right] \\
 &= \frac{\exp(-x)}{m(x)\pi} \left[\frac{\exp(\theta) + \exp(\theta)\theta^2 - 2\theta \exp(\theta)}{(1+\theta^2)^2} \right] \\
 &= \frac{\exp(-x) \exp(\theta)}{m(x)\pi} \left[\frac{1 + \theta^2 - 2\theta}{(1+\theta^2)^2} \right] \\
 &= \frac{\exp(-x) \exp(\theta)}{m(x)\pi} \left[\frac{(\theta-1)^2}{(1+\theta^2)^2} \right].
 \end{aligned}$$

Y como esta derivada siempre es positiva, $\pi(\theta|x)$ es creciente para $\theta \leq x$. De aquí se sigue que $\pi(\theta|x)$ se maximiza en $\hat{\theta} = x$, el cual es la probabilidad máxima generalizada estimada de θ .

Otros estimadores bayesianos comunes de θ son la media y la mediana de $\pi(\theta|x)$.

3.1.1. Error de estimación

Cuando se presenta un estimador estadístico, usualmente es necesario indicar la precisión de dicha estimación. La medida usual de la precisión de la estimación (en una dimensión) es la varianza a posteriori de la estimación, la cual se define como:

Definición 3.2 Si θ es un parámetro de valor real con distribución a posteriori $\pi(\theta|x)$, y δ es el estimador de θ , entonces la varianza a posteriori de δ es

$$V_{\delta}^{\pi}(x) = E^{\pi(\theta|x)}[(\theta - \delta)^2].$$

Cuando δ es la media a posteriori

$$\delta = \mu^{\pi}(x) = E^{\pi}(\theta)$$

entonces $V^{\pi} = V_{\delta}^{\pi}$ y simplemente se le llamará varianza a posteriori. La desviación estándar a posteriori es $\sqrt{V^{\pi}(x)}$. Es costumbre usar $\sqrt{V_{\delta}^{\pi}}$ como la desviación estándar de la estimación δ .

Para propósitos de cálculos, es útil usar la siguiente fórmula:

$$\begin{aligned}
 V_{\delta}^{\pi} &= E^{\pi(\theta|x)} \left[(\theta - \delta)^2 \right] \\
 &= E \left[\left(\theta - \mu^{\pi}(x) + \mu^{\pi}(x) - \delta \right)^2 \right] \\
 &= E \left[\left(\theta - \mu^{\pi}(x) \right)^2 \right] + E \left[2 \left(\theta - \mu^{\pi}(x) \right) \left(\mu^{\pi}(x) - \delta \right) \right] + E \left[\left(\mu^{\pi}(x) - \delta \right)^2 \right] \\
 &= V^{\pi}(x) + 2 \left(\mu^{\pi}(x) - \delta \right) \left(E(\theta) - \mu^{\pi}(x) \right) + \left(\mu^{\pi}(x) - \delta \right)^2 \\
 &= V^{\pi}(x) + \left(\mu^{\pi}(x) - \delta \right)^2.
 \end{aligned} \tag{3.1}$$

Obsérvese que de (3.1) la media a posteriori $\mu^\pi(x)$, minimiza $V_\delta^\pi(x)$ (sobre todas las delta), y por tanto es la estimación con error estándar mas pequeño. Por esta razón, es costumbre usar $\mu^\pi(x)$ para estimar θ y reportar $\sqrt{V^\pi(x)}$ como el error estandar.

Ejemplo 3.3 Continuación del ejemplo 2.4. Es claro que

$$V^\pi(x) = \rho^{-1} = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2},$$

por tanto, en el ejemplo de la prueba de inteligencia, el niño con $x = 115$ podría estimarse que tiene un IQ de $\mu^\pi(115) = 110.39$, con un error estandar

$$\sqrt{V^\pi(115)} = \sqrt{69.23} = 8.32$$

3.1.2. Estimación multivariada

La estimación bayesiana de un vector $(\theta_1, \dots, \theta_p)^t$ es también relativamente fácil de obtener. La pobabilidad máxima generalizada estimada (la moda a posteriori) es a menudo una razonable estimación, aunque la existencia y la unicidad no siempre se puede garantizar en el caso multivariado.

La media a posteriori $\mu^\pi(x) = (\mu_1^\pi(x), \dots, \mu_p^\pi(x))^t = E^{\pi(\theta|x)}[\theta]$ es una estimación bayesiana muy atractiva, y la presición puede ser descrita por la matriz de covarianzas a posteriori

$$V^\pi(x) = E^{\pi(\theta|x)} \left[\left(\theta - \mu^\pi(x) \right) \left(\theta - \mu^\pi(x) \right)^t \right].$$

Para esta instancia, el error de estimación estandar de $\mu_i^\pi(x)$ de θ_i sería $\sqrt{V_{ii}^\pi(x)}$ donde $V_{ii}^\pi(x)$ es el elemento (i, i) de $V^\pi(x)$. El análogo de (3.1), para una estimación general δ de θ , es

$$\begin{aligned} V_\delta^\pi &= E^{\pi(\theta|x)} \left[(\theta - \delta)(\theta - \delta)^t \right] \\ &= V^\pi(x) + \left(\mu^\pi(x) - \delta \right) \left(\mu^\pi(x) - \delta \right)^t. \end{aligned}$$

Otra vez es claro que la media a posteriori minimiza V_δ^π .

3.2. Conjuntos creíbles

Otro método común de la inferencia clásica son los intervalos de confianza para θ . La estimación análoga en estadística bayesiana son los conjuntos creíbles, definidos de la siguiente manera:

Definición 3.3 Un conjunto creíble C de Θ al $100(1 - \alpha)\%$ es aquel que:

$$1 - \alpha \leq P(C|x) = \begin{cases} \int_C \pi(\theta|x) d\theta & \text{caso continuo.} \\ \sum_{\theta \in C} \pi(\theta|x) & \text{caso discreto.} \end{cases}$$

Como la distribución a posteriori es la distribución actual de θ así se puede hablar de la probabilidad de $\theta \in C$. Este es un claro contraste con los métodos clásicos de intervalos de confianza, los cuales solo pueden ser interpretados en términos de una probabilidad de cobertura (la probabilidad de que la variable aleatoria X este en tal intervalo de confianza $C(X)$). Para la elección de los conjuntos creíbles para θ , es común elegir aquel cuya longitud sea mínima. Para hacer esto, se necesita incluir los puntos con mayor probabilidad a posteriori, es decir los valores más probables de θ .

Definición 3.4 Un conjunto creíble de máxima densidad a posteriori (HPD por sus siglas en inglés Highest Posterior Density) para θ al $100(1 - \alpha)\%$ (el conjunto HPD mantiene la más alta densidad a posteriori), es un subconjunto C de Θ de la forma:

$$C = \{\theta \in \Theta : \pi(\theta|x) \geq k(\alpha)\}$$

donde $k(\alpha)$ es la constante más grande tal que

$$P(C|x) \geq 1 - \alpha.$$

Ejemplo 3.4 Continuación del ejemplo 2.4. Como la densidad a posteriori de θ dada x es $N(\mu(x), \rho^{-1})$ la cual es unimodal y simétrica cerca de $\mu(x)$, entonces, el conjunto creíble HPD al $100(1 - \alpha)\%$, está dado por

$$C = \left(\mu(x) + z\left(\frac{\alpha}{2}\right)\rho^{-1/2}, \mu(x) - z\left(\frac{\alpha}{2}\right)\rho^{-1/2} \right),$$

donde $z(\alpha)$ es el α -cuantil de una $N(0, 1)$.

En el ejemplo del IQ, cuando el niño obtenía una puntuación de 115 en la prueba de inteligencia, tenía como distribución a posteriori $N(110.39, 69.23)$ para θ , luego un conjunto creíble HPD al 95% para θ es

$$\left(110.39 + (-1.96)(69.23)^{1/2}, 110.39 + (1.96)(69.23)^{1/2} \right) = (94.08, 126.70).$$

Por otro lado un intervalo de confianza (clásico) al 95% para θ sería :

$$\left(115 - (1.96)(10), 115 + (1.96)(10) \right) = (95.4, 134.6)$$

el cual tiene una longitud mayor al que se calculó bayesianamente.

Los conjuntos creíbles bayesianos son usualmente más fáciles de calcular que sus contrapartes clásicas, particularmente cuando no existen los estadísticos suficientes. La idea general detrás del cálculo de los conjuntos creíbles HPD, cuando $\pi(\theta|x)$ es continua en θ , se puede implementar en un programa a través de las siguientes líneas:

1. Crear una subrutina la cual, dado k , encuentre todas las soluciones a la ecuación $\pi(\theta|x) = k$. El conjunto $C(k) = \{\theta : \pi(\theta|x) \geq k\}$, puede ser fácilmente calculado a partir de esas soluciones. Por ejemplo, si Θ es un intervalo infinito en \mathfrak{R} , y solo tiene dos soluciones, $\theta_1(k)$ y $\theta_2(k)$, entonces

$$C(k) = (\theta_1(k), \theta_2(k)).$$

2. Crear una subrutina la cual calcule

$$P^{\pi(\theta|x)}(C(k)) = \int_{C(k)} \pi(\theta|x)d\theta.$$

3. Ahora solo hay que resolver numéricamente la ecuación

$$P^{\pi(\theta|x)}(C(k)) = 1 - \alpha$$

llamando las subrutinas 1 y 2, variando k adecuadamente.

En ocasiones puede pasar que el conjunto creíble HPD sea la unión de intervalos disjuntos, pero en estas situaciones se toma el conjunto creíble HPD como un conjunto convexo, por las siguientes razones:

1. Los intervalos disjuntos aparecen cuando existe un choque en la información, esto es, la a priori dice una cosa mientras que los datos dicen otra.
2. Las a priori conjugadas son a menudo unimodales y también sus a posteriori, y por tanto se pretende cubrir los choques de información (mencionados anteriormente).

Para θ multivariado, la definición de conjunto creíble HPD se sigue manteniendo.

3.2.1. Diferencias entre los conjuntos creíbles y los intervalos de confianza

- En la teoría bayesiana un intervalo creíble C al $(1 - \alpha)100\%$ tiene el siguiente significado:
 - La probabilidad de que θ este en C dados los datos observados es al menos $(1 - \alpha)$.
- En la teoría clásica un intervalo de confianza al $(1 - \alpha)100\%$ tiene el siguiente significado:
 - Si se pudiera recalcular C para una gran base de datos de conjuntos de muestras aleatorias X , cerca del $(1 - \alpha)100\%$ de ellos cubrirían el valor real de θ .

Es claro ver que un conjunto creíble da la certeza de que dada esa muestra aleatoria tal conjunto contendría al parámetro desconocido θ con al menos una probabilidad $1 - \alpha$, mientras que el intervalo de confianza dice que si generas 100 conjuntos de muestras aleatorias, y calculas el intervalo de confianza para cada caso (los cuales no tienen que ser necesariamente iguales), al menos $(1 - \alpha)100$ cubrirían a θ . La desventaja principal que tiene los intervalos de confianza clásicos es que no siempre se pueden generar tantas muestras aleatorias.

3.3. Contraste de hipótesis

El contraste de hipótesis constituye una técnica inferencial cuyo objetivo es comprobar la plausibilidad de dos hipótesis estadísticas contrapuestas mediante la formulación e interpretación de una regla de decisión.

Desde el enfoque clásico o frecuentista esta regla de decisión está basada en las características del estadístico de contraste, en las de su distribución muestral, y la evaluación de las hipótesis se realiza en términos de las probabilidades de dos tipos de error. Estas probabilidades de error representan la “casualidad” de que se observe una muestra para la cual el contraste acepte la hipótesis errónea. Desde el enfoque bayesiano la tarea de evaluación de las hipótesis resulta conceptualmente más sencilla. La regla de decisión se basa en el cálculo de las probabilidades a posteriori de las hipótesis contrastadas y su evaluación depende de los resultados obtenidos. La ventaja de este último análisis es que dichas probabilidades a posteriori son las verdaderas (subjetivas) probabilidades de las hipótesis que reflejan los datos observados y la distribución a priori.

Aunque las probabilidades a posteriori de las hipótesis son las principales medidas en los problemas de contraste, en este capítulo se define, entre otros, un concepto de gran interés: el factor de Bayes.

Además se abordarán, desde la perspectiva bayesiana, la resolución de varios tipos de contrastes cuyas hipótesis se refieren a parámetros de distribución.

Sea $\{f(x, \theta), \theta \in \Theta\}$ una familia de funciones de densidad, $\Theta \subseteq \mathfrak{R}$. Supóngase que se desea contrastar la hipótesis

$$H_0 : \theta \in \Theta_0 \text{ frente a } H_1 : \theta \in \Theta_1.$$

Desde la perspectiva bayesiana, para contrastes de hipótesis se seleccionarían las distribuciones a priori $\pi_i = \pi(\Theta_i)$, $i = 0, 1$, y se calcularían las probabilidades a posteriori $\alpha_i = \pi(\Theta_i|x)$, $i = 0, 1$, para los parámetros de las hipótesis nula, H_0 , y la hipótesis alternativa, H_1 , respectivamente. La tarea de decidir entre ambas hipótesis depende de dichos resultados.

Definición 3.5 Se denomina razón a priori de H_0 respecto de H_1 al cociente π_0/π_1 donde $\pi_i = \pi(\Theta_i)$, y análogamente, se denomina razón a posteriori al cociente α_0/α_1 con $\alpha_i = \pi(\Theta_i|x)$. Además

$$B = \frac{\text{razón a posteriori}}{\text{razón a priori}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

es denominado **factor de Bayes** a favor de Θ_0 .

Aunque el problema anterior de contraste de hipótesis se ha enunciado mediante un lenguaje bayesiano estándar, la mayoría de las personas trabajan mejor con razones que con probabilidades, por lo que a menudo resulta conveniente resumir la evidencia del contraste en términos de su razón a posteriori (concluir que $\alpha_0/\alpha_1 = 10$ claramente conduce a que la hipótesis nula es 10 veces más probable de ser cierta que la hipótesis alternativa). En ocasiones el factor de Bayes puede ser interpretado como “la razón dada por los datos para H_0 respecto de H_1 ”. Esta interesante interpretación es, claramente, válida cuando las hipótesis en estudio son simples.

Sea $\Theta = \{\theta_0, \theta_1\}$, X una variable aleatoria con función de densidad $f(x|\theta)$ y se desean contrastar las hipótesis $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ respectivamente. En este caso,

$$\alpha_0 = \pi(\theta_0|x) = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}$$

$$\alpha_1 = \pi(\theta_1|x) = \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}$$

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)}$$

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

En otras palabras, B representa la razón de verosimilitudes de la hipótesis nula H_0 respecto de la alternativa H_1 . Este cociente se interpreta comúnmente, incluso por muchos no bayesianos[1], como la razón que arrojan los datos para H_0 respecto de H_1 . En general, el factor de Bayes dependerá de la distribución a priori inicial por lo que será conveniente escribirla como:

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{si } \theta \in \Theta_0 \\ \pi_1 g_1(\theta) & \text{si } \theta \in \Theta_1 \end{cases}$$

donde g_0 y g_1 son densidades (propias) que describen como la masa a priori se extiende sobre las hipótesis. Entonces, si definimos $dF^{g(\theta)}(\theta) = g(\theta)d\theta$ se tiene que

$$\begin{aligned} \frac{\alpha_0}{\alpha_1} &= \frac{\int_{\Theta_0} dF^{\pi(\theta|x)}(\theta)}{\int_{\Theta_1} dF^{\pi(\theta|x)}(\theta)} \\ &= \frac{\int_{\Theta_0} f(x|\theta)\pi_0 dF^{g_0}(\theta)/m(x)}{\int_{\Theta_1} f(x|\theta)\pi_1 dF^{g_1}(\theta)/m(x)} \\ &= \frac{\pi_0 \int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\pi_1 \int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)}, \end{aligned}$$

y por la definición del factor de Bayes se tiene

$$B = \frac{\int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)}, \tag{3.2}$$

que representa la razón de verosimilitudes “ponderadas” (por g_0 y g_1) de Θ_0 respecto a Θ_1 . Como consecuencia de la introducción de las densidades g_0 y g_1 , B no podrá interpretarse, en general, como una medida que depende solamente de los datos (esto sólo ocurrirá cuando la elección de tales densidades no influya en B).

Ejemplo 3.5 Continuación del ejemplo 2.4. Supónganse que se desea contrastar la hipótesis

$$H_0 : \theta \leq 100 \text{ contra } H_1 : \theta > 100.$$

También supóngase que la puntuación del niño fue de 115. Como $\pi(\theta|x) = N(110.39, 69.23)$, se tiene que:

$$\alpha_0 = \pi(\theta \leq 100|x) = .106 \text{ y } \alpha_1 = \pi(\theta > 100|x) = .894$$

$$\frac{\alpha_0}{\alpha_1} = 1/8.44$$

Es decir H_1 es 8.44 veces más probable que H_0 . Además puesto que la distribución a priori es $N(100, 225)$, se tiene que:

$$\pi_0 = \pi(\theta \leq 100) = .5, \quad \pi_1 = \pi(\theta > 100) = .5$$

Por tanto el factor de Bayes es:

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = 1/8.44$$

Es decir, la puntuación dada por los datos para H_1 frente H_0 es 8.44. A vista de este resultado se podría concluir que el coeficiente intelectual del individuo en estudio es mayor o igual al de la media poblacional, o lo que es lo mismo se rechazaría H_0 .

Es interesante notar que no es conveniente el uso de distribuciones a priori impropias en problemas de contraste de hipótesis. Supóngase que se toma una distribución a priori impropia

$$\pi(\theta) = \begin{cases} c_0 g_0(\theta) & \text{si } \theta \in \Theta_0 \\ c_1 g_1(\theta) & \text{si } \theta \in \Theta_1, \end{cases}$$

donde g_0 y g_1 son funciones cuya integral diverge por lo que las constantes normalizadoras c_i , $i = 0, 1$, no están determinadas. En este caso las probabilidades a priori π_i , $i = 0, 1$, serán infinitas por lo que la razón a priori π_0/π_1 no estará definida y el factor Bayes será de la forma (3.2)

$$B = \frac{c_0 \int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{c_1 \int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)}, \quad (3.3)$$

el cual está definido salvo la constante (c_0/c_1) , que es arbitraria.

3.3.1. Contraste de hipótesis unilateral

Sea $\{f(x, \theta), \theta \in \Theta\}$ una familia de funciones de densidad, $\Theta \subseteq \mathfrak{R}$. El contraste de hipótesis unilateral será de la forma

$$H_0 : \theta \leq \theta_0 \text{ frente } H_1 : \theta > \theta_0,$$

o su dual

$$H_0 : \theta \geq \theta_0 \text{ frente } H_1 : \theta < \theta_0.$$

Desde el punto de vista bayesiano este tipo de contraste no presenta ninguna característica especial, sin embargo, representa una de las pocas situaciones de contraste en las que el enfoque clásico, en particular el uso del P-valor, tiene justificación bayesiana.

Considérese el siguiente ejemplo.

Ejemplo 3.6 Sea X una v.a. con distribución $N(\theta, \sigma^2)$, con σ^2 conocida, y considérese la distribución a priori no informativa uniforme, $\pi(\theta) = 1$. Supóngase que se desea realizar el contraste:

$$H_0 : \theta \leq \theta_0 \text{ frente } H_1 : \theta > \theta_0.$$

En este caso se tiene que

$$h(x, \theta) = \pi(\theta)f(x|\theta) = f(x|\theta)$$

$$m(x) = \int_{-\infty}^{\infty} f(x|\theta)d\theta = (2\pi)^{-1/2}(\sigma^{-1}) \int_{-\infty}^{\infty} \exp\left[\frac{-(x-\theta)^2}{2\sigma^2}\right] = 1$$

y

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = (2\pi^{-1/2})\sigma^{-1} \exp\left[\frac{-(x-\theta)^2}{2\sigma^2}\right].$$

Por lo tanto la distribución a posteriori de θ dado x es una $N(x, \sigma^2)$. Tipificando este resultado se tiene que

$$\alpha_0 = \pi(\theta \leq \theta_0|x) = \Phi((\theta_0 - x)/\sigma),$$

donde Φ es la función de distribución acumulada de una $N(0, 1)$.

Desde el enfoque clásico el P-valor contra H_0 es la probabilidad, cuando $\theta = \theta_0$, de observar un X “más extremo” que el dato observado x . Y puesto que $X \sim N(\theta, \sigma^2)$ tipificando el resultado, el P-valor será

$$\text{P-valor} = P(X \geq x) = 1 - \Phi((x - \theta_0)/\sigma).$$

Como consecuencia de la simetría de la distribución normal, se sigue que α_0 es igual al P-valor contra H_0 . Véase pues, que en este caso los resultados obtenidos mediante ambas inferencias coinciden, lo cual es realmente positivo.

El uso de la distribución a priori uniforme en el ejemplo puede parecer a primera vista desconcertante, ya que al ser impropia otorga masa infinita a cada hipótesis, y poco conveniente pues en este caso el factor Bayes es de la forma (3.3) definido salvo una constante multiplicativa.

No obstante, en este caso es posible justificar su uso como una distribución a priori no informativa cuando se pretende expresar la idea de que se tiene poco conocimiento a priori sobre el parámetro en estudio. Además, en gran número de contrastes de este tipo la distribución a priori por defecto da lugar a resultados en los que la probabilidades a posteriori son similares a los P-valores.

Sin embargo, no siempre se obtienen estos resultados en todos los problemas de contraste de hipótesis unilateral. Por ejemplo, el contraste

$$H_0 : \theta = 0 \text{ (} H_0 : 0 \leq \theta \leq \epsilon \text{)} \quad \text{contra} \quad H_1 : \theta > 0 \text{ (ó } H_1 : \theta > \epsilon \text{)}$$

es unilateral y los P-valores como las probabilidades a posteriori tienden a ser diferentes[1].

3.3.2. Contraste de hipótesis nula puntual

Sea $\{f(x, \theta), \theta \in \Theta\}$ una familia de funciones de densidad, $\Theta \subseteq \mathfrak{R}$. Es muy común en la estadística clásica plantear contrastes de la forma

$$H_0 : \theta = \theta_0 \text{ frente } H_0 : \theta \neq \theta_0,$$

estos contrastes se denominan de hipótesis nula puntual. Como se verá a lo largo de esta subsección, este tipo de contrastes son muy interesantes, en parte porque desde la perspectiva bayesiana contienen algunas características nuevas, pero principalmente porque

las respuestas bayesianas difieren radicalmente de las respuestas clásicas, lo que pone de manifiesto un grave conflicto entre ambas inferencias.

Antes de discutir estas cuestiones, hay que contemplar otros aspectos relativos a la naturaleza de este tipo de contraste. En primer lugar, este contraste se plantea comúnmente en situaciones inapropiadas, pues nunca se dará el caso en el que se abrigue seriamente la posibilidad de que $\theta = \theta_0$ exactamente. Una hipótesis nula más razonable será $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, donde $b > 0$ es una constante elegida de tal manera que todo $\theta \in \Theta_0$ pueda considerarse indistinguible de θ_0 . Un ejemplo en el que podría presentarse esta situación sería el intento de analizar un producto químico a través de la reacción de alguna de sus características, θ , con otro producto químico conocido. Por ejemplo, se podría plantear el estudio de que el producto químico desconocido posee una componente específica con una potencia de reacción, θ_0 , con una precisión b , entonces sería razonable un contraste de la forma

$$H_0 : \theta \in (\theta_0 - b, \theta_0 + b) \quad \text{frente a} \quad H_1 : \theta \notin (\theta_0 - b, \theta_0 + b).$$

Por supuesto, existen muchos problemas de decisión que dan lugar a una hipótesis nula de la forma anterior con un b grande, en estos problemas la aproximación por el contraste de hipótesis nula puntual será raramente satisfactoria. Puesto que realmente se debería contrastar $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, sería necesario saber cuando es lícita la aproximación de H_0 por $H_0 : \theta = \theta_0$.

Desde la perspectiva bayesiana tal aproximación resultará razonable cuando la probabilidad a posteriori de H_0 sea prácticamente igual en ambos contrastes. Además, otra condición importante para esta cuestión, será que la función de probabilidad de las observaciones sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$.

Pese a que desde la perspectiva bayesiana resultará mucho más sencillo tratar directamente con la hipótesis referente del intervalo, que estimar la efectividad de su aproximación por una hipótesis nula puntual, a continuación se explicará el contraste de hipótesis puntual.

Es importante notar que en el contraste bayesiano de hipótesis nula puntual ($H_0 : \theta = \theta_0$) no es posible utilizar densidades a priori continuas pues en ese caso dichas distribuciones a priori (al igual que las distribuciones a posteriori) otorgarían probabilidad cero a θ_0 . Una aproximación razonable será dar a θ_0 una probabilidad positiva π_0 y a $\theta \neq \theta_0$ la distribución a priori $\pi_1 g_1(\theta)$ donde $\pi_1 = 1 - \pi_0$ y g_1 es propia. Uno puede pensar en π_0 como la masa que se le asignaría a la hipótesis nula real, $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, si no se hubiera preferido aproximar por la hipótesis nula puntual.

Teniendo en cuenta esta puntualización la densidad marginal de X será

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta = f(x|\theta_0)\pi_0 + (1 - \pi_0)m_1(x),$$

donde

$$m_1(x) = \int_{(\theta \neq \theta_0)} f(x|\theta)g_1(\theta)d\theta$$

es la densidad marginal de X con respecto a g_1 . Por lo tanto, la probabilidad a posteriori para $\theta = \theta_0$ será

$$\begin{aligned}
 \pi(\theta_0|x) &= \frac{f(x|\theta_0)\pi_0}{m(x)} \\
 &= \frac{f(x|\theta_0)\pi_0}{f(x|\theta_0)\pi_0 + (1 - \pi_0)m_1(x)} \\
 &= \left[1 + \frac{1 - \pi_0}{\pi_0} \frac{m_1(x)}{f(x|\theta_0)} \right]^{-1}.
 \end{aligned} \tag{3.4}$$

Recordando que $\pi_1 = 1 - \pi_0$, la razón a posteriori resulta

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi(\theta_0|x)}{1 - \pi(\theta_0|x)} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 m_1(x)}, \tag{3.5}$$

por lo que el factor de Bayes para H_0 frente H_1 será

$$B = \frac{f(x|\theta_0)}{m_1(x)}. \tag{3.6}$$

Ejemplo 3.7 Sea X_1, \dots, X_n una m.a. de una $N(\theta, \sigma^2)$ donde σ es conocida y se desea contrastar

$$H_0 : \theta = \theta_0 \text{ frente a } H_1 : \theta \neq \theta_0.$$

Simplificando los cálculos mediante el uso de la media muestral, \bar{X} , es un estadístico suficiente para la muestra, y se obtiene la siguiente función de probabilidad

$$f(\bar{x}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp \left[-\frac{n}{2\sigma^2} (\theta - \bar{x})^2 \right].$$

Si se supone que g_1 se distribuye como una $N(\mu, \tau^2)$ en $\theta \neq \theta_0$ (normalmente $\mu = \theta_0$ pues a priori se considerará más probable que esté cerca) entonces, como vimos en el ejemplo 2.4, $m_1(\bar{x})$ tendrá una densidad $N(\mu, \tau^2 + \sigma^2/n)$, donde el punto $\theta = \theta_0$ no interviene en la integración puesto que g_1 es una densidad continua. Sustituyendo los datos en (3.4) se tiene

$$\pi(\theta_0|x) = \left\{ 1 + \frac{1 - \pi_0}{\pi_0} \frac{\left[2\pi(\tau^2 + \sigma^2/n) \right]^{-1/2} \exp \left\{ -(\bar{x} - \mu)^2 / [2(\tau^2 + \sigma^2/n)] \right\}}{\left[2\pi\sigma^2/n \right]^{-1/2} \exp \left[-(\bar{x} - \theta_0)^2 / (2\sigma^2/n) \right]} \right\}^{-1}. \tag{3.7}$$

En el caso especial en que $\mu = \theta_0$ el resultado (3.4) se reduce a

$$\begin{aligned} \alpha_0 &= \left\{ 1 + \frac{1 - \pi_0}{\pi_0} \frac{\exp \left\{ -(\bar{x} - \theta_0)^2 n \tau^2 / [2\sigma^2(\tau^2 + \sigma^2/n)] \right\}}{\left[1 + n\tau^2\sigma^2 \right]^{1/2}} \right\}^{-1} \\ &= \left\{ 1 + \frac{1 - \pi_0}{\pi_0} \frac{\exp \left\{ \frac{1}{2} z^2 \left[1 + \sigma^2 / (n\tau^2) \right]^{-1} \right\}}{\left(1 + n\tau^2 / \sigma^2 \right)^{1/2}} \right\}^{-1}, \end{aligned} \quad (3.8)$$

donde

$$z = \sqrt{n} |\bar{x} - \theta_0| / \sigma,$$

es el estadístico clásico utilizado para el contraste $H_0 : \theta = \theta_0$. Para calcular el factor de Bayes, solo resta sustituir en (3.6).

$$\begin{aligned} B &= \frac{\frac{1}{\sqrt{2\pi\sigma^2/n}} \exp \left[-\frac{n(\theta_0 - \bar{x})^2}{2\sigma^2} \right]}{\frac{1}{\sqrt{2\pi(\tau^2 + \sigma^2/n)}} \exp \left[-\frac{(\theta_0 - \bar{x})^2}{2(\tau^2 + \sigma^2/n)} \right]} \\ &= \frac{\sqrt{n\tau^2 + \sigma^2}}{\sigma} \exp \left[\frac{-n^2(\theta_0 - \bar{x})^2\tau^2}{2\sigma^2(n\tau^2 + \sigma^2)} \right]. \end{aligned}$$

Capítulo 4

Simulación e integración Monte Carlo

4.1. Integración Monte Carlo

En problemas de estadística se tiene la situación de evaluar:

$$\int f(x|\theta)\pi(\theta)d\theta.$$

La idea básica del método de Monte Carlo consiste en escribir la integral requerida como el valor esperado de alguna función con respecto a alguna distribución de probabilidad. Esto es:

$$h(x) = E_{\pi}[f(x|\theta)] = \int f(x|\theta)\pi(\theta)d\theta. \quad (4.1)$$

Donde $\theta \sim \pi(\theta)$. A partir de esta idea, Monte Carlo propone un conjunto de métodos para calcular tales integrales, los cuales se clasifican como:

- Métodos directos
- Métodos indirectos

De estos dos métodos sólo se describirá en la tesis el método directo, ya que este es el que utilizará el software desarrollado.

4.1.1. Métodos directos

Muestreo directo

Uno de los métodos para evaluar la integral (4.1) es el muestreo directo. Para este método se supone $\theta \sim \pi(\theta|x)$, y se busca el valor de

$$\gamma = E[g(\theta)] = \int g(\theta)\pi(\theta|x)d\theta. \quad (4.2)$$

Entonces, se generan posibles variables aleatorias $\theta_1, \dots, \theta_m$ *i.i.d.* de $\pi(\theta|x)$ (un muestreo) y se tiene la aproximación de (4.2) como el promedio muestral:

$$\gamma = \frac{1}{m} \sum_{i=1}^m g(\theta_i) \quad (4.3)$$

que converge a $E[g(\theta)]$ con probabilidad 1, cuando $m \rightarrow \infty$, según la ley fuerte de los grandes números. A continuación se demostrará la ley débil de los grandes números y dos teoremas auxiliares, después se enunciará la ley fuerte de los grandes números.

Teorema 4.1 (Desigualdad de Markov) Supóngase X una variable aleatoria tal que $P(X \geq 0) = 1$. Entonces, para cualquier número $t > 0$

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Demostración Supóngase que X tiene una distribución discreta cuya función de densidad de probabilidad es f . La demostración para una distribución continua o un tipo de distribución más general es análoga. Para una distribución discreta

$$E(X) = \sum_x xf(x) = \sum_{x < t} xf(x) + \sum_{x \geq t} xf(x),$$

como X probabilísticamente puede tomar únicamente valores no negativos, todos los términos de la suma son no negativos. Luego

$$E(X) \geq \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} tf(x) = tP(X \geq t).$$

Por tanto

$$\frac{E(X)}{t} \geq P(X \geq t).$$

□

Teorema 4.2 (Desigualdad de Chebyshev) Sea X una variable aleatoria para la cual $\text{Var}(X)$ existe. Entonces para cualquier número concreto $\epsilon > 0$

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

Demostración Sea $Y = [X - E(X)]^2$. Entonces $P(Y \geq 0) = 1$ y $E(Y) = \text{Var}(X)$. Aplicando a Y la desigualdad de Markov

$$P(|X - E(X)| \geq \epsilon) = P(Y \geq \epsilon^2) \leq \frac{E(Y)}{\epsilon^2} = \frac{\text{Var}(X)}{\epsilon^2}.$$

Por lo tanto

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

□

Teorema 4.3 (Ley débil de los grandes números) Supóngase que X_1, \dots, X_n constituyen una m.a. cuya media es μ y sea \bar{X}_n la media muestral. Entonces

$$P \lim_{n \rightarrow \infty} \bar{X}_n = \mu.$$

Esto quiere decir, la sucesión X_1, \dots, X_n de variables aleatorias converge en probabilidad a μ si para cualquier número dado $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

Demostración Supóngase que la distribución de donde proviene la muestra aleatoria tiene una media finita μ y una varianza finita σ^2 . De la desigualdad de Chebyshev, para cualquier número dado $\epsilon > 0$

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \epsilon) &\leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\text{Var}(X)}{n\epsilon^2}, \\ 1 - P(|\bar{X}_n - \mu| < \epsilon) &\leq \frac{\text{Var}(X)}{n\epsilon^2}, \\ 1 - \frac{\text{Var}(X)}{n\epsilon^2} &\leq P(|\bar{X}_n - \mu| < \epsilon) \end{aligned}$$

entonces

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1,$$

$$\therefore P \lim_{n \rightarrow \infty} \bar{X}_n = \mu.$$

□

Es decir, la media muestral siempre converge en probabilidad a la media de la distribución de la que se seleccionó la muestra aleatoria.

Teorema 4.4 (Ley fuerte de los grandes números) Sea X_1, \dots, X_n una m.a. cuya media es μ y sea \bar{X}_n la media muestral. Entonces

$$P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1.$$

Es decir la sucesión X_1, \dots, X_n converge a μ con probabilidad 1.

En este caso, $\pi(\theta|x)$ es una distribución a posteriori y γ es la media a posteriori de $g(\theta)$. Aquí el cálculo de la esperanza a posteriori solo requiere una muestra de tamaño m de la distribución a posteriori.

Note que la calidad de la aproximación en (4.3) mejora cuando se incrementa m , el tamaño de la muestra generada por el método Monte Carlo (que nosotros elegimos) y se aproxima a n , el tamaño del conjunto de datos que está típicamente más allá de nuestro control. Otro contraste con los métodos asintóticos es que la estructura de (4.3) también

nos permite evaluar su precisión para cualquier m fijo. Dado que $\hat{\gamma}$ es en si mismo una media muestral de observaciones independientes, se tiene que

$$\begin{aligned} \text{Var}(\hat{\gamma}) &= \text{Var}\left[\frac{1}{m}\sum_{i=1}^m g(\theta_i)\right] = \frac{1}{m^2}\text{Var}\left[\sum_{i=1}^m g(\theta_i)\right] \\ &= \frac{m}{m^2}\text{Var}[g(\theta)] = \frac{1}{m}\text{Var}[g(\theta)], \end{aligned}$$

pero $\text{Var}[g(\theta)]$ puede estimarse mediante la varianza muestral de los valores $g(\theta_i)$,

$$\begin{aligned} \text{Var}[g(\theta)] &= E[(g(\theta) - \hat{\gamma})^2] \\ &= \int (g(\theta) - \hat{\gamma})^2 \pi(\theta|x) d\theta \\ &\approx \frac{1}{m} \sum_{i=1}^m (g(\theta_i) - \hat{\gamma})^2, \end{aligned}$$

entonces el error estandar estimado de γ esta dado por

$$\hat{s}e(\hat{\gamma}) = \sqrt{\frac{1}{m(m-1)} \sum_{i=1}^m (g(\theta_i) - \hat{\gamma})^2}. \quad (4.4)$$

La simulación Monte Carlo, proporciona un ejemplo donde la simulación es claramente apropiada. Note que

$$p = P\{a < g(\theta) < b|x\} = \int_a^b g(\theta) d\theta = E\{g(\theta) I_{(a,b)}|x\},$$

luego un estimador de p es simplemente

$$\hat{p} = \frac{\text{número de } \theta_i \in (a, b)}{m} \quad (4.5)$$

donde $\theta_i \in \{\theta_1, \dots, \theta_m | \theta_j\}$ es una simulación de $g(\theta)$. Y esta integral tiene un error estándar dado por (4.4).

4.2. Métodos de simulación Monte Carlo

4.2.1. Antecedentes

En la práctica suceden procesos que evolucionan con el tiempo de una manera aleatoria. Estos son llamados procesos estocásticos.

Definición 4.1 Un **proceso estocástico** es una familia de variables aleatorias $\{X_t\}_{t \in T}$ indexada por un conjunto T que representa el tiempo y para cada tiempo $t_0 \in T$, X_{t_0} es el estado del proceso en el tiempo t_0 . A todo el conjunto de estados posibles se le llama **espacio de estados**.

Definición 4.2 Una **cadena de Markov** es un tipo especial de proceso estocástico, que se describe como sigue. En cualquier instante de tiempo t dado, cuando el estado actual X_t y todos los estados previos X_1, \dots, X_{t-1} del proceso son conocidos, las probabilidades de los estados futuros X_j ($j > t$) dependen solamente del estado actual X_t y no dependen de los estados anteriores X_1, \dots, X_{t-1} . Es decir, una cadena de Markov es un proceso estocástico tal que para $t = 1, 2, \dots$ y para cualquier sucesión posible de estados x_1, x_2, \dots, x_{t+1} ,

$$P(X_{t+1} = x_{t+1} | X_1 = x_1, X_2 = x_2, \dots, X_t = x_t) = P(X_{t+1} = x_{t+1} | X_t = x_t).$$

Definición 4.3 Una cadena de Markov para la cual existe sólo un número finito k de estados posibles $S = \{s_1, \dots, s_k\}$ (donde S es el espacio de estados) y en cualquier instante de tiempo la cadena debe estar en uno de estos k estados, se conoce como **cadena de Markov finita**.

Definición 4.4 La probabilidad condicional $P(X_{t+1} = s_j | X_t = s_i) = p_{ij}$ de que la cadena de Markov esté en el estado s_j en el instante de tiempo $t + 1$ si está en el estado s_i en el instante de tiempo t se llama **probabilidad de transición**.

Definición 4.5 Una cadena de Markov tiene **probabilidades de transición estacionarias** si, para cualquier par de estados s_i y s_j , existe una probabilidad de transición p_{ij} tal que

$$P(X_{t+1} = s_j | X_t = s_i) = p_{ij} \quad t = 1, 2, 3, \dots$$

Se denota con $\mu_j(t)$ a la probabilidad de que la cadena esté en el estado j al tiempo t . Y $\mu(t)$ denota el vector fila del espacio de estados de probabilidades al paso t . Se inicia la cadena por especificación de un vector inicial $\mu(0)$, frecuentemente todas las componentes de $\mu(0)$ son cero excepto una que es 1, correspondiente al inicio del proceso en un estado particular. La probabilidad de que la cadena tenga un valor en el estado s_i al tiempo $t + 1$ está dado por la igualdad de Chapman-Kolmogorov.

Teorema 4.5 (Igualdad de Chapman-Kolmogorov) La suma sobre todas las probabilidades de estar en un estado particular en el paso actual y la probabilidad de transición de que el estado esté en el estado s_i , están relacionadas por la fórmula recursiva

$$\mu_i(t + 1) = \sum_k p_{ki} \mu_k(t).$$

Demostración Para verificar la fórmula se aplica la probabilidad total

$$\begin{aligned} \mu_i(t + 1) &= P(X_{t+1} = s_i) \\ &= \sum_k P(X_{t+1} = s_i | X_t = s_k) P(X_t = s_k) \\ &= \sum_k p_{ki} \mu_k(t). \end{aligned}$$

□

Sucesivas iteraciones de la igualdad de Chapman-Kolmogorov describen la evolución de la cadena, y se puede escribir más compactamente en forma de matriz mediante la igualdad de Chapman-Kolmogorov.

Definición 4.6 La matriz de probabilidad de transición P de una cadena de Markov, se define como la matriz cuya entrada o componente i, j es p_{ij} , la probabilidad de moverse del estado i al estado j , (esto implica que las filas suman uno, $\sum_j p_{ij} = 1$).

La igualdad de Chapman-Kolmogorov se reescribe como:

$$\mu(t+1) = \mu(t)P.$$

Usando la forma matricial, se tiene que:

$$\mu(t) = \mu(t-1)P = (\mu(t-2)P)P = \mu(t-2)P^2,$$

continuando, se tiene que

$$\mu(t) = \mu(0)P^t.$$

Definiendo para el n -ésimo paso la probabilidad de transición $p_{ij}^{(n)}$ como la probabilidad de que en el n -ésimo paso, el proceso esté en el estado j dado que éste inició en el estado i , esto es,

$$p_{ij}^{(n)} = P(X_{t+n} = s_j | X_t = s_i),$$

que es justamente la componente ij -ésima de P^n . La evolución de una cadena es descrita por sus probabilidades de transición, $P(X_{t+1} = s_j | X_t = s_i)$, lo cual puede ser un poco complicado, dado que sus probabilidades dependen de tres cantidades t, j e i . Cuando las probabilidades no dependen de t (sólo de i y j) se tiene la siguiente definición.

Definición 4.7 La cadena $\{X_t\}$ es llamada **homogénea** si

$$P(X_{t+1} = s_j | X_t = s_i) = P(X_1 = s_j | X_0 = s_i).$$

Definición 4.8 Una cadena de Markov se dice que es **irreducible** si existe un número positivo n_{ij} , tal que $p_{ij}^{(n_{ij})} > 0$ para todo i, j , esto es, todos los estados se comunican con los otros estados.

Definición 4.9 El período $d(i)$ del estado $i \in S$ es definido por

$$d(i) = \text{mcd} \{n \geq 0 : p_{ij}^n > 0\},$$

el máximo común divisor (*mcd*) del estado para el cual regresar es posible. Se le llama al estado i periódico si $d(i) > 1$, Y aperiódico si $d(i) = 1$.

Definición 4.10 Una cadena de Markov es **aperiódica** si $\forall i \in S, i$ es aperiódico.

Definición 4.11 Una cadena de Markov se encuentra en una **distribución estacionaria** π^* , cuando el vector de probabilidades es, en particular, un estado independientemente de la condición inicial. La distribución estacionaria satisface

$$\pi^* = \pi^* P.$$

Las condiciones para una distribución estacionaria es que la cadena sea **irreducible** y **aperiódica**. Una condición suficiente para una única distribución estacionaria es que la cadena de Markov sea **reversible**, esto es

$$P(j \rightarrow k)\pi_j^* = P(k \rightarrow j)\pi_k^*$$

que es la condición de reversibilidad. Para entender más el concepto de cadena de Markov observe el siguiente ejemplo.

Ejemplo 4.1 Supóngase que el espacio de estados consta de tres posibilidades: lluvioso, soleado y nublado, y que el clima sigue un proceso de Markov. Esto es, la probabilidad del clima de mañana depende del clima de hoy, y no de cualquier otro día previo. Supóngase además que la probabilidad de transición dado que hoy está lloviendo es:

$$\begin{aligned} P(\text{mañana llueve—hoy llueve}) &= 0.5, \\ P(\text{mañana soleado—hoy llueve}) &= 0.25, \\ P(\text{mañana nublado—hoy llueve}) &= 0.25. \end{aligned}$$

La primera fila de la matriz de transición de probabilidad es (0.5 0.25 0.25). Supóngase que el resto de la matriz de transición está dada por

$$P = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

Esta cadena de Markov es irreducible, pues $p_{ij}^k > 0$ para $k \geq 2$. Además es fácil ver que

$$d(i) = 1 \quad i = 1, 2, 3$$

con lo cual cada estado es aperiódico y por tanto la cadena de Markov es aperiódica.

Si se supone que hoy es soleado. ¿Cuál es la probabilidad de que el clima se encuentre en alguno de los estados dentro de dos días? ¿Y en siete días?. Aquí $\mu(0) = (0 \ 1 \ 0)$, entonces

$$\begin{aligned}\mu(2) &= \mu(0)P^2 \\ &= (0\ 1\ 0) \begin{pmatrix} 0.4375 & 0.1875 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.1875 & 0.4375 \end{pmatrix} \\ &= (0.375\ 0.25\ 0.375)\end{aligned}$$

y

$$\begin{aligned}\mu(7) &= \mu(0)P^7 \\ &= (0\ 1\ 0) \begin{pmatrix} 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \end{pmatrix} \\ &= (0.4\ 0.2\ 0.4)\end{aligned}$$

Ahora si se supone que hoy está lloviendo, entonces $\mu(0) = (1\ 0\ 0)$. La probabilidad de que el clima se encuentre en alguno de los estados está dado por

$$\mu(2) = (0.4375\ 0.1875\ 0.375) \quad y \quad \mu(7) = (0.4\ 0.2\ 0.4)$$

Note que después de un tiempo suficientemente largo, el clima esperado es independiente del valor inicial. En otras palabras la cadena ha alcanzado una distribución estacionaria.

La idea de una cadena de Markov en estado discreto se puede generalizar para un proceso de Markov en estado continuo, teniendo una probabilidad que satisface

$$\int P(x, y)dy = 1,$$

y la igualdad de Chapman-Kolomogrov está dada por

$$\mu_t(y) = \int \mu_{t-1}(x)P(x, y)dy,$$

para equilibrar, la distribución estacionaria satisface

$$\mu^*(y) = \int \mu^*(x)P(x, y)dy.$$

Ahora se analizarán los métodos de simulación Monte Carlo vía cadenas de Markov.

4.2.2. Motivación

Se presenta un grave problema cuando se desea calcular

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta)f(x|\theta)d\theta},$$

pues el denominador no adopta una forma funcional conocida y no se puede hacer la evaluación de forma analítica. Se hace necesario el tratamiento numérico, que se agrava en muchos casos porque la dimensión del espacio paramétrico es mayor que uno, es decir, $\theta = (\theta_1, \dots, \theta_d)$.

Si fuera posible generar directamente muestras independientes de $\pi(\theta|x)$ mediante algún método aleatorio de simulación, conduciría a la obtención de la cantidad a posteriori de interés. Pero en muchos casos no es posible simular directamente muestras independientes para $\pi(\theta|x)$. Sin embargo, puede ser factible simular muestras con algún tipo de dependencia, que converjan (bajo ciertas condiciones de regularidad) a la distribución de interés $\pi(\theta|x)$.

Se presenta un problema al aplicar integración Monte Carlo para obtener muestras de la misma complejidad que la distribución objetivo. Intentar resolver este problema es la raíz de los métodos Monte Carlo vía cadenas de Markov.

Los métodos basados en simulación Monte Carlo mediante cadenas de Markov (MCMC) permiten muestrear la distribución a posteriori, siempre y cuando se conozca la forma funcional de la distribución aunque la constante de normalización sea desconocida, gracias a la construcción de una cadena de Markov cuya distribución estacionaria es precisamente la distribución a posteriori. A la distribución a posteriori se le llama **distribución objetivo**.

La clave de la simulación con cadenas de Markov es crear una cadena de Markov cuya distribución estacionaria es $\pi(\theta|x)$. Se corre la simulación un tiempo suficientemente largo tal que la distribución generada en la iteración actual sea lo suficientemente cerca de la distribución estacionaria. La convergencia hacia tal distribución objetivo la proporciona el siguiente teorema.

Teorema 4.6 Sea $\theta_1, \theta_2, \dots$ una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados Θ y distribución de equilibrio $\pi(\theta|x)$. Entonces conforme $t \rightarrow \infty$

$$\begin{aligned} (i) \quad & \theta_t \xrightarrow{D} \theta, \quad \text{donde } \theta \sim \pi(\theta|x) \\ (ii) \quad & \frac{1}{t} \sum_{i=1}^t g(\theta_i) \rightarrow E[g(\theta)|x]. \end{aligned}$$

Un método de simulación Monte Carlo vía cadenas de Markov, mas utilizado es el algoritmo de Metropolis-Hastings, el cual se analiza a continuación.

4.2.3. Algoritmo de Metropolis-Hastings

El algoritmo de Metropolis-Hastings es un método Monte Carlo vía cadenas de Markov que ha sido ampliamente usado en la física matemática y restauración de imágenes por

muchos años, pero recientemente descubierto por los estadísticos. El método en su forma más simple lo publica Metropolis, el cual se explicará a continuación. Supóngase que la verdadera distribución a posteriori conjunta de un parámetro θ (posiblemente vector de valores) tiene densidad $\pi(\theta|x)$. Se escoge una función auxiliar $q(\theta_{t+1}|\theta_t)$, tal que $q(\cdot|\theta_t)$ es una función de densidad de probabilidad para toda θ_t , además es simétrica, es decir:

$$q(\theta^*|\theta) = q(\theta|\theta^*).$$

La función q es llamada **distribución de densidad propuesta o candidato-generadora**. Entonces, el algoritmo de Metropolis genera una cadena de Markov como sigue:

Entrada: Conocer la forma funcional de $\pi(\theta|x)$.

Dar un valor arbitrario $\theta_0 \in \text{sop}^1(\pi(\theta|x))$, tal que $\pi(\theta_0|x) > 0$.

T tiempo al cual se alcanza la convergencia.

$q(\theta^*|\theta)$ distribución de densidad candidato-generadora.

Salida: $(\theta_0, \theta_1, \dots, \theta_k, \dots)$.

1.- Para $t = 0$ hasta T .

2.- Generar una observación θ^* de $q(\theta^*|\theta_t)$.

3.- Generar una variable $u \sim U(0, 1)$.

4.- Dado el punto candidato θ^* , calcular

$$\alpha(\theta^*, \theta_t) = \min\left(\frac{\pi(\theta^*|x)}{\pi(\theta_t|x)}, 1\right).$$

5.- Si $u \leq \alpha(\theta^*, \theta_t)$, hacer $\theta_{t+1} = \theta^*$,
en caso contrario, hacer $\theta_{t+1} = \theta_t$.

6.- Salida $(\theta_0, \theta_1, \dots, \theta_k, \dots)$.

El algoritmo anterior genera una cadena de Markov $(\theta_0, \theta_1, \dots, \theta_k, \dots)$, donde la probabilidad de transición de θ_t a θ_{t+l} depende sólo de θ_t y no de $\theta_0, \dots, \theta_{t-l}$. Y cuya distribución de transición es

$$P(\theta_{t+1}|\theta_t) = \alpha(\theta_{t+1}, \theta_t)q(\theta_{t+1}|\theta_t).$$

La cadena se aproxima a una distribución estacionaria y las muestras del vector $(\theta_{k+1}, \dots, \theta_{k+n})$ son muestras de la densidad objetivo $\pi(\theta|x)$.

Note que la densidad a posteriori conjunta π es necesaria sólo en una constante de proporcionalidad. Dado que en aplicaciones bayesianas

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta),$$

¹El soporte (*sop*) de una distribución es el conjunto cerrado más pequeño cuyo complemento tiene probabilidad cero.

es una forma típicamente disponible.

Hastings generaliza el algoritmo de Metropolis usando una función arbitraria de probabilidad de transición (candidato-generadora) y aceptando la probabilidad de un punto candidato como

$$\alpha(\theta^*, \theta_t) = \min \left(\frac{\pi(\theta^*|x)q(\theta_t|\theta^*)}{\pi(\theta_t|x)q(\theta^*|\theta_t)}, 1 \right),$$

este es el algoritmo de Metropolis-Hastings. Si la distribución propuesta es simétrica, se recupera el algoritmo de Metropolis original.

La idea de una distribución candidato en el algoritmo Metropolis-Hastings es muestrear puntos candidatos de la distribución objetivo, esto es

$$q(\theta^*|\theta) = \pi(\theta^*|x) \quad \forall \theta.$$

Pero como las simulaciones se aplican a problemas que no se puede muestrear directamente de $\pi(\theta^*|x)$, entonces una buena distribución candidato debe tener las siguientes propiedades.

- Para cualquier θ , es fácil muestrear de $q(\theta^*|\theta)$.
- Es fácil calcular el cociente

$$\frac{\pi(\theta^*|x)q(\theta_t|\theta^*)}{\pi(\theta_t|x)q(\theta^*|\theta_t)}$$

- En el algoritmo de Metropolis-Hastings se debe cumplir que $\text{sop}(\pi(\theta|x)) \subset \bigcup_{\theta \in \text{sop}(\pi(\theta|x))} \text{sop}(q(\cdot|\theta))$.

Ejemplo 4.2 Supóngase que la distribución a posteriori es una $Ga(\theta|\alpha, \beta)$, es decir

$$\pi(\theta|\vec{x}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta)$$

y que

$$q(\theta|\theta^*) = Ex(\theta|\theta^*) = \theta^* \exp(-\theta^*\theta),$$

entonces se tiene que

$$\begin{aligned} \frac{\pi(\theta^*|\vec{x})q(\theta_t|\theta^*)}{\pi(\theta_t|\vec{x})q(\theta^*|\theta_t)} &= \frac{\left[\frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{*\alpha-1} \exp(-\beta\theta^*) \right] \left[\theta^* \exp(-\theta^*\theta_t) \right]}{\left[\frac{\beta^\alpha}{\Gamma(\alpha)} \theta_t^{\alpha-1} \exp(-\beta\theta_t) \right] \left[\theta_t \exp(-\theta_t\theta^*) \right]} \\ &= \left(\frac{\theta^*}{\theta_t} \right)^\alpha \exp[-\beta(\theta^* - \theta_t)]. \end{aligned}$$

Ahora según el algoritmo de Metropolis-Hastings, dado que la función objetivo no es simétrica entonces

$$\begin{aligned}\alpha(\theta^*, \theta_t) &= \min \left(\frac{\pi(\theta^*|\vec{x})q(\theta_t|\theta^*)}{\pi(\theta_t|\vec{x})q(\theta^*|\theta_t)}, 1 \right) \\ &= \min \left(\left(\frac{\theta^*}{\theta_t} \right)^\alpha \exp[-\beta(\theta^* - \theta_t)], 1 \right).\end{aligned}$$

Así el algoritmo de Metropolis-Hastings queda :

Entrada: θ_0 es un número aleatorio en $(0,1)$.

- 1.- Para $t = 0$ hasta T .
- 2.- Generar una observación θ^* de una $Ex(\theta_t)$.
- 3.- Generar una variable $u \sim U(0, 1)$.
- 4.- Dado el punto candidato θ^* , calcular

$$\alpha(\theta^*, \theta_t) = \min \left(\left(\frac{\theta^*}{\theta_t} \right)^\alpha \exp[-\beta(\theta^* - \theta_t)], 1 \right).$$

- 5.- Si $u \leq \alpha(\theta^*, \theta_t)$, hacer $\theta_{t+1} = \theta^*$,
en caso contrario, hacer $\theta_{t+1} = \theta_t$.

- 6.- Salida $(\theta_0, \theta_1, \dots, \theta_T)$.

Ejemplo 4.3 Considérese la distribución a posteriori como una $Be(\theta|\alpha, \beta)$, es decir

$$\pi(\theta|\vec{x}) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

y que

$$q(\theta|\theta^*) = U(0, 1).$$

La cual es claramente simétrica, entonces se tiene que

$$\begin{aligned}\alpha(\theta^*, \theta_t) &= \min \left(\frac{\pi(\theta^*|\vec{x})}{\pi(\theta_t|\vec{x})}, 1 \right) \\ &= \min \left(\left(\frac{\theta^*}{\theta_t} \right)^{\alpha-1} \left(\frac{1 - \theta^*}{1 - \theta_t} \right)^{\beta-1}, 1 \right).\end{aligned}$$

Así el algoritmo de Metropolis-Hastings queda :

Entrada: θ_0 es un número aleatorio en $(0,1)$.

- 1.- Para $t = 0$ hasta T .
- 2.- Generar una observación $\theta^* \sim U(0, 1)$.

3.- Generar una variable $u \sim U(0, 1)$.

4.- Dado el punto candidato θ^* , calcular

$$\alpha(\theta^*, \theta_t) = \min \left(\left(\frac{\theta^*}{\theta_t} \right)^{\alpha-1} \left(\frac{1-\theta^*}{1-\theta_t} \right)^{\beta-1}, 1 \right).$$

5.- Si $u \leq \alpha(\theta^*, \theta_t)$, hacer $\theta_{t+1} = \theta^*$,
en caso contrario, hacer $\theta_{t+1} = \theta_t$.

6.- Salida $(\theta_0, \theta_1, \dots, \theta_T)$.

Capítulo 5

Software

La función principal que tiene ANESBA es la de realizar inferencia bayesiana, aunque también se le ha agregado un apartado para el análisis descriptivo de los datos, específicamente las funciones que realiza son las siguientes:

- Inferencia bayesiana
 - Estimación
 - Puntual
 - Por intervalos
 - Contraste de hipótesis
- Estadística descriptiva
 - Diagramas de caja
 - Histogramas
 - Cálculo de cuantiles
 - Cálculo de la media y la varianza muestral
- Análisis de sensibilidad

El software está realizado en Visual Basic Net 2005, esto con la intención de brindarle al usuario un ambiente de trabajo visual y totalmente amigable.

5.1. Diseño e Implementación

Las funciones principales que realiza el software se pueden presentar de forma muy general mediante un diagrama de carpetas (5.1).

A su vez mediante el uso de un diagrama de flujo (5.2) se muestra como se interrelacionan las funciones principales del software.

Para un mayor entendimiento del diagrama de flujo a continuación se explicará en que consiste cada una de las funciones en dicho diagrama.

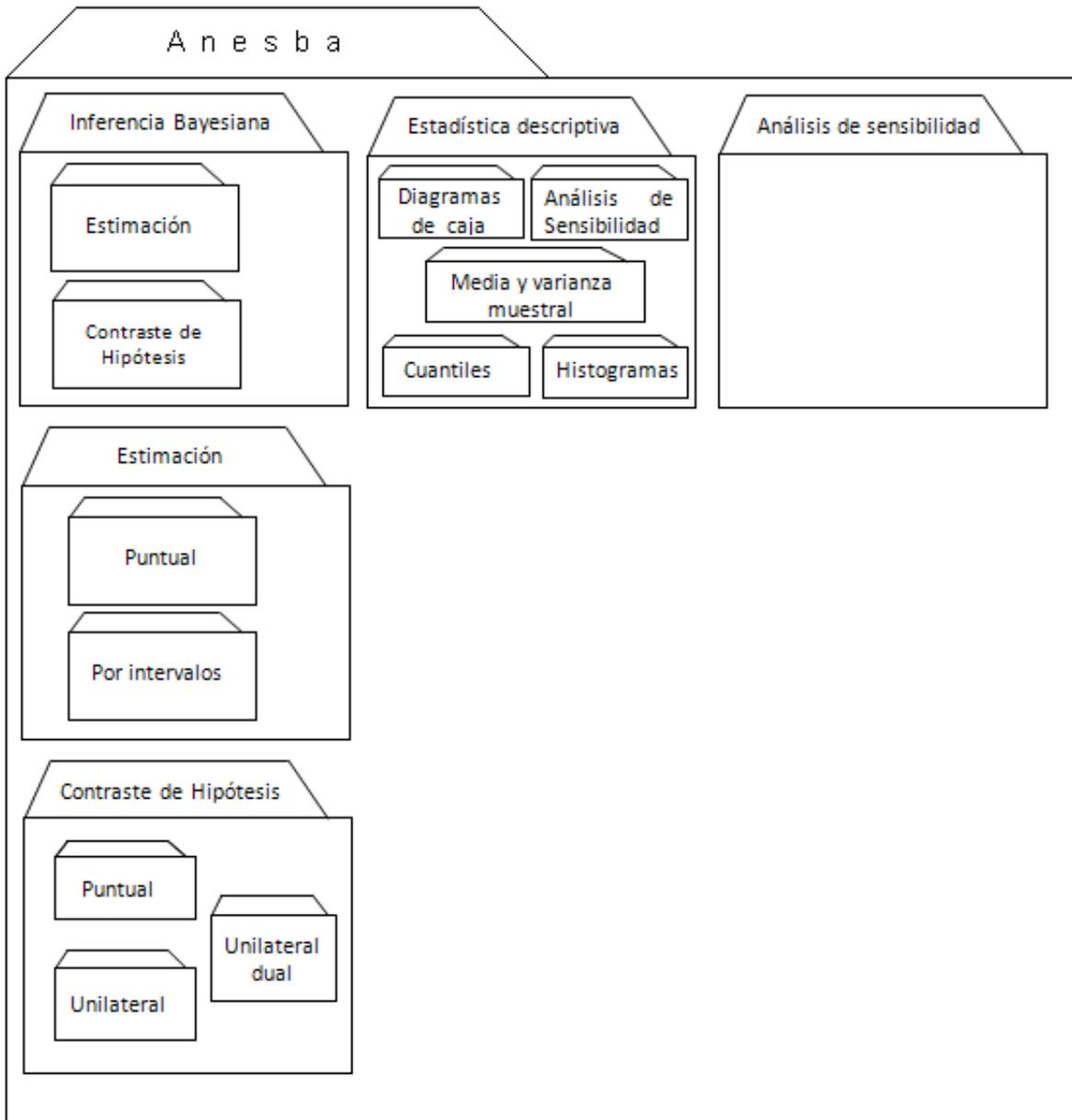


Figura 5.1: Diagrama de carpetas.

Abrir datos

Esta función le permite al usuario ingresar una muestra aleatoria desde un archivo de texto¹, los datos ingresados serán cargados a la memoria del programa.

¹Véase el manual anexo a esta tesis.

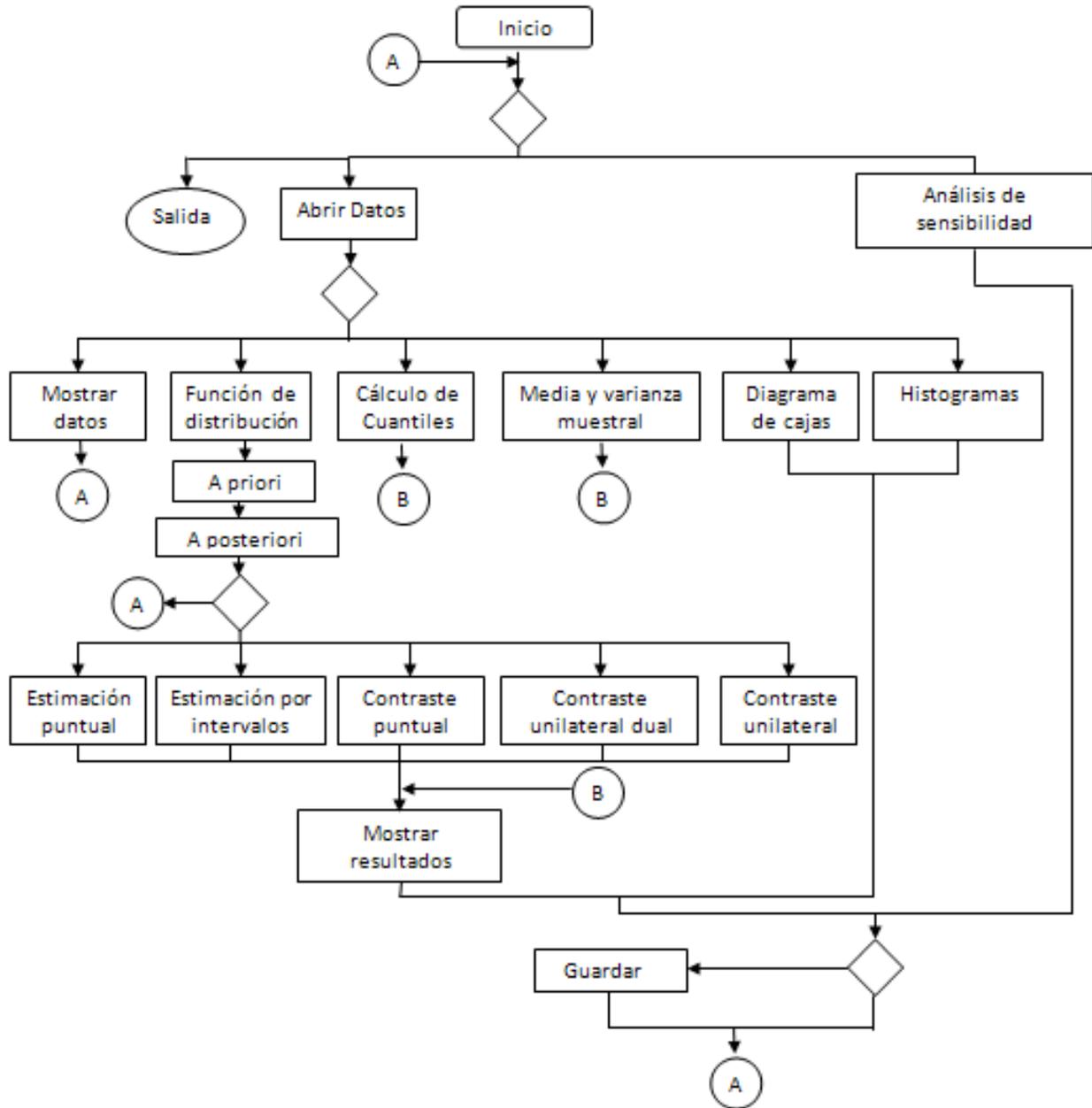


Figura 5.2: Diagrama de flujo.

Mostrar datos

Esta función le permite al usuario visualizar la muestra aleatoria que esté cargada en la memoria del programa.

Mostrar resultados

Esta función muestra los resultados obtenidos después de realizar cualquier función que sea parte de la inferencia bayesiana o del análisis descriptivo de los datos.

Guardar

Esta función permite al usuario guardar los resultados obtenidos en un archivo de texto o una imagen (como es el caso de los diagramas de caja, histogramas y el análisis de sensibilidad), después de realizar cualquier función que sea parte de la inferencia bayesiana o del análisis descriptivo de los datos.

El resto de las funciones que intervienen en el diagrama de flujo, serán explicadas en las siguientes secciones.

En cada sección se incluirá un poco de teoría e incluso algunos algoritmos desarrollados, cuando sea necesario para mostrar la manera en que cada función realiza su tarea.

5.1.1. Funciones de distribución

Las funciones de distribución que ANESBA le permitirá al usuario asignarle a los datos serán las siguientes:

1. Bernoulli (con parámetro desconocido θ)
2. Poisson (con parámetro desconocido λ)
3. Exponencial (con parámetro desconocido θ)
4. Normal (con media conocida y parámetro desconocido λ)
5. Normal (con varianza conocida y parámetro desconocido θ)

5.1.2. Funciones de distribución a priori

Para cada función de distribución el usuario podrá asignarle al parámetro desconocido alguno de los siguientes tipos de función de distribución a priori:

1. Distribución a priori conjugada
2. Distribución a priori no informativa

5.1.3. Función de distribución a posteriori

Después de que el usuario ingrese la distribución a priori del parámetro desconocido, esta función se encarga de obtener la densidad a posteriori del parámetro desconocido.

Dada una observación de una muestra aleatoria, $\vec{X} = (X_1, X_2, \dots, X_n)$ donde $X_i \sim f(x_i|\theta)$ (función de densidad) y una distribución a priori $\pi(\theta)$ (conjugada o por el método de Jeffreys), a continuación se presenta el cálculo de la distribución a posteriori. Para realizar el cálculo de la a posteriori se usará el lema (2.7), mediante un estimador

suficiente² $S(\vec{X}) = S$ del parámetro y su función de distribución $g(s|\theta)$.

A continuación se procederá a calcular la función de distribución a priori por método de Jeffeys y después la a posteriori, para las funciones de densidad de distribución usadas por el software.

Modelo Bernoulli/Beta

Sea

$$\begin{aligned} f(x_i|\theta) &= \theta^{x_i}(1-\theta)^{1-x_i} \\ S(\vec{X}) &= T = \sum_{i=1}^n X_i \\ g(t|\theta) &= Bi(t|\theta, n), \end{aligned}$$

por (2.1)

$$\begin{aligned} I(\theta) &= -E_{x|\theta} \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right] = - \sum_{x=0}^1 f(x|\theta) \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \\ &= - \sum_{x=0}^1 \theta^x (1-\theta)^{1-x} \frac{\partial}{\partial \theta} \left[\frac{x}{\theta} - \frac{(1-x)}{1-\theta} \right] \\ &= - \sum_{x=0}^1 \theta^x (1-\theta)^{1-x} \left[-\frac{x}{\theta^2} - \frac{(1-x)}{(1-\theta)^2} \right] \\ &= \sum_{x=0}^1 \theta^x (1-\theta)^{1-x} \left[\frac{x}{\theta^2} + \frac{(1-x)}{(1-\theta)^2} \right] \\ &= (1-\theta)(1-\theta)^{-2} + \theta(\theta^{-2}) = (1-\theta)^{-1} + \theta^{-1} \\ &= \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{\theta + 1 - \theta}{\theta(1-\theta)} \\ &= \theta^{-1}(1-\theta)^{-1}. \end{aligned}$$

Así, utilizando (2.2)

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}} = \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \propto Be\left(\theta \mid \frac{1}{2}, \frac{1}{2}\right).$$

²La demostración de que cada estadístico que se tomará a continuación es suficiente puede verse en [11].

Ahora se calculará la a posteriori usando el lema (2.7)

$$\begin{aligned}
 \pi(\theta|\vec{x}) &\propto \pi(\theta)g(t|\theta) \\
 &= Be\left(\theta\left|\frac{1}{2}, \frac{1}{2}\right.\right) Bi(t|\theta, n) \\
 &= \left[\frac{\Gamma(1)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}\right] \left[\binom{n}{t}\theta^t(1-\theta)^{n-t}\right] \\
 &\propto \theta^{t-\frac{1}{2}}(1-\theta)^{n-t-\frac{1}{2}} \\
 &= \theta^{(t+\frac{1}{2})-1}(1-\theta)^{(n-t+\frac{1}{2})-1} \\
 &\propto Be\left(\theta\left|t+\frac{1}{2}, n-t+\frac{1}{2}\right.\right). \\
 \\
 \therefore \pi(\theta|\vec{x}) &= Be\left(\theta\left|t+\frac{1}{2}, n-t+\frac{1}{2}\right.\right).
 \end{aligned}$$

Modelo Poisson/Gamma

Dada

$$\begin{aligned}
 f(x_i|\lambda) &= \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \\
 S(\vec{X}) &= T = \sum_{i=1}^n X_i \\
 g(t|\theta) &= Pn(t|n\lambda),
 \end{aligned}$$

por la ecuación (2.3)

$$\pi(\lambda) \propto \lambda^{-\frac{1}{2}},$$

luego usando el lema (2.7) la a posteriori queda como

$$\begin{aligned}
 \pi(\lambda|\vec{x}) &\propto \pi(\lambda)g(t|\lambda) \propto \left(\lambda^{-\frac{1}{2}}\right) \left[\frac{\exp(-n\lambda)}{t!}(n\lambda)^t\right] \propto \lambda^{t-\frac{1}{2}} \exp(-n\lambda) \\
 &= \lambda^{(t+\frac{1}{2})-1} \exp(-n\lambda) \\
 &\propto Ga\left(\lambda\left|t+\frac{1}{2}, n\right.\right).
 \end{aligned}$$

$$\therefore \pi(\lambda|\vec{x}) = Ga\left(\lambda\left|t+\frac{1}{2}, n\right.\right).$$

Modelo Exponencial/Gamma

Supóngase que

$$\begin{aligned} f(x_i|\theta) &= \theta \exp(-\theta x_i) \\ S(\vec{X}) &= T = \sum_{i=1}^n X_i \\ g(t|\theta) &= Ga(t|n, \theta), \end{aligned}$$

así

$$\begin{aligned} I(\theta) &= -E_{x|\theta} \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right] \\ &= - \int_0^\infty f(x|\theta) \frac{\partial^2}{\partial \theta^2} \left\{ \log [f(x|\lambda)] \right\} dx \\ &= \int_0^\infty \theta \exp(-\theta x) \left\{ - \frac{\partial^2}{\partial \theta^2} \log[\theta \exp(-\theta x)] \right\} dx \\ &= - \int_0^\infty \theta \exp(-\theta x) \frac{\partial^2}{\partial \theta^2} (\log \theta - \theta x) dx \\ &= \int_0^\infty \frac{\theta \exp(-\theta x)}{\theta^2} dx \\ &= \theta^{-1} \int_0^\infty \exp(-\theta x) dx = \theta^{-1} \theta^{-1} \\ &= \theta^{-2}, \end{aligned}$$

entonces por la ecuación (2.2)

$$\pi(\theta) \propto [I(\lambda)]^{\frac{1}{2}} = \theta^{-1},$$

luego usando el lema (2.7)

$$\begin{aligned} \pi(\theta|\vec{x}) &\propto \pi(\theta) Ga(t|n, \theta) \\ &= \theta^{-1} \left[\frac{\theta^n}{\Gamma(n)} t^{n-1} \exp(-\theta t) \right] \\ &\propto \theta^{n-1} \exp(-\theta t) \\ &\propto Ga(\theta|n, t). \end{aligned}$$

$$\therefore \pi(\theta|\vec{x}) = Ga(\theta|n, t).$$

Modelo Normal (varianza conocida)/Normal

Sea

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[- \frac{1}{2} \left(\frac{x - \theta}{\sigma} \right)^2 \right]$$

nótese que

$$\begin{aligned}\log f(x|\theta) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2 \\ \frac{\partial \log f(x|\theta)}{\partial \theta} &= \frac{1}{\sigma}\left(\frac{x-\theta}{\sigma}\right) \\ \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} &= -\frac{1}{\sigma^2}.\end{aligned}$$

$$\therefore I(\theta) = -E_{x|\theta}\left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right] = \frac{1}{\sigma^2},$$

luego por (2.2) se tiene que

$$\begin{aligned}\pi(\theta) &\propto [I(\theta)]^{\frac{1}{2}} = \frac{1}{\sigma} \\ &\propto 1.\end{aligned}$$

Así

$$\begin{aligned}\pi(\theta|\vec{x}) &\propto f(\vec{x}|\theta)\pi(\theta) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\theta \sum_{i=1}^n x_i + n\theta^2\right)\right] \quad \text{si} \quad T = \sum_{i=1}^n X_i \\ &\propto \exp\left[-\frac{1}{2\sigma^2}(-2\theta + n\theta^2)\right] \\ &= \exp\left[-\frac{1}{2\sigma^2} \left(\sqrt{n}\theta - \frac{t}{\sqrt{n}}\right)^2 - \frac{t^2}{n}\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2} \left(\sqrt{n}\theta - \frac{t}{\sqrt{n}}\right)^2\right] \\ &= \exp\left[-\frac{1}{\frac{2\sigma^2}{n}} \left(\theta - \frac{t}{n}\right)^2\right] \\ &\propto N\left(\theta \left| \frac{t}{n}, \frac{\sigma^2}{n}\right.\right).\end{aligned}$$

$$\therefore \pi(\theta|\vec{x}) = N\left(\theta \left| \frac{t}{n}, \frac{\sigma^2}{n}\right.\right).$$

Modelo Normal (media conocida)/Gamma

Dada

$$f(x_i|\sigma) = N(x_i|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(x-\theta)^2}{\sigma^2}\right], \theta \text{ conocida.}$$

La distribución a priori de Jeffreys es

$$\begin{aligned}
 I(\sigma) &= -E_{x|\sigma} \left[\frac{\partial^2 \log f(x|\sigma)}{\partial \sigma^2} \right] \\
 \log f(x|\sigma) &= \log \frac{1}{(2\pi)^{\frac{1}{2}}} - \log \sigma - \frac{t}{2\sigma^2}, \quad \text{donde, } T = (X - \theta)^2, \\
 \frac{\partial \log f(x|\sigma)}{\partial \sigma} &= -\sigma^{-1} + t\sigma^{-3} \\
 \frac{\partial^2 \log f(x|\sigma)}{\partial \sigma^2} &= \frac{1}{\sigma^2} - \frac{3t}{\sigma^4} \\
 -E_{x|\sigma} \left(\frac{1}{\sigma^2} - \frac{3t}{\sigma^4} \right) &= -E_{x|\sigma} \left(\frac{1}{\sigma^2} \right) + \frac{3}{\sigma^4} (x - \theta)^2 \\
 &= -\sigma^{-2} + 3\sigma^{-4} E[(x - \theta)^2] \\
 &= -\sigma^{-2} + 3\sigma^{-4} \text{Var}(x) \\
 &= -\sigma^{-2} + 3\sigma^{-4} \sigma^2 \\
 &= -\sigma^{-2} + 3\sigma^{-2} \\
 &= 2\sigma^{-2}.
 \end{aligned}$$

Entonces la a priori queda como

$$\pi(\sigma) \propto I(\sigma)^{\frac{1}{2}} \propto \sigma^{-1},$$

por otro lado

$$\begin{aligned}
 \pi(\sigma|\vec{x}) &\propto f(\vec{x}|\sigma)\pi(\sigma)\sigma^{-1} \quad \text{si se toma } T = \sum_{i=1}^n (X_i - \theta)^2 \\
 &\propto \frac{1}{\sigma^n} \exp \left(-\frac{t}{2\sigma^2} \right) \\
 &= \sigma^{n-1} \exp \left[-\sigma^{-2} \left(\frac{t}{2} \right) \right].
 \end{aligned}$$

Nótese que no es distinguible el tipo de distribución que se tiene, para evitar esto se utiliza la siguiente forma equivalente de la función de densidad normal

$$f(x_i|\lambda) = N(x_i|\theta, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \exp \left[-\frac{1}{2} \lambda (x_i - \theta)^2 \right],$$

donde λ es el parámetro desconocido, así, se obtiene

$$\begin{aligned}
 I(\lambda) &= -E_{x|\lambda} \left[\frac{\partial^2 \log f(x|\lambda)}{\partial \lambda^2} \right] \\
 \log f(x|\lambda) &= \log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \log \lambda - \frac{1}{2} \lambda (x - \theta)^2 \\
 \frac{\partial \log f(x|\lambda)}{\partial \lambda} &= \left(\frac{1}{2} \right) \left(\frac{1}{\lambda} \right) - \frac{(x - \mu)^2}{2} \\
 \frac{\partial^2 \log f(x|\lambda)}{\partial \lambda^2} &= -\frac{1}{2\lambda^2} \\
 I(\lambda) &= -E_{x|\lambda} \left(-\frac{1}{2\lambda^2} \right) = \frac{1}{2\lambda^2},
 \end{aligned}$$

entonces usando (2.2)

$$\pi(\lambda) \propto \sqrt{\frac{1}{2\lambda^2}} \propto \frac{1}{\lambda} = \lambda^{-1}.$$

Además se tiene que $T = \sum_{i=1}^n (X_i - \theta)^2$ es estimador suficiente de λ con $T \sim Ga(t|\frac{n}{2}, \frac{\lambda}{2})$ por la ecuación (2.7) se tiene que

$$\begin{aligned}
 \pi(\lambda|\vec{x}) &\propto Ga\left(t|\frac{n}{2}, \frac{\lambda}{2}\right) \pi(\lambda) = \left\{ \frac{\left(\frac{\lambda}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} t^{\frac{n}{2}-1} \exp\left[-\left(\frac{\lambda}{2}\right)t\right] \right\} \lambda^{-1} \\
 &\propto \lambda^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}t\lambda\right) \\
 &\propto Ga\left(\lambda\left|\frac{n}{2}, \frac{t}{2}\right.\right).
 \end{aligned}$$

$$\therefore \pi(\lambda|\vec{x}) = Ga\left(\lambda\left|\frac{n}{2}, \frac{t}{2}\right.\right).$$

A continuación se procederá a calcular la función de distribución a posteriori, para las funciones de densidad de distribución antes mencionadas. El cálculo de la distribución a posteriori se realiza mediante el lema (2.7).

Modelo Bernoulli/Beta

Sea

$$\begin{aligned}
 f(x_i|\theta) &= Br(x_i|\theta) \\
 \pi(\theta) &= Be(\theta|\alpha, \beta) \\
 S(\vec{X}) &= T = \sum_{i=1}^n X_i \\
 g(t|\theta) &= Bi(t|\theta, n).
 \end{aligned}$$

Entonces por el lema (2.1)

$$\begin{aligned}
 \pi(\theta|\vec{x}) &\propto Be(\theta|\alpha, \beta)Bi(t|\theta, n) \\
 &= \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \right] \left[\binom{n}{t} \theta^t (1-\theta)^{n-t} \right] \\
 &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \theta^t (1-\theta)^{n-t} \\
 &= \theta^{(\alpha+t)-1} (1-\theta)^{(\beta+n-t)-1} \\
 &\propto Be\left(\theta \middle| \alpha + t, \beta + n - t\right). \\
 \\
 \therefore \pi(\theta|\vec{x}) &= Be\left(\theta \middle| \alpha + t, \beta + n - t\right).
 \end{aligned}$$

Modelo Poisson/Gamma

Dada

$$\begin{aligned}
 f(x_i|\lambda) &= Pn(x_i|\lambda) \\
 \pi(\lambda) &= Ga(\lambda|\alpha, \beta) \\
 S(\vec{X}) &= T = \sum_{i=1}^n X_i \\
 g(t|\lambda) &= Pn(t|n\lambda).
 \end{aligned}$$

Entonces

$$\begin{aligned}
 \pi(\lambda|\vec{x}) &\propto Ga(\lambda|\alpha, \beta)Pn(t|n\lambda) \\
 &= \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \right] \left[\frac{(n\lambda)^t}{t!} \exp(-n\lambda) \right] \\
 &\propto \lambda^{\alpha-1} \exp(-\beta\lambda) \lambda^t \exp(-n\lambda) \\
 &= \lambda^{(\alpha+t)-1} \exp[-(\beta+n)\lambda] \propto Ga\left(\lambda \middle| t + \alpha, \beta + n\right). \\
 \\
 \therefore \pi(\lambda|\vec{x}) &= Ga\left(\lambda \middle| t + \alpha, \beta + n\right).
 \end{aligned}$$

Modelo Exponencial/Gamma

Supóngase que

$$\begin{aligned}
 f(x_i|\theta) &= Ex(x_i|\theta) \\
 \pi(\theta) &= Ga(\theta|\alpha, \beta) \\
 S(\vec{X}) &= T = \sum_{i=1}^n X_i \\
 g(t|\theta) &= Ga(t|n, \theta).
 \end{aligned}$$

Luego por (2.7)

$$\begin{aligned}
 \pi(\theta|\vec{x}) &\propto Ga(\theta|\alpha, \beta)Ga(t|n, \theta) \\
 &= \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \right] \left[\frac{\theta^n}{\Gamma(n)} t^{n-1} \exp(-\theta t) \right] \\
 &\propto \theta^{\alpha-1} \exp(-\beta\theta) \theta^n \exp(-\theta t) \\
 &= \theta^{(\alpha+n)-1} \exp[-(\beta+t)\theta] \\
 &\propto Ga\left(\theta \middle| \alpha+n, \beta+t\right).
 \end{aligned}$$

$$\therefore \pi(\theta|\vec{x}) = Ga\left(\theta \middle| \alpha+n, \beta+t\right).$$

Modelo Normal (varianza conocida)/Normal

Si $X \sim N(\theta, \sigma^2)$, donde σ es desconocido y $\pi(\theta)$ es una $N(\mu, \tau^2)$. Entonces usando la ecuación (2.8) se tiene que

$$\begin{aligned}
 \pi(\theta|\vec{x}) &= N\left(\theta \middle| \frac{(\sigma^2/n)\mu + \tau^2\bar{x}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}\right) \\
 &= N\left(\theta \middle| \frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).
 \end{aligned}$$

$$\therefore \pi(\theta|\vec{x}) = N\left(\theta \middle| \frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

Modelo Normal (media conocida)/Normal

Sea

$$\begin{aligned}
 f(x_i|\lambda) &= N(x_i|\theta, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(x_i - \theta)^2\right] \\
 \pi(\lambda) &= Ga(\lambda|\alpha, \beta) \\
 S(\vec{X}) &= T = \sum_{i=1}^n (X_i - \theta)^2 \\
 g(t|\lambda) &= Ga\left(t \middle| \frac{n}{2}, \frac{\lambda}{2}\right).
 \end{aligned}$$

Entonces

$$\begin{aligned}
 \pi(\lambda|\vec{x}) &\propto \pi(\lambda)g(s|\lambda) \\
 &= Ga\left(\lambda\middle|\alpha, \beta\right)Ga\left(t\middle|\frac{n}{2}, \frac{\lambda}{2}\right) \\
 &= \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)\right]\left[\left(\frac{\lambda}{2}\right)^{\frac{n}{2}}\frac{1}{\Gamma(\frac{n}{2})}t^{\frac{n}{2}-1}\exp\left(-\frac{\lambda}{2}t\right)\right] \\
 &\propto \lambda^{\alpha-1}\exp(-\beta\lambda)\lambda^{\frac{n}{2}}\exp\left(-\frac{\lambda}{2}t\right) \\
 &= \lambda^{(\alpha+\frac{n}{2})-1}\exp\left[-\lambda\left(\beta+\frac{t}{2}\right)\right] \\
 &\propto Ga\left(\lambda\middle|\alpha+\frac{n}{2}, \beta+\frac{t}{2}\right). \\
 \\
 \therefore \pi(\lambda|\vec{x}) &= Ga\left(\lambda\middle|\alpha+\frac{n}{2}, \beta+\frac{t}{2}\right).
 \end{aligned}$$

5.1.4. Estimación puntual

Una vez calculada la distribución a posteriori del parámetro desconocido, esta función se encarga de encontrar la moda o estimación puntual de dicho parámetro. Las funciones a posteriori que el software utilizará son la gamma, la beta y la normal. A continuación se presenta el procedimiento analítico para encontrar tal estimación.

Distribución a posteriori gamma

Sea $\pi(\theta|\vec{x}) = Ga(\theta|\alpha, \beta)$, $\theta > 0$, para encontrar $\hat{\theta}$, la moda de la distribución, es necesario maximizar $\pi(\theta|\vec{x})$. A continuación se desarrollan las condiciones necesarias para encontrar el punto $x_0 > 0$ donde la función $Ga(x|\alpha, \beta)$ alcanza su máximo (es claro que el estimador es $\hat{\theta} = x_0$).

Se deriva $\pi(\theta|\vec{x})$ con respecto a x , se iguala a cero la derivada y se encuentre la solución, x_0 , a esa ecuación.

$$\begin{aligned}
 \frac{dGa(x|\alpha, \beta)}{dx} &= \frac{d}{dx}[cx^{\alpha-1}\exp(-\beta x)] \quad \text{donde } c = \frac{\beta^\alpha}{\Gamma(\alpha)} \\
 &= c\frac{d}{dx}[x^{\alpha-1}\exp(-\beta x)] \\
 &= c[(\alpha-1)x^{\alpha-2}\exp(-\beta x) - \beta x^{\alpha-1}\exp(-\beta x)] \\
 &= 0.
 \end{aligned}$$

Entonces

$$\begin{aligned}
 (\alpha-1)x_0^{\alpha-2}\exp(-\beta x_0) &= \beta x_0^{\alpha-1}\exp(-\beta x_0) \\
 \frac{\alpha-1}{\beta} &= x_0 \quad \text{el cual es un punto crítico.}
 \end{aligned}$$

Recuérdese que $x_0 = \frac{\alpha-1}{\beta} > 0$ luego necesariamente se tendrá que $\alpha > 1$ de lo contrario no existe estimador.

Supóngase $\alpha > 1$

Para saber si x_0 maximiza a $Ga(x|\alpha, \beta)$, aplicamos el criterio de la segunda derivada, es decir ver que $\frac{d^2}{dx^2}Ga(x|\alpha, \beta)$ sea negativa en x_0 .

$$\begin{aligned}
 \frac{d^2}{dx^2}Ga(x|\alpha, \beta) &= c \frac{d}{dx} \left[(\alpha - 1)x^{\alpha-2} \exp(-\beta x) - \beta x^{\alpha-1} \exp(-\beta x) \right] \\
 &= c \left\{ (\alpha - 1)[(\alpha - 2)x^{\alpha-3} \exp(-\beta x) - \beta x^{\alpha-2} \exp(-\beta x)] \right. \\
 &\quad \left. - \beta[(\alpha - 1)x^{\alpha-2} \exp(-\beta x) - \beta x^{\alpha-1} \exp(-\beta x)] \right\} \\
 &= c \exp(-\beta x) \left[(\alpha - 1)(\alpha - 2)x^{\alpha-3} - \beta(\alpha - 1)x^{\alpha-2} \right. \\
 &\quad \left. - \beta(\alpha - 1)x^{\alpha-2} + \beta^2 x^{\alpha-1} \right] \\
 &= c \exp(-\beta x) \left[(\alpha - 1)(\alpha - 2)x^{\alpha-3} - 2\beta(\alpha - 1)x^{\alpha-2} + \beta^2 x^{\alpha-1} \right] \\
 &= c x^{\alpha-1} \exp(-\beta x) \left[(\alpha - 1)(\alpha - 2)x^{-2} - 2\beta(\alpha - 1)x^{-1} + \beta^2 \right].
 \end{aligned}$$

Evaluando en x_0

$$\begin{aligned}
 \left. \frac{d^2}{dx^2}Ga(x|\alpha, \beta) \right|_{x_0} &= c x_0^{\alpha-1} \exp(-\beta x) \left[\frac{(\alpha - 1)(\alpha - 2)\beta^2}{(\alpha - 1)^2} - \frac{2\beta(\alpha - 1)\beta}{(\alpha - 1)} + \beta^2 \right] \\
 &= c x_0^{\alpha-1} \exp(-\beta x) \beta^2 \left[\frac{(\alpha - 2)}{(\alpha - 1)} - 1 \right],
 \end{aligned}$$

entonces

$$Ga(x|\alpha, \beta) \text{ tiene un máximo en } x_0 \Leftrightarrow \frac{\alpha - 2}{\alpha - 1} - 1 < 0 \quad \Leftrightarrow \quad \alpha - 2 < \alpha - 1 \Leftrightarrow \quad 1 < 2.$$

Luego se puede afirmar que:

- Si $\alpha > 1$

$$x_0 = \hat{\theta} = \frac{\alpha - 1}{\beta} \text{ siempre maximiza } Ga(x|\alpha, \beta)$$

- Si $\alpha \leq 1$

no existe estimador.

Distribución a posteriori beta

Análogamente para encontrar el estimador puntual $\hat{\theta}$ de la función beta hay que encontrar el valor x_0 que maximiza a

$$Be(x|\alpha, \beta) = cx^{\alpha-1}(1-x)^{\beta-1} \quad x \in (0, 1) \quad c = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad \alpha, \beta > 0$$

y claramente se tendrá que $\hat{\theta} = x_0$.

Igual que antes

$$\frac{d}{dx}Be(x|\alpha, \beta) = c \left[(\alpha - 1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta - 1)x^{\alpha-1}(1-x)^{\beta-2} \right] = 0,$$

luego

$$x^{\alpha-1}(1-x)^{\beta-1} \left[(\alpha - 1)x^{-1} - (\beta - 1)(1-x)^{-1} \right] = 0,$$

como $x \in (0, 1)$ entonces

$$(\alpha - 1)x_0^{-1} - (\beta - 1)(1 - x_0)^{-1} = 0.$$

Luego se tiene que

$$\frac{(\alpha - 1)}{x_0} = \frac{(\beta - 1)}{(1 - x_0)}. \quad (5.1)$$

Aquí se presentan dos casos:

- si $\alpha + \beta \neq 2$.
De la ecuación (5.1)

$$x_0 = \frac{\alpha - 1}{\beta - 2 + \alpha} = \frac{1 - \alpha}{2 - \alpha - \beta}.$$

Como $x_0 \in (0, 1)$ entonces

$$0 < \frac{1 - \alpha}{2 - \alpha - \beta} < 1. \quad (5.2)$$

Nótese que

$$Si \quad \alpha + \beta > 2 \quad \Rightarrow \quad \alpha > 1 \quad y \quad \beta > 1. \quad (5.3)$$

Esto es fácil de ver pues si $\alpha + \beta > 2$ así $0 > 2 - \alpha - \beta$ luego de la ecuación (5.2) $0 > 1 - \alpha > 2 - \alpha - \beta$ entonces $\alpha > 1$ y $\beta > 1$.

Ahora dado que se cumple lo siguiente

$$x_0 \text{ es máxima a } Be(x|\alpha, \beta) \text{ si y solo si } \left. \frac{d^2}{dx^2}Be(x|\alpha, \beta) \right|_{x_0} < 0.$$

Ahora bien,

$$\begin{aligned}
& \frac{d^2}{dx^2} c \left[(\alpha - 1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta - 1)x^{\alpha-1}(1-x)^{\beta-2} \right] \\
= & c \left\{ (\alpha - 1) \left[(\alpha - 2)x^{\alpha-3}(1-x)^{\beta-1} - x^{\alpha-2}(\beta - 1)(1-x)^{\beta-2} \right] \right. \\
& \left. - (\beta - 1) \left[(\alpha - 1)x^{\alpha-2}(1-x)^{\beta-2} - x^{\alpha-1}(\beta - 2)(1-x)^{\beta-3} \right] \right\} \\
= & c \left[(\alpha - 1)(\alpha - 2)x^{\alpha-3}(1-x)^{\beta-1} - (\alpha - 1)(\beta - 1)x^{\alpha-2}(1-x)^{\beta-2} \right. \\
& \left. - (\beta - 1)(\alpha - 1)x^{\alpha-2}(1-x)^{\beta-2} + (\beta - 1)(\beta - 2)x^{\alpha-1}(1-x)^{\beta-3} \right] \\
= & cx^{\alpha-2}(1-x)^{\beta-2} \left[(\alpha - 1)(\alpha - 2)x^{-1}(1-x) - (\alpha - 1)(\beta - 1) \right. \\
& \left. - (\beta - 1)(\alpha - 1) + (\beta - 1)(\beta - 2)x(1-x)^{-1} \right] \\
= & cx^{\alpha-2}(1-x)^{\beta-2} \left[(\alpha - 1)(\alpha - 2)x^{-1}(1-x) - 2(\alpha - 1)(\beta - 1) \right. \\
& \left. + (\beta - 1)(\beta - 2)x(1-x)^{-1} \right].
\end{aligned}$$

Evaluando la segunda derivada de $Be(x|\alpha, \beta)$ en $x_0 \in (0, 1)$, se tiene

$$\begin{aligned}
\left. \frac{d^2}{dx^2} Be(x|\alpha, \beta) \right|_{x_0} &= cx_0^{\alpha-2}(1-x_0)^{\beta-2} \left[(\alpha - 1)(\alpha - 2) \frac{(2 - \alpha - \beta)}{1 - \alpha} \frac{(1 - \beta)}{2 - \alpha - \beta} \right. \\
&\quad \left. - 2(\alpha - 1)(\beta - 1) + (\beta - 1)(\beta - 2) \frac{(1 - \alpha)}{2 - \alpha - \beta} \frac{(2 - \alpha - \beta)}{1 - \beta} \right] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [-(\alpha - 2)(1 - \beta) - 2(\alpha - 1)(\beta - 1) - (\beta - 2)(1 - \alpha)] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [(\alpha - 2)(\beta - 1) - 2(\alpha - 1)(\beta - 1) - (\beta - 2)(1 - \alpha)] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [(\beta - 1)[\alpha - 2 - 2\alpha + 2] - (\beta - 2)(1 - \alpha)] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [(\beta - 1)[- \alpha] - (\beta - 2)(1 - \alpha)] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [-\alpha\beta + \alpha + (2 - \beta)(1 - \alpha)] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [-\alpha\beta + \alpha + 2 - 2\alpha - \beta + \alpha\beta] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [-\alpha + 2 - \beta] \\
&= cx_0^{\alpha-2}(1-x_0)^{\beta-2} [2 - \alpha - \beta],
\end{aligned}$$

luego

$$x_0 \text{ maximiza a } Be(x|\alpha, \beta) \quad \Leftrightarrow \quad 2 - \alpha - \beta < 0 \quad \Leftrightarrow \quad 2 < \alpha + \beta,$$

y por la ecuación (5.3) se puede afirmar lo siguiente

$$x_0 \text{ maximiza a } Be(x|\alpha, \beta) \quad \Leftrightarrow \quad 2 < \alpha + \beta, \quad \alpha > 1 \quad \text{y} \quad \beta > 1.$$

- Si $\alpha + \beta = 2$
De la ecuación (5.1)

$$\frac{d}{dx} Be(x|\alpha, \beta) = 0 \iff \frac{(\alpha - 1)}{x_0} = \frac{(\beta - 1)}{1 - x_0}$$

entonces

$$\frac{d}{dx} Be(x|\alpha, \beta) = 0 \iff \alpha = 1 \quad y \quad \beta = 1$$

luego se tiene que

$$x \in (0, 1) \text{ maximiza a } Be(x|\alpha, \beta) \iff \alpha = \beta = 1.$$

Así, en general se tiene el siguiente resultado para la función beta:

- Si $\alpha + \beta = 2$

$$\hat{\theta} = x_0 \in (0, 1) \text{ maximiza a } Be(x|\alpha, \beta) \iff \alpha = \beta = 1.$$

- Si $\alpha + \beta \neq 2$

$$\hat{\theta} = x_0 = \frac{\alpha - 1}{\beta - 2 + \alpha} \text{ maximiza a } Be(x|\alpha, \beta) \iff 2 < \alpha + \beta, \quad \alpha > 1 \quad y \quad \beta > 1.$$

Distribución a posteriori normal

Si se tiene una distribución a posteriori $N(\theta|\mu, \sigma^2)$ es claro que el estimador puntual $\hat{\theta}$ es μ .

5.1.5. Intervalos creíbles de máxima densidad a posteriori (HPD)

Dada una distribución a posteriori $\pi(\theta|\vec{x})$ para encontrar su HPD al $100(1 - \alpha)\%$ ANESBA considera dos casos:

- i) Cuando $\pi(\theta|\vec{x})$ es una $Ga(\theta|\alpha, \beta)$ ó $Be(\theta|\alpha, \beta)$.

Entonces se utiliza el siguiente algoritmo:

Algoritmo de Chen-Shao para la estimación del HPD[4].

- 1.- Obtener una muestra aleatoria usando el algoritmo de Metropolis-Hastings $\{\theta_i, i = 1, \dots, n\}$ de $\pi(\theta|\vec{x})$.
- 2.- Ordenar $\{\theta_i, i = 1, \dots, n\}$ para obtener los valores ordenados

$$\theta_{(1)} \leq \theta_{(2)} \leq \dots \leq \theta_{(n)}$$

- 3.- Calcular los intervalos creíbles al $100(1 - \alpha)\%$

$$R_j(n) = (\theta_{(j)}, \theta_{(j + \lfloor (1 - \alpha)n \rfloor)})$$

para $j = 1, 2, \dots, n - \lfloor (1 - \alpha)n \rfloor$.

4.- El HPD al $100(1 - \alpha)\%$ es aquel $R_j(n)$ con la longitud mas pequeña.

ii) Si

$$\pi(\theta|\vec{x}) = N(\theta|\mu_n, \lambda_n) = \frac{1}{\sqrt{2\pi\lambda_n}} \exp\left(-\frac{1}{2\lambda_n}(x - \mu_n)^2\right).$$

Dado que $\pi(\theta|\vec{x})$ es simétrica hay que tomar el siguiente HPD

$$\left(\mu_n + \sqrt{\lambda_n} z_{\frac{\alpha}{2}}, \mu_n - \sqrt{\lambda_n} z_{\frac{\alpha}{2}}\right),$$

donde $z_{\frac{\alpha}{2}}$ es el numero tal que $Pr(Z \leq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ con $Z \sim N(0, 1)$.

Ahora se realizará un algoritmo para encontrar $z_{\frac{\alpha}{2}}$.

Dado que $0 < \alpha < 1 \Rightarrow 0 < \frac{\alpha}{2} < \frac{1}{2}$ entonces $z_{\frac{\alpha}{2}} < 0$.

Sea $limInf$ un número tal que:

$$F(limInf) \approx 0$$

con $F(x) = F(Z \leq x)$, $Z \sim N(0, 1)$ y también sea $tol > 0$ un número pequeño al cual se le considera como 0. Entonces el algoritmo quedaría:

Salida: $Z_{\frac{\alpha}{2}}$

- 1.- $a \leftarrow limInf$
- 2.- $b \leftarrow 0$
- 3.- Mientras true
 - 4.- $m = \frac{a+b}{2}$
 - 5.- Si $|F(a) - \frac{\alpha}{2}| < tol$ entonces Retornar a
 - 6.- Si $|F(b) - \frac{\alpha}{2}| < tol$ entonces Retornar b
 - 7.- Si $|F(m) - \frac{\alpha}{2}| < tol$ entonces Retornar m
 - 8.- De lo contrario
 - 9.- Si $F(m) < \frac{\alpha}{2}$ entonces $a = m$
 - 10.- De lo contrario $b = m$
- 11.- Fin mientras.

5.1.6. Contraste de hipótesis

Dada la distribución a posteriori del parámetro desconocido, a continuación se presentan las funciones que realizan los diferentes tipos de contraste de hipótesis permitidos por el software.

Contraste unilateral

Sea x_1, \dots, x_n una observación de una muestra aleatoria donde $X_i \sim f(x_i|\theta)$, además θ tiene una distribución a priori $\pi(\theta)$ y a posteriori $\pi(\theta|\vec{x})$ con $\vec{x} = (x_1, \dots, x_n)$.

La función de contraste unilateral, implementada en el software, permite realizar el siguiente contraste:

$$H_0 : \theta \leq \theta_0 \quad \text{contra} \quad H_1 : \theta > \theta_0.$$

Se necesita calcular $\alpha_0, \alpha_1, \pi_0, \pi_1$ y el factor de Bayes (B) (Hay que recordar que para que exista π_0, π_1 y B , $\pi(\theta)$ necesita ser propia).

ANESBA tiene implementadas sus funciones de densidad acumulada para las distribuciones Gama($x|\alpha, \beta$), Beta($x|\alpha, \beta$) y N($x|\theta, \sigma^2$), y con ellas realizará el contraste unilateral, ya que al final de cuentas este tipo de contrastes se reduce a calcular probabilidades acumuladas.

Para calcular la función de distribución acumulada de una función de densidad de distribución $f(x)$, la cuál puede ser Gama($x|\alpha, \beta$) o Beta($x|\alpha, \beta$), es necesario calcular la siguiente integral

$$\int_0^a f(x)dx, \tag{5.4}$$

debe resaltarse que dicha integral no puede aproximarse por los métodos numéricos que se han revisado, debido a que $f(x)$ no es siempre diferenciable, lo cual es una condición necesaria para asegurar la convergencia de tales métodos numéricos. Ante esta imposibilidad se usa el método de Monte Carlo para calcular (5.4).

Función de distribución acumulada de una distribución normal

El problema se reduce a encontrar la función acumulada de una normal estandar, es decir hallar

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}} I(x)$$

donde

$$I(x) = \int_{-\infty}^x \exp\left(-\frac{x^2}{2}\right) dx.$$

Para encontrar $I(x)$ se ha diseñado el siguiente algoritmo:

$$Si \quad x = 0 \quad \Rightarrow \quad I(0) = .5\sqrt{2\pi}$$

$$Si \quad x > 0 \quad \text{entonces}$$

$$I(x) = I(0) + \int_0^x \exp\left(-\frac{x^2}{2}\right) dx.$$

$$Si \quad x < 0$$

$$I(x) = I(0) - \int_0^{|x|} \exp\left(-\frac{x^2}{2}\right) dx.$$

Donde

$$\int_0^a \exp\left(-\frac{x^2}{2}\right) dx,$$

se calcula usando el método numérico de Romberg (véase apéndice B).

Función de distribución acumulada de una distribución gamma

Dada una función de distribución $f(\theta) = Ga(\theta|\alpha, \beta)$ se desea encontrar

$$F(u) = P(\theta \leq u) = \int_0^u f(\theta) d\theta.$$

Para esto se toma en la ecuación (4.5) $a = 0$, $b = u$, $g(\theta) = f(\theta)$. Y para conseguir $\theta_0, \dots, \theta_m$ se corre m veces el algoritmo de Metrópolis-Hastings desarrollado en el ejemplo 4.2, y por cada corrida se toma θ_T , generando así la muestra de tamaño m .

Función de distribución acumulada de una distribución beta

Dada una función de distribución $f(\theta) = Be(\theta|\alpha, \beta)$ se quiere hallar

$$F(u) = P(\theta \leq u) = \int_0^u f(\theta) d\theta.$$

Para ello solo hay que hacer en la ecuación (4.5) $a = 0$, $b = u$, $g(\theta) = f(\theta)$. Para obtener $\theta_0, \dots, \theta_m$ se corre m veces el algoritmo de Metrópolis-Hastings desarrollado en el del ejemplo 4.3, y por cada corrida se toma θ_T , generando así la muestra de tamaño m .

Contraste unilateral dual

Sea x_1, \dots, x_n una observación de una m.a. donde $X_i \sim f(x_i|\theta)$, además θ tiene una distribución a priori $\pi(\theta)$ y a posteriori $\pi(\theta|\vec{x})$ con $\vec{x} = (x_1, \dots, x_n)$.

La función de contraste unilateral dual, implementada en ANESBA, permite realizar el siguiente contraste:

$$H_0 : \theta \geq \theta_0 \quad \text{contra} \quad H_1 : \theta < \theta_0.$$

La forma en que esta función realiza este tipo de contrastes es semejante al contraste unilateral explicado anteriormente.

Contraste puntual

Esta función, implementada en ANESBA, permite llevar a cabo el contraste de hipótesis puntual

$$H_0 : \theta = \theta_0 \quad \text{frente} \quad H_1 : \theta \neq \theta_0,$$

primero se necesita que el usuario brinde el valor de π_0 (véase el capítulo de inferencia bayesiana), además este tipo de contraste sólo se realizará cuando la a priori sea propia.

Se calcula la

$$\text{razón a posteriori} = \frac{\alpha_0}{\alpha_1} = \frac{\pi_0}{(1 - \pi_0)} \frac{f(\vec{x}|\theta_0)}{m_1(x)},$$

donde

$$m_1(x) = \int_{\theta \neq \theta_0} f(\vec{x}|\theta)g_1(\theta)d\theta = \int f(\vec{x}|\theta)\pi(\theta)d\theta,$$

y también se calcula el factor de Bayes

$$B = \frac{f(\vec{x}|\theta_0)}{m_1(x)}.$$

El único problema que se presenta en esta parte es el de encontrar

$$m_1(\vec{x}) = \int f(\vec{x}|\theta)\pi(\theta)d\theta$$

el cual ANESBA lo divide en dos casos:

- a) Si $\pi(\theta)$ es una función gamma o beta, entonces se usa el siguiente algoritmo :
 Paso 1: Generar variables aleatorias $\theta_1, \dots, \theta_m$ i.i.d. de $\pi(\theta)$ (a priori).
 Paso 2: Presentar la aproximación a $m_1(\vec{x})$ dada por la ecuación (4.3).

$$\gamma = E[f(\vec{x}|\theta)] = \int f(\vec{x}|\theta)\pi(\theta)d\theta = \frac{1}{m} \sum_{i=1}^m f(\vec{x}|\theta_i).$$

- b) Si $\pi(\theta) = N(\theta|\mu, \tau^2)$, entonces este caso se presenta cuando $X_i \sim N(X_i|\theta, \sigma^2)$ y $\pi(\theta|\vec{x}) = N(\theta|\mu_n, \lambda_n)$, donde

$$\begin{aligned} \mu_n &= \frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2} \\ \lambda_n &= \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}. \end{aligned}$$

Como

$$\pi(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta)}{\int f(\vec{x}|\theta)\pi(\theta)d\theta} = \frac{f(\vec{x}|\theta)\pi(\theta)}{m_1(x)},$$

entonces

$$\begin{aligned} m_1(\vec{x}) &= \frac{f(\vec{x}|\theta)\pi(\theta)}{\pi(\theta|\vec{x})} \quad \text{el cual no depende de } \theta, \text{ si se toma } \theta = 0 \\ m_1(\vec{x}) &= \frac{f(\vec{x}|0)\pi(0)}{\pi(0|\vec{x})} = \frac{\left(\prod_{i=1}^n N(x_i|0, \sigma^2) \right) N(0|\mu, \tau^2)}{N(0|\mu_n, \lambda_n)}. \end{aligned}$$

La cual es una forma mas fácil que evaluar la integral.

5.1.7. Estadística descriptiva

En esta sección se explicarán todas las funciones utilizadas en el software para lograr el análisis descriptivo de los datos y también se incluirá un poco de teoría para la mejor comprensión de los algoritmos usados.

Definición 5.1 (Mediana para datos no agrupados)

Sea x_1, x_2, \dots, x_n una muestra de observaciones dispuestas en orden creciente. La mediana muestral es la observación central si n es impar y es el promedio de dos observaciones centrales si n es par. La mediana muestral se denota por q_2 o \tilde{X} .

Cálculo del rango intercuartil muestral

- 1.- Determinar la localización de la mediana $\frac{n+1}{2}$, donde n es el tamaño de la muestra.
- 2.- Truncar la mediana al entero inmediato inferior ($\lfloor \frac{n+1}{2} \rfloor$).
- 3.- Calcular el cuartil C_q con la fórmula

$$C_q = \frac{\text{mediana truncada} + 1}{2}.$$

- 4.- Determinar q_1 al contar en sentido creciente desde el punto de datos mínimo hasta el punto C_q . Si C_q no es entero, entonces q_1 es el promedio de los datos en las posiciones $C_q - \frac{1}{2}$ y $C_q + \frac{1}{2}$. Aproximadamente el 25 % de los datos es menor o igual que q_1 .
- 5.- Calcular q_3 al contar en sentido decreciente desde el punto de datos máximo hasta la posición C_q , como en el paso 4. Aproximadamente el 75 % de los datos es menor o igual que q_3 .
- 6.- Definir el rango intercuartílico

$$q = q_3 - q_1.$$

Diagrama de cajas

El siguiente algoritmo es el utilizado por ANESBA para elaborar los diagramas de caja.

Entrada: un vector de datos, q , cuartiles q_1 , q_2 (mediana), q_3 .

- 1.- Dibujar y marcar un eje de medida horizontal.
- 2.- Construir un rectángulo (caja) cuyo borde izquierdo está arriba de q_1 y el borde derecho está arriba de q_3 .
- 3.- Dibujar un segmento de recta vertical dentro de la caja arriba de la mediana.
- 4.- Prolongar rectas desde cada extremo de la caja hasta las observaciones mas lejanas que estén todavía a menos de $1.5q$ de distancia (distancia menor estricta) de los bordes correspondientes.

- 5.- Dibujar un círculo abierto para identificar a cada observación de la muestra que caiga entre (incluso los extremos) $1.5q$ y $3q$ del borde al cual esté mas cercano (los llamados puntos inusuales suaves).
- 6.- Dibujar un círculo lleno para identificar cada observación que caiga a mas de $3q$ del borde mas cercano (llamados puntos inusuales extremos).

Cálculo de histogramas y cuantiles

Construcción de tablas de frecuencias para datos agrupados

En esta parte se mostrará la forma para generar las tablas de frecuencias, las cuales son utilizadas en la realización de los histogramas.

Definición 5.2 La frecuencia absoluta es el número de veces que se repite un dato particular o fenómeno.

Definición 5.3 Una tabla de frecuencias es un arreglo tabular donde se muestra las veces que ocurre una característica, la cual puede estar determinada por una cualidad o por su pertenencia a un intervalo, llamado intervalo de clase.

Sea x_1, x_2, \dots, x_n un conjunto de datos, se define:

$$\begin{aligned} \text{Rango o Amplitud} &= \text{dato mayor} - \text{dato menor} \\ \text{Número de intervalos de clase} &= [1 + 3.322 \text{Log}_{10}(n)] \\ \text{Longitud de clase} &= \frac{\text{Rango}}{\text{Número de intervalos de clase}} \\ &\text{(Se toma el entero mas cercano)} \end{aligned}$$

Para mostrar como realizar una tabla de frecuencias se hará un breve ejemplo. Supóngase que se tiene la siguiente muestra:

53, 58, 55, 57, 56, 59, 57, 59, 58, 65,
63, 65, 63, 69, 67, 60, 61, 64, 65, 64,
68, 66, 60, 62, 64, 65, 64, 68, 66, 61,
62, 63, 65, 63, 67, 66, 61, 62, 64, 64,
63, 69, 67, 66, 60, 61, 62, 72, 71, 70.

Entonces

$$\begin{aligned} n &= 50 \\ \text{Rango o Amplitud} &= 72 - 53 = 19 \\ \text{Número de intervalos de clase} &= 7 \\ \text{Longitud de clase} &= 19/7 = 2.71 \approx 3 \end{aligned}$$

Luego la tabla queda de la siguiente forma:

Tabla 2: Tabla de frecuencias para datos agrupados.

Clase	Int. de clase	Int. real de clase	F. absoluta	F. relativa	F. acumulada
1	53-55	[52.5,55.5)	2	2/50	2
2	56-58	[55.5,58.5)	5	5/50	7
3	59-61	[58.5,61.5)	9	9/50	16
4	62-64	[61.5,64.5)	15	15/50	31
5	65-67	[64.5,67.5)	12	12/50	43
6	68-70	[67.5,70.5)	5	5/50	48
7	71-73	[70.5,73.5)	2	2/50	50

La forma de rellenar la tabla es muy fácil, primero defínase:

noMaxDecimales= el número máximo de decimales que contiene la muestra aleatoria, en este caso es cero pues todos los datos son enteros.

Ahora se mostrará columna a columna (donde C_i es la i -ésima columna) como obtener la tabla.

C₁ Solo hay que enumerar del 1 hasta (número de intervalos de clase).

C₂ Cada renglón de esta columna es de la forma

$$\text{límite inferior} - \text{límite superior}$$

En el primer renglón se tiene que

$$\text{límite inferior} = \text{dato menor}$$

$$\text{límite superior} = \text{dato menor} + (\text{longitud de clase}) - 1$$

En general para cualquier renglón i ($2, \dots$, número de intervalos de clase) se tendrá que

$$\text{límite inferior} = (\text{límite superior del renglón } (i - 1)) + 1$$

$$\text{límite superior} = (\text{límite inferior del renglón } i)$$

$$+(\text{longitud de clase}) - 1.$$

C₃ Cada renglón de esta columna es de la forma

$$\text{límite real inferior} - \text{límite real superior}$$

Y la forma de calcularlos es la siguiente:

$$\text{límite real inferior} = \text{límite inferior del mismo renglón}$$

pero de la columna 3

$$-(5 * 10^{-(\text{noMaxDecimales}+1)})$$

$$\text{límite real superior} = \text{límite superior del mismo renglón}$$

pero de la columna 3

$$+(5 * 10^{-(\text{noMaxDecimales}+1)})$$

- C_4 Para el renglón i es el número de datos que se encuentran contenidos en cada intervalo real de clase (dado por la columna 3 en el renglón i).
- C_5 Para el renglón i es el valor de la tabla en el renglón i de la columna 4 entre n .
- C_6 Para el primer renglón es el valor que posee el primer renglón de la columna 4. Para el renglón i ($2, \dots$, número de intervalos de clase) se define como: la cantidad del renglón i pero de la columna 4 mas la cantidad del renglón $(i-1)$ de la columna 6.

Para generar el histograma se utilizan las frecuencias de la columna 5 donde cada una corresponde a el intervalo de clase de cada renglón.

Cuantiles

La fórmula que se utilizará para el cálculo de los cuantiles para datos agrupados es la siguiente:

$$q_k = L_k + \frac{k\left(\frac{n}{TC}\right) - F_{k-1}}{f_k} * l$$

donde

- L_k = Límite inferior de la clase real del cuantil k
- n = Número total de observaciones
- F_{k-1} = Frecuencia acumulada de la clase que antecede a la clase del cuantil k
- f_k = Frecuencia absoluta de la clase del cuantil k
- TC = $\begin{cases} 4, & \text{para Cuartil;} \\ 10, & \text{para Decil;} \\ 100, & \text{para Percentil.} \end{cases}$
- k = Número de cuantil a calcular
- = $\begin{cases} 2, 3 & \text{para Cuartil;} \\ 2, 3, \dots, 9 & \text{para Decil;} \\ 2, 3, \dots, 99 & \text{para Percentil.} \end{cases}$

5.2. Requerimientos

Los requerimientos mínimos necesarios para poder usar el software ANESBA son los siguientes:

- Un editor de texto para generar los archivos para la muestra aleatoria
- 256MB de memoria RAM
- 500MB de espacio en disco duro
- Un procesador Pentium a 600MHz
- Windows 2000/XP/Server/Vista

Capítulo 6

Resultados y trabajo a futuro

6.1. Resultados

La principal herramienta que se usó para comparar si las salidas del software desarrollado son correctas fué el programa Mathematica 5.1, el cual es de uso comercial. Cabe mencionar que mientras se realizaban dichas pruebas se encontraron errores en Mathematica, para ilustrar esto se pone el siguiente ejemplo (Figura A.1).

Tomemos $\alpha = 1113$, $\beta = 1121$ y obsérvese los siguientes cálculos realizados en Mathematica.

$$\int_0^{.9} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx = 2.11 * 10^{913},$$

$$\int_0^{.93} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx = -1.51 * 10^{938}.$$

Recuérdese que lo que se esta integrando es la función de densidad de distribución beta, y por lo tanto se debe cumplir que

$$0 \leq \int_a^b \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \leq 1$$

para a, b en $[0, 1]$. Esto muestra que el software Mathematica ha arrojado resultados erróneos.

Por otra parte con el propósito de mostrar que ANESBA funciona correctamente se utilizará para realizar el ejemplo 2.4, ya que se conocen los resultados de este, los cuales estan redondeados a su primer decimal. Resumiendo la información de este ejemplo se tiene que la variable aleatoria X , que representa la puntuación obtenida por un niño en una prueba de inteligencia, tiene una distribución $N(\theta, 100)$ y además $\theta \sim N(100, 225)$. También se tiene información de que el niño obtuvo una puntuación en la prueba de 115, lo que indica que la m.a. es de tamaño uno, y $x = 115$.

Para comenzar a usar el software primero se abre un archivo de nombre ejemplo1.txt en el escritorio, el cual solo contendrá el número 115.

Ahora dada la pantalla del menú principal de ANESBA, elegir **Archivo-Abrir datos** (Figura A.2) y seleccione su muestra aleatoria.

Elección de la distribución de los datos

Como paso siguiente en el menú principal de ANESBA se debe ir a **Analizar-Elección de densidades-Función de Distribución** (Figura A.3). Al hacer clic se abrirá una nueva ventana (Figura A.4).

Se elige y acepta **Normal (varianza conocida)** la cual mostrará la ventana (Figura A.5) donde se ingresa la cantidad 100 y se presiona Aceptar. Con ello se ha elegido para los datos una distribución normal con una varianza de 100.

Elección de la función de distribución a priori del parámetro desconocido θ

Del menú principal de ANESBA, se elige **Analizar-Elección de densidades-A priori**, después aparecerá una nueva ventana de la cual se debe seleccionar la opción Conjugada (Figura A.6).

Ahora en la siguiente ventana, se debe ingresar en los cuadros de texto el valor de la media 100 y la varianza 225 como se muestra en la figura A.7. Con ello se ha asignado a θ una distribución $N(100, 225)$.

Estimación puntual

En el menú principal se elige **Analizar-Inferencia Estadística-Estimación Puntual** y aparecerá la ventana de resultados (Figura A.8).

Esto indica que $\hat{\theta} = 110.384615384615$, nótese que anteriormente en el ejemplo 2.4 se había encontrado que $\hat{\theta} = 110.39$.

Cálculo del HPD

Del menú principal se elige a **Analizar-Inferencia Estadística-Estimación por intervalos**, y aparecerá la una nueva ventana (Figura A.9), dado que se quiere un HPD al 95% se tomará $\alpha = .05$, así se debe insertar el valor de .05 en el cuadro de texto; al Aceptar aparecerá una nueva ventana (Figura A.10), donde se muestran los resultados de los cálculos.

Así se obtiene, por el software, que el intervalo de máxima densidad de probabilidad al 95% de confianza es:

$$(94.0767288463275, 126.692501922903)$$

El cual encierra un área de 0.950000006125208. Mientras que en el ejemplo se había encontrado el HPD como:

$$(94.08, 126.70)$$

Contraste de Hipótesis

Ahora se realizará el siguiente contraste de hipótesis en el software ANESBA.

$$H_0 : \theta \leq 100 \text{ contra } H_1 : \theta > 100.$$

Del menú principal de ANESBA se elige **Analizar-Inferencia Estadística-Contraste de Hipótesis** y aparecerá una nueva ventana (Figura A.11), se debe escoger la opción $H_0: \theta \leq \theta_0$.

Al presionar calcular aparecerá la figura A.12, donde debe colocarse en el cuadro de texto el valor de θ_0 , en este caso 100 y después se mostrará la ventana de soluciones

(Figura A.13). Luego el software arrojará:

La razón a priori es 1.

La razón a posteriori es $0.118570379042875 \approx 1/8.44$.

El factor de Bayes es $0.118570379042875 \approx 1/8.44$.

Dado que en el ejercicio desarrollado en los capítulos 2 y 3, los resultados se han redondeado al primer decimal, entonces los resultados obtenidos por ANESBA son los mismos, así con el desarrollo de este ejemplo se puede apreciar el buen funcionamiento de ANESBA, y también su fácil manejo en la inferencia bayesiana, con lo cual se cumplen las expectativas iniciales de crear un software en español, confiable y amigable.

6.2. Trabajo a futuro

Durante el desarrollo del programa, surgieron muchas ideas acerca de nuevas funciones que se le podrían agregar a ANESBA para un mejor funcionamiento, y aunque se puso un gran esfuerzo en la elaboración de este software, desafortunadamente por cuestión de tiempo y dado que en un principio muchas de tales funciones no eran el objetivo de la tesis, no se pudieron integrar todas esas ideas al software, por ello a continuación sugiero algunas de estas funciones para implementarlas en ANESBA, esperando que en un trabajo futuro sean agregadas.

- Agregar una parte que realice la predicción bayesiana.
- Implementar funciones para realizar regresión lineal.
- Ampliar el número de las funciones de distribución usadas por el software.
- Crear una ayuda integrada al programa y otra en línea.
- Mejorar la presentación de los reportes

Apéndice A

Apéndice de figuras

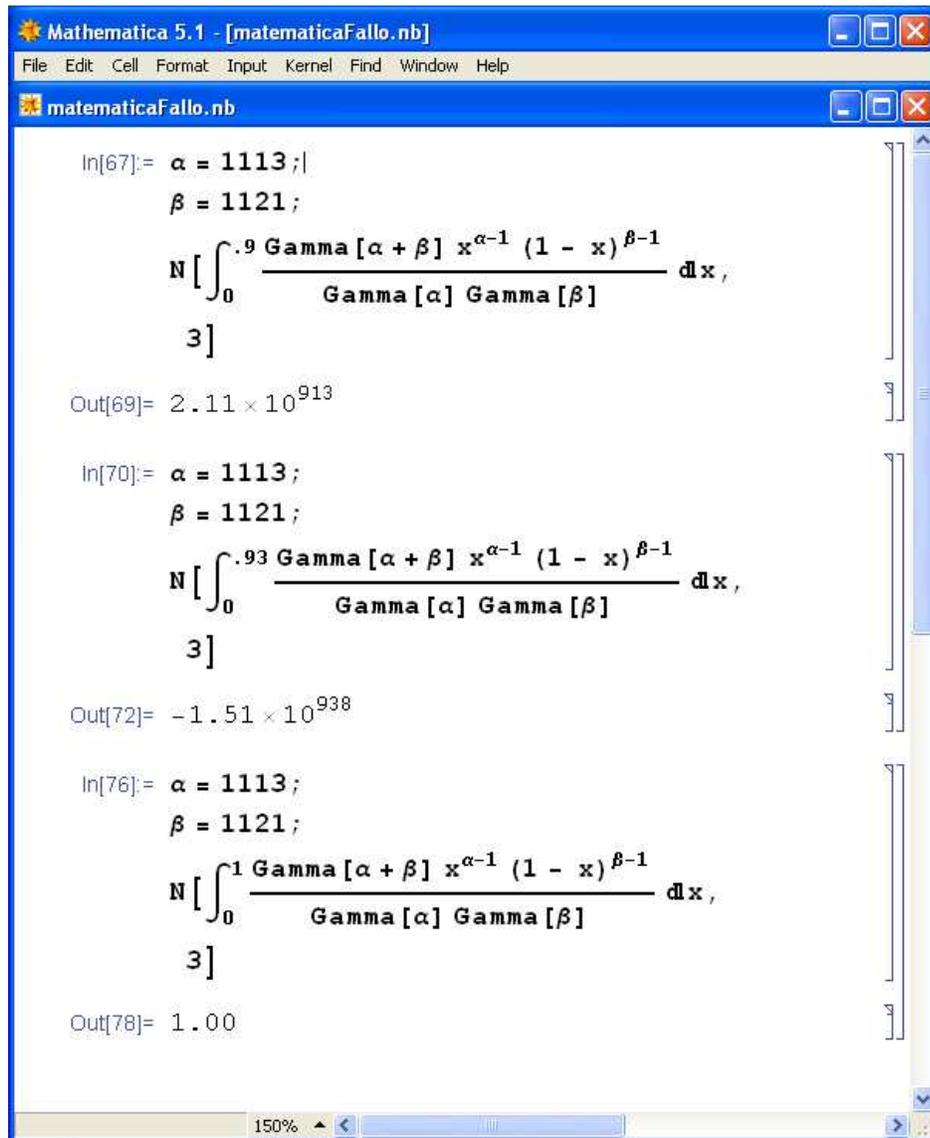


Figura A.1: Evaluación errónea de una integral en Mathematica.

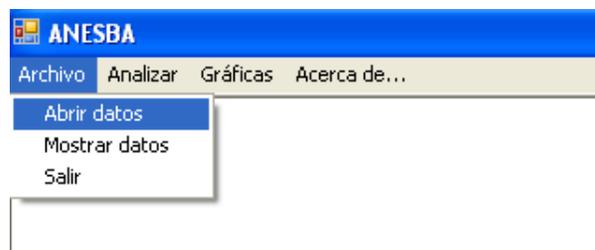


Figura A.2: Menú principal para elegir una m.a.

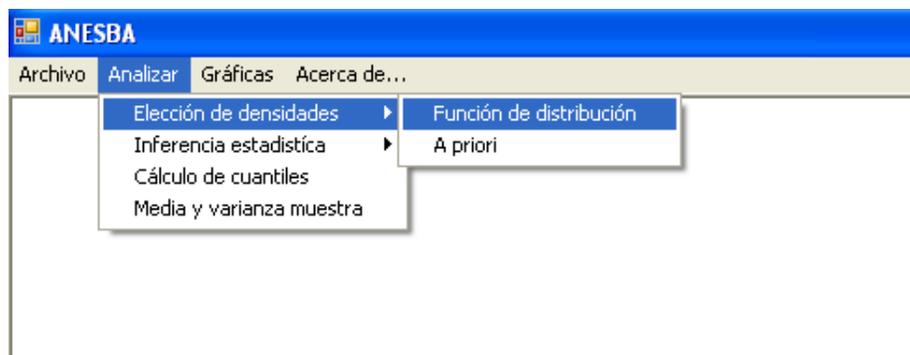


Figura A.3: Elección de la función de densidad de distribución para la m.a.

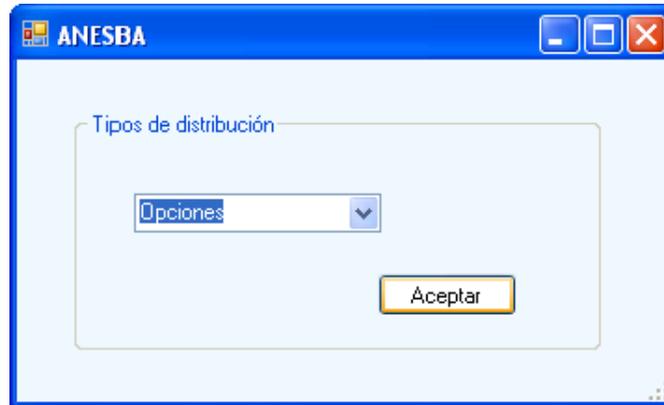


Figura A.4: Opciones de las posible densidades de distribución.

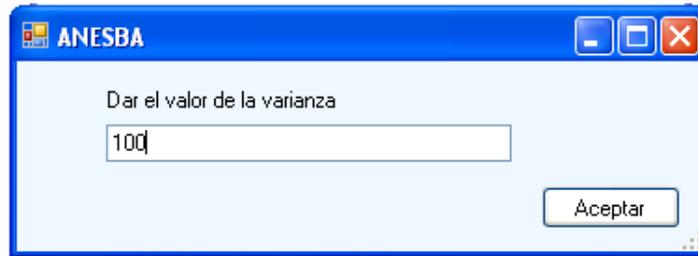


Figura A.5: Se introduce la varianza, para la distribución normal.



Figura A.6: Menú para elegir la distribución a priori.

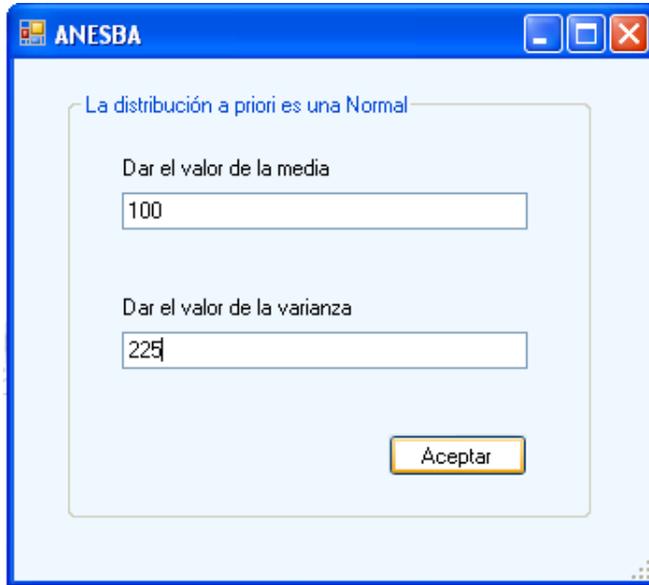


Figura A.7: Se introduce la media y la varianza para la distribución a priori elegida.

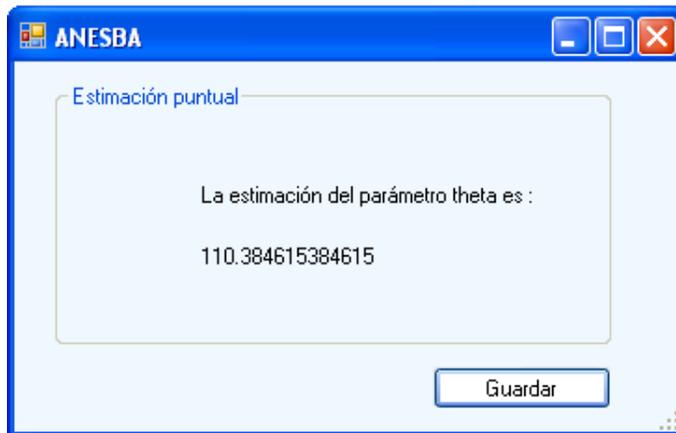


Figura A.8: Resultado de la estimación puntual realizada.

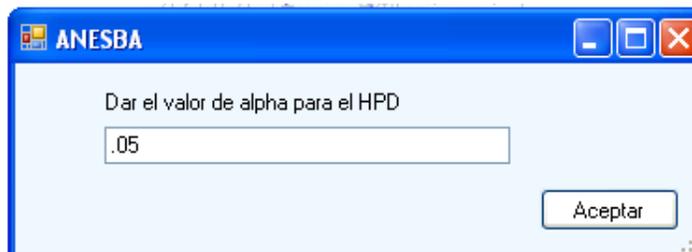


Figura A.9: Elección del nivel de significancia para calcular el HPD.

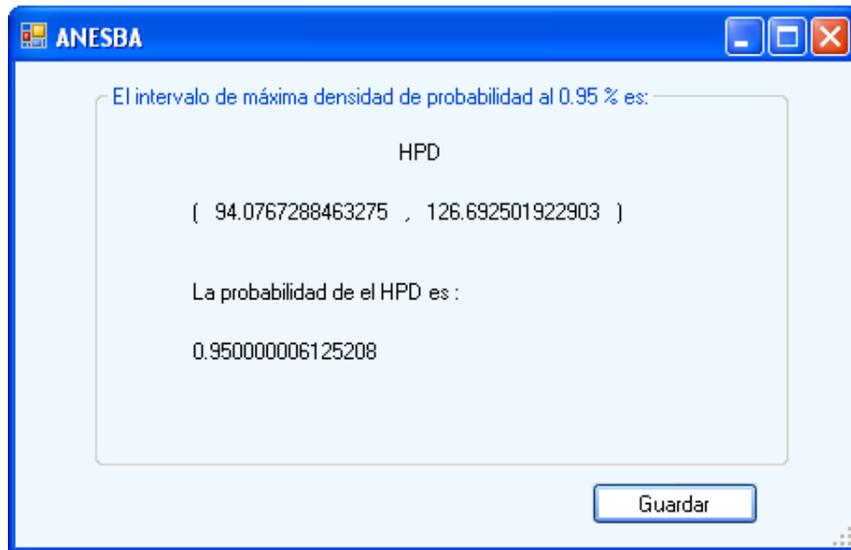


Figura A.10: Se muestra el HPD y el área que determina.

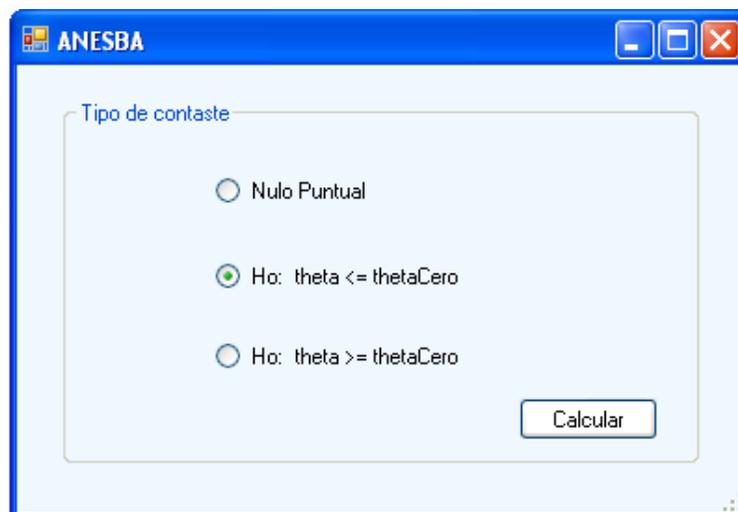


Figura A.11: Opciones de los tipos de contrastes de hipótesis.

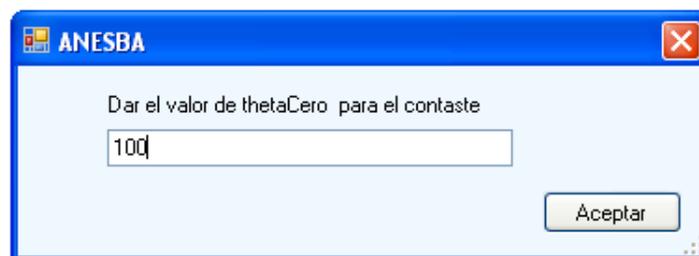


Figura A.12: Ingreso del valor θ_0 para realizar el contraste de hipótesis.



Figura A.13: Resultados del contraste unilateral realizado.

Apéndice B

Apéndice

B.1. Algoritmo de Romberg

Para aproximar $I = \int_a^b f(x)dx$, seleccionar un entero $n > 0$.

ENTRADA: Los puntos de los extremos a, b ; el entero n .

SALIDA: La aproximación a I ($R_{2,n}$).

1.- Tomar $h=b-a$

$$R_{1,1} = h(f(a) + f(b))/2.$$

2.- Para $i = 2, \dots, n$ seguir Pasos 3-6

3.- Tomar

$$R_{2,1} = \frac{1}{2} \left[R_{1,1} + h \sum_{k=1}^{2^{i-2}} f(a + (k - 0.5)h) \right].$$

4.- Para $j = 2, \dots, i$

$$\text{tomar } R_{2,j} = \frac{4^{j-1}R_{2,j-1} - R_{1,j-1}}{4^{j-1} - 1}.$$

5.- Tomar $h = h/2$.

6.- Para $j = 1, 2, \dots, i$ tomar $R_{1,j} = R_{2,j}$.

7.- Parar

La única suposición que pide este algoritmo es que $f \in C^{2n+2}[a, b]$.

B.2. Funciones de verosimilitud

Sea X_1, \dots, X_n , con $X_i \sim f(x_i|\theta)$ a continuación se obtendrá $f(\vec{x}|\theta)$ (función de verosimilitud) con $\vec{x} = (x_1, \dots, x_n)$.

Función de distribución bernoulli

Si $X_i \sim Br(x_i|\theta)$ entonces

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i},$$

por lo que la función de verosimilitud es

$$f(\vec{x}|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Función de distribución poisson

Supóngase que $X_i \sim Pn(\lambda)$ entonces

$$f(x_i|\lambda) = \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda),$$

luego la función de verosimilitud es

$$f(\vec{x}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \exp(-n\lambda).$$

Función de distribución exponencial

Dado que $X_i \sim Ex(x_i|\theta)$

$$f(x_i|\theta) = \theta \exp(-\theta x_i),$$

así

$$f(\vec{x}|\theta) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right).$$

Función de distribución normal, media conocida (normal modificada)

Si $X_i \sim N(x_i|\theta, \lambda)$

$$f(x_i|\lambda) = \frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(x_i - \theta)^2\right],$$

entonces

$$f(\vec{x}|\lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\lambda}{2} \sum_{i=1}^n (x_i - \theta)^2\right].$$

Función de distribución normal, varianza conocida

Dado que $X_i \sim N(x_i|\theta, \sigma^2)$

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2\right],$$

luego

$$f(\vec{x}|\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right].$$

Apéndice C

Notación

$Bi(x n, \theta)$	Distribución binomial con parámetros $0 < \theta < 1$, $n = 1, 2, \dots$ y $x = 0, 1, 2, \dots, n$.
$Be(x \alpha, \beta)$	Distribución beta con parámetros $\alpha, \beta > 0$ y $0 < x < 1$.
$Br(x \theta)$	Distribución bernoulli con parámetros $0 < \theta < 1$, y $x = 0, 1$.
$Ex(x \theta)$	Distribución exponencial con parametro $\theta > 0$ y $x > 0$.
$Ga(x \alpha, \beta)$	Distribución gamma con parámetros $\alpha, \beta > 0$ y $0 < x$.
$N(x \mu, \sigma^2) = N(\mu, \sigma^2)$	Distribución normal con parametro $\sigma > 0$, $\mu \in \mathfrak{R}$ y $x \in \mathfrak{R}$.
$Pn(x \lambda)$	Distribución Poisson con parámetro $\lambda > 0$ y $x = 0, 1, \dots$
i.i.d.	Independiente e idénticamente distribuida.
<i>d.o.f</i>	De otra forma.
m.a.	Muestra aleatoria.
\propto	Proporcional.
$[a]$	Parte entera del número a .
f.d.p.	Función de densidad de probabilidad.

Bibliografía

- [1] Berger, James O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Segunda edición, Springer-Verlag.
- [2] Bernardo, José M. y Smith, Adrian F. M. (1994) *Bayesian Theory*. John Wiley & Sons Ltd.
- [3] Bradley P. Carlin and Thomas A. Louis (2000) *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC
- [4] Chen Ming-Hui, Shao Qi-Man, Ibrahim Joseph G. (2000) *Monte Carlo Methods in Bayesian Computation*. Springer.
- [5] Christian P. Robert, George Casella (1999) *Monte Carlo Statistical Methods*. Springer.
- [6] Edward Arnold (1989) *Bayesian Statistics: An Introduction*. Great Britain.
- [7] Evangelos Petroutsos (2006) *La biblia de Visual Basic 2005*. ANAYA, multimedia.
- [8] Gutiérrez Peña, Eduardo *Métodos Computacionales en la inferencia Bayesiana*. II-MAS, UNAM.
- [9] Hastings W. K. (1970) *Monte Carlo sampling methods using Markov chains and their applications*. *Biometrika* 57, 97-109.
- [10] J. Susan Milton, Jesse C. Arnold (2004) *Probabilidad y estadística*. Mc Graw Hill.
- [11] Mood Alexander M., Graybill Franklin A., Boes Duane C. *Introduction to the theory of statistics*. Mc.
- [12] Perez Perez Olivia.(1996) *Introducción a la teoría de la decisión estadística*. Tesis de licenciatura de la Universidad Nacional Autónoma de México.
- [13] R. L. Wolpert (1991) *Monte Carlo Integration in Bayesian Statistical Analysis*. *Contemporary Mathematics* 115,101-116.
- [14] Richard L. Burden, J. Douglas Faires (1985) *Análisis Numérico*. Grupo Editorial Iberoamérica.
- [15] Zacarías Santiago Adriana (2006) *Métodos de integración y simulación Monte Carlo en la teoría bayesiana*. Tesis de licenciatura de la Universidad Tecnológica de la Mixteca.