Introducción a la estadística bayesiana: notas de clase

Book · January 2018

DOI: 10.22430/9789585414242

CITATION
1

READS
14,033

2 authors:

Juan Carlos Correa
National University of Colombia-Medellín Colombia
113 PUBLICATIONS **442** CITATIONS
SEE PROFILE

Carlos J Barrera-Causil
Instituto Tecnológico Metropolitano
17 PUBLICATIONS **62** CITATIONS
SEE PROFILE

TEXTOS ACADÉMICOS

Introducción a la Estadística Bayesiana

Juan Carlos Correa Morales
Carlos Javier Barrera Causil

FONDO EDITORIAL ITM

INTRODUCCIÓN A LA ESTADÍSTICA BAYESIANA

INTRODUCCIÓN A LA ESTADÍSTICA BAYESIANA

Juan Carlos Correa Morales

Carlos Javier Barrera Causil

Correa Morales, Juan Carlos
Introducción a la Estadística Bayesiana / Juan Carlos Correa Morales, Carlos Javier Barrera Causil. – 1a ed. – Medellín:
Instituto Tecnológico Metropolitano, 2018
222 p. – (Textos Académicos)

Incluye referencias bibliográficas
ISBN 978-958-5414-24-2

1. Estadística bayesiana I. Barrera Causil, Carlos Javier II. Título III. Serie

519.542 SCDD Ed.21

Catalogación en la publicación - Biblioteca ITM
Introducción a la Estadística Bayesiana

Edición: enero 2018
ISBN: 978-958-5414-24-2
Publicación electrónica para consulta gratuita

Autores
JUAN CARLOS CORREA MORALES
CARLOS JAVIER BARRERA CAUSIL

Editado en Medellín, Colombia
Instituto Tecnológico Metropolitano
Sello editorial Fondo Editorial ITM
Calle 73 No. 76A 354
Tel.: (574) 440 5197 • Fax: 440 5382
www.itm.edu.co

La incertidumbre está en todas partes
y tú no puedes escapar de ella.
Dennis Lindley

El azar no es, sin embargo, una loca fantasía;
responde a su vez a leyes.
Los dados obedecen a la gravedad
y sólo tienen seis caras.
Juan José Sebreli
Comediantes y mártires: ensayo contra los mitos

Prefacio

La estadística bayesiana es un campo que ha tenido un desarrollo impresionante en los últimos años, en especial desde la introducción de la parte computacional. Muchas ideas han estado circulando desde hace tiempo, pero su imposibilidad práctica hacía que se miraran con cierto pesar, ya que eran muy atractivas pero inaplicables. Esto afortunadamente ha cambiado. Es lamentable que muchos de los libros básicos en estadística no hagan una presentación de los elementos básicos de esta aproximación para la solución de problemas estadísticos. Libros en estadística bayesiana han aparecido en las últimas dos décadas por cantidades apreciables. Antes de los años 90 se tenían libros más enfocados en la teoría de la decisión ([1]; [2]; [3]; [4]; [5]; [6]; [7]; [8]), en aspectos teóricos de la probabilidad subjetiva ([9]; [10]) y algunos pocos a la estadística bayesiana aplicada ([11]; [12]; [13]; [14]). En las últimas dos décadas esto se ha revertido y encontramos libros aplicados de estadística bayesiana en muchas áreas: generales ([15]; [16]; [17]; [18]; [19]; [20]; [21]; [22]; [23]; [24]; [25]), pronósticos [26], econometría [27], bioestadística ([28]; [29]; [30]), ciencias sociales ([31]; [32]), confiabilidad ([33]; [34]), mercadeo [35], aplicaciones en ingeniería civil [36] y otros dedicados a la parte computacional ([37]; [38]; [39]; [40]; [41]; [42]). Samaniego [43] presenta una extensa comparación entre los métodos frecuentistas y los métodos bayesianos. La estadística bayesiana no ha tenido un camino fácil en el mundo del trabajo aplicado. Qin [44] presenta un recuento histórico del uso, discusiones y reservas, de la estadistica bayesiana en econometría, historia que puede ser similar en diferentes áreas de investigación. La Inteligencia Artificial es un área de un fuerte desarrollo tanto teórico y aplicado de gran importancia que hace uso extenso de métodos bayesianos [45].

Aquí vamos a presentar una aproximación eminentemente práctica, esto es, el lector puede aplicar de forma casi inmediata los métodos a problemas reales. El software que se utilizará es de dominio público como el *R* [46] y el *OpenBUGS*. Se requiere familiaridad con el primer programa, al menos a un nivel operativo básico. Haremos énfasis en la parte de construcción de la distribución a priori que resuma el conocimiento previo del experto. Esta parte generalmente no es considerada en los textos de estadística bayesiana moderna, pero consideramos que es la esencia misma del análisis bayesiano y constituye el aporte de este trabajo. Consideramos que si la estadística bayesiana se diferencia en algo de la estadística tradicional (clásica) es en permitirle al usuario incorporar información disponible de una manera transparente y directa.

El programa y lenguaje estadístico R [46] se ha vuelto uno de los estándares para realizar trabajo estadístico, tanto aplicado como el desarrollo de nuevas metodologías. La estadística bayesiana se ha beneficiado enormemente de la flexibilidad y el potencial de este programa, el cual permite crear fácilmente librerías y ponerlas en la red de tal forma que usuarios a nivel mundial puedan usarlas y validarlas, retroalimentando a sus creadores de tal forma que, en muy poco tiempo se tengan subprogramas de gran eficiencia y calidad. Hay ahora muchas librerías que han sido creadas para resolver problemas de tipo general, como son modelos lineales y lineales generalizados, o más generales aún que permiten a un usuario resolver problemas propios mediante el uso de muestreadores proporcionados en estas librerías, como ejemplo tenemos MCMCpack [47] [48], la cual permite ajustar muchos modelos útiles en el trabajo aplicado de una manera simple y directa como se hace en R. Creemos que el éxito de R ha venido en el detrimento de programas como el $WinBUGS$, ya que un investigador prefiere crear programas que por un lado sean más transparentes y, por otro lado, que lleguen a un público más amplio, aunque los estadísticos bayesianos dicen que son complementarios.

Este texto está dirigido a investigadores, estudiantes de pregrado y posgrado en estadística, ingeniería y ciencias, que tengan familiaridad con los métodos estadísticos a un nivel operativo, al menos. Conocimiento de inferencia a un nivel de un texto básico de estadística matemática del estilo de Hogg y Craig [49] o Mood, Graybill y Boes [50] ayuda bastante.

Índice general

Parte I

Elementos básicos

Capítulo 1

Introducción

El problema fundamental del progreso científico, y uno fundamental en la vida diaria, es el de aprender de la experiencia. El conocimiento obtenido de esta manera es parcialmente una descripción de lo que ya hayamos observado, pero una parte consiste en la realización de inferencias de la experiencia pasada para predecir la experiencia futura[9].

La escuela bayesiana en estadística ha tomado fuerza en los últimos años, debido a su potencial para resolver problemas que no se pueden atacar con otros métodos y porque permite incorporar naturalmente información que es útil en la solución del problema enfrentado. Nadie niega que ante un problema en particular debemos utilizar toda la información disponible acerca del mismo o de sucesos similares. Para nuestro caso estadístico, la incertidumbre sobre parámetros poblacionales se resume por medio de distribuciones de probabilidad, que antes de recoger información muestral relevante para ellos, se conoce como 'distribución a priori.' El problema está en la forma de cuantificar esta información sin generar alguna contradicción.

La aproximación bayesiana es una herramienta fundamental en situaciones donde la recolección de información muestral sea muy difícil, por ejemplo en tópicos de alta sensibilidad social, tales como el consumo de drogas ilícitas, o extremadamente costosos o imposibles, como sería el caso de la determinación del riesgo de falla de una nueva nave espacial o cuál es la probabilidad de que haya vida inteligente en nuestra galaxia.

Un problema que se ha planteado cuando se habla de la escuela bayesiana es que dos personas enfrentadas ante un problema y una decisión a tomar, y asumiendo que tengan la misma información muestral, pueden llegar a dos decisiones opuestas si su información adicional es diferente. Greenland [51] afirma que «los epidemiólogos perciben la especificación de la distribución a priori como no práctica y además pocos epidemiólogos emplearían métodos que no están disponibles en paquetes estadísticos líderes». Dienes [52] discute en detalle las posiciones de ambas escuelas.

En estadística realizamos y tratamos de responder preguntas con respecto a las características de una o varias poblaciones. En la aproximación bayesiana tenemos:

- La información sobre un parámetro (puede ser un vector) que se tiene se debe resumir en una distribución de probabilidad, esta será llamada la *distribución a priori*.
- Los parámetros son considerados variables aleatorias (esto no es aceptable en la estadística clásica).
- La información a priori puede provenir de:
  - Estudios previos.
  - Información subjetiva de expertos (la cuantificación de esta información es lo que llamamos *elicitación*).

Albert [53] presenta las siguientes razones por las cuales se debería enseñar estadística desde el punto de vista bayesiano:

- El paradigma bayesiano es un medio natural de implementar el método científico donde la distribución a priori representa sus creencias iniciales acerca del modelo, usted recoge los datos adecuados, y la distribución posterior representa sus creencias actualizadas después de ver los datos.
- Si la incertidumbre acerca de los modelos es expresada utilizando probabilidad subjetiva, entonces la regla de Bayes es la única receta que uno necesita para realizar inferencias de los datos.
- Las afirmaciones inferenciales bayesianas son más fáciles de entender que las basadas en la inferencia tradicional basadas en muestreo repetido. La probabilidad de que un parámetro caiga dentro de un intervalo calculado es igual a 0.95. También, en contraste con los procedimientos tradicionales de pruebas de hipótesis, tiene sentido hablar acerca de la probabilidad de que una hipótesis estadística sea cierta.
- Por el principio de condicionalidad, los únicos datos relevantes para ejecutar inferencias son los datos realmente observados. Uno puede ignorar otros resultados de un espacio muestral que no son observados.
- Los problemas de predicción no son más difíciles que los problemas de estimación de parámetros. Parámetros y observaciones futuras son cantidades desconocidas que son modeladas subjetivamente.

Western y Jackman [54] hacen un recuento de las críticas que dos famosos estadísticos hacen de la aproximación bayesiana (Fisher y Efron). Una de las críticas es la introducción de información subjetiva a priori que haría que los prejuicios de los analistas fueran introducidas en los análisis, dañando los resultados. Tanto Fisher como Efron argumentan que con la inclusión de información subjetiva no es posible realizar un análisis justo de los datos. A lo cual Western y Jackman replican diciendo:

> En la práctica, sin embargo, la información a priori entra en la mayoría de los análisis a través de decisiones de codificación, transformaciones y búsquedas

no reportadas en conjuntos de variables exploratorias para obtener resultados que parezcan significativos en el sentido de caer dentro de un rango de valores esperados. Mientras todos los analistas de datos usan creencias previas, los bayesianos hacen la forma de volver estas aprioris explícitas e integrarlas sistemáticamente en el análisis. Y, reparafrasean a de Finetti quien dijo que el reconocimiento de la subjetividad es el camino a la objetividad.

Kyburg, Jr. [55] nos presenta esta reflexión sobre la incertidumbre:

> Hay dos clases de ignorancias que considero: la clase más simple de ignorancia es la que hace que las loterías sean excitantes; la otra es la que hace que las carreras de caballos sean excitantes.
>
> Una lotería es excitante debido que aunque sepamos exactamente que uno de los números de los posibles se obtendrá, y aunque sepamos que todo lo posible haya sido hecho para garantizar que ninguno estos números tenga ventaja sobre los otros, no sabemos cuál saldrá. Esto es generalmente expresado diciendo que la probabilidad de un estado es la misma que la de cualquier otro estado: o, en el caso particular de la lotería, que la probabilidad de que un tiquete gane es igual a la de cualquier otro tiquete. Esto no es el caso típico en las carreras de caballos; uno no puede organizar las carreras de caballos de tal forma que cada caballo en la carrera tenga (mediante algún consenso general) la misma probabilidad de ganar. Existe una gran cantidad de información acerca de cada caballo que determina la probabilidad que ese caballo gane (si es que tal probabilidad existe del todo), y no hay una forma aceptable de cambiar esas circunstancias- lastimándolo, digamos- tal que las probabilidades sean iguales. Uno encontraría difícil, quizá, elaborar una distinción clara y precisa entre estas dos clases de situaciones que no pudiera ser atacada como artificial; y aún así ellas parecen diferir de una forma importante. Yo consideraría la primera situación como una incertidumbre estadística; y la última, donde las probabilidades dependen fuertemente del conocimiento, como una incertidumbre epistemológica.

1.1. Ejemplos típicos

Ejemplo 1.1. Cálculo de la edad de una persona. En nuestra sociedad es considerado como una forma de mala educación preguntar la edad de una persona. El día que conoce a alguien, usted más o menos puede calcular la edad de esta persona. Este proceso se hace de una manera inconsciente y usualmente llega a un número que aproxima sus creencias sobre la posible edad. Para esto usa la información recolectada previamente sobre ella, por ejemplo, si esta persona tiene una apariencia determinada, si se viste de cierta forma, si se graduó del colegio en cierta época, etc. Si dos personas tienen que calcular la edad de este sujeto, puede que ellas no coincidan en sus resultados, pero no se puede decir cuál de los dos está equivocado (o si los dos lo están), solo hasta el momento en que se conozca la verdadera edad de la persona en cuestión. La incertidumbre que usted tiene acerca de la edad de una persona la podemos expresar en términos probabilísticos con ayuda de la siguiente plantilla (ver Figura 1.1).

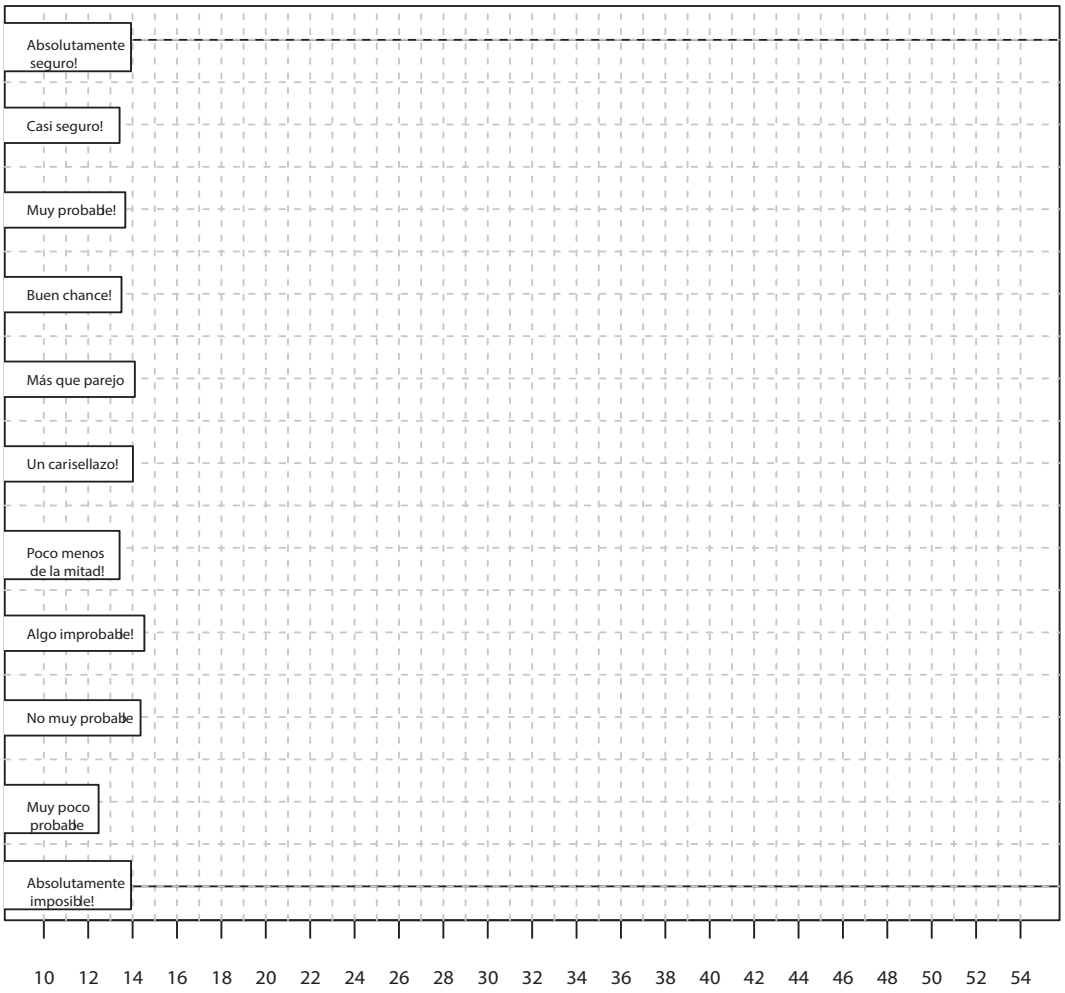

Figura 1.1: *mediante la ayuda de la plantilla podemos 'elicitar' la distribución de probabilidad que nos refleja la incertidumbre que tenemos sobre la edad de una persona. Nota: todas las figuras y tablas del texto son de elaboración propia del autor*

Ejemplo 1.2. La lotería que jugó anoche. Suponga que a usted un amigo le ofrece un billete de lotería, pero con el problema que la lotería jugó anoche. Su amigo, que ha demostrado ser una persona honesta le informa que él no sabe el resultado de la lotería, y usted tampoco. En una situación como esta podemos pensar en una probabilidad de que el billete sea el ganador es la misma que el billete tenía antes de que se jugara la lotería, ¿no lo piensa así?

Ejemplo 1.3. Estatura de los colombianos. Si pensamos en la estatura promedio de los hombres colombianos podemos pensar seriamente que este valor no es mayor que 180 cm., ni menor que 160 cm. Es claro que si conocemos muchos hombres colombianos nuestra información puede utilizarse en un proceso inferencial, pero confiaríamos más si la información sobre la estatura proviene de algún estudio previo realizado sobre el mismo tema.

Ejemplo 1.4. La nota esperada. A un estudiante que acaba de presentar un examen se le puede preguntar cuál es su nota esperada. Con base en su propio

conocimiento de su capacidad y de su preparación, de cómo respondió el examen, él puede tener una idea sobre la nota que espera obtener al ser calificado su examen. Obviamente la nota exacta no la conoce ya que existen múltiples factores que entran en una evaluación, pero puede proporcionar un rango dentro del cual se sienta muy seguro.

Ejemplo 1.5. Sobre una proporción. Un estudiante universitario que visite con frecuencia los distintos campus puede intentar estimar el porcentaje de mujeres que estudian en ésta. Él puede establecer valores entre los cuales, cree, cae el porcentaje de mujeres que estudian en la universidad.

Ejemplo 1.6. Porcentaje de estudiantes que consumen una sustancia psicoactiva. Si queremos determinar el porcentaje de estudiantes que consumen un tipo de sustancia psicoactiva podemos utilizar la información que se haya recogido en estudios pasados.

Ejemplo 1.7. Tasa de estudiantes que ejercen la prostitución. Si queremos determinar el porcentaje de estudiantes que ejercen la prostitución en nuestra universidad, no parece fácil resolver esto mediante una simple encuesta, aunque es posible utilizar procedimientos como el de la respuesta aleatorizada, el hecho de enfrentar un encuestador puede llevar a dar respuestas socialmente aceptables.

1.2. Probabilidad personal o subjetiva

Las ideas iniciales de la probabilidad surgieron relacionadas con los juegos de azar y su conceptualización e interpretación son básicamente frecuentistas. Esta formulación frecuentista trabaja bien en muchas situaciones, pero no en todas.

Entre otras, destacamos las tres diferentes interpretaciones que Kyburg, Jr. [55] señala que pueden considerarse respecto a la probabilidad:

1. Interpretación empírico-frecuentista. Esta es la interpretación más común de la probabilidad y hace relación al comportamiento (real o hipotético) de ciertos objetos.

2. Interpretación lógica. Esta interpretación no es común entre los estadísticos y está más bien reservada al mundo de los lógicos. De acuerdo con esta interpretación, hay una relación lógica entre una afirmación (considerada como una hipótesis) y otra afirmación (considerada como evidencia), en virtud de la cual la primera tiene cierta probabilidad relativa a la segunda. Probabilidad lógica es el grado de creencia en proposiciones, que asocian un conjunto de premisas con un conjunto de conclusiones. En la probabilidad lógica , esta relación es única. Fue De Morgan quien primero definió la probabilidad en términos de «grados de creencia»[56].

   Bajo la influencia de Bertrand Russell, Keynes adoptó una *proposición* (en lugar de un evento) «como eso que puede llevar el atributo de la probabilidad». Keynes dice que la probabilidad es *relación lógica* indefinible entre (1)

una proposición y un cuerpo de conocimiento, (2) entre una afirmación y otra afirmación (es) que representa evidencia, una relación asociada con el *grado de creencia* racional en la proposición [56]. Un concepto de probabilidad lógica es empleado cuando uno dice, basado en la evidencia real, que la teoría de un universo permanentemente estable es menos probable que la teoría del Big Bang o que la culpabilidad de un acusado está probada más allá de una duda razonable no es completamente cierta. Qué tan probable es una hipótesis, dada una evidencia, determina el grado de creencia que es racional tener en esa hipótesis, si toda esa evidencia que uno tiene es relevante para ella [57].

Sivia [58] discute sobre cómo la definición frecuentista de la probabilidad más que ser objetiva, esconde dificultades mayores y que en términos generales va en contravía del quehacer científico. Nadie parte en ciencia de un desconocimiento total ni ejecuta experimentos en forma repetida, por ejemplo.

3. Interpretación subjetivista. Esta es una versión más débil de la interpretación lógica. Es más del tipo sicológico que lógico. El grado de creencia es el concepto fundamental de la interpretación: las afirmaciones probabilísticas representan los grados de creencias de los individuos (estos no son más que individuos idealizados).

Una característica distintiva de la estadística bayesiana es que tiene en cuenta de forma explícita la información previa y se involucra en el análisis en forma de distribución, llamada distribución a priori. La teoría clásica la considera básicamente para determinar tamaños muestrales y el diseño de experimentos y, a veces, como forma de crítica de los resultados hallados.

La expresión de la información previa en forma cuantitativa puede ser un proceso complejo y delicado, aunque se han hallado soluciones que pueden llegar a parecer extrañas, como lo puede ser el uso de lo que se conoce como distribuciones no informativas, pero que se utilizan extensamente en el trabajo bayesiano aplicado.

Fuentes tradicionales para la construcción de la distribución a priori son:

- *Estudios previos similares.* La utilización de estudios previos sobre unos pocos parámetros específicos ha dado origen a un área conocida como *metaanálisis*, la cual puede trabajarse desde el punto clásico y bayesiano.
- *Opinión de expertos.* La utilización de expertos es casi obligatoria en situaciones completamente nuevas donde experimentar puede ser muy costoso o imposible, por ejemplo en la implementación de políticas a nivel macroeconómico.

Wallsten y Budescu [59] presentan las condiciones para que un proceso de elicitación produzca una distribución apropiada. En teoría de la medición se menciona con frecuencia las dificultades que tienen los procesos de cuantificación de sentimientos en relación con eventos externos y la determinación del error. El proceso de codificación debe garantizar condiciones básicas, como confiabilidad y validez. La confiabilidad

se mide como la correlación al cuadrado entre los valores observados de las probabilidades y las verdaderas probabilidades, $\rho^2_{p\pi}$. La probabilidad subjetiva es una variable aleatoria, p, que puede ser descompuesta como la verdadera probabilidad fija π y un error, e:

$$p = \pi + e.$$

Los siguientes supuestos son estándares para este modelo:

1. $E(e) = 0$
2. $\rho_{\pi e} = 0$
3. Para cualquier par de mediciones independientes los errores son incorrelacionados: $\rho_{e_i e_j} = 0$ para $i \neq j$
4. $\rho_{\pi_i e_j} = 0$ para $i \neq j$

Sea σ^2_e la varianza del error. La raíz cuadrada de esta cantidad se conoce como el error estándar de la medición. Del modelo y los supuestos anteriores se tiene:

$$\sigma^2_p = \sigma^2_\pi + \sigma^2_e$$

Así, el coeficiente de confiabilidad será:

$$0 \leq \rho^2_{\pi p} = \frac{\sigma^2_\pi}{\sigma^2_p} = 1 - \frac{\sigma^2_e}{\sigma^2_p} \leq 1$$

La confiabilidad está inversamente relacionado con el error de medición y es perfecta cuando $\sigma^2_e = 0$. Ya que π no puede ser observado directamente, no podemos determinar σ^2_π. Esto puede resolverse parcialmente a través de métodos indirectos, por ejemplo, usando varios métodos de cuantificación.

La validez se define como la correlación entre dos procedimientos de cuantificación independientes, digamos ρ_{xy}.

Ayyub [60] presenta una clasificación de la ignorancia que es importante considerar cuando se determina la claridad de un experto. La ignorancia puede ser consciente o ciega. La ignorancia ciega incluye conocimiento irrelevante que puede estar conformado por un conocimiento relevante y que es descartado o no considerado intencionalmente y por un conocimiento no confiable (prejuicios) o que no aplica al problema de interés.

Un elicitador subjetivo está bien calibrado si para cada probabilidad p, en la clase de todos los eventos en los cuales asigna una probabilidad subjetiva, la frecuencia relativa de ocurrencia es igual a p.

A pesar de que el concepto anterior es atractivo, en la práctica puede ser difícil de verificar, excepto en ciertas circunstancias donde el elicitador permanentemente

construye distribuciones de probabilidad subjetivas, como es el caso de los meteorólogos. En estos caso se pueden construir pruebas de tipo estadístico para medir el nivel de concordancia de lo elicitado con lo observado. A nivel experimental se pueden establecer ambientes controlados en los cuales los elicitadores pueden ser evaluados, sin embargo no hay garantía que el resultado de un laboratorio pueda extrapolarse a un ambiente real.

1.3. Aproximaciones al análisis bayesiano

Una clasificación de las diversas aproximaciones que podemos realizar cuando consideramos el enfoque bayesiano es la siguiente [61]:

1. *Análisis bayesiano objetivo:* esta posición se caracteriza por la utilización de distribuciones no informativas.

2. *Análisis bayesiano subjetivo:* la utilización de distribuciones a priori subjetivas es a menudo disponible como alternativa en algunos problemas.

3. *Análisis bayesiano robusto:* esta posición asume que es imposible especificar completamente la distribución a priori o el modelo, en cuyo caso es mejor trabajar dentro de clases donde haya un nivel de incertidumbre sobre esta distribución o modelo.

4. *Análisis bayesiano-frecuentista:* hay problemas en los cuales la aproximación frecuentista produce resultados satisfactorios, como en los métodos no paramétricos, y al bayesiano le toca aceptarlos como soluciones seudobayesianas.

5. *Análisis cuasibayesiano:* esta aproximación utiliza distribuciones a priori seleccionadas de una forma que acomoden a la solución «bonita»del problema, ajustando estas distribuciones a priori de diversas formas, por ejemplo seleccionando distribuciones a priori vagas, o ajustando los parámetros.

1.4. Problemas con la aproximación clásica

Desde otras disciplinas puede parecer extraño el mundo estadístico como lo expresa el físico Loredo [62].

> Para un extraño, la estadística puede tener la apariencia de ser una simple 'industria' donde métodos estadísticos son inventados sin un claro criterio racional, y luego son evaluados por una masa de datos simulados y se analiza el comportamiento promedio a largo término. Como resultado, a menudo hay varios métodos disponibles para un asunto estadístico particular, y cada uno da una respuesta algo diferente de los otros, sin ningún criterio determinante para escoger entre ellos.

Esta queja es válida y permanente, como más adelante lo ilustramos con referencia a la construcción de intervalos de confianza para algunos problemas típicos.

Sawyer y Peter [63] señalan:

«Debido a que los investigadores toman muchas decisiones que pueden influenciar enormemente la probabilidad de rechazar la hipótesis nula, es equivocado considerar el proceso de pruebas de significancia estadística como objetivo solamente debido a la objetividad de las matemáticas».

La estadística clásica (Fisher y Neyman-Pearson) ha utilizado como bandera en contra de la estadística bayesiana el concepto de objetividad, aunque Savage [10] discute ampliamente este hecho y muestra cómo la única forma de hacer una estadística coherente es vía probabilidades subjetivas. Sawyer y Peter [63] ilustran diversas partes del proceso de probar estadísticamente una hipótesis que pueden generar problemas entre los usuarios, usualmente debido a interpretaciones equivocadas (valores p, tamaño de la prueba y potencia) o a problemas técnicos involucrados en el proceso (tamaños muestrales o diferencias significativas desde el punto de vista práctico). Labovitz [64], Sawyer y Peter [63], Lecoutre, Lecoutre y Poitevineau [65] y Harrell [66] han presentado algunos de los problemas que ocurren con la aproximación clásica a varios problemas estadísticos. Quizás un factor principal que contribuye al valor percibido de las pruebas de significancia estadística es la ilusión de que ellas son completamente objetivas. La probabilidad de rechazar la hipótesis nula es una función de cinco factores: si es una prueba de una o de dos colas, del nivel, de la desviación estándar, del tamaño de la desviación verdadera y del número de observaciones.

1. En pruebas de hipótesis:

   - Un experimento proporciona los elementos para una posibilidad de rechazar la hipótesis nula. Considere el siguiente ejemplo presentado por Berger y Wolpert [67]:

     Ejemplo 1.8. Berger y Wolpert [67]. Suponga que $X = 1, 2$ ó 3 y $\theta = 0, 1$, con $P_\theta(x)$ dada en la siguiente tabla

Hipótesis	1	2	3
P_0	0.009	0.001	0.99
P_1	0.001	0.989	0.01

     La prueba, que aceptaría P_0 si $x = 3$ y a P_1 en otro caso, es una prueba más potente con ambas probabilidades de error igual a 0.01. Luego, desde el punto de vista frecuentista se puede afirmar que el tamaño de la prueba es 0.01. Esto parece equivocado, ya que si se observa $x = 1$, la razón de verosimilitud es 9 a 1 en favor de P_0, que con la prueba se estaría rechazando. Esto nos permite presentar el principio establecido por ellos:

     Principio de aptitud (adequacy). Un concepto de evidencia estadística es (muy) inepto si no distingue evidencia de (muy) diferentes fortalezas.

   - El rechazo de una hipótesis nula es diferente de su rechazo lógico.

- Una hipótesis contradicha por los datos (un valor-p pequeño) significa que un evento improbable ha ocurrido, o que la hipótesis nula es falsa, o ambas. Un ejemplo que aparece en la literatura [20] y que se debe a Lindley y Phillips es el siguiente: asuma que se desea verificar la hipótesis que una moneda tiene una probabilidad de caer en cara $H_0 : \pi = 1/2$ contra la alternativa $H_1 : \pi > 1/2$. Se tienen 12 lanzamientos, de los cuales 9 corresponden a caras. El problema que surge es que el modelo de muestreo que genera los datos se vuelve determinante en el cálculo del valor-p. Si el modelo es binomial (n se fija de antemano) el valor-p se calcula como

$$P_{H_0}(X \geq 9) = \sum_{j=9}^{12} \left(\begin{array}{c} 12 \\ j \end{array} \right) \pi^j (1-\pi)^{12-j} = 0.075,$$

  y si el modelo es binomial-negativo (se determina observar 9 caras de antemano y se cuenta el número de fracasos hasta obtener las 9 caras) el valor-p será

$$P_{H_0}(X \geq 9) = \sum_{j=9}^{\infty} \left(\begin{array}{c} 2+j \\ j \end{array} \right) \pi^j (1-\pi)^3 = 0.0325$$

  ¡A un nivel de 0.05 con un esquema no rechazamos H_0 y con el otro sí!
- ¿Qué hacer si la hipótesis nula no es rechazada?
- De acuerdo con Fisher una hipótesis nula nunca es aceptada.
- ¿Cuál estadístico de prueba utilizar?
  - No hay una regla general sobre cuál estadístico de prueba utilizar. Existe un consenso implícito en el uso de prueba de tipo asintótico, que son usadas a la ligera por muchos usuarios sin darse cuenta de las restricciones que estas imponen con respecto a tamaños muestrales. Pawitan [68] presenta una situación donde uno de los estadísticos más usados, el estadístico de Wald, puede fallar inesperadamente en ciertos problemas.
  - Diferentes estadísticos pueden llevar a diferentes conclusiones del mismo análisis. Con datos categóricos se trabajan muchas veces pruebas asintóticas usando χ^2 de Pearson o vía LRT (Likelihood ratio test) que también se distribuye χ^2, con resultados similares muchas veces, pero otras veces contradictorios. Savage [69] realza la falta de integridad de la estadística clásica.
  - Se pueden obtener conclusiones inconsistentes de manera lógica, por ejemplo colapsando tablas de contingencia y realizando pruebas χ^2.
- En la teoría de Neyman-Pearson una prueba estadística de hipótesis (H_0) no está sola sino contra teorías competidoras (H_1). Se pueden cometer dos tipos de errores y la idea es tener probabilidades de ambos errores tan pequeñas como sean posibles. El problema es de interpretación: ¿Qué significa aceptar o rechazar?

- En ambas escuelas no hay probabilidades de que las teorías sean correctas.
- Problemas con los valores-p:
  - Solo pueden ser utilizados como evidencia contra una hipótesis, no proporcionan evidencia a favor de una hipótesis. Un ejemplo extremo es el siguiente: Queremos verificar, para el caso de una población Poisson, $H_0 : \lambda = 1$ versus $H_0 : \lambda = 1.5$. Suponga una muestra de tamaño 100 y una media muestral $\bar{x} = 5$. El valor-p en este caso es prácticamente cero, pero bajo la alternativa también es cero!
  - Valores-p iguales no proporcionan igual evidencia acerca de una hipótesis [70].
  - Si usamos valor-$p < 0.05$ como un evento binario, la evidencia es mayor en estudios más grandes.
  - Si usamos el valor-p real, la evidencia es mayor en estudios más pequeños.
- Muchos resultados pueden ser estadísticamente significativos debido a un n grande y no a una diferencia significativa desde un punto de vista práctico.

Rosenkranz [71] propone construir algunos indicadores para percibir mejor los resultados de pruebas de hipótesis, en especial el valor-p, pero en el fondo termina con una interpretación bayesiana.

2. En estimación:

   - Los intervalos de confianza son a menudo malinterpretados. La mayoría de los usuarios tienden a pensar que el nivel de confianza es una forma de probabilidad y es a veces frustrante cuando se le dice al usuario que el verdadero parámetro está en el intervalo obtenido con probabilidad uno o con probabilidad cero. Solo la interpretación probabilística del nivel se le puede dar a un proceso imaginario de muestreo repetitivo (que nunca se realizará).

Winkler [72] presenta argumentos similares en el área de la salud. Sus puntos son:

1. *Los métodos frecuentistas tienden a responder las preguntas equivocadas.* Por ejemplo, en lugar de buscar la probabilidad que el efecto de un tratamiento sea positivo o que esté por encima de cierto nivel, ellos dan la probabilidad de los datos, dado los valores del efecto del tratamiento. Las interpretaciones de intervalos de confianza son realizadas con frecuencia en términos probabilísticos.

2. *Los métodos frecuentistas violan a menudo principios importantes, principalmente el principio de verosimilitud.* La función de verosimilitud da la posibilidad del valor observado dados diferentes valores del parámetro[1], y el principio de verosimilitud significa que cualquier otro resultado diferente al observado es irrelevante para el proceso inferencial. Si se observa el valor p, sin embargo, este involucra las verosimilitudes de otros resultados diferentes a los observados y por lo tanto viola el principio de verosimilitud.

[1] Bayarri, DeGroot y Kadane [73] presentan una discusión extensa sobre este concepto y algunas variantes.

3. *Los resultados de los análisis frecuentistas a menudo son simplistas y pueden confundir.* Un ejemplo típico es el uso generalizado de valores tales como 0.05 para α, la probabilidad del error de tipo I en pruebas de hipótesis. Tales valores son usado a menudo sin pensar seriamente acerca de la seriedad de los dos tipos de error y sin ninguna consideración de la probabilidad del error tipo II, que consiste en no rechazar una hipótesis nula falsa.

4. *Parte del problema es del abuso del marco de prueba de hipótesis.* Muchos problemas son forzados a caer en el marco de la dicotomía acepto-rechazo que pudiera no ser apropiado.

5. *La salida de los análisis frecuentistas no son muy útiles en el proceso de toma de decisiones.* Lo que usualmente se necesita en un proceso de toma de decisiones son probabilidades, bien sea probabilidades de futuros resultados (probabilidades predictivas) o probabilidades de valores de los parámetros (probabilidades posteriores). La estadística clásica no admite tales probabilidades.

6. *Los métodos frecuentistas se pueden aplicar demasiado fácilmente sin pensar.* Es notable el uso extenso de pruebas de hipótesis realizadas a modelos que los programas estadísticos sacan por defecto y que el usuario añade a sus análisis sin que realmente se tenga una idea clara de la justificación o relevancia de las mismas. Muchos programas estadísticos producen demasiados resultados sin que el usuario los solicite (un ejemplo típico es el PROC UNIVARIATE de SAS, el cual produce por defecto una prueba de normalidad) y que terminan siendo interpretados, a veces de forma equivocada.

   Una línea de trabajo que se ha desarrollado últimamente es el desarrollo de programas de análisis inteligentes de datos. La idea es que ante el acceso a megabases de datos, el usuario tenga un conjunto de rutinas que hagan los análisis. Esto tiene su lado positivo, pero no le quita al analista la responsabilidad de entender qué es lo que el programa hace en su interior con sus datos.

7. *Contrario a lo que se dice, los métodos frecuentistas no son objetivos.* La forma en que el problema es planteado, la manera en que los datos son recolectados, y los diferentes supuestos de los modelos (por ejemplo, la selección de un modelo lineal, un conjunto particular de variables explicativas, la distribución normal para los errores) son selecciones subjetivas. Las distribuciones muestrales y las funciones de verosimilitud no están en algún lugar para que sean usadas; ellas son resultados de las selecciones de los modelos hechas por el analista.

Diferencias entre la teoría clásica y la teoría bayesiana

Característica	*Teoría clásica*	*Teoría bayesiana*
Parámetros de interés	Constantes desconocidas	Variables aleatorias
Distribución a priori	No existe	Existe y es explícita
Modelo muestral	Se asume	Se asume
Distribución posterior	No existe	Existe y se deriva
Razonamiento	Inductivo	Deductivo

Utilizaremos la siguiente notación:

$\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$	Vector de parámetros
$x_1, \ldots, x_n$	Observaciones muestrales (i.i.d.)
$\xi(\boldsymbol{\theta})$	Distribución a priori conjunto de $\boldsymbol{\theta}$
$f(x_i\|\boldsymbol{\theta})$	Distribución de x_i dado $\boldsymbol{\theta}$

Capítulo 2

Probabilidad subjetiva «a priori»

El trabajo estadístico descansa en el concepto de probabilidad. La definición matemática es clara: es una función aditiva no negativa, cuyo máximo valor es la unidad [74]. El problema fundamental está en la forma como se determine esa función. Ashby [75] comenta «tres interpretaciones se le pueden dar a las distribuciones a priori: como distribuciones de frecuencia basadas quizá en datos previos, como representaciones normativas y objetivas de lo que es racional creer acerca de un parámetro o como una medida subjetiva de los que un individuo particular realmente cree».

2.1. Clasificación de las distribuciones a priori

$$\text{Distribuciones a priori} = \begin{cases} \text{Propias} \\ \text{Impropias} \end{cases}$$

Definición 2.1 (Distribución a priori propia)**.** *Es una distribución que asigna pesos no negativos y que suman o integran hasta uno, a todos los valores posibles del parámetro.*

Así, una distribución propia satisface las condiciones de función de densidad de probabilidad. Una distribución impropia es la que suma o integra a un valor diferente de uno, digamos K. Si K es finito, entonces la distribución impropia induce una distribución propia normalizando la función. Si K es infinito, entonces la distribución tiene un papel de ponderación o de herramienta técnica para llegar a una distribución posterior.

$$\text{Distribuciones a priori} = \begin{cases} \text{Informativas} \\ \text{No informativas} \end{cases}$$

Definición 2.2 (Distribución a priori no informativa)**.** *Decimos que una distribución a priori es no informativa cuando refleja una ignorancia total o un conocimiento muy limitado sobre el parámetro de interés.*

$$\text{Distribuciones a priori} = \begin{cases} \text{Conjugadas} \\ \text{No conjugadas} \end{cases}$$

Definición 2.3 (Distribución a priori conjugada)**.** *Decimos que una distribución a priori es conjugada, si al proceder a su actualización mediante la información muestral, la distribución a posteriori es igual a la a priori, excepto en los hiperparámetros, es decir, en parámetros distintos al los del modelo muestral.*

2.2. Distribuciones a priori no informativas

En muchas ocasiones sabemos nada o muy poco acerca del parámetro de interés o no queremos involucrar en nuestro estudio información previa, sino más bien dejar que sean los datos los que «hablen por ellos mismos». En este caso la distribución debe reflejar nuestro total desconocimiento de los valores posibles del parámetro. Esta es un área de trabajo que ha crecido enormemente.

Bernardo y Ramón [76] señalan:

> Desde un punto de vista *fundamental*, la derivación de una posterior no subjetiva deberá verse como una parte de un *análisis de sensibilidad* necesario, diseñado para analizar los cambios en la posterior de interés inducidos por los cambios en la a priori: una posterior no subjetiva trata de dar una respuesta a una pregunta sobre qué puede decirse acerca de una cantidad de interés, si el conocimiento a priori de uno estuviera dominado por los datos. Cuando la información subjetiva a priori es especificada, la correspondiente posterior pudiera ser comparada entonces con la a posteiori no informativa para determinar la importancia relativa de las opiniones iniciales en la inferencia final. José Bernardo ha sido un gran investigador junto con sus colaboradores en el área de aprioris no informativas.

Un argumento adicional presentado por Bernardo y Ramón [76] sobre la bondad de usar aprioris no informativas es que la automatización de procesos bayesianos en software estadístico que utiliza ciertos procesos numéricos, estilo MCMC, se vuelven mucho más complejos e imprácticos si no se usan aprioris no informativas.

Evans [77] afirma:

> No es claro para mí por qué los bayesianos son tan insistentes con las aprioris ignorantes. Al menos en ingeniería, la idea debe ser cuantificar lo que el ingeniero conoce, no lo que él no sabe. Muchos artículos en bayesiana comienzan con una o más aprioris ignorantes. Si usted no sabe nada de antemano, qué tiene la bayesiana para ofrecer? Igualmente, si los datos van a tumbar la a priori, qué tiene la bayesiana para ofrecer?...En una situación en ingeniería, rara vez hay una excusa para escoger una a priori ignorante. Si los ingenieros no tienen una buena idea razonablemente de cómo las cosas pueden resultar, todos ellos deben ser despedidos y cada línea de producción debió pararse hace tiempo.

2.2.1. Distribuciones a priori informativas

Una de las mayores dificultades en la ejecución de un análisis bayesiano concierne con la identificación, de la selección y la justificación de la distribución a priori. Las siguientes preguntas deben ser resueltas sin lugar a dudas:

- ¿Qué clase de distribución a priori debemos utilizar?
- ¿Qué tipos de datos están disponibles para seleccionar el modelo a priori?
- ¿Cómo cuantificamos la información subjetiva?
- ¿Cómo ajustamos la distribución a priori con los datos subjetivos disponibles?

2.2.2. Probabilidad personal

Horowitz [78] define la probabilidad como:

> La probabilidad no es sino un número índice entre 0 y 1, que expresa un pensamiento del individuo sobre la posibilidad del resultado, relativo, de una experiencia. Debemos por tanto, reconocer que podemos evaluar la probabilidad, bien cuando el suceso es único o se trata de un suceso de carácter repetitivo, que pueda presentarse en varias pruebas. El hecho de que el suceso vaya a ocurrir una vez no impide que un individuo pueda formar un juicio acerca de lo probable que suceda respecto a otros posibles resultados; es decir, puede asignar probabilidades a cada uno de los posibles resultados.

Para Poirier [79]:

> Para un subjetivista, la probabilidad es interpretada como un grado de creencia fundamentalmente interna de un individuo como opuesto a alguna característica del mundo externo. La probabilidad subjetiva mide una relación entre el observador y los eventos (no necesariamente 'repetitivos') del mundo exterior, expresando la incertidumbre personal del observador acerca de esos eventos.

2.3. Probabilidad subjetiva, apuestas y loterías

Shafer [80] dice:

> Laplace, escribiendo a comienzos del siglo diecinueve, definió la probabilidad numérica como la medida de la 'razón que tenemos de creer.' Pero a mediados del siglo diecinueve, muchos estudiante de la probabilidad estaban buscando una definición más empírica. Ellos hallaron esta definición en la idea de la frecuencia, y ellos procedieron a rechazar aquellas aplicaciones de la teoría de la probabilidad que no pudiera estar basada en las frecuencias observadas. En particular, rechazaron el método de Laplace de calcular la probabilidad de las causas, que es un caso especial de la estrategia de probabilidad condicional.

> La filosofía frecuentista restringe severamente el dominio de aplicaciones de la probabilidad numérica, y aquellos quienes querían usar probabilidades numéricas más generalmente fueron forzados a buscar una fundamentación filosófica para la estrategia de la probabilidad condicional que ajustaría la posición positivista. Tal fundamentación filosófica fue finalmente establecida en el siglo veinte por Ramsey, de Finetti y especialmente Savage. Estos autores concibieron la idea que a la probabilidad subjetiva se le debe dar una interpretación de comportamiento y por lo tanto positivista. Las probabilidades de una persona deben ser deducibles de sus elecciones. Ellos formulan postulados para lo que llaman un comportamiento racional, postulados que aseguran que las elecciones de una persona determinan probabilidades numéricas. Y ellos argumentan que es normativo que se sigan estos postulados y por lo tanto normativo tener probabilidades subjetivas. Poirier [79] afirma que los frecuentistas interpretan probabilidad como una propiedad del mundo externo.

Cox [81] afirma:

> En una discusión bayesiana completamente personal, de la forma que lo entiendo, las afirmaciones probabilísticas no pueden estar equivocadas: ellas son lo que son y asumiendo que son consistentes, esto es, consistentes internamente, que es todo lo que se requiere. A un nivel más pragmático pudieran presentarse dificultades si los datos están en conflicto con *cualquier* modelo de las formas propuestas o si los datos y la a priori discrepan. Así una posible muestra de una distribución Poisson podría tener una varianza aproximadamente igual a su media pero una media bien metida en las colas de la distribución a priori propuesta. Eso podría significar que los datos tienen algún error sistemático o que la a priori está basada en conceptos errónes (o ambos). Si la a priori se supone inocua, o seleccionada con una mirada soslayada de los datos, esta dificultad no aparece. Sin embargo, como una atracción del formalismo bayesiano, la a priori es un intento serio de incluir información adicional, el tema de la consistencia de la a priori y los datos parece en principio importante y rara vez discutida. En tanto la a priori pueda considerarse como seudo-datos, los chequeos son ciertamente posibles, aunque ellos están por fuera de la formulación bayesiana.

Las creencias pueden ser expresadas en términos de apuestas, esto se hace mucho en la práctica, y esto puede ser utilizado como una forma general de hacer las creencias relativas explícitas [82]. Hay condiciones naturales a ser impuestas sobre las apuestas:

- La apuesta debe ser reversible y que ninguna apuesta pueda ser elaborada tal que uno pierda o gane con certeza. Esta condición obliga al sujeto a asignar las apuestas consistentemente con sus creencias.

- La anterior también obliga a aceptar la segunda condición: una vez él ha fijado los odds, él debe estar preparado para apostar en cualquier dirección. Esta coherencia juega dos papeles importantes:

1. Es moral y obliga a la gente a ser honesta.
2. Otro formal, que permite que las reglas básicas de la probabilidad sean derivadas como teoremas.

Una probabilidad puede pensarse en términos de las cantidades que se involucren en una apuesta y la porción que el sujeto estuviera dispuesto a arriesgar. Por ejemplo, si usted dice, «Las posibilidades que el equipo de fútbol A gane el torneo son de uno entre 10», esto implica que,

$$P\left(\text{A gane el torneo}\right) = \frac{1}{10},$$

y

$$P\left(\text{A no gane el torneo}\right) = \frac{9}{10}$$

Así, para el equipo A es nueve veces más posible que no gane el torneo, a que gane el torneo, según su opinión. En términos de apostadores se dice que las apuestas están 9 a 1. Esto en otras palabras significa que si usted apuesta un peso por el evento «A gana el torneo», si este evento ocurre, usted ganaría nueve pesos.

Es evidente que «subjetividad»no se puede confundir con «arbitrariedad», ya que todos los elementos para la asignación de probabilidades deben tenerse en cuenta, incluyendo el conocimiento que otros puedan asignar a las diferentes posibilidades de los mismos eventos.

Winkler [12] presenta el siguiente ejemplo sobre como funcionaría un proceso de elicitación usando loterías (la construcción formal desde un punto de vista constructivista de la definición de probabilidad subjetiva mediante el uso de loterías ha sido desarrollada por Anscombe y Aumann [83]):

> Suponga que una persona debe escoger entre la Lotería A y la Lotería B.
>
> *Lotería A* La persona gana \$100 con probabilidad $\frac{1}{2}$.
> La persona gana \$0 con probabilidad $\frac{1}{2}$.
>
> *Lotería B* La persona gana \$100 si llueve mañana.
> La persona gana \$0 si no llueve mañana.
>
> Dado que el premio es el mismo en ambas loterías, la persona preferiría la lotería que le dé la mayor posibilidad de ganar el premio. Así, si el individuo escoge la Lotería B, entonces debe sentir que la probabilidad que llueva mañana es mayor que $\frac{1}{2}$; si esta escoge la Lotería A, entre las dos loterías, entonces sentiría que esta probabilidad es menor que $\frac{1}{2}$; si la persona es indiferente entre las dos loterías, entonces siente que la probabilidad que llueva mañana es igual a $\frac{1}{2}$. Ahora, consideremos las mismas loterías, excepto que las probabilidades en la Lotería A se cambiaron a $\frac{1}{4}$ y $\frac{3}{4}$. Si el individuo aún prefiere la Lotería A a la B, lo que implica que siente que tiene una mayor oportunidad de ganar con A que con B, entonces su probabilidad subjetiva de lluvia es menor que $\frac{1}{4}$. Presumiblemente la persona puede estar cambiando las probabilidades en

la Lotería A hasta que sea indiferente a la selección entre la Lotería A y la Lotería B; si esto sucede cuando sus probabilidades sean 0.1 y 0.9, entonces su probabilidad subjetiva de lluvia es 0.1. De forma similar, puede determinar su probabilidad subjetiva de cualquier evento.

Una definición formal de probabilidad subjetiva puede darse en términos de loterías como sigue. Su probabilidad subjetiva $P(E)$ del evento E es el número $P(E)$ que hace que el interesado esté indiferente entre las siguientes dos loterías.

Lotería A	La persona gana X con probabilidad $P(E)$. La persona gana Y con probabilidad $1 - p(E)$.
Lotería B	La persona gana X si ocurre E. La persona gana Y si no ocurre E.

Aquí X y Y son dos «premios». La única restricción sobre X y Y es que uno debe ser preferido sobre el otro; si el individuo es indiferente entre X y Y, entonces será indiferente entre las dos loterías sin importar la escogencia de $P(E)$.

Un problema con esta forma de elicitación es que es altamente demandante tanto para el elicitador como para la persona que está proporcionando la información. Esto genera agotamiento y a la larga la persona elicitada termina dando información a la ligera [84].

Un problema que ha sido señalado para las apuestas es que realmente se asume que las personas involucradas son "agentes racionales" [85] cuyo fin es maximizar su propia función de utilidad[1], que por simplicidad se asume lineal con respecto al dinero involucrado. Obviamente esto no es cierto para cualquier rango de dinero, ya que no todo el mundo tiene la misma percepción sobre el valor del dinero a ciertos

[1] Una función de utilidad U es una función de valor real definida en R si tiene la siguiente propiedad: Sean P_1 y P_2 dos funciones de probabilidad tal que $E_{P_1}[U]$ y $E_{P_2}[U]$ existan. Entonces P_1 será a lo más tan preferido como P_2, denotado por $P_1 \precsim P_2$, si, y solo sí, $E_{P_1}[U] \leq E_{P_2}[U]$. Para cada premio $r \in R$, el número $U(r)$ es llamada la utilidad de r. El $E_P[U]$ se conoce como la utilidad de P. Los siguientes dos resultados son consecuencias de la definición de función de utilidad:

1. Si r_1 y r_2 son dos premios en R, entonces $r_1 \precsim r_2$, si, y solo sí, $U(r_1) \leq U(r_2)$.
2. Si $\wp$ es el conjunto de distribuciones de probabilidad de interés, entonces las funciones de probabilidad se pueden comparar. Esto es, podemos ordenarlas.

Los siguientes supuestos son establecidos en teoría de la utilidad:

- Si P_1, P_2 y P son distribuciones en la clase de referencia $\wp$ y si α es un número cualquiera en $(0, 1)$. Entonces $P_1 \prec P_2$ si, y solo sí, $\alpha P_1 + (1 - \alpha)P \prec \alpha P_2 + (1 - \alpha)P$.
- Si P_1, P_2 y P son distribuciones en la clase de referencia $\wp$ tal que $P_1 \prec P \prec P_2$. Entonces existen números α y β en $(0, 1)$ tal que $P \prec \alpha P_2 + (1 - \alpha)P_1$ y $P \succ \beta P_2 + (1 - \beta)P_1$.

Un resultado importante que permite justificar la construcción de distribuciones a priori vía loterías es el siguiente: Sean r_1, r_2 y r tres premios cualesquiera en R tal que $r_1 \prec r_2$ y $r_1 \precsim r \precsim r_2$. Entonces existe un único número v $(0 \leq v \leq 1)$ tal que $r \sim v\, r_2 + (1 - v)r_1$.

niveles. Un ejemplo es si una lotería involucra cantidades en miles de pesos, las personas actúan diferente a si se refieren a millones de pesos. Aun para dos personas, la misma cantidad puede tener utilidades diferentes, suponga el costo de una bolsa de leche, que puede ser muy marginal para mucha gente, pero esta cantidad puede apreciarse diferente desde el punto de vista de una madre pobre con un bebé a quien alimentar.

Otra alternativa similar consiste en el uso de un *contrato de referencia* [12]. En esta situación se asume la existencia de un premio muy atractivo que depende de la ocurrencia de un evento de interés. Por ejemplo, suponga que a un individuo le ofrecen un contrato que le paga un millón de pesos si su carro es robado en menos de un año (asumamos que su carro cuesta esta cantidad). Cuál es la máxima cantidad de dinero que el individuo estaría dispuesto a pagar por este contrato? Si él está dispuesto a pagar cien mil pesos, entonces su probabilidad subjetiva es $100.000/1.000.000 = 0.1$. Este argumento es simular al caso de las apuestas y la idea detrás es la maximización de la utilidad del individuo. Kadane y Winkler [86] discuten la relación entre elicitación de probabilidades y utilidades y otras alternativas.

Capítulo 3

Análisis preposterior

Dadas las características del proceso bayesiano, es sano tener una posición crítica con relación a cualquier a priori que obtengamos por cualquier método, deberíamos realizar pruebas que permitan determinar de alguna manera la calidad de la distribución a priori. Martz y Waller [33] recomiendan lo siguiente para garantizar la realización de un buen análisis bayesiano:

1. Una justificación y análisis detallado de la distribución a priori seleccionada, con un claro entendimiento de las implicaciones matemáticas de la a priori. Por ejemplo, la selección de una familia normal para representar nuestro conocimiento a priori sobre un parámetro nos restringe a una clase de distribuciones unimodales.

2. Una documentación completa de las fuentes de datos utilizados en la identificación y selección de la a priori.

3. Un análisis preposterior de la distribución a priori con resultados de prueba hipotéticas.

4. Una distribución a posteriori claramente definida para los parámetros de interés.

5. Un análisis de sensibilidad de las inferencias bayesianas para el modelo a priori seleccionado.

Cada uno de estos puntos debe desarrollarse cuidadosamente tanto en el proceso de elicitación como en el de validación de la información recogida. El segundo paso de la lista de Martz y Waller es ejecutado casi simultáneamente en la misma elicitación. Al experto se le alienta a dar respuestas, pero el elicitado debe proporcionar una retroalimentación oportuna, lo que en términos prácticos significa que es inmediata. Para llevar a cabo esta tarea, se puede generar muestras de la población bajo diferentes condiciones planteadas en la a priori, por ejemplo, podemos tomar la mediana, la media, la moda y cuartiles para generar estas muestras. Si el sujeto se siente cómodo con los valores que las muestras están presentado y él cree que no aparecen valores muy raros, o no observa valores que considera deberían estar presentes en la población, esto no obliga a revisar la elicitación.

Es necesario documentar detalladamente el proceso en sí mismo, bien sea con grabaciones o registro fílmico, para su posterior análisis. Cuando la distribución a priori es obtenida a partir de estudios o datos previos, es necesario justificar plenamente la pertinencia de esos estudios o datos en el problema que estamos enfrentando. Un estudio realizado sobre consumo de drogas en Medellín-Colombia puede no ser útil para uno que se realice en alguna ciudad del exterior, pero ser pertinente para un estudio similar en Cali-Colombia. En otros casos puede darse la pertinencia debido a que se estudian fenómenos biológicos que puedan de alguna manera ser generalizados a poblaciones de muchos lugares, por ejemplo, si queremos determinar la duración promedio de un embarazo.

La selección de la distribución a priori puede realizarse de muchas maneras, unas más fáciles que otras. De cualquier manera es recomendable usar varios procedimientos para su elicitación, los cuales deben coincidir de alguna manera. La distribución elicitada a mano alzada debería ser muy parecida a algún modelo paramétrico ajustado, por ejemplo.

El análisis preposterior debe realizarse con la distribución a priori para verificar la confianza de los resultados obtenidos como resultantes del proceso elicitado. Por ejemplo, si se elicita el número promedio de goles realizados en un partido de fútbol, y además se asume que el número de goles marcados en un partido se puede modelar mediante la distribución de Poisson, valores de la probabilidad de observar ciertos marcadores, condicionados en valores tomados de la a priori sobre la media, deben ser consistentes con lo que el sujeto piensa. Este análisis es equivalente a un análisis exploratorio de datos (EDA), en el cual se buscan características y problemas con la distribución a priori especificada. Luego de este análisis uno podría responder afirmativamente a la siguiente pregunta: ¿estaría confiado, bajo el supuesto que no pueda obtener información muestral, para usar esta distribución para realizar todo el trabajo inferencial?

Otro punto que se analiza en esta etapa es la calidad de la a priori. ¿Es creíble? ¿Se puede utilizar fácilmente a nivel computacional? ¿Está completamente definida sobre el espacio parametral? ¿Cambia radicalmente si se obtienen unos pocos datos?

Una de los asuntos más delicados es la obtención de dos o más aprioris que sean muy diferentes por parte de diferentes expertos o por diferentes métodos de elicitación. En caso de diferencias irreconciliables, es sano trabajar paralelamente con varias aprioris. Hoy en día esto no es una gran problema, debido a la disponibilidad de recursos computacionales.

3.1. Distribución predictiva a priori

La construcción de la distribución predictiva a priori es una de las herramientas que tiene el analista para determinar la calidad de la distribución a priori elicitada, además de entregar información que puede ser útil para determinar los procedimientos de análisis posteriores. Si $\xi(\boldsymbol{\theta})$ representa la distribución a priori sobre $\boldsymbol{\theta}$ elicitada

y si se conoce (obviamente hasta cierto nivel) la distribución que genera los datos, siendo $f(y|\boldsymbol{\theta})$ la distribución de las observaciones futuras, entonces la distribución predictiva a priori se define como,

$$p(y) = \int f(y|\boldsymbol{\theta})\,\xi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

Si los datos generados por esta distribución se «acomodan»a lo que el analista cree deben ser, entonces puede, de cierta manera pensar que la a priori concuerda con lo que piensa y proceder al proceso posterior.

Este análisis se puede realizar con diferentes distribuciones a priori de tal forma que se puede mirar la robustez de los resultados ante cambios de la a priori. Observemos que este análisis no es posible de realizarse con distribuciones a priori no informativas.

Ejemplo 3.1. Número promedio de goles en el fútbol colombiano. Supongamos que elicitamos el número promedio de goles que se marcan en un partido del fútbol profesional en Colombia a un experto y obtenemos que $\theta \sim N(2.5, (0.20)^2)$. Además si pensamos que el número de goles que se marcan en un partido del rentado colombiano es $Poisson(\theta)$, la distribución predictiva a priori será

$$p(y) = \int_0^\infty \frac{\theta^y \exp(-\theta)}{y!} \frac{1}{\sqrt{2\pi}0.20} \exp\left(-\frac{1}{2\times 0.20^2}(\theta - 2.5)^2\right) d\theta$$

```
# Simulación de la Distribución Predictiva a priori
# Nro. promedio de goles del campeonato colombiano
# A priori: Normal(2.5, 0.20^2)
# Dist. muestral: Poisson(theta)
# función que muestrea de la predictiva
dist.predictiva<-function(Nsim=10000,media=2.5,dt=0.20){
# genera valores de la Poisson con parámetros de la normal
res<-rpois(Nsim,rnorm(Nsim,mean=media,sd=dt))
# tabla de frecuencias para los valores generados
tabla<-table(res)
print(tabla/sum(tabla))
barplot(tabla)  # diagrama de barras para valores generados
print(summary(res)) # resumen estadístico
print(var(res)) }   # fin de la función

dist.predictiva()
title(main='Distribución predictiva\n campeonato colombiano',
      xlab='Número de Goles', ylab='Frecuencia')
legend(7,2000,c('Media a priori=2.5','Desv. Est. a priori=0.20')
       ,bty='n')
```

Los resultados de la anterior simulación son los siguientes:

Tabla 3.1: *distribución predictiva a priori para el número de goles marcados en el fútbol colombiano*

y	0	1	2	3	4	5	6	7	8	9	10
$p(y)$	0.0849	0.2002	0.2592	0.2101	0.1320	0.0701	0.0288	0.0103	0.0032	0.0010	0.0002

Tabla 3.2: *resumen estadístico de los datos simulados de la predictiva.*

mín	Cuartil 1	Mediana	Media	Cuartil 3	máx	Varianza
0.000	1.000	2.000	2.509	3.000	10.000	2.545

A priori, notamos que la mediana del número de goles para un nuevo partido del rentado colombiano sería de 2 goles.

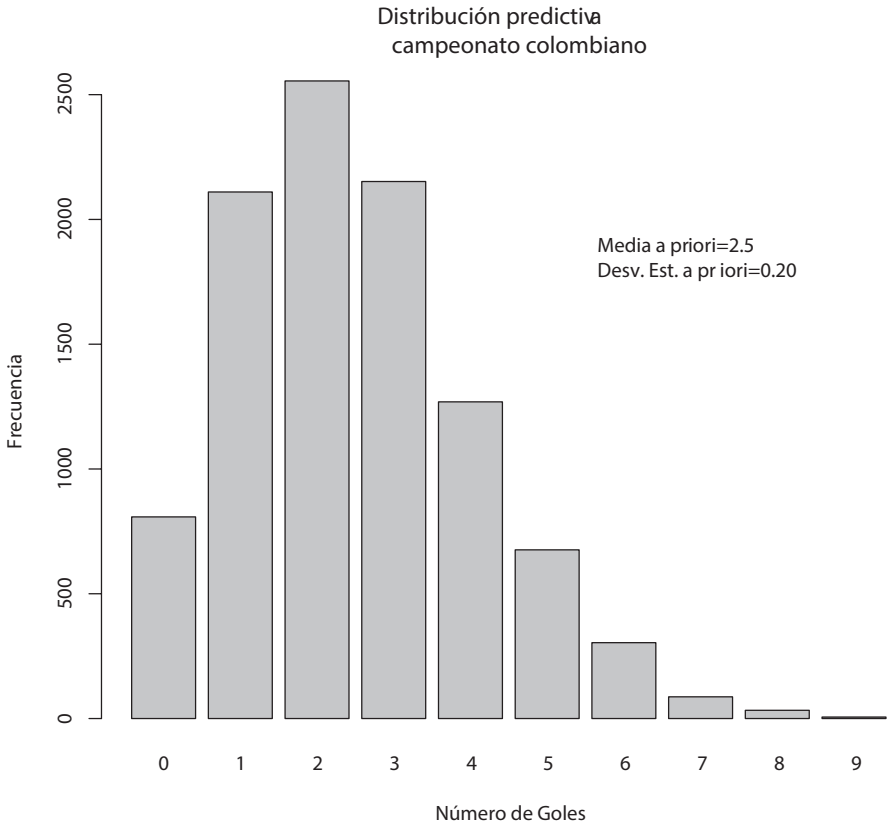

Figura 3.1: *distribución predictiva a priori. Esta distribución no es una distribución Poisson debido a nuestra propia incertidumbre sobre la media del proceso, la cual la expresamos mediante una distribución Normal*

Capítulo 4

Teorema de Bayes

El teorema de Bayes es ahora una de las piedras fundamentales del trabajo estadístico y sigue siendo de cierta discusiones, tanto de sus orígenes como de sus implicaciones filosóficas [87]. Este teorema fue publicado varios años después de la muerte de reverendo Thomas Bayes por un amigo.

Teorema 4.1. (Teorema de Bayes) *Sean $B_1, B_2, \cdots, B_k$ eventos mutuamente excluyentes y exhaustivos. Para cualquier evento nuevo A, tenemos:*

$$P\left(B_i|A\right) = \frac{P\left(B_i \bigcap A\right)}{P(A)} = \frac{P\left(A|B_i\right) P\left(B_i\right)}{\sum_{i=1}^{k} P\left(A|B_i\right) P\left(B_i\right)}, \quad si\ P(A) \neq 0,\ P(B_i) \neq 0,\ i = 1, 2, \ldots, k$$

Prueba: Ejercicio.

Teorema 4.2. (Teorema de Bayes para variables aleatorias) *Sean X y θ variables aleatorias con fdp's $f\left(x|\theta\right)$ y $\xi(\theta)$.*

$$\xi\left(\theta|x\right) = \frac{f\left(x|\theta\right)\xi(\theta)}{\int_{\Theta} f\left(x|\theta\right)\xi(\theta)\, d\theta}$$

Dentro del marco bayesiano tenemos que:

- X : datos (escalar o vector o matriz)
- θ: parámetro desconocido (escalar o vector o matriz)
- $f\left(x_1, \ldots, x_n|\theta\right)$: verosimilitud de los datos dado el parámetro (desconocido) θ.
- $\xi(\theta)$: distribución a priori de θ.

Por el teorema anterior,

$$\xi\left(\theta|x_1, \ldots, x_n\right) = \frac{f\left(x_1, \ldots, x_n|\theta\right)\xi(\theta)}{\int_{\Theta} f\left(x_1, \ldots, x_n|\theta\right)\xi(\theta)\, d\theta}$$

Esta es llamada la **distribución posterior**. La inferencia bayesiana se deriva de esta distribución. En la práctica, el denominador de la expresión anterior no necesita ser calculado en general, y la regla de Bayes se escribe como:

$$\xi\left(\theta|x_1,\ldots,x_n\right) \propto f\left(x_1,\ldots,x_n|\theta\right)\xi(\theta).$$

Por lo tanto, solo necesitamos conocer la distribución posterior hasta una constante de normalización. Muchas veces somos capaces de identificar la distribución posterior de θ mirando solamente este numerador. El teorema de Bayes lo que hace es una «actualización»de $\xi(\theta)$ a $\xi\left(\theta|x_1,\ldots,x_n\right)$.

Nota: El aprendizaje bayesiano será:

$$\begin{aligned}
\xi\left(\theta|x_1\right) &\propto f\left(x_1|\theta\right)\xi(\theta), \\
\xi\left(\theta|x_1,x_2\right) &\propto f\left(x_2|\theta\right) f\left(x_1|\theta\right)\xi(\theta), \\
&\propto f\left(x_2|\theta\right)\xi\left(\theta|x_1\right).
\end{aligned}$$

Por lo tanto el teorema de Bayes nos muestra cómo el conocimiento acerca del estado de la naturaleza representada por θ es continuamente modificada a medida que nuevos datos son adquiridos.

Ejemplo 4.1. Distribución a priori uniforme truncada. Muchas veces somos capaces en un problema binomial de especificar claramente en qué región es imposible que esté el parámetro, pero somos incapaces de especificar mejor nuestro conocimiento sobre él. Podemos pensar en utilizar una distribución a priori que refleje esta ignorancia, para ello consideremos una uniforme truncada, es decir,

$$\pi \sim U\left(\pi_0,\pi_1\right)$$

Esto es,

$$\xi\left(\pi|\pi_0,\pi_1\right) = \frac{1}{\pi_1-\pi_0} \qquad 0 \leq \pi_0 < \pi < \pi_1 \leq 1$$

La distribución posterior de π dado $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ es

$$\xi\left(\pi|\boldsymbol{x},\pi_0,\pi_1\right) = \frac{\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^{(y+1)-1}(1-\pi)^{(n-y+1)-1}}{\int_{\pi_0}^{\pi_1}\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\pi^{(y+1)-1}(1-\pi)^{(n-y+1)-1}d\pi},$$

donde $y = \sum_{i=1}^{n} x_i$.

Notemos que el denominador de la función es la $P\left(\pi_0 < W < \pi_1|y+1, n-y+1\right)$, donde $W \sim Beta(y+1, n-y+1)$, y esto se calcula fácilmente en programas como R.

Es fácil hallar la media y la varianza a posteriori. Ellas son:

$$E\left(\pi|\boldsymbol{x},\pi_0,\pi_1\right) = \frac{y+1}{n+2}\frac{P\left(\pi_0 < W < \pi_1|y+2, n-y+1\right)}{P\left(\pi_0 < W < \pi_1|y+1, n-y+1\right)},$$

y

$$\begin{aligned} Var\left(\pi|\boldsymbol{x},\pi_0,\pi_1\right) &= \frac{(y+2)(y+1)}{(n+3)(n+2)}\frac{P\left(\pi_0<W<\pi_1|y+3,n-y+1\right)}{P\left(\pi_0<W<\pi_1|y+1,n-y+1\right)} \\ &\quad -\left(\frac{(y+1)}{(n+2)}\frac{P\left(\pi_0<W<\pi_1|y+2,n-y+1\right)}{P\left(\pi_0<W<\pi_1|y+1,n-y+1\right)}\right)^2 \end{aligned}$$

Ejemplo 4.2. Aplicación numérica del caso anterior. Suponga que creemos que el porcentaje de mujeres que actualmente estudia en cierta universidad está entre el 35 % y el 70 %, o sea,

$$\begin{aligned} \xi(\pi) &= \frac{1}{0.70-0.35} \qquad \text{para } \pi\in(0.35\,,\,0.70) \\ &= 0 \qquad \text{en otro caso.} \end{aligned}$$

Asumamos, además, que tomamos una muestra al azar de la población de 10 estudiantes y encontramos que 6 son hombres y 4 mujeres, o sea, $y=4$ y $n=10$. El intervalo de confianza clásico (clásico porque la mayoría de los textos básicos es el único que presentan) basado en el teorema central del límite, y a pesar de que el tamaño muestral sea pequeño, se puede aplicar, dada la casi simetría de la distribución poblacional, será:

$$\hat{\pi}\pm 1.96\sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}},$$

lo que produce $(0.0963\,,\ 0.704)$ a un nivel del 95 %.

La aproximación bayesiana nos da una distribución posterior,

$$\xi\left(\pi|n=10,y=4,\pi_0=0.35,\pi_1=0.70\right)=\frac{\frac{\Gamma(12)}{\Gamma(5)\Gamma(5)}\pi^4(1-\pi)^6}{K(0.70\,;\,5\,,\,7)-K(0.35\,;\,5\,,\,7)},$$

donde:

$$K(z;\alpha,\beta)=\int_0^z\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$$

Entonces,

$$E\left(\pi|n=10,y=4,\pi_0=0.35,\pi_1=0.70\right)=0.4823673,$$

y un intervalo de credibilidad del 95 % es $(0.3561442\,,\ 0.6680237)$. Este último se encuentra resolviendo:

$$\int_{\pi_*}^{\pi^*}\frac{\frac{\Gamma(12)}{\Gamma(5)\Gamma(5)}\pi^4(1-\pi)^6}{K(0.70\,;\,5\,,\,7)-K(0.35\,;\,5\,,\,7)}d\pi=0.95,$$

y formando el intervalo $(\pi_*\,,\,\pi^*)$. Note que, este intervalo representa una actualización de la creencia inicial en el problema, en el cual se aprecia una mayor precisión al intervalo inicial.

Definición 4.1 (Consistencia posterior)**.** *La distribución posterior, $\xi(\theta\,|Datos)$, se dice que es consistente en un valor dado θ_0 si para cualquier vecindad V de θ_0, $\xi(\theta \notin V\,|Datos) \to 0$ (en probabilidad) cuando $n \to \infty$, siendo θ_0 el verdadero valor del parámetro.*

La consistencia posterior equivale a decir que bajo condiciones muy generales, no importa qué a priori se use, en el fondo si el tamaño muestral crece indefinidamente, la a priori no tendrá efecto y lo que es fundamental es el experimento que genera los datos, ya que se presupone que es un experimento «insesgado».

4.1. Usos de la función de verosimilitud en análisis bayesiano

Berger, Liseo y Wolpert [88] presenta diferentes usos para la función de verosimilitud, $L(\theta)$:

1. *Reporte científico:* se considera una buena práctica de reporte presentar separadamente $L(\theta)$ y $\xi(\theta|x)$, a menudo gráficamente, para indicar el efecto de la distribución a priori. Esto le permite a otros investigadores utilizar sus propias distribuciones a priori.

2. *Análisis de sensibilidad:* es importante estudiar la sensibilidad a $\xi(\theta)$, y tener disponible $L(\theta)$ para este propósito es valioso.

3. *Costo de elicitación:* obtener distribuciones a priori subjetivas es a menudo muy costoso, tanto en tiempo como en esfuerzo. Es a menudo efectivo a nivel de costos eliminar los parámetros de molestia de una forma básica, produciendo $L(\theta)$, y concentrar la elicitación subjetiva a $\xi(\theta)$.

4. *Objetivismo:* aunque la «objetividad»no se puede garantizar en ningún estudio, el presentar $L(\theta)$ ayuda a darle esta impresión a muchos investigadores.

5. *Combinación de verosimilitudes:* si se obtiene información sobre θ de diferentes fuentes independientes, y vienen con sus respectivas verosimilitudes, digamos $L_i(\theta)$, podemos resumir toda esta información como $\prod_i L_i(\theta)$. Esta es la base del meta-análisis. De hecho, no se pueden multiplicar aposterioris de esta forma.

6. *Aprioris impropias:* se reducen los peligros de utilizar aprioris impropias.

Definición 4.2 (El Principio de Verosimilitud (PV))**.** *Considere dos experimentos $E_1 = \{Y_1, \theta, f_1(y_1|\theta)\}$ y $E_2 = \{Y_2, \theta, f_2(y_2|\theta)\}$ que involucran el mismo parámetro θ. Suponga que para realizaciones particulares y_1 y y_2 de los datos, $L_1(\theta; y_1) = cL_2(\theta; y_2)$ para alguna constante c que no depende de θ. Entonces las funciones de verosimilitud proporcionan la misma información, y por tanto, la misma inferencia sobre θ.*

Una de las más espectaculares implicaciones del PV es que las reglas de parada en la recolección de datos de las cuales dependan los datos reales recolectados, pero no de θ, son irrelevantes para el propósito de determinar la evidencia sobre θ proporcionada por los datos.

Birnbaum en 1962 mostró que el PV es una consecuencia del Principio de Suficiencia (un estadístico suficiente resume toda la evidencia de un experimento) y del Principio de Condicionalidad (los experimentos no realizados son irrelevantes para la conclusión) [67].

Edwards [89] presenta una reflexión interesante entre hipótesis y modelo:

> Un marco suficiente para obtener inferencias inductivas es proporcionado mediante los conceptos de un *modelo estadístico* y una *hipótesis estadística*. Los dos conceptos conjuntamente proporcionan una descripción, en términos probabilísticos, del proceso por el cual se supone se generan las observaciones. Por *modelo* entendemos aquella parte de la descripción que no está presente en la pregunta, y que puede darse como dada, y por *hipótesis estadística* entendemos la asignación de valores particulares a los parámetros desconocidos del modelo, o de las cualidades particulares de las entidades desconocidas, estos parámetros o entidades en cuestión son la materia de investigación. No hay una distinción absoluta entre las dos partes de una descripción estadística, por lo que una ocasión puede ser considerada como dada, y por lo tanto una parte del modelo, y en otra ocasión, puede ser parte de la disputa, y por lo tanto parte de la hipótesis.

Poirier [79] propone cinco principios fundamentales:

Principio Pragmático de Construcción de Modelos 1 (PPMB 1): dada la verosimilitud, las inferencias sobre los parámetros desconocidos deben realizarse condicionadas en los datos observados en lugar de promediar sobre todos los posibles conjuntos de datos como lo requieren los procedimientos frecuentistas convencionales.

Principio Pragmático de Construcción de Modelos 2 (PPMB 2): la utilización de creencias a priori debe acompañarse por un análisis de sensibilidad local, y extenderlo hasta donde sea posible a un análisis de sensibilidad global.

Principio Pragmático de Construcción de Modelos 3 (PPMB 3): las verosimilitudes y los parámetros son más útiles vistos como artefactos de la mente en lugar de características intrínsecas del mundo observable. Ellos proporcionan ventanas útiles a través de las cuales los investigadores pueden ver el mundo observable e involucrarse en la predicción de observables futuras basadas en lo que han visto en el pasado. En lo que las pruebas de hipótesis y la estimación por intervalo son importantes hasta donde faciliten estas predicciones.

Principio Pragmático de Construcción de Modelos 4 (PPMB 4): dada una ventana (i.e. una verosimilitud) a través de la cual ver el mundo, la mayor tarea

que enfrenta el investigador es llevar a cabo un análisis de sensibilidad de la distribución predictiva con respecto a un amplio rango de aprioris de interés profesional como sea posible. Con respecto a la simplificación y las actividades de selección del modelo y adherencia al Principio de Verosimilitud, ellas juegan un papel importante. La búsqueda de verdades probando hipótesis, no importa qué tan nobles puedan aparecer, ellas tienen un papel relativamente menor que jugar en el análisis.

Principio Pragmático de Construcción de Modelos 5 (PPMB 5): la regla de Cromwell) Nunca asigne una probabilidad de uno a la ventana a través de la cual usted escoge ver el mundo.

Edwards [89] apunta

> Dado el modelo, podemos referirnos a las *consecuencias* de una hipótesis estadística, la parte que juega el modelo como es entendido... Una característica adicional de una hipótesis estadística es que el sujeto de las hipótesis no pueden ser observadas directamente, y las inferencias acerca de ellas pueden realizarse a través del conocimiento que una consecuencia particular ha ocurrido

Definición 4.3 (Intercambiabilidad)**.** *Se dice que las variables aleatorias* Y_i*,* $i = 1, \ldots, n$ *son intercambiables, si las distribuciones de* $(Y_1, \ldots, Y_n)$ *y* $(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ *son las mismas para todas las permutaciones* $(\pi(1), \ldots, \pi(n))$ *[90]).*

Las creencias de un individuo con respecto a unas cantidades aleatorias observables $Z_1, Z_2, \ldots, Z_n$ descritas por una distribución de probabilidad conjunta se dice que son *intercambiables* si, y solo sí, la distribución es invariante bajo todas las permutaciones de los subíndices $1, 2, \ldots, n$. Creencias con respecto a una sucesión infinita se dice que son intercambiables si, y solo sí, las creencias inducidas son intercambiables para cada subconjunto finito.

Intercambiabilidad e «Independientes e Idénticamente Distribuídas (IID)»no son lo mismo: IID implica intercambiabilidad, y variables intercambiables Y_i tienen idénticas distribuciones marginales, pero ellas no son necesariamente independientes.

La intercambiabilidad impone una forma de «simetría»sobre las sucesiones de observables. Si en el lanzamiento de una moneda varias veces, no importa el orden en que aparezcan los resultados, la a priori no debe cambiar bajo este principio. Intercambiabilidad es un término que en el sentido usado por de Finetti es limitado y Draper et al. [91] presentan una extensa discusión sobre el significado y alcances de este término. La idea central tras este término es el de similaridad entre las unidades observacionales.

Intercambiabilidad es una restricción más débil que el de independencia. Si las creencias subjetivas sobre $Z_1, Z_2, \ldots$ son independientes, entonces no puede haber aprendizaje de la experiencia. Por el contrario, la intercambiabilidad de las creencias en $Z_1, Z_2, \ldots$ implica una visión isomórfica del mundo en el cual existe la variable ficticia θ tal que $Z_1, Z_2, \ldots$ son independientes condicionales en θ (esto es, condicionalmente independientes).

Capítulo 5

Distribuciones conjugadas

Dada la magnitud de la tarea de determinar una distribución a priori que refleje de una manera clara nuestra información bayesiana, intuitivamente se piensa en limitar la búsqueda a familias de distribuciones a priori que posean ciertas características, tales como:

1. Tratabilidad analítica:
   - *a*) Facilidad de determinación de la distribución posterior de la muestra y de la a priori.
   - *b*) Facilidad para obtener características de interés, por ejemplo, valores esperados.
   - *c*) La a priori y a posteriori deben ser miembros de la misma familia (cerrada).
2. Flexibilidad y riqueza: debe permitir modelar una gran variedad de información a priori y creencias.
3. Interpretabilidad: los parámetros deben ser de tal forma que el analista pueda relacionarlos fácilmente con sus creencias e información.

Dickey [92] define este proceso de búsqueda de aprioris que sean convenientes de alguna forma como *distribuciones a priori operacionales o subrogadas*. Otro criterio discutido es el de *distribuciones a posteriori operacionales*.

Raiffa y Schlaifer en 1961 formalizaron el concepto de familias conjugadas [93]. La definición y la construcción de una familia conjugada depende de la existencia e identificación de estadísticos suficientes de dimensión finita para una función de verosimilitud dada. Si existe este estadístico suficiente entonces la dimensionalidad puede ser reducida. Cuando existe el estadístico suficiente, entonces existe una familia conjugada.

«Una a priori conjugada natural tiene la propiedad adicional de tener la misma forma funcional de la verosimilitud. Esta propiedad significa que la información a priori puede ser interpretada de la misma manera que la información en la función

de verosimilitud. En otras palabras, la a priori puede ser interpretada como si surgiera de un conjunto de datos ficticios obtenidos del mismo proceso que generó los datos reales»[27].

Las distribuciones conjugadas juegan un papel importante en los métodos bayesianos, ya que su uso puede simplificar el procedimiento de integración requerido para la marginalización. Ya que al pertenecer la a priori y la a posteriori a la misma familia, el proceso de actualización de parámetros se simplifica [94], lo cual es una gran ventaja para los sistemas inteligentes.

La conjugación nos limita a la selección de una clase de aprioris limitada y la información a priori solo puede utilizarse para la selección de los hiperparámetros. Si la clase es lo suficientemente grande esto puede no ser un gran problema. Robert [18] afirma que la automatización de la selección a priori es una ventaja y una desventaja, ya que por un lado se facilita el proceso de actualización, en especial cuando esto se hace en un proceso dinámico, pero en muchas ocasiones limita el proceso de representación de la distribución a priori y hace referencia a un experimento planteado por Diaconis y Ylvisaker sobre dejar caer una moneda que se tiene parada sobre su borde, en forma perpendicular a una superficie horizontal, ellos dicen que la experiencia muestra que la distribución es bimodal con modas en $1/3$ y $2/3$.

Si $x_1, \ldots, x_n$ son v.a.'s i.i.d. de un proceso definido por $f(x|\theta)$, donde θ puede ser un escalar o un vector de parámetros desconocidos de interés. Asumimos que existe un familia conjugada para este proceso, donde $\xi(\theta|\phi)$, cuyos miembros están indexados por el hiperparámetro ϕ. Ya que existe la familia conjugada, por lo tanto es posible factorizar la verosimilitud $L(\theta|x_1, \ldots, x_n)$ de la siguiente manera:

$$L(\theta|x_1, \ldots, x_n) = u(x_1, \ldots, x_n)\, v(T(x_1, \ldots, x_n), \theta),$$

donde $u(\cdot)$ no depende de θ y $v(T(x_1, \ldots, x_n), \theta)$ es una función del parámetro y del estadístico suficiente. Asumamos que estamos interesado en una transformación biyectiva de los datos. Sean $y_1, \ldots, y_n$ los datos transformados tal que,

$$y_i = h(x_i)$$

Dado que la transformación $h(x)$ es biyectiva, su inversa, $h^{-1}(x)$, existe.

Si el proceso es continuo, tenemos entonces:

$$L(\theta|y) = f\left(h^{-1}(y)|\theta\right)\left|\frac{d}{dy}h^{-1}(y)\right|,$$

y para el conjunto de datos transformado la función de verosimilitud es,

$$L(\theta|y_1, \ldots, y_n) = f\left(h^{-1}(y_1), \ldots, h^{-1}(y_n)|\theta\right)|J|,$$

donde J es el jacobiano de la transformación. Ya que los datos transformados se distribuyen en forma independiente entonces:

$$J = \prod_{i=1}^{n} \frac{d}{dy_i}h^{-1}(y_i)$$

La función de verosimilitud de los datos transformados será:

$$L\left(\theta\left|y_1,\ldots,y_n\right.\right)=u\left(h^{-1}\left(y_1\right),\ldots,h^{-1}\left(y_n\right)\right)v\left(T\left(h^{-1}\left(y_1\right),\ldots,h^{-1}\left(y_n\right)\right),\theta\right)\left|J\right|$$

Ya que $|J|$ es una función de las x_i's, tenemos:

$$\mu\left(x_1,\ldots,x_n\right)=u\left(x_1,\ldots,x_n\right)\left|J\right|$$

Por lo tanto,

$$L\left(\theta\left|y_1,\ldots,y_n\right.\right)=\mu\left(x_1,\ldots,x_n\right)v\left(T\left(x_1,\ldots,x_n\right),\theta\right)$$

Esta expresión de la verosimilitud transformada es el producto de una función de los datos que no involucra el parámetro y el kernel de la verosimilitud sin transformar. Por lo tanto, la a priori conjugada del proceso sin transformar es la misma que la del proceso transformada.

A continuación se presentarán las distribuciones conjugadas para diferentes modelos de probabilidad. Algunas de estas son ilustradas usando datos reales.

5.1. Distribución binomial

Teorema 5.1. *Suponga que $X_1,\ldots,X_n$ es una muestra aleatoria de una distribución Bernoulli con parámetro π, donde el valor de π es desconocido. También supongamos que la distribución a priori de π es una beta con parámetros $\alpha(>0)$ y $\beta(>0)$. Entonces la distribución posterior de π cuando $X_i=x_i$, para $i=1,\ldots,n$ es una beta con parámetros $\alpha+\sum_{i=1}^{n}x_i$ y $\beta+n-\sum_{i=1}^{n}x_i$.*

Sean $X_1,\ldots,X_n$ variables aleatorias independientes Bernoulli(π). La verosimilitud es:

$$L(\pi)\propto\pi^{\sum_i X_i}(1-\pi)^{n-\sum_i X_i}$$

El parámetro π es univariable, y restringido al intervalo $[0,1]$. La distribución conjugada será:

$$\xi(\pi)\propto\pi^{\alpha-1}(1-\pi)^{\beta-1},\ \text{con}\ \alpha,\beta>0,$$

donde α y β son llamados *hiperparámetros*. Esta palabra se utiliza para distinguirlos del parámetro modelo muestral π. Si comparamos la a priori con la verosimilitud vemos que $\alpha-1$ puede asociarse con $\sum_i X_i$ y $\beta-1$ con $n-\sum_i X_i$. Por lo tanto el experto que debe expresar su información a priori puede realizar la tarea mental de extraer una muestra imaginaria de 0's y 1's de tamaño $\alpha+\beta-2$ y distribuir tanto los ceros y los unos como su imaginación se lo dicte. El tamaño de esta muestra imaginaria puede asociarse con el nivel de confianza subjetiva que el experto tenga en sus asignaciones. Esta distribución a priori se puede resumir mediante:

$$\begin{aligned} E(\pi) &= \frac{\alpha}{\alpha+\beta}, \\ Moda &= \frac{\alpha-1}{\alpha+\beta-2}, \\ Varianza &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{E(\pi)(1-E(\pi))}{\alpha+\beta+1} \end{aligned}$$

La esperanza a priori $E(\pi)$ corresponde a la probabilidad marginal de tener un éxito antes de obtener cualquier observación:

$$E(\pi) = \int \pi\xi(\pi)\,d\pi = \int p(Y=1|\pi)\xi(\pi)\,d\pi = p(X=1)$$

Ya que la varianza de π es una función decreciente de $\alpha+\beta$ para una media dada, la suma de los hiperparámetros $\alpha+\beta$ es también llamada la *precisión* de la distribución.

La distribución posterior es:

$$\xi(\pi|X_1,\ldots,X_n) \propto \pi^{\alpha+\sum_i X_i-1}(1-\pi)^{\beta+n-\sum_i X_i-1},$$

la cual es una distribución beta con hiperparámetros $\alpha+\sum_i X_i$ y $\beta+n-\sum_i X_i$. Por lo tanto, la precisión posterior se incrementa por el tamaño muestral n.

La media a posteriori se puede expresar como,

$$\frac{\alpha+\sum_{i=1}^n X_i}{\alpha+\beta+n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{\sum_{i=1}^n X_i}{n}\right),$$

lo que es una media ponderada,

$$E\left(\pi|X_1,\ldots,X_n,\alpha,\beta\right) = w\cdot E(\pi|\alpha,\beta) + (1-w)\cdot\frac{\sum_{i=1}^n X_i}{n},$$

donde $w=(\alpha+\beta)/(\alpha+\beta+n)$.

Ejemplo 5.1. Entradas a un hospital

Este ejemplo es desarrollado por Draper [95] y hace referencia a entradas de pacientes a un hospital universitario con Ataque Agudo del Miocardio (AAM). Se considera la tasa de mortalidad de los pacientes en los 30 días siguientes a la admisión al hospital. Se conoce que en Inglaterra esta tasa es del 15 % (no necesariamente para este hospital la tasa sea igual). Para elicitar la distribución a priori sobre la proporción de pacientes con AAM que muere en lo 30 días siguientes, se utiliza esta información como, digamos el promedio. Ahora se necesita un poco más de información y el analista, tal vez usando el Teorema Central del Límite, piensa que el 95 % de las posibles tasas de mortalidad para este hospital deben estar entre 5 % y 30 %. Debemos buscar por lo tanto una distribución $Beta(\alpha,\beta)$ que tenga una media de 0.15 y el área bajo la curva entre los límites (0.05, 0.30) debe ser igual a 0.95. Mediante ensayo y error se encuentra que $\alpha=4.5$ y $\beta=25.5$ se tiene una distribución con las características deseadas.

Escrito esto en forma jerárquica el modelo es:

$$\begin{aligned}(\alpha, \beta) = &\quad (4.5,\ 25.5) \quad &\text{(Hiperparámetros)}\\ \pi|\alpha, \beta \sim &\quad Beta(\alpha, \beta) \quad &\text{(A priori)}\\ X_1, \ldots, X_n \sim &\quad Bernoulli(\pi) \quad &\text{(Verosimilitud)}\end{aligned}$$

La función de verosimilitud de los datos es:

$$L(\pi) = p\left(X_1, \ldots, X_n|\pi\right) = \pi^S(1-\pi)^{n-S} \propto Beta(S+1, n-S+1),$$

donde $S = \sum_{i=1}^{n} X_i$. Si hemos observado 400 personas con AAM en el hospital, de los cuales 72 fallecieron en los siguientes 30 días, lo cual produce una verosimilitud proporcional a una $Beta(73\,,\,329)$. La distribución posterior será, por lo tanto,

$$\xi(\pi|S = 72, n = 400) \propto Beta(76.5,\ 353.5)$$

La información muestral equivalente en la distribución a priori se puede asociar con $n^* = \alpha + \beta$, en este caso es $n^* = 4.5 + 25.5 = 30$. La información muestral es muy grande con relación a la a priori 400/30 es más de 13 a 1.

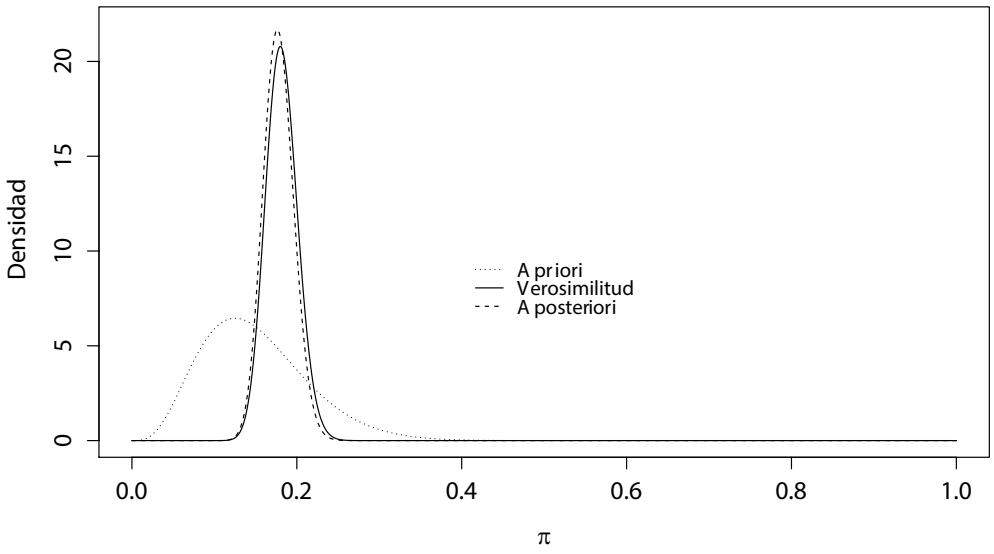

Figura 5.1: *podemos ver que la distribución a posteriori es poco influenciada por la a priori*

5.2. Distribución binomial negativa

Teorema 5.2. *Suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución binomial negativa con parámetros r y π, donde r tiene un valor específico $(r > 0)$ y el valor de π es desconocido. También supongamos que la distribución a priori de π es una beta con parámetros $\alpha(> 0)$ y $\beta(> 0)$. Entonces la distribución posterior de π cuando $X_i = x_i$, para $i = 1, \ldots, n$ es una beta con parámetros $\alpha + rn$ y $\beta + \sum_{i=1}^{n} x_i$.*

Prueba: (Ejercicio).

5.3. Distribución geométrica

Otra distribución de conteo popular es la geométrica, la cual cuenta el número de fracasos antes de obtener el primer éxito. Su función de probabilidad está dada por:

$$P(X = x) = \pi(1 - \pi)^x \qquad x = 0, 1, 2, \ldots$$

Su media es $(1 - \pi)/\pi$ y su varianza $(1 - \pi)/\pi^2$.

Si suponemos que la distribución a priori de π es una beta con parámetros $\alpha(> 0)$ y $\beta(> 0)$. La distribución posterior es:

$$\xi\left(\pi \,|X_1, \ldots, X_n\right) \propto \pi^n(1 - \pi)^{\sum_{i=1}^n x_i}\pi^{\alpha-1}(1 - \pi)^{\beta-1} = \pi^{\alpha+n-1}(1 - \pi)^{\beta+\sum_{i=1}^n x_i-1},$$

lo cual es una $Beta(\alpha + n\,,\, \beta + \sum_{i=1}^n x_i)$.

5.4. Distribución multinomial

La distribución multinomial juega un papel fundamental en el trabajo aplicado, siendo la generalización multivariable de la distribución binomial.

Definición 5.1 (Distribución Dirichlet)**.** *El vector aleatorio* $\boldsymbol{X} = (X_1, \ldots, X_k)'$ *se distribuye como una Dirichlet con vector de parámetros* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ *con* $\alpha_i > 0$; $i = 1, \ldots, k$*, si la p.d.f.* $f(\boldsymbol{x}|\boldsymbol{\alpha})$ *para* $\boldsymbol{x} = (x_1, \ldots, x_k)$ *y* $\sum_{i=1}^k x_i = 1$ *está dada por:*

$$f(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1-1} \cdots x_k^{\alpha_k-1}$$

La media y la varianza de X_i son respectivamente:

$$E(X_i) = \frac{\alpha_i}{\alpha_0},$$

$$var(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)},$$

donde $\alpha_0 = \sum_{i=1}^k \alpha_i$. La covarianza entre X_i y X_j es,

$$Cov(X_i, X_j) = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)}, \text{ para } (i \neq j)$$

Teorema 5.3. *Suponga que* $\boldsymbol{Y} = (Y_1, \ldots, Y_k)'$ *tiene una distribución multinomial con parámetros n (fijo) y* $\boldsymbol{W} = (W_1, \ldots, W_k)'$*, desconocidos. Suponga también que la distribución a priori de* $\boldsymbol{W}$ *es una Dirichlet con vector de parámetros* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ *con* $\alpha_i > 0$; $i = 1, \ldots, k$*. Entonces la distribución posterior de* $\boldsymbol{W}$ *cuando* $Y_i = y_i$*,* $i = 1, \ldots, k$*, es una distribución Dirichlet con vector de parámetros* $\boldsymbol{\alpha}^* = (\alpha_1 + y_1, \ldots, \alpha_k + y_k)'$.

Prueba: Ejercicio.

El parámetro α_k puede ser interpretado como el conteo a priori, antes de ver los datos, que esperaríamos ver en la celda k. Un valor grande para este parámetro muestran un gran conocimiento previo acerca de la distribución, mientras que valores pequeños corresponden a poco conocimiento.

Ejemplo 5.2. Tipos de sangre. La siguiente tabla presenta los datos sobre el tipo de sangre en una muestra de personas de la región central y oriental de Antioquia-Colombia.

	Tipo de Sangre			
	O	A	AB	B
Frecuencia	474	246	11	59

Si no tenemos un conocimiento a priori sobre las diversas proporciones, digamos π_O, π_A, π_{AB} y π_{AB}, entonces podemos escoger como a priori una $Dirichlet(1, 1, 1, 1)$. Entonces la a posteriori será $Dirichlet(474 + 1, 246 + 1, 11 + 1, 59 + 1)$.

A continuación, mostramos las distribuciones posterior marginales para cada proporción y la distribución conjunta de las dos proporciones más frecuentes.

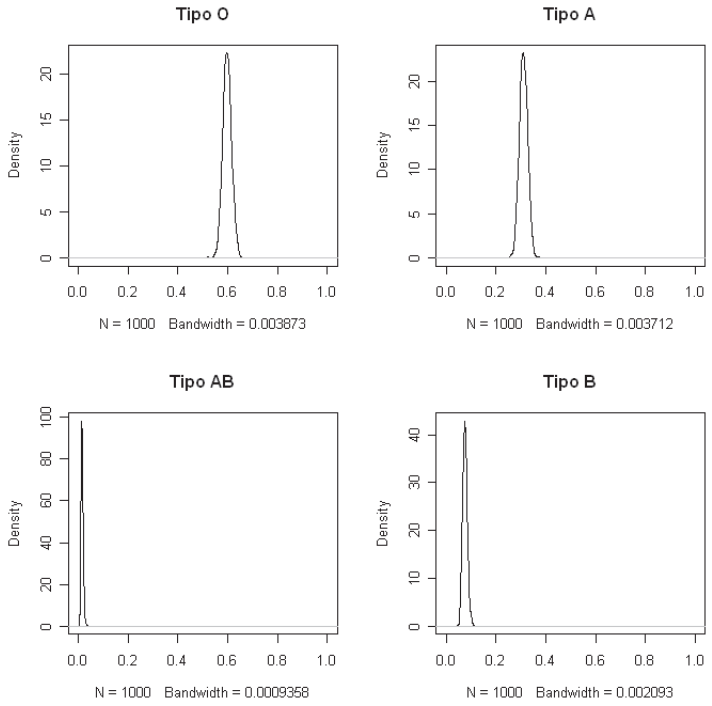

Figura 5.2: *distribución posterior marginal para cada una de las proporciones del tipo de sangre*

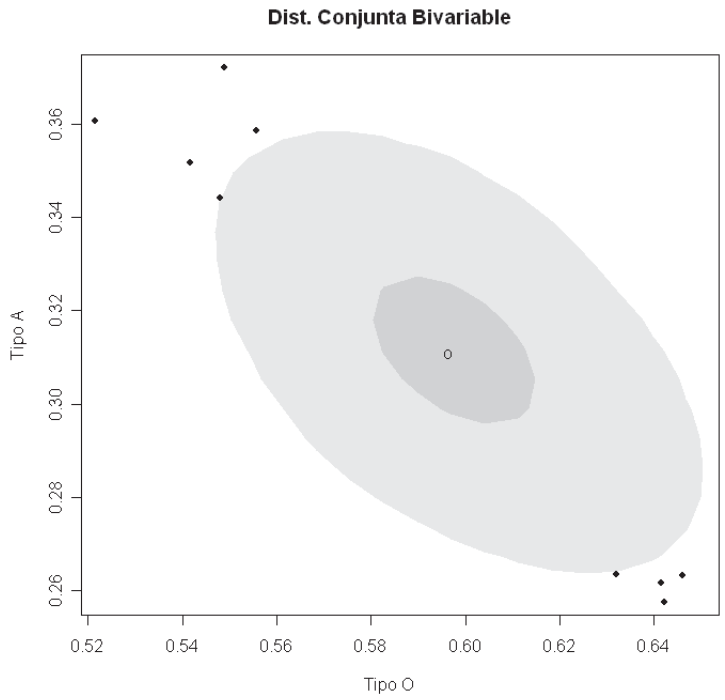

Figura 5.3: *distribución conjunta entre un par de los parámetros considerado en el problema de los tipos de sangre*

5.5. Distribución Poisson

El modelo de conteo más utilizado es el modelo Poisson, ya que su desarrollo teórico es claro y muchos problemas reales pueden modelarse muy bien de esta forma. Decimos que una variable aleatoria de conteo X se distribuye $Poisson(\theta)$ si su función de probabilidad está dada por:

$$f(x) = \frac{\theta^x \exp(-\theta)}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

Para esta distribución se tiene que $E(X) = Var(X) = \theta$.

Teorema 5.4. *Suponga que $X_1, \ldots, X_n$ es una muestra de una distribución Poisson con media desconocida θ. También supóngase que la distribución a priori de θ es una gamma con parámetros $\alpha(> 0)$ y $\beta(> 0)$. Entonces la distribución posterior de θ cuando $X_i = x_i$, para $i = 1, \ldots, n$ es una gamma con parámetros $\alpha + \sum_{i=1}^{n} x_i$ y $\beta + n$.*

Prueba: Si $X_1, \ldots, X_n$ es una muestra de una distribución Poisson con media desconocida θ, entonces la verosimilitud será:

$$L\left(\theta \,|X_1, \ldots, X_n\right) = \prod_{i=1}^{n} \frac{\theta^{X_i} \exp(-\theta)}{X_i!} \propto \theta^{\sum_{i=1}^{n} X_i} \exp(-n\theta)$$

Si la priori es $Gamma(\alpha, \beta)$ su densidad será:

$$\xi(\theta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$$

La a posteriori será entonces,

$$\xi\left(\theta\,|X_1,\ldots,X_n\right) \propto \theta^{\alpha-1}\exp(-\beta\theta)\theta^{\sum_{i=1}^n X_i}\exp(-n\theta) = \theta^{\alpha+\sum_{i=1}^n X_i-1}\exp\left(-(\beta+n)\theta\right)$$

Esto muestra el resultado.

Ejemplo 5.3. Caso de accidentalidad. En la presentación de la Alcaldía de Medellín-Colombia llamada «Geo-referenciación de la accidentalidad en los principales tramos y avenidas de Medellín. Año 2008», se dice que el número de accidentes de tránsito en la ciudad con muertos fue de 315 para ese año. Si asumimos que el número de accidentes con muertes sigue una distribución Poisson con parámetro θ y si asumimos una Gamma a priori poco informativa, digamos $\alpha_0 = 0.001$ y $\beta_0 = 0.001$, la a posteriori será Gamma con $\alpha_1 = 315.001$ y $\beta_1 = 1.001$. La media a posteriori será 314.6863, la cual es bastante parecida al valor obtenido en la muestra de tamaño 1 que tenemos.

Elicitación de la distribución a priori conjugada para el parámetro de la Poisson

Suponga que deseamos estudiar el número de goles marcados por los equipos locales en el torneo profesional colombiano. Asumimos que el número de goles marcados por el equipo local se puede modelar mediante la Poisson.

El parámetro θ en la Poisson es la media. ¿Cómo escogemos la $Gamma(\alpha,\beta)$ que represente adecuadamente nuestro conocimiento del problema?

Vamos a presentar una aproximación usando una forma predictiva.

Probabilidad	0	1	2	3	4	5	6 o más.
$\theta^x\exp(-\theta)/x!$	π_0	π_1	π_2	π_3	π_4	π_5	π_{6+}

¡Si miramos con cuidado el problema lo convertimos en multinomial!

Debemos determinar el vector $(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_{6+})$. ¿Cómo lo hacemos?

Le decimos al experto que nos responda algo como esto: si se observan 1000 partidos de fútbol, ¿en cuántos esperaría que el local no hiciera goles?, ¿En cuántos esperaría que hiciera un gol? ¿dos goles? ¿tres? ¿cuatro? ¿cinco? ¿seis o más goles? Esto nos da una tabla como la siguiente:

Goles	0	1	2	3	4	5	6 ó más.
Nro. de juegos	n_0	n_1	n_2	n_3	n_4	n_5	n_{6+}

$$\sum_i n_i = 1000.$$

Resultado del experto:

Goles	0	1	2	3	4	5	6 ó más.
Nro. de juegos	170	250	300	180	60	35	5

Ahora, generamos N muestras de tamaño 1000 de una multinomial con probabilidades

$$(170/1000, 250/1000, 300/1000, 180/1000, 60/1000; 35/1000, 5/1000)$$

Para cada muestra multinomial, calculamos la probabilidad de cada celda, o sea, dividimos cada muestra por 1000. Digamos:

$$\left(\pi_0^j, \pi_1^j, \pi_2^j, \ldots, \pi_{6+}^j\right), \quad j = 1, \ldots, N$$

Usando estas probabilidades, calculamos la media de la distribución Poisson, teniendo en cuenta que la última celda corresponde a un truncamiento.

$$\theta_j = \sum_{i=0}^{\infty} i \cdot \pi_i^j \approx \sum_{i=0}^{6} i \cdot \pi_i^j$$

La aproximación siempre es por debajo del verdadero valor, ya que se reemplazan todos los valores mayores que 6 por 6. Podríamos realizar alguna corrección.

```
error<-NA; acumulado<-NA; media.sin<-NA; media.corr<-NA
medias<-seq(0.5,4,length=20)

for(i in medias){
 proba<-dpois(0:20,i)
 acumu<-1-sum(proba[1:6])
 media<-sum((0:5)*proba[1:6])+6*(1-sum(proba[1:6]))
 media.sin<-c(media.sin,media)
 error<-c(error,i-media)
 acumulado<-c(acumulado,acumu)
}

acumulado<-acumulado[-1]; error<-error[-1]; media.sin<-media.sin[-1]

# Relación entre la media y el error
plot(medias,error,xlab=expression(theta),ylab='Error')
title(main='Error que se comete con el truncamiento \n en la
estimación de la media')

# El problema es que hay que conocer la verdadera media
# pero si usamos el porcentaje de observaciones hasta el punto
# de truncamiento podemos aproximar la corrección.

plot(acumulado,error)
acumulado2<- acumulado^2
summary(modelo<-lm(error~acumulado+acumulado2))
```

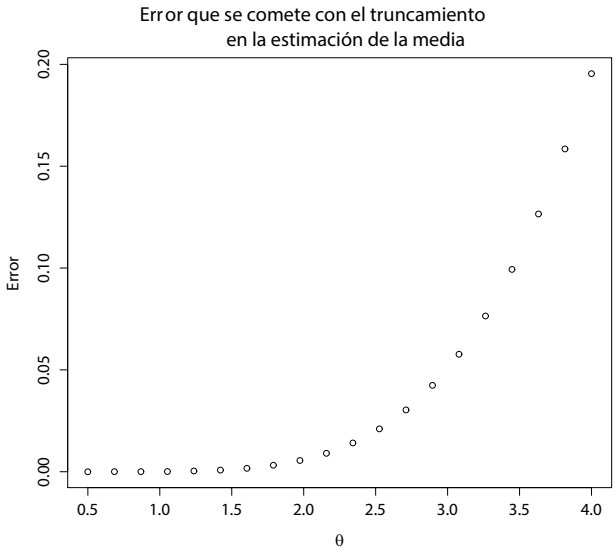

Figura 5.4: *observamos que el error absoluto es cuadrático con respecto a la media de la Poisson*

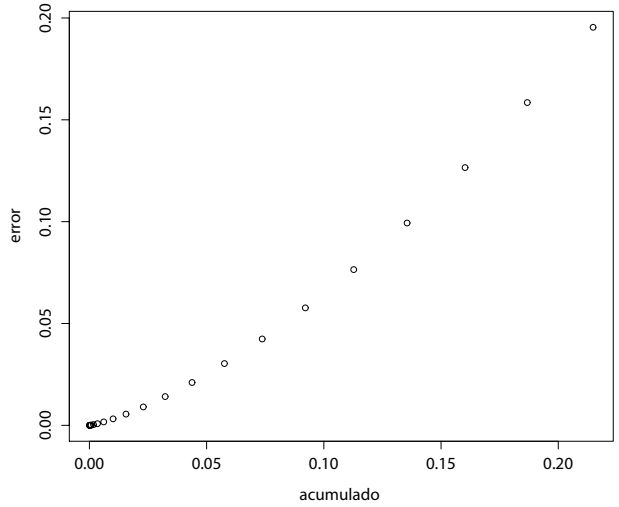

Figura 5.5: *consideramos el porcentaje de observaciones hasta el punto de truncamiento. Podemos aproximar esta relación mediante una función cuadrática. Con datos reales la corrección la obtenemos calculando el porcentaje de observaciones bajo el punto de truncamiento*

Como resultado del resumen estadístico del modelo, tenemos lo siguiente:

```
#Call:
#lm(formula = error ~ acumulado + acumulado2)
#
#Residuals:
#       Min         1Q     Median         3Q        Max
#-0.0010180 -0.0006270  0.0001269  0.0006803  0.0008440
```

```
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept) -0.0007817  0.0002460  -3.178  0.00551 **
#acumulado    0.4164908  0.0082370  50.563  < 2e-16 ***
#acumulado2   2.3313949  0.0427119  54.584  < 2e-16 ***
#
#Residual standard error: 0.0007206 on 17 degrees of freedom
#Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
#F-statistic: 6.387e+04 on 2 and 17 DF,  p-value: < 2.2e-16
```

Podemos hallar la media de la Poisson haciendo la corrección mediante el ajuste cuadrático:

```
calcula.theta<-function(proba){
acumu<-1-sum(proba[1:6])
media<-sum((0:5)*proba[1:6])+6*(1-sum(proba[1:6]))
media.cor<-media-0.0007817+ 0.4164908*acumu + 2.3313949*acumu^2
return(media.cor) }

# Generación de la multinomial
temp<-c(170,250,300,180,60,35,5)

res.multi<-rmultinom(2000,1000,temp)/1000
thetas<-apply(res.multi,2,calcula.theta)

hist(thetas,freq=F,xlab=expression(theta),
   main='Distribución a priori',ylab='Frecuencia')

summary(thetas)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.701   1.812   1.837   1.838   1.863   1.967
```

Notamos que la media a priori del número promedio de goles marcados por el local es de 1.8 goles por partido. Esta distribución puede ser ajustada a una Gamma usando la función `fitdistr` de R.

```
require(MASS)

fitdistr(thetas,'gamma')
     shape         rate
  2040.68620   1110.41968
 (  64.37553) (  35.03361)
```

Los parámetros de la gamma a priori serán:

$$\begin{aligned} \alpha &= 2040.68620 \\ \beta &= 1110.41968 \end{aligned}$$

```
# Gráfico de la a priori elicitada
xx<-seq(1.5,2.2,length=100)
yy<-dgamma(xx,2040.68620,rate=1110.41968)
points(xx,yy,type='l',col='red')
```

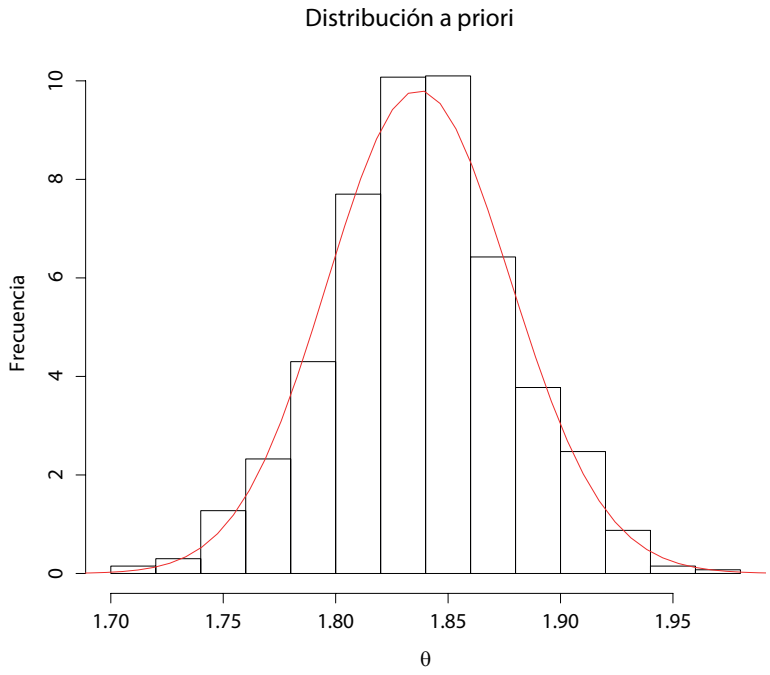

Figura 5.6: *distribución a priori para la Poisson*

5.6. Distribución exponencial

La distribución exponencial tiene función de densidad de probabilidad dada por:

$$f(x) = \lambda e^{-\lambda x}, \qquad \text{con } \lambda > 0 \text{ y } x \in (0, \infty)$$

Teorema 5.5. *Suponga que $X_1, \ldots, X_n$ es una muestra de una distribución exponencial con parámetro desconocido λ. También supóngase que la distribución a priori de λ es una gamma con parámetros $\alpha(>0)$ y $\beta(>0)$. Entonces la distribución posterior de λ cuando $X_i = x_i$, para $i = 1, \ldots, n$ es una gamma con parámetros $\alpha + n$ y $\beta + \sum_{i=1}^{n} x_i$.*

La prueba es directa:

$$\begin{aligned} \xi(\lambda) &\propto \lambda^{\alpha-1} \exp\left(-\beta\lambda\right), \\ L\left(\lambda \,|Datos\right) &\propto \lambda^{n} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right), \\ \xi\left(\lambda \,|Datos\right) &\propto \lambda^{(\alpha+n)-1} \exp\left(-\lambda \left[\beta + \sum_{i=1}^{n} x_i\right]\right) \end{aligned}$$

5.6.1. Caso especial: se observa solo el primer estadístico de orden

Si solo tenemos el valor del primer estadístico de orden, o sea el menor valor de la muestra de tamaño n, nuestra verosimilitud será proporcional a la densidad del

primer estadístico de orden. Si $X_{(1)}$ denota el mínimo valor de la muestra de una distribución absolutamente continua, entonces la función de distribución de $X_{(1)}$, está dada por:

$$F_1\left(x_{(1)}\right) = 1 - \left[1 - F\left(x_{(1)}\right)\right]^n,$$

y la densidad está dada por:

$$f_1\left(x_{(1)}\right) = nf\left(x_{(1)}\right)\left[1 - F\left(x_{(1)}\right)\right]^{n-1},$$

donde F es la función de distribución de X. En el caso exponencial:

$$f_1\left(x_{(1)}\right) = n\lambda \exp\left(-n\lambda x_{(1)}\right)$$

Si la a priori de λ es una $Gamma(\alpha, \beta)$, la a posteriori sería:

$$\begin{aligned} \xi\left(\lambda \left| x_{(1)}, n\right.\right) &\propto \lambda \exp\left(-n\lambda x_{(1)}\right) \lambda^{\alpha-1} \exp\left(-\beta\lambda\right), \\ &\propto \lambda^{(\alpha+1)-1} \exp\left(-\lambda\left(\beta + nx_{(1)}\right)\right) \end{aligned}$$

Esta corresponde a una $Gamma(\alpha + 1, \beta + nx_{(1)})$.

5.6.2. Caso especial: se observa solo el n-ésimo estadístico de orden

Si solo tenemos el valor del n-ésimo estadístico de orden, o sea el mayor valor de la muestra de tamaño n, nuestra verosimilitud será proporcional a la densidad del n-ésimo estadístico de orden. Si $X_{(n)}$ denota el máximo valor de la muestra de una distribución absolutamente continua, su función de distribución está dada por:

$$F_n\left(x_{(n)}\right) = \left[F\left(x_{(n)}\right)\right]^n,$$

y la densidad está dada por:

$$f_n\left(x_{(n)}\right) = nf\left(x_{(n)}\right)\left[F\left(x_{(n)}\right)\right]^{n-1}$$

En el caso exponencial,

$$f_n\left(x_{(n)}\right) = n\lambda \exp\left(-\lambda x_{(n)}\right)\left(1 - \exp\left(-\lambda x_{(n)}\right)\right)^{n-1}$$

Si la a priori de λ es una $Gamma(\alpha, \beta)$, la a posteriori sería:

$$\begin{aligned} \xi\left(\lambda \left| x_{(n)}, n\right.\right) &\propto \lambda \exp\left(-\lambda x_{(n)}\right)\left(1 - \exp\left(-\lambda x_{(n)}\right)\right)^{n-1} \lambda^{\alpha-1} \exp\left(-\beta\lambda\right), \\ &\propto \lambda^{(\alpha+1)-1} \exp\left(-\lambda\left(\beta + x_{(n)}\right)\right)\left(1 - \exp\left(-\lambda x_{(n)}\right)\right)^{n-1} \end{aligned}$$

Observe cómo en este caso la distribución posterior no pertenece a la familia gamma.

5.6.3. Caso especial: se observan algunos datos censurados en el punto x_0

Si asumimos una a priori $Gamma(\alpha, \beta)$ y tenemos n_0 observaciones que no han fallado en el tiempo x_0, de las n observaciones que se disponen (digamos que n_1 sí se observaron completamente con mediciones $x_1, x_2, \ldots, x_{n_1}$), su verosimilitud será:

$$\begin{aligned} L(\lambda \,|Datos) &\propto \lambda^{n_1} \exp\left(-\lambda \sum_{i=1}^{n_1} x_i\right) \left(P(X > x_0 \,|\lambda)\right)^{n_0}, \\ &\propto \lambda^{n_1} \exp\left(-\lambda \sum_{i=1}^{n_1} x_i\right) \exp\left(-n_0 \lambda x_0\right), \\ &\propto \lambda^{n_1} \exp\left(-\lambda \left(n_0 x_0 + \sum_{i=1}^{n_1} x_i\right)\right) \end{aligned}$$

La a posteriori, que corresponde a una $Gamma\left(\alpha + n_1, \beta + n_0 x_0 + \sum_{i=1}^{n_1} x_i\right)$ será:

$$\xi(\lambda \,|Datos) \propto \lambda^{\alpha + n_1 - 1} \exp\left(-\lambda \left(\beta + n_0 x_0 + \sum_{i=1}^{n_1} x_i\right)\right)$$

5.6.4. Caso especial: se observan todos los datos censurados en el punto x_0

Si asumimos una a priori $Gamma(\alpha, \beta)$ y tenemos n observaciones que no han fallado en el tiempo x_0, su verosimilitud será:

$$L(\lambda \,|Datos) \propto \left(P(X > x_0 \,|\lambda)\right)^n = \exp\left(-n\lambda x_0\right)$$

La a posteriori, que corresponde a una $Gamma\left(\alpha, \beta + n x_0\right)$ será:

$$\xi(\lambda \,|Datos) \propto \lambda^{\alpha - 1} \exp\left(-\lambda \left(\beta + n x_0\right)\right)$$

5.6.5. Aumentación (data augmentation)

Aumentación[1] de datos (Data Augmentation) hace referencia a una familia de métodos computacionales que típicamente adicionan nuevos datos latentes que son identificados parcialmente por los datos. Por 'parcialmente identificados' se quiere decir que existe alguna información sobre estas nuevas variables, pero que a medida que el tamaño muestral se incrementa, la cantidad de información acerca de cada variable no se incrementa. Ejemplos de aplicaciones son datos censurados, indicadores de pertenencia a poblaciones en mezclas y variables latentes para regresiones discretas.

[1] Esta palabra no existe en el español, pero podemos usarla como una adecuación fonética de su equivalente en inglés, y aun siendo incorrecta la usamos por lo parecida. Correcto podría se «Datos Aumentados».

El algoritmo de aumento de datos (que llamamos aquí aumentación) fue propuesto por Tanner y Wong en 1987 [96]. Este es un proceso iterativo para calcular la densidad posterior completa. La idea de ellos fue la de incrementar (aumentar) los datos observados y por una cantidad z, que son conocidos como datos latentes.La aumentación de datos está diseñada para simplificar las simulaciones en espacios mayores de 'datos completos' que en el espacio original. Se asume que dados y y z se puede calcular o muestrear de la distribución posterior aumentada $\xi(\theta|y,z)$. Para obtener la distribución posterior observada $\xi(\theta|y)$, se generan múltiples valores de z de la distribución predictiva $p(z|y)$ y entonces se calcula el promedio de $\xi(\theta|y,z)$ sobre las imputaciones, ya que $p(z|y)$ depende de $\xi(\theta|y)$.

El algoritmo

1. Dada la aproximación actual $g_i(\theta)$ de $\xi(\theta|y)$,

   a) Generar θ de g_i

   b) Generar z de $p(z|\phi,y)$, donde ϕ es el valor generado en el paso anterior.

2. Los dos pasos anteriores son repetidos m veces para obtener $z^{(1)},\ldots,z^{(m)}$.

3. La actualización se obtiene como:

$$g_{i+1}(\theta)=\frac{1}{m}\sum_{j=1}^{m}\xi\left(\theta\left|z^{(j)},y\right.\right)$$

Bajo condiciones de regularidad (Serfling, 1980) el algoritmo converge a la verdadera distribución posterior.

Expansión de parámetros hace referencia a la adición de nuevos parámetros que no son identificados (en términos bayesianos significa que tienen distribuciones a priori impropias). Un ejemplo es remplazar un parámetro θ por un producto $\phi\varphi$, así que la inferencia puede hacerse con respecto al producto pero no con respecto a los parámetros individuales.

Tanto aumentación de datos como expansión de parámetros pueden ser utilizadas para incrementar la simplicidad y la velocidad de los cálculos computacionales. Estos métodos no cambian la verosimilitud, cambian la parametrización.

Desde el punto de vista bayesiano, las nuevas parametrizaciones llevan a nuevas aprioris y así a nuevos modelos. Una forma en la que ocurre a menudo es si un parámetro es *condicionalmente conjugado*, esto es, conjugado en la distribución posterior condicional, dados los datos y todos los otros parámetros del modelo.

Gelman [97] presenta la siguiente ilustración que es útil para mostrar el potencial de estas aproximaciones. Suponga que se tiene una báscula que solo puede pesar objetos hasta 200 libras. De 200 animales pesados, , $n=91$ fueron pesados exitosamente, sus pesos fueron $y_1,\ldots,y_{91}$. Se asumió que las pesos se distribuyen $N(\mu,\sigma^2)$. Él considera dos escenarios:

1. Considerando el escenario de «datos truncados», donde se considera que N es desconocido, la distribución posterior de μ y σ^2 está dada por:

$$\xi\left(\mu,\sigma\left|\boldsymbol{y}\right.\right)\propto\xi\left(\mu,\sigma\right)\prod_{i=1}^{91}\left\{\frac{\phi\left(y_i\left|\mu,\sigma^2\right.\right)}{\left[1-\Phi\left(\frac{\mu-200}{\sigma}\right)\right]}\right\}$$

   Considerar este caso, a partir de la descripción planteada por Gelman, no parece lógico, ya que N obviamente es conocido, y si se considera el caso donde la muestra efectiva fueran los 91 animales, se desperdiciaría la información entregada por los datos efectivamente mayores a 200.

2. El caso donde se tienen *datos censurados*, N es conocido y la distribución posterior:

$$\xi\left(\mu,\sigma\left|\boldsymbol{y}\right.\right)\propto\xi\left(\mu,\sigma\right)\left[\Phi\left(\frac{\mu-200}{\sigma}\right)\right]^{N-91}\prod_{i=1}^{91}\phi\left(y_i\left|\mu,\sigma^2\right.\right)$$

Gelman presenta el modelo de datos truncados como censurados pero con un número desconocido de datos censurados.

$$\xi\left(\mu,\sigma,N\left|\boldsymbol{y}\right.\right)\propto\xi\left(N\right)\xi\left(\mu,\sigma\right)\left(\begin{array}{c}N\\91\end{array}\right)\left[\Phi\left(\frac{\mu-200}{\sigma}\right)\right]^{N-91}\prod_{i=1}^{91}\phi\left(y_i\left|\mu,\sigma^2\right.\right)$$

La distribución posterior marginal de μ y σ se halla como:

$$\xi\left(\mu,\sigma\left|\boldsymbol{y}\right.\right)\propto\sum_{N=91}^{\infty}\xi\left(N\right)\xi\left(\mu,\sigma\right)\left(\begin{array}{c}N\\91\end{array}\right)\left[\Phi\left(\frac{\mu-200}{\sigma}\right)\right]^{N-91}\prod_{i=1}^{91}\phi\left(y_i\left|\mu,\sigma^2\right.\right)$$

$$=\xi\left(\mu,\sigma\right)\prod_{i=1}^{91}\phi\left(y_i\left|\mu,\sigma^2\right.\right)\sum_{N=91}^{\infty}\xi\left(N\right)\left(\begin{array}{c}N\\91\end{array}\right)\left[\Phi\left(\frac{\mu-200}{\sigma}\right)\right]^{N-91}$$

Si se escoge $\xi(N)\propto 1/N$, entonces la forma dentro de la sumatoria es la de una binomial negativa con $\theta=N-1$, $\alpha=91$ y

$$\frac{1}{\beta+1}=\Phi\left(\frac{\mu-200}{\sigma}\right)$$

Esta expresión dentro de la sumatoria es proporcional a:

$$\left(1-\Phi\left(\frac{\mu-200}{\sigma}\right)\right)^{-1}$$

5.7. Distribución normal

La distribución normal es la más ampliamente conocida y utilizada distribución en el trabajo estadístico. Hay básicamente dos razones para ello:

- Muchas poblaciones pueden ser modeladas aproximadamente por esta distribución.
- Como resultados límites se llega a ella en muchas situaciones.

Su función de densidad es:

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

con soporte $x \in (-\infty, \infty)$. Su función de distribución acumulada se denota $\Phi(x)$, su media es μ y su varianza σ^2. Esta distribución posee dos parámetros, lo cual nos lleva a considerar diferentes situaciones. La precisión es el inverso de la varianza.

5.7.1. Inferencia sobre la media: precisión conocida

Teorema 5.6. *Suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal con un valor desconocido de la media μ y un valor especificado de la precisión r $(r > 0)$.*

- **Distribución a priori:** $\mu \sim N(\mu_0, \tau_0)$ *donde τ_0 es la precisión, tal que $-\infty < \mu_0 < \infty$ y $\tau_0 > 0$.*
- **Distribución posterior:**

  $$(\mu | \boldsymbol{X} = \boldsymbol{x}) \sim N(\mu_1, \tau_1),$$

  donde,

  $$\begin{aligned} \mu_1 &= \frac{\tau_0\mu_0 + nr\bar{x}}{\tau_0 + nr} \\ \tau_1 &= \tau_0 + nr \text{ es la precisión,} \end{aligned}$$

  y $\bar{x}$ es la media muestral.

Prueba: La prueba es elemental.

Observe que la media posterior se puede expresar como:

$$\mu_1 = \frac{\tau_0\mu_0 + nr\bar{x}}{\tau_0 + nr} = \frac{nr}{\tau_0 + nr}\bar{x} + \frac{\tau_0}{\tau_0 + nr}\mu_0$$

Se ve claramente que la media posterior es una media ponderada de la media a priori y la media muestral.

5.7.2. Inferencia sobre la precisión: media conocida

Este tipo de problema surge en control de calidad cuando lo que interesa controlar es la variabilidad de un proceso determinado.

Teorema 5.7. *Suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal con un valor conocido de la media m $(-\infty < m < \infty)$ y un valor desconocido de la precisión W $(W > 0)$.*

- **Distribución a priori:** $W \sim Gamma\ (\alpha_0, \beta_0)$ *donde* $\alpha_0 > 0$ *y* $\beta_0 > 0$.[2]
- **Distribución posterior:**

$$(W|\boldsymbol{X} = \boldsymbol{x}) \sim Gamma\,(\alpha_1, \beta_1)\,,$$

donde,

$$\begin{aligned} \alpha_1 &= \alpha_0 + \frac{n}{2}, \\ \beta_1 &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - m)^2 \end{aligned}$$

Prueba: La prueba es directa.

5.7.3. Media y precisión desconocidas

Este caso, a pesar de lo simple que puede parecer, muestra la complejidad a la que puede llegar a enfrentar el estadístico ante la presencia de varios parámetros.

Teorema 5.8. *Suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal con un valor desconocido de la media μ y un valor desconocido de la precisión R $(R > 0)$.*

- **Distribución a priori conjunta de μ y R:**
  1. *La distribución condicional de μ cuando $R = r$ es $\mu \sim N\,(\mu_0, \tau_0 r)$ donde $\tau_0 r$ es la precisión, tal que $-\infty < \mu_0 < \infty$ y $\tau_0 > 0$,*
  2. *La distribución marginal de R es $Gamma\,(\alpha_0, \beta_0)$ donde $\alpha_0 > 0$ y $\beta_0 > 0$.*
- **Distribución posterior conjunta de μ y R cuando $\boldsymbol{X} = \boldsymbol{x}$:**
  1. *La distribución condicional de μ cuando $R = r$ es:*

$$(\mu|\boldsymbol{X} = \boldsymbol{x}) \sim N\,(\mu_1, \tau_1)\,,$$

[2] Asumimos una gamma de la forma:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

donde,

$$\begin{aligned} \mu_1 &= \frac{\tau_0\mu_0 + n\bar{x}}{\tau_0 + n}, \\ \tau_1 &= (\tau_0 + n)r, \end{aligned}$$

y $\bar{x}$ es la media muestral.

2. *La distribución marginal de R es $\Gamma(\alpha_1, \beta_1)$, donde:*

$$\begin{aligned} \alpha_1 &= \alpha_0 + \frac{n}{2}, \\ \beta_1 &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\tau n\left(\bar{x} - \mu_0\right)^2}{2(\tau + n)} \end{aligned}$$

Prueba:
Recuerde que,

$$f(x|y) = \frac{f(x,y)}{f(y)} \Rightarrow f(x,y) = f(x|y)\, f(y)$$

Si $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal con un valor desconocido de la media μ y un valor desconocido de la precisión τ ($\tau > 0$) la verosimilitud será:

$$\begin{aligned} L(\mu, \tau|\, Datos) &= \prod_{i=1}^{n} \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right) \\ &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned}$$

Ahora,

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ &= (n-1)S^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

$$\begin{aligned} L(\mu, \tau|\, Datos) &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\left((n-1)S^2 + n(\bar{x} - \mu)^2\right)\right) \\ &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}(n-1)S^2\right) \exp\left(-\frac{n\tau}{2}(\bar{x} - \mu)^2\right) \end{aligned}$$

La a priori es:

$$\begin{aligned}\xi(\mu,\tau) &= \xi(\mu|\tau)\,\xi(\tau)\\ &\propto \tau^{1/2}\exp\left(-\frac{\tau_0\tau}{2}(\mu-\mu_0)^2\right)\tau^{\alpha_0-1}\exp(-\beta_0\tau)\end{aligned}$$

La a posteriori será:

$$\begin{aligned}\xi(\mu,\tau) &\propto \tau^{n/2}\exp\left(-\frac{\tau}{2}(n-1)S^2\right)\exp\left(-\frac{n\tau}{2}(\bar{x}-\mu)^2\right)\\ &\quad\times\tau^{1/2}\exp\left(-\frac{\tau_0\tau}{2}(\mu-\mu_0)^2\right)\tau^{\alpha_0-1}\exp(-\beta_0\tau)\\ &\propto \tau^{n/2+1/2}\exp\left(-\frac{\tau}{2}\left[n(\bar{x}-\mu)^2+\tau_0(\mu-\mu_0)^2\right]\right)\\ &\quad\times\tau^{\alpha_0-1}\exp\left(-\tau\left(\frac{(n-1)S^2}{2}+\beta_0\right)\right)\end{aligned}$$

Ahora,

$$\begin{aligned}\left[n(\bar{x}-\mu)^2+\tau_0(\mu-\mu_0)^2\right] &= n(\mu-\bar{x})^2+\tau_0(\mu-\mu_0)^2\\ &= n\mu^2-2n\mu\bar{x}+n\bar{x}^2+\tau_0-2\tau_0\mu\mu_0+\tau_0\mu_0^2\\ &= (n+\tau_0)\mu^2-2\mu(n\bar{x}+\tau_0\mu_0)+n\bar{x}^2+\tau_0\mu_0^2\\ &= (n+\tau_0)\left[\mu^2-2\mu\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right]+n\bar{x}^2+\tau_0\mu_0^2\\ &= (n+\tau_0)\left[\mu^2-2\mu\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}+\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)^2}\right]-\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)}+n\bar{x}^2+\tau_0\mu_0^2\\ &= (n+\tau_0)\left(\mu-\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right)^2-\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)}+n\bar{x}^2+\tau_0\mu_0^2\end{aligned}$$

Luego la a posteriori queda:

$$\begin{aligned}\xi(\mu,\tau) &\propto \exp\left(-\frac{\tau}{2}\left[(n+\tau_0)\left(\mu-\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right)^2-\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)}+n\bar{x}^2+\tau_0\mu_0^2\right]\right)\\ &\quad\times\tau^{n/2+1/2}\tau^{\alpha_0-1}\exp\left(-\tau\left(\frac{(n-1)S^2}{2}+\beta_0\right)\right)\\ &\propto \exp\left(-\frac{\tau(n+\tau_0)}{2}\left(\mu-\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right)^2\right)\times\exp\left(-\frac{\tau}{2}\left[-\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)}+n\bar{x}^2+\tau_0\mu_0^2\right]\right)\\ &\quad\times\tau^{n/2+1/2}\tau^{\alpha_0-1}\exp\left(-\tau\left(\frac{(n-1)S^2}{2}+\beta_0\right)\right)\end{aligned}$$

Ahora,

$$\begin{aligned}-\frac{(n\bar{x}+\tau_0\mu_0)^2}{(n+\tau_0)}+n\bar{x}^2+\tau_0\mu_0^2 &= \frac{n^2\bar{x}^2+n\tau_0\mu_0^2+n\tau_0\bar{x}^2+\tau_0^2\mu_0^2-n^2\bar{x}^2-2n\bar{x}\tau_0\mu_0-\tau_0^2\mu_0^2}{n+\tau_0}\\ &= \frac{n\tau_0\mu_0^2+n\tau_0\bar{x}^2-2n\bar{x}\tau_0\mu_0}{(n+\tau_0)}\\ &= \frac{n\tau_0\left(\mu_0^2+\bar{x}^2-2\bar{x}\mu_0\right)}{(n+\tau_0)}\\ &= \frac{n\tau_0(\mu_0-\bar{x})^2}{(n+\tau_0)}\end{aligned}$$

Entonces:

$$\begin{aligned}\xi(\mu,\tau) &\propto \exp\left(-\frac{\tau(n+\tau_0)}{2}\left(\mu-\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right)^2\right)\\ &\times \exp\left(-\frac{\tau}{2}\left[\frac{n\tau_0(\mu_0-\bar{x})^2}{(n+\tau_0)}\right]\right)\\ &\times \tau^{n/2+1/2}\tau^{\alpha_0-1}\exp\left(-\tau\left(\frac{(n-1)S^2}{2}+\beta_0\right)\right)\\ &\propto \tau^{1/2}\exp\left(-\frac{\tau(n+\tau_0)}{2}\left(\mu-\frac{(n\bar{x}+\tau_0\mu_0)}{(n+\tau_0)}\right)^2\right)\\ &\times \tau^{\alpha_0+n/2-1}\exp\left(-\tau\left(\frac{(n-1)S^2}{2}+\beta_0+\frac{n\tau_0(\mu_0-\bar{x})^2}{2(n+\tau_0)}\right)\right)\end{aligned}$$

Con esto queda demostrado el resultado.

5.8. Distribución gamma

La distribución gamma ha sido ampliamente aplicada en confiabilidad y en pruebas de vida. Decimos que la variable aleatoria X tiene una distribución gamma con parámetros β y α si su densidad es:

$$f(x|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)\qquad x>0,\alpha>0,$$

donde α denota el parámetro de forma y β es el recíproco de un parámetro de escala. Si $x_1,x_2,\ldots,x_n$ es una muestra aleatoria de esta distribución, entonces la función de densidad conjunta es:

$$\prod_{i=1}^n f(x_i|\alpha,\beta)=\frac{\beta^{n\alpha}}{[\Gamma(\alpha)]^n}p^{\alpha-1}\exp(-s\beta),$$

donde,

$$\begin{aligned}s &= \sum_{i=1}^n x_i,\\ p &= \prod_{i=1}^n x_i.\end{aligned}$$

Miller (1980) usa una clase conjugada muy general definida por la conjunta:

$$\xi(\alpha,\beta)\propto\frac{\beta^{\nu'\alpha-1}}{[\Gamma(\alpha)]^{n'}}(p')^{\alpha-1}\exp(-s'\beta),$$

donde $\alpha>0$, $\beta>0$, $n'>0$, $\nu'>0$, $s'>0$ y $p'>0$, tal que $n'(p')^{1/n'}/s'<1$. La distribución posterior es proporcional a:

$$\xi(\alpha,\beta|\boldsymbol{x})\propto\frac{\beta^{\nu''\alpha-1}}{[\Gamma(\alpha)]^{n''}}p''^{\alpha-1}\exp(-s''\beta),$$

donde $\nu''=\nu'+n$, $p''=p'p$, $s''=s'+s$ y $n''=n'+n$.

La distribución condicional de β dado α es una $Gamma\,(\nu'', s'')$, y la distribución marginal posterior de α es proporcional a:

$$\frac{\Gamma\,(\nu''\alpha)}{[\Gamma(\alpha)]^{\nu''}}\left(\frac{r''}{\nu''}\right)^{\nu''\alpha},$$

donde,

$$\frac{r''}{n''} = \frac{\sqrt[\nu'']{p''}}{s''} = \frac{(p')^{1/(\nu'+n)}(r/n)^{n/(\nu'+n)}s^{n/(\nu'+n)}}{s'+s}$$

Damsleth [98] propone el uso de las distribuciones gamma con Tipo I y Tipo II para el caso de observaciones que provienen de una población $Gamma(\alpha, \beta)$. El primer caso par cuando β es conocido y fijo y el segundo caso para cuando se desconocen ambos parámetros.

5.9. Conjugadas en tramos

A pesar de las ventajas que se tiene de usar distribuciones conjugadas para realizar el proceso bayesiano, no siempre es posible hallar una distribución en la familia que refleje el conocimiento previo. Una relajación a este problema en clases de distribuciones conjugadas, es trabajar con mezclas de distribuciones dentro de las familias conjugadas, que algunos autores argumentan, pueden representar casi cualquier conocimiento previo ya que ellas pueden aproximar casi cualquier distribución a priori [99].

Meeden [99] propuso el uso de distribuciones conjugadas por tramos. Para ilustrar esto considere la proporción, π, como el parámetro de interés. Como hemos visto, la familia conjugada en este caso es la Beta. Un ejemplo del una conjugada en tramos tenemos:

$$\xi(\pi) \propto \begin{cases} \pi^{\alpha_1-1}\,(1-\pi)^{\beta_1-1} & \text{para } 0 < \pi \leq \lambda \\ k\pi^{\alpha_2-1}\,(1-\pi)^{\beta_2-1} & \text{para } \lambda < \pi < 1, \end{cases}$$

donde,

$$k = \lambda^{\alpha_1-\alpha_2}(1-\lambda)^{\beta_1-\beta_2}.$$

Con esta selección de k, la densidad ξ es continua en λ.

Como el espacio parametral fue particionado en dos regiones, Meeden ha llamado esta distribución de orden 2. Uno podría ajustar splines cúbicas restringidas también como una alternativa.

Capítulo 6

Distribuciones a priori no informativas

El uso de distribuciones a priori no informativas buscan que ellas tengan un impacto mínimo sobre la distribución posterior del parámetro de interés y que sea relativamente plana con relación a la verosimilitud. Esto busca que sean los datos los que tengan un claro dominio en la distribución posterior, y por tanto, en todas las inferencias que de ellas se obtengan. También se conocen como vagas, difusas, planas o de referencia. El área de las distribuciones no informativas es grande y polémica. Kadane, Schervish y Seidenfeld [100] comentan, «algunos estadísticos usan distribuciones impropias, especialmente distribuciones uniformes, como una representación de nuestra ignorancia. Otros consideran esto como la pérdida de la oportunidad que proporcionan las distribuciones a priori para modelar las opiniones del cliente». Existen diferentes posiciones sobre cómo reflejar ignorancia mediante una distribución.

Estas distribuciones no informativas se reúnen en dos grupos:

Propias: cuando la distribución de probabilidad integra a una constante finita, se dice que es propia. Por ejemplo, para el caso de la distribución binomial, su parámetro π, que denota el porcentaje de éxitos en la población, podemos asumir como a priori la $U(0,1)$, lo cual refleja nuestra ignorancia total, al asumir que cualquier valor en este intervalo es igualmente posible como valor.

Impropias: una distribución a priori $\xi(\theta)$ es impropia si,

$$\int_{\Theta} \xi(\theta)\, d\theta = \infty$$

Winkler [101] dice:

> Los términos difuso y no-difuso son relativos en este contexto, no términos absolutos. Cuando decimos que nuestra información es difusa realmente queremos decir que es difusa relativa a la información muestral. También queremos decir que es localmente difusa (i.e., difusa solo dentro dentro de un cierto rango). Así, 'difuso' puede depender no solo de la precisión de la información muestral sino también de los valores específicos de la información muestral. En

muchos casos el uso de distribuciones a priori difusas por parte del bayesiano puede ser sicológicamente iluminador, bien sea para otros o para él mismo, aún si su distribución a priori no es difusa.

Notas:

1. Una distribución a priori impropia puede terminar en una a posteriori impropia y por lo tanto no se podrán hacer inferencias.

2. Una distribución a priori impropia puede llevar a una a posteriori propia.

Ejemplo 6.1. Asumamos que $y_1, \ldots, y_n|\theta$ son variables distribuidas normal e independientemente con media θ y con varianza conocida σ^2. Asumamos que $\xi(\theta) \propto 1$ es la distribución a priori uniforme (impropia) sobre los números reales. La verosimilitud es:

$$L\left(\theta|\boldsymbol{y}\right) \propto \exp\left(-\frac{n}{2}\frac{\left(\bar{y}-\theta\right)^2}{\sigma^2}\right),$$

y la distribución posterior es:

$$\theta|\boldsymbol{y} \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right),$$

la cual es una distribución propia.

Yang y Berger [102] presentan varias razones por las cuales es importante considerar las distribuciones no informativas. Tenemos entre ellas:

- Con frecuencia la elicitación de las distribuciones a priori es imposible, por múltiples razones, por ejemplo, limitaciones de costo o tiempo, o resistencia o falta de entrenamiento de los clientes.

- El análisis estadístico debe aparecer como «objetivo».

- La elicitación subjetiva puede producir malas distribuciones subjetivas, por ejemplo si la elicitación es sesgada.

- En problemas de alta dimensión, lo más que se puede esperar es obtener buenas distribuciones subjetivas para algunos pocos parámetros, y a los parámetros de perturbación se les asignan distribuciones no informativas.

- El análisis bayesiano con distribuciones no informativas puede utilizarse para obtener procedimientos clásicos buenos.

Aún cuando un investigador tenga creencias a priori fuertes, puede ser más convincente analizar los datos utilizando una a priori de referencia dominada por la verosimilitud. Además podemos automatizar el proceso de hallar aprioris. Yang y Berger [102] proporcionan un amplio catálogo de distribuciones no informativas que es útil en el trabajo aplicado.

Robert [18] señala que en muchas situaciones una distribución impropia es el límite de distribuciones propias. Pueden interpretarse así como casos extremos donde la información a priori ha desaparecido completamente.

6.1. El principio de la razón insuficiente de Laplace

Si el espacio parametral es finito se puede utilizar una distribución a priori uniforme para reflejar ignorancia total.

$$\xi(\theta) \propto 1 \text{ para } \theta \in \Theta$$

Bhattacharya [103] dice: «Esta distribución ha causado mucha controversia entre los estadísticos bayesianos ya que no puede interpretarse como una densidad de probabilidad en el sentido tradicional. Claramente, en casos de un espacio parametral no acotado, una densidad uniforme asigna una medida infinita al espacio».

6.2. A priori de Jeffreys

La distribución a priori de Jeffreys[1] satisface la propiedad local de uniformidad para distribuciones a priori no informativas, es decir, es uniforme sobre el espacio parametral. Esta a priori está basada en la matriz de información de Fisher. Jeffreys la propuso como una «regla general»para determinar la distribución a priori [105].

Definición 6.1. *Sea $f(x|\theta)$ la densidad de x dado θ. La información de Fisher es definida como:*

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 \log\left(f\left(x|\theta\right)\right)}{\partial\theta^2}\right]$$

Si $\boldsymbol{\theta}$ es un vector de p componentes, entonces,

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \log\left(f\left(x|\boldsymbol{\theta}\right)\right)}{\partial\theta_i\,\partial\theta_j}\right]_{p\times p},$$

y entonces $\mathcal{I}(\boldsymbol{\theta})$ será una matriz de dimensión $p \times p$.

Definición 6.2. *La distribución a priori de Jeffreys se define como:*

$$\xi(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2}$$

La distribución a priori de Jeffreys es localmente uniforme y por lo tanto no informativa. Esta propiedad es importante ya que nos proporciona un esquema automatizado para hallar distribuciones a priori no informativas para cualquier modelo paramétrico [106]. Esta distribución es impropia para muchos modelos.

Ejemplo 6.2. Asumamos que $y_1, \ldots, y_n$ son variables distribuidas independientemente Bernoulli(π). Encontremos la distribución a priori de Jeffreys para π.

La densidad para una variable Bernoulli(π) es:

$$p(y|\pi) = \pi^y(1-\pi)^{1-y}.$$

[1] Jeffreys fue un notable académico con contribuciones de gran importancia en estadística y algunos la comparan con la de Fisher y es una lástima el poco conocimineto que tiene la comunidad científica de él [104].

Entonces tenemos,

$$\begin{aligned}
\log\left(p(y|\pi)\right) &= y\log(\pi) + (1-y)\log(1-\pi), \\
\frac{\partial}{\partial\pi}\log\left(p(y|\pi)\right) &= \frac{y}{\pi} - \frac{1-y}{1-\pi}, \\
\frac{\partial^2}{\partial\pi^2}\log\left(p(y|\pi)\right) &= -\frac{y}{\pi^2} - \frac{1-y}{(1-\pi)^2}, \\
\mathcal{I}(\pi) &= -E\left[\frac{\partial^2}{\partial\pi^2}\log\left(p(y|\pi)\right)\right] \\
&= \frac{E(y)}{\pi^2} + \frac{1-E(y)}{(1-\pi)^2} = \frac{1}{\pi} + \frac{1-\pi}{(1-\pi)^2} \\
&= \frac{1}{\pi} + \frac{1}{1-\pi} = \frac{1}{\pi(1-\pi)}.
\end{aligned}$$

Por lo tanto, la distribución a priori de Jeffreys es:

$$\begin{aligned}
\xi(\pi) &\propto \mathcal{I}(\pi)^{1/2} \\
&= \left(\frac{1}{\pi(1-\pi)}\right)^{1/2} \\
&= \pi^{-1/2}(1-\pi)^{-1/2} \\
&= \pi^{1/2-1}(1-\pi)^{1/2-1}
\end{aligned}$$

Así $\pi \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$. Por lo que vemos en este caso la distribución a priori de Jeffreys es propia.

Tuyl, Gerlach y Mengersen [107] discuten el caso donde en la muestra no se tienen éxitos y comparan la a priori de Laplace y la de Jeffreys. Cuando este es el caso, la distribución de Jeffreys puede ser muy informativa, y selecciones de una familia $Beta(\alpha, \beta)$ que sea informativa también pueden ser excesivamente informativas, por ejemplo para valores $\alpha < 1$, sobrepasando la información de la muestra.

Ejemplo 6.3. Asumamos que $y_1, \ldots, y_n|\mu$ son variables distribuidas normal e independientemente con media μ y con precisión τ desconocidas. Calculemos la distribución a priori de Jeffreys para (μ, τ).

$$\begin{aligned}
f\left(x|\mu,\tau\right) &= \frac{\tau^{1/2}}{\sqrt{2\pi}}\exp\left(-\frac{\tau}{2}\left(x-\mu\right)^2\right), \\
\log\left(f\left(x|\mu,\tau\right)\right) &= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{\tau}{2}\left(x-\mu\right)^2, \\
\frac{\partial\log\left(f\left(x|\mu,\sigma\right)\right)}{\partial\mu} &= \tau(x-\mu), \\
\frac{\partial^2\log\left(f\left(x|\mu,\tau\right)\right)}{\partial\mu^2} &= -\tau, \\
\frac{\partial\log\left(f\left(x|\mu,\tau\right)\right)}{\partial\tau} &= \frac{1}{2\tau} - \frac{1}{2}(x-\mu)^2, \\
\frac{\partial^2\log\left(f\left(x|\mu,\tau\right)\right)}{\partial\tau^2} &= -\frac{1}{2\tau^2}, \\
\frac{\partial^2\log\left(f\left(x|\mu,\tau\right)\right)}{\partial\mu\partial\tau} &= (x-\mu)
\end{aligned}$$

Tomando la esperanza obtenemos

$$\mathcal{I}\left(\begin{pmatrix}\mu\\ \tau\end{pmatrix}\right) = \begin{bmatrix} -\tau & 0 \\ 0 & -\frac{1}{2\tau^2} \end{bmatrix}$$

Así la distribución a priori será:

$$\begin{aligned} \xi(\mu,\tau) &\propto \left|\mathcal{I}\left(\begin{pmatrix}\mu\\ \tau\end{pmatrix}\right)\right|^{1/2} \\ &= \left(\tau \times \frac{1}{\tau^2}\right)^{1/2} \\ &\propto \tau^{-1/2} \end{aligned}$$

Esta distribución a priori de Jeffreys es impropia.

La distribución a priori de Jeffreys tiene la propiedad de invarianza, ya que para cualquier otra transformación uno a uno sigue siendo no informativa. Esto surge de la relación,

$$\mathcal{I}(\theta) = \mathcal{I}(\psi(\theta)) \left(\frac{d\psi(\theta)}{d\theta}\right)^2,$$

donde $\psi(\theta)$ es una transformación uno a uno de θ. Así:

$$(\mathcal{I}(\theta))^{1/2} = (\mathcal{I}(\psi(\theta)))^{1/2} \left|\frac{d\psi(\theta)}{d\theta}\right|$$

Note que $\left|\frac{d\psi(\theta)}{d\theta}\right|$ es el valor absoluto del jacobiano de la transformación de θ a $\psi(\theta)$. Así

$$(\mathcal{I}(\theta))^{1/2}\, d\theta = (\mathcal{I}(\psi))^{1/2}\, d\psi$$

La a priori de Jeffreys preserva la escala en parametrizaciones.

Ejemplo 6.4. Supongamos $x \sim N(\mu, 1)$. La distribución a priori de Jeffreys para μ es $\xi(\mu) \propto 1$. Sea $\psi(\mu) = e^{\mu}$. Esta es una transformación uno a uno en μ. La correspondiente a priori de jeffreys para $\psi(\mu)$ es:

$$\begin{aligned} (\mathcal{I}(\psi(\mu)))^{1/2} &= (\mathcal{I}(\mu))^{1/2} \left|\frac{d\psi(\mu)}{d\mu}\right|^{-1} \\ &= 1 \times e^{-\mu} \\ &= e^{-\mu} \end{aligned}$$

Así la distribución a priori de Jeffreys para $\psi(\mu) = e^{\mu}$ es:

$$\xi(\mu) \propto e^{-\mu}, \qquad -\infty < \mu < \infty$$

La propiedad de invarianza significa que si tenemos una distribución a priori localmente uniforme en θ, y si $\psi(\theta)$ es una función uno a uno de θ, entonces $\xi\left(\psi(\theta)\right)$ es una distribución a priori localmente uniforme para $\psi(\theta)$.

Ejemplo 6.5. A priori de Jeffreys para una binomial y una binomial negativa. Según el principio de verosimilitud no existe diferencias entre la información proporcionada por los dos esquemas de muestreo (ver ejemplo 1.8). Sin embargo si se escoge una distribución no informativa de Jeffreys para el caso binomial, esta es:

$$\xi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

Para el caso de la distribución binomial negativa, la distribución a priori de Jeffreys es:

$$\xi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1}$$

Esto produce, usando esta distribución a priori, diferentes resultados inferenciales con la a posteriori [67]. Este es un problema que tiene esta aproximación para construir aprioris en forma automatizada, donde el esquema de muestreo es el que nos determina la distribución a priori.

Papathomas y Hocking [108] discuten sobre la condicionalización de Jeffreys. Si se tienen dos proposiciones q_1 y q_2 y supongamos que las creencias para ellas están basadas sobre alguna información denotada por h. Si $P_h(q_1, q_2)$ denota la probabilidad que ambas proposiciones sean ciertas y si se obtiene información adicional $g(q_2)$ la cual cambia nuestra creencia de $P_h(q_2)$ a $P_{h,g(q_2)}(q_2)$. ¿Cómo afecta esto nuestra creencia sobre q_1? La nueva información sobre q_2 no cambia nuestra creencia condicional para q_1, o sea,

$$P_{h,g(q_2)}\left(q_1 \,|q_2\right) = P_h\left(q_1 \,|q_2\right)$$

La regla de actualización de nuestras creencias para q_1 y q_2, dada la información adicional $g(q_2)$, es:

$$P_{h,g(q_2)}\left(q_1, q_2\right) = P_{h,g(q_2)}\left(q_2\right) P_h\left(q_1 \,|q_2\right)$$

Esta forma de actualización es llamada la condicionalización de Jeffreys.

Ejercicio

1. Encuentre la distribución a priori de Jeffreys para el caso de una $N(\mu, \mu)$ con $\mu > 0$.

2. Si $X_1, X_2, \ldots, X_n$ es una m.a. de la distribución anterior, calcule la distribución posterior.

6.3. Otras alternativas para las a priori

Definición 6.3 (Distribución A priori de Máxima Entropía)**.** *Cuando θ es univariable y puede tomar cualquier valor sobre la recta real, y la media y la varianza a priori están especificadas, la distribución a priori de máxima entropía es la Normal con la media y la varianza especificadas.*

En caso de soporte finito la distribución de máxima entropía es la uniforme, sin embargo cuando existe alguna información previa, por ejemplo una media, entonces se construye esta a priori teniendo en cuenta esta restricción. Consideremos el siguiente ejemplo, donde el soporte es finito, digamos $\theta_1, \ldots, \theta_k$, con $E[\theta] = C$, entonces debemos maximizar:

$$H = -\sum_{i=1}^{k} p(\theta_i) \log(p(\theta_i)),$$

sujeto a la restricción,

$$\sum_{i=1}^{k} \theta_i p(\theta_i) = C,$$

y que,

$$\sum_{i=1}^{k} p(\theta_i) = 1$$

Los $p(\theta_i)$ que resuelvan este problema forman la distribución de máxima entropía.

Como una ilustración consideremos el siguiente caso [109]: el espacio parametral es $\Theta = \{1, 2, 3, 4, 5\}$ con $E(\theta) = 2$. La solución aproximada es $p_1 = 0.459, p_2 = 0.261, p_3 = 0.148, p_4 = 0.084, p_5 = 0.048$. Esto puede hallarse vía algoritmos genéticos.

Kass y Wasserman [105] presentan la definición planteada por Novick y Hall:

Definición 6.4 (Distribución a priori indiferente)**.** *Se define una distribución a priori indiferente si identificando una clase de conjugadas se selecciona una a priori de esta clase que satisfaga:*

- *La a priori debe ser impropia*
- *Una «muestra mínima necesaria»debe inducir una posterior propia.*

Un ejemplo de la anterior definición es claro en el problema binomial, con la clase conjugada de las Betas, la distribución a priori $\{\pi(1-\pi)\}^{-1}$ es una a priori indiferente. Esta distribución a priori se conoce como la a priori de Haldane. Esta distribución es impropia. Si se trabaja con una distribución de Laplace para el $logit(\pi) = \log(\pi/1-\pi)$, entonces la distribución sobre π será ésta [110].

Bernardo [111][112] propone la distribución a priori de referencia en la cual se busca una a priori no informativa que maximice la información muestral de la distribución posterior. Esto se hace en términos de distancia entre distribuciones, la a priori y la a posteriori. Esta distancia corresponde más bien a la entropía relativa esperada o distancia de Kullback-Leibler entre la distribución a priori y la correspondiente a posteriori. En muchos caso esta distribución coincide con la no informativa de Jeffreys. Clarke y Sun [113] proponen la definición de aprioris de referencia en términos de bondad de ajuste en lugar de información, donde la distancia Chi-cuadrado

entre densidades p y q se define como $\chi^2(p,q) = \int (p-q)^2/q$. Para una a priori $\xi(\theta)$ y la correspondiente a posteriori $\xi(\theta\,|\mathbf{X})$ se examina $E\left(\chi^2\left(\xi(\theta\,|\mathbf{X}),\,\xi(\theta)\right)\right)$ y se busca la a priori que maximice el anterior valor esperado.

Box y Tiao [13] proponen el uso de distribuciones *a priori localmente uniformes*, las cuales consideran el comportamiento local de la a priori en una región donde la verosimilitud es apreciable, pero la a priori no se asume grande por fuera de esa región.

Ejemplo 6.6. Distribución Poisson.
Si se utiliza una distribución a priori no informativa tenemos varias alternativas:

- **A priori no informativa uniforme**

$$\pi(\lambda) \propto 1$$

- **A priori de Jeffreys**

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

Consideremos los goles del local marcados en cada torneo del fútbol colombiano. Si usamos una a priori no informativa de Laplace como a priori inicial y si cada distribución posterior sirve como a priori del torneo siguiente y asumiendo que el número de goles marcados por el local sigue una distribución Poisson con parámetro λ, la a posteriori será Gamma.

Tabla 6.1: *número de goles marcados por el local en el fútbol colombiano desde el año 2000 hasta el 2009. α y β corresponden a los parámetros estimados de la distribución a posteriori. Datos del fútbol colombiano*

	Número de Goles								Número	Total	Promedio		
Torneo	0	1	2	3	4	5	6	7	Partidos	Goles	de Goles	α	β
2000-1	28	60	51	26	6	3	2	0	176	291	1.6534	291	176
2000-2	38	55	39	33	9	2	0	0	176	278	1.5795	569	352
2001-1	35	55	53	24	5	3	1	0	176	274	1.5568	843	528
2001-2	37	57	45	27	7	2	1	0	176	272	1.5455	1115	704
2002-1	47	65	64	13	6	2	1	0	198	272	1.3737	1387	902
2002-2	38	75	45	33	7	0	0	0	198	292	1.4747	1679	1100
2003-1	28	59	53	17	4	0	1	0	162	238	1.4691	1917	1262
2003-2	31	63	40	23	5	0	0	0	162	232	1.4321	2149	1424
2004-1	31	58	46	17	8	2	0	0	162	243	1.5000	2392	1586
2004-2	34	62	38	21	5	1	1	0	162	232	1.4321	2624	1748
2005-1	36	62	39	21	2	2	0	0	162	221	1.3642	2845	1910
2005-2	31	56	45	17	10	2	0	0	161	247	1.5342	3092	2071
2006-1	38	58	42	13	8	1	2	0	162	230	1.4198	3322	2233
2006-2	26	61	39	27	8	1	0	0	162	257	1.5864	3579	2395
2007-1	27	54	56	16	7	1	1	0	162	253	1.5617	3832	2557
2007-2	35	65	34	17	7	1	1	0	160	223	1.3938	4055	2717
2008-1	30	60	38	23	8	2	0	1	162	254	1.5679	4309	2879
2008-2	35	62	40	16	7	2	0	0	162	228	1.4074	4537	3041
2009-1	34	59	41	21	6	0	1	0	162	234	1.4444	4771	3203
2009-2	22	60	43	22	12	3	0	0	162	275	1.6975	5046	3365

Tabla 6.2: *resumen estadístico de la distribución posterior en cada torneo*

	Media	Varianza	Perc. 0.05	Mediana	perc. 0.95
1	1.6534	0.0094	1.4973	1.6515	1.8160
2	1.6165	0.0046	1.5067	1.6155	1.7295
3	1.5966	0.0030	1.5072	1.5960	1.6881
4	1.5838	0.0022	1.5066	1.5833	1.6626
5	1.5377	0.0017	1.4704	1.5373	1.6062
6	1.5264	0.0014	1.4656	1.5261	1.5881
7	1.5190	0.0012	1.4624	1.5188	1.5765
8	1.5091	0.0011	1.4560	1.5089	1.5631
9	1.5082	0.0010	1.4578	1.5080	1.5593
10	1.5011	0.0009	1.4533	1.5010	1.5497
11	1.4895	0.0008	1.4439	1.4894	1.5358
12	1.4930	0.0007	1.4491	1.4928	1.5374
13	1.4877	0.0007	1.4455	1.4875	1.5304
14	1.4944	0.0006	1.4535	1.4942	1.5357
15	1.4986	0.0006	1.4590	1.4985	1.5387
16	1.4925	0.0005	1.4541	1.4923	1.5312
17	1.4967	0.0005	1.4594	1.4966	1.5344
18	1.4919	0.0005	1.4557	1.4918	1.5286
19	1.4895	0.0005	1.4542	1.4894	1.5252
20	1.4996	0.0004	1.4650	1.4995	1.5344

En las Tablas 6.1 y 6.2, observamos la gran variabilidad de las medias muestrales comparadas con la mediana de las distribuciones aposterioris.

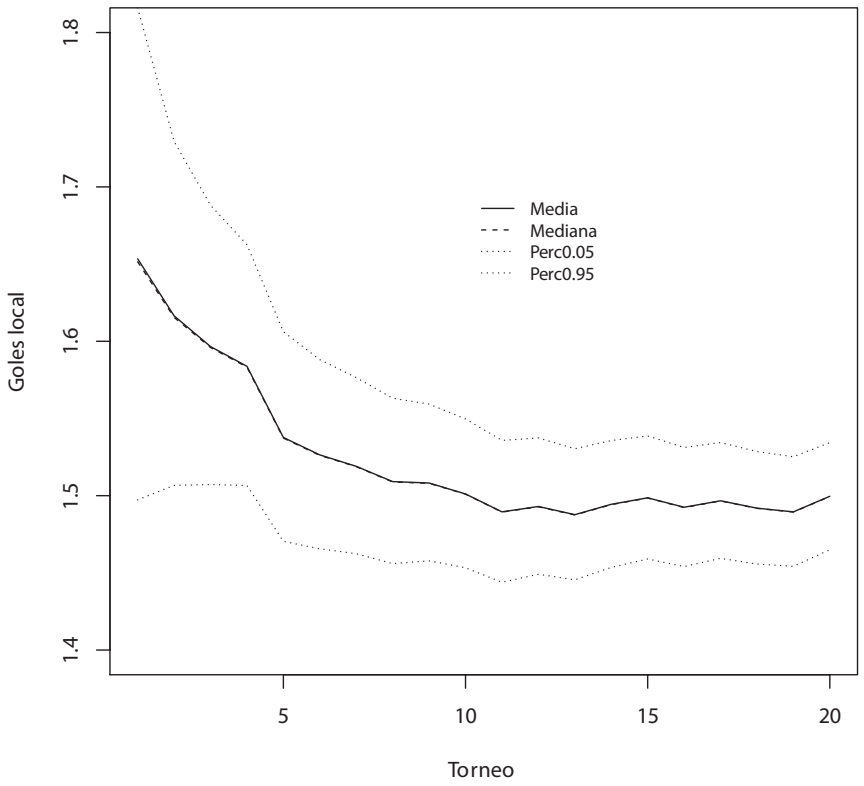

Figura 6.1: *se usan distribuciones a priori no informativa Laplace para iniciar el proceso. Las aposterioris son Gamma(α, β) para el problema de los goles del local bajo el supuesto que se distribuye Poisson con parámetro λ. Notamos como disminuye el número promedio de goles a medida que el torneo es más reciente*

Capítulo 7

Marginalización

Mucho del trabajo estadístico aplicado se centra sobre ciertos parámetros que son de interés primario por parte del investigador, por ejemplo los parámetros de localización. Un ingeniero en control de calidad puede tener interés en la variabilidad de un proceso, por lo tanto los demás parámetros pasan a ser secundarios. La marginalización es un concepto fundamental en el trabajo bayesiano [114]. Los parámetros de molestia (nuisance parameters) han recibido atención en la estadística clásica durante mucho tiempo, llevando a diferentes soluciones sobre las cuales no hay total acuerdo. En la estadística bayesiana esto no resulta ser un problema, pues la marginalización se reduce a un problema de integración (lo cual no quiere decir que sea fácil) o de sumar para eliminar los parámetros de molestia.

Ejemplo 7.1. Eliminando un término de molestia. En muchas situaciones tenemos un vector de parámetros X, pero solo estamos interesados realmente en unos pocos componentes. Debemos por lo tanto proceder a «eliminar» aquellos términos de molestia. Esto lo hacemos mediante la marginalización. Suponga que $x_1, \ldots, x_n$ es una muestra aleatoria de una $N(\mu, \sigma^2)$, donde (μ, σ^2) son desconocidos. Sea $\tau = 1/\sigma^2$. Suponga que especificamos una a priori no informativa de Jeffreys

$$\xi\left(\mu, \sigma^2\right) \propto \tau^{-1/2}$$

Ahora,

$$\xi\left(\mu, \tau \,|\, \boldsymbol{x}\right) \propto \tau^{\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right\}$$

Así, para eliminar el término nuisance τ marginalizamos

$$\xi\left(\mu \,|\, \boldsymbol{x}\right) \propto \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right\} d\tau$$

No es difícil llegar a:

$$\xi\left(\mu \,|\, \boldsymbol{x}\right) \propto \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right\} \exp\left\{-\frac{n\tau}{2}\left(\bar{x} - \mu\right)^2\right\} d\tau$$

Sea:

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

Entonces,

$$\begin{aligned}
\xi(\mu \,|\, \boldsymbol{x}) &\propto \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}\left(ns^2 + n(\mu-\bar{x})^2\right)\right\} d\tau \\
&\propto \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left\{-\frac{n\tau}{2}\left(s^2 + (\mu-\bar{x})^2\right)\right\} d\tau \\
&\propto \frac{\left(\frac{n}{2}\left(s^2 + (\mu-\bar{x})^2\right)\right)^{(n+1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} \\
&\quad \times \int_0^\infty \frac{\Gamma\left(\frac{n+1}{2}\right)}{\left(\frac{n}{2}\left(s^2 + (\mu-\bar{x})^2\right)\right)^{(n+1)/2}} \tau^{\frac{n+1}{2}-1} \\
&\quad \times \exp\left\{-\frac{n\tau}{2}\left(s^2 + (\mu-\bar{x})^2\right)\right\} d\tau \\
&\propto \left(s^2 + (\mu-\bar{x})^2\right)^{-(n+1)/2} \\
&\propto \left(1 + \frac{(\mu-\bar{x})^2}{s^2}\right)^{-(n+1)/2} \\
&\propto \left(1 + \left(\frac{n-1}{n-1}\right)\frac{(\mu-\bar{x})^2}{s^2}\right)^{-(n+1)/2} \\
&\propto \left(1 + \left(\frac{1}{n-1}\right) t_{n-1}^2\right)^{-(n+1)/2}, \text{ donde } t_{n-1} = \frac{\sqrt{n-1}(\mu-\bar{x})}{s}
\end{aligned}$$

Así,

$$\mu|\boldsymbol{x} \sim t\left(n-1, \bar{x}, \frac{s^2}{n}\right)$$

Por lo tanto,

$$\frac{\mu - \bar{x}}{s/\sqrt{n}} \sim t_{(n-1)}$$

A pesar de haber llegado a un resultado que es de uso común en la estadística clásica, la interpretación aquí es diferente.

Ejemplo 7.2. Eliminando otro término de molestia En el ejemplo anterior supongamos que el término de molestia es μ. Debemos por lo tanto hallar $\xi(\tau|\boldsymbol{x})$. Procedemos de manera similar:

$$\begin{aligned}
\xi(\tau|\boldsymbol{x}) &\propto \int_{-\infty}^{\infty} \tau^{\frac{n-1}{2}} \exp\left\{-\frac{\tau}{2}\left(ns^2 + n(\mu-\bar{x})^2\right)\right\} d\mu \\
&\propto \tau^{\frac{n-1}{2}} \exp\left\{-\frac{ns^2\tau}{2}\right\} \tau^{-1/2} \int_{-\infty}^{\infty} \tau^{1/2} \exp\left\{-\frac{n\tau}{2}(\mu-\bar{x})^2\right\} d\mu \\
&\propto \tau^{\frac{n}{2}-1} \exp\left\{-\frac{ns^2\tau}{2}\right\}
\end{aligned}$$

Así,

$$\tau|\boldsymbol{x} \sim Gamma\left(\frac{n}{2}, \frac{ns^2}{2}\right)$$

De lo anterior obtenemos que,

$$ns^2\tau \sim \chi^2_{n-1}$$

Albert [115] presenta conceptos tales como la distribución posterior perfilada. Si tenemos un problema donde exista un parámetro de molestia (θ, ν) y el parámetro de interés es θ, a ν se le conoce como un parámetro de molestia (nuisance) y la distribución posterior es $\xi(\theta, \nu\,|x\,)$, si la a priori que se ha usado es una distribución no informativa, esta posterior será proporcional a la verosimilitud. La marginalización equivaldría a lo que se conoce como una verosimilitud integrada. Otra alternativa es considerar la posterior perfilada, en la cual el parámetro de molestia se elimina remplazando este parámetro por el valor $\hat{\nu}(\theta)$ que maximiza la verosimilitud conjunta. Esta posterior perfilada será:

$$\xi^P(\theta\,|x\,) \propto \xi(\theta, \hat{\nu}(\theta)\,|x\,)$$

Si la a priori es una uniforme, entonces esto se conoce como la verosimilitud perfilada. Desde el punto bayesiano es preferible trabajar marginalizando, sin embargo en algunos casos pudiera ser mejor considerar la perfilada por cuestiones computacionales.

Una ayuda es considerar transformaciones que produzcan una matriz de varianzas y covarianzas entre θ y ν cercana a una matriz diagonal. El concepto de independencia es cercano al de ortogonalidad en la estadística clásica. Si $l(\theta, \nu)$ denota la log-verosmilitud de (θ, ν), entonces la matriz de información observada se define como:

$$I(\theta,\nu) = -\begin{bmatrix} \frac{\partial^2 l(\theta,\nu)}{\partial\theta^2} & \frac{\partial^2 l(\theta,\nu)}{\partial\theta\partial\nu} \\ \\ \frac{\partial^2 l(\theta,\nu)}{\partial\theta\partial\nu} & \frac{\partial^2 l(\theta,\nu)}{\partial\nu^2} \end{bmatrix}$$

evaluada en el estimador de máxima verosimilitud $\left(\hat{\theta}, \hat{\nu}\right)$ de (θ, ν). Los parámetros se dice que son ortogonales si la matriz anterior evaluada en $\left(\hat{\theta}, \hat{\nu}\right)$ es diagonal.

Capítulo 8

Inferencia bayesiana

8.1. Estimación puntual

Dada una distribución sobre un parámetro particular, digamos θ, requerimos seleccionar un mecanismo para escoger un «buen»un estimador $\hat{\theta}$. Supongamos que θ_0 es el verdadero parámetro, desconocido. Sea d nuestra adivinanza de este valor. Debemos de alguna forma medir el error que cometemos (digamos que esto puede ser una multa o un pago) al adivinar a θ_0 mediante d. Esto puede ser medido por $(d-\theta_0)^2$ o por $|d-\theta_0|$ o mediante alguna otra función.

Un problema estadístico puede resumirse como (S, Ω, D, L), donde,

S: Es el espacio muestral de un experimento relevante que tiene asociada una variable aleatoria X cuya distribución de probabilidad está parametrizada por un elemento de Ω.

Ω: Espacio parametral (en un sentido amplio)

D: Un espacio de decisiones

L: Una función de pérdida

Una vez un problema estadístico ha sido especificado, el problema de inferencia estadística es seleccionar un procedimiento (estadístico), a veces llamado una función de decisión, que nos describe la forma de tomar una decisión una vez un resultado muestral ha sido obtenido.

Definición 8.1. *Una función de decisión o procedimiento estadístico es una función o estadístico d que mapea de S a D.*

Definición 8.2. *Sea D un espacio arbitrario de decisiones. Una función no negativa L que mapea de $D \times \Omega$ a $\mathbb{R}$ es llamada una función de pérdida.*

Definición 8.3. *El valor esperado de $L(d(X), \theta)$ es llamado la función de riesgo:*

$$R(d, \theta) = E_\theta \left[L\left(d(X), \theta\right) \right] = \int L\left(d(x), \theta\right) \, dP_\theta(x)$$

Función de pérdida cuadrática:

$$L(d,\theta) = (d-\theta)^2$$

Miremos el riesgo para esta función de pérdida. Sea:

$$b = E_{\xi(\theta|\boldsymbol{x})}(\theta) = \int \theta\, \xi\left(\theta|\boldsymbol{x}\right)\, d\theta,$$

el promedio de la distribución a posteriori. Entonces,

$$\begin{aligned} E\left[L(d,\theta)\right] &= \int L(d,\theta)\, \xi\left(\theta|\boldsymbol{x}\right)\, d\theta \\ &= \int (d-b+b-\theta)^2\, \xi\left(\theta|\boldsymbol{x}\right)\, d\theta \\ &= (d-b)^2 + \int (b-\theta)^2\, \xi\left(\theta|\boldsymbol{x}\right)\, d\theta \\ &\geq \int (b-\theta)^2\, \xi\left(\theta|\boldsymbol{x}\right)\, d\theta, \end{aligned}$$

para cualquier valor de d. La desigualdad anterior se convierte en igualdad cuando $d = b$. El estimador bayesiano bajo una función de pérdida cuadrática es la media de la distribución posterior.

Función de pérdida error absoluto:

$$L(d,\theta) = |d-\theta|$$

El riesgo es minimizado tomando d como la mediana de la distribución posterior, digamos d^*. O sea, la mediana es el estimador bayesiano cuando la función de pérdida es el valor absoluto. Para mostrar esto supongamos otra decisión tal que $d > d^*$. Entonces,

$$|\theta-d| - |\theta-d^*| = \begin{cases} d^*-d & \text{si } \theta \geq d, \\ d+d^*-2\theta & \text{si } d^* < \theta < d, \\ d-d^* & \text{si } \theta \leq d^*. \end{cases}$$

Ya que $(d+d^*-2\theta) > (d^*-d)$ cuando $d^* < \theta < d$, entonces el siguiente resultado se consigue,

$$\begin{aligned} E(|\theta-d| - |\theta-d^*|) &\geq (d^*-d)P(\theta \geq d) + (d^*-d)P(d^* < \theta < d) \\ &\quad +(d-d^*)P(\theta \leq d^*) \\ &= (d-d^*)\left[P(\theta \leq d^*) - P(\theta > d^*)\right] \geq 0 \end{aligned}$$

Esta última desigualdad sigue del hecho que d^* es la mediana de la distribución de θ. La primera desigualdad en este conjunto de ecuaciones será una igualdad si, y solo si, $P(d^* < \theta < d) = 0$. La desigualdad final será una igualdad si, y solo sí,

$$P(\theta \leq d^*) = P(\theta > d^*) = \frac{1}{2}$$

Estas condiciones implican que d es también una mediana. Por lo tanto, $E(|\theta - d|) \geq E(|\theta - d^*|)$, y la igualdad se cumple si, y solo si, d es también mediana.

Una prueba similar puede hacerse si $d < d^*$.

Función de pérdida error absoluto asimétrica:

$$\begin{aligned} L(d,\theta) &= (1-p)\,|d-\theta| \text{ si } d < \theta \\ &= p\,|d-\theta| \text{ si } d \geq \delta, \end{aligned}$$

donde $0 < p < 1$. Bajo esta función de pérdida el estimador bayesiano se encuentra resolviendo la siguiente ecuación:

$$\int_{-\infty}^{d} \xi\left(\theta\,|\boldsymbol{x}|\right)\,d\theta = p$$

Función de pérdida para un espacio parametral discreto:

$$\begin{aligned} L(d,\theta) &= 0 \text{ si } d = \theta \\ &= 1 \text{ si } d \neq \theta \end{aligned}$$

El riesgo es minimizado maximizando la a posteriori, es decir, el estimador bayesiano es la moda posterior.

Función de pérdida escalonada:

$$\begin{aligned} L(d,\theta) &= 0 \text{ si } |d-\theta| \leq \delta \\ &= 1 \text{ si } |d-\theta| > \delta, \end{aligned}$$

donde δ es un número predeterminado, usualmente pequeño.

$$\begin{aligned} E\left[L(d,\theta)\right] &= \int_{\Theta} I\left(|d-\theta| > \delta\right)\xi\left(\theta|\boldsymbol{x}\right)\,d\theta \\ &= \int_{\Theta} I\left(1-\left(|d-\theta| \leq \delta\right)\right)\xi\left(\theta|\boldsymbol{x}\right)\,d\theta \\ &= 1 - \int_{d-\delta}^{d+\delta} \xi\left(\theta|\boldsymbol{x}\right)\,d\theta \\ &\approx 1 - 2\delta\xi\left(d|\boldsymbol{x}\right) \end{aligned}$$

Para minimizar el riesgo es necesario maximizar $\xi\left(d|\boldsymbol{x}\right)$ con respecto a d y el estimador bayesiano es el maximizador. Por lo tanto, el estimador bayesiano será el que maximiza la posterior, esto es, el valor modal. Este estimador es llamado el estimador máximo-a posteriori (MAP).

Propiedad de invarianza de los estimadores de máximo-a posteriori

Sea $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ el parámetro k-dimensional y Θ el espacio parametral. Se desea hallar el estimador de máximo a posteriori de $\boldsymbol{g} = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta}))$, para $1 \leq r \leq k$. Sea G que denota el espacio parametral inducido por la transformación de Θ. G es un espacio r-dimensional. Definamos:

$$\xi^*(\boldsymbol{g}|x_1, \ldots, x_n) = \sup_{\{\boldsymbol{\theta}: g(\boldsymbol{\theta})=\boldsymbol{g}\}} \xi(\boldsymbol{\theta}|x_1, \ldots, x_n),$$

ξ^* es algunas veces llamada la a posteriori inducida por $\boldsymbol{g}$. Cuando estimamos $\boldsymbol{\theta}$ maximizamos la función de verosimilitud $\xi(\boldsymbol{\theta}|x_1, \ldots, x_n)$ como función de $\boldsymbol{\theta}$ para valores fijos de la muestra. Cuando estimamos $g(\boldsymbol{\theta}) = \boldsymbol{g}$ maximizamos la a posteriori inducida por la función g, ξ^*, como una función de $\boldsymbol{g}$ manteniendo fija la muestra. Así el estimador de MAP de $g(\boldsymbol{\theta}) = \boldsymbol{g}$, denotada por $\hat{\boldsymbol{g}}$, es cualquier valor que maximice la función a posteriori inducida para la muestra fija; esto es, $\hat{\boldsymbol{g}}$ es tal que,

$$\xi^*(\hat{\boldsymbol{g}}|x_1, \ldots, x_n) \leq \xi^*(\boldsymbol{g}|x_1, \ldots, x_n), \quad \forall \boldsymbol{g} \in G$$

Teorema 8.1. Propiedad de invarianza *Sea* $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\right)$*, el estimador MAP en la a posteriori* $\xi(\boldsymbol{\theta}|Datos)$*. Si* $g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta}))$*, para* $1 \leq r \leq k$*, es una transformación del espacio parametral* Θ*, entonces un estimador MAP en la densidad inducida a posteriori es* $g(\hat{\boldsymbol{\theta}})$*.*

Prueba:
Sea $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\right)$ el MAP. Es suficiente mostrar que,

$$\xi^*\left(g(\hat{\boldsymbol{\theta}})|x_1, \ldots, x_n\right) \leq \xi^*(g(\boldsymbol{\theta})|x_1, \ldots, x_n),$$

para todo $\boldsymbol{\theta} \in \Theta$, lo cual sigue inmediatamente de la desigualdad,

$$\begin{aligned}
\xi^*(\boldsymbol{g}|x_1, \ldots, x_n) &= \sup_{\{\boldsymbol{\theta}: g(\boldsymbol{\theta})=\boldsymbol{g}\}} \xi(\boldsymbol{\theta}|x_1, \ldots, x_n) \\
&\leq \sup_{\boldsymbol{\theta} \in \Theta} \xi(\boldsymbol{\theta}|x_1, \ldots, x_n) \\
&= \xi\left(\hat{\boldsymbol{\theta}}|x_1, \ldots, x_n\right) \\
&= \sup_{\left\{\boldsymbol{\theta}: g(\boldsymbol{\theta})=g(\hat{\boldsymbol{\theta}})\right\}} \xi(\boldsymbol{\theta}|x_1, \ldots, x_n) \\
&= \xi^*\left(g\left(\hat{\boldsymbol{\theta}}\right)|x_1, \ldots, x_n\right)
\end{aligned}$$

Una estimación que puede ser utilizada en una o más dimensiones, especialmente cuando la función de pérdida no ha sido definida explícitamente, es el valor del parámetro en el cual se maximiza la distribución posterior. Para cualquier observación de x, sea $\psi(\cdot|x)$ que denota la distribución posterior de W en el espacio parametral Ω. Sea $\hat{w}(x)$ el valor de w que satisface la relación.

Ejemplo 8.1. Estimación puntual de la media de una población normal con varianza conocida.

- **Datos**: $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Asumimos que $y_i \sim N(\theta, \sigma^2)$, para todo $i = 1, 2, \ldots, n$. y la varianza es conocida.
- **Distribución A priori** para θ:

$$\theta \sim N\left(\mu_o,\, \sigma_o^2\right),$$

o

$$\xi(\theta) \propto \exp\left(-\frac{1}{2}\frac{(\theta-\mu_o)^2}{\sigma_o^2}\right)$$

- **Verosimilitud**:

$$\begin{aligned} f\left(\boldsymbol{y}|\theta,\sigma^2\right) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\sum_{i=1}^{n}\frac{(y_i-\theta)^2}{2\sigma^2}\right) \end{aligned}$$

- **Distribución a posteriori**: se aplica la regla de Bayes

$$\begin{aligned} \xi\left(\theta|\boldsymbol{y}\right) &\propto \xi(\theta)\cdot f(\boldsymbol{y}|\theta) \\ &\propto \xi(\theta)\cdot L\left(\theta|\right) \\ &\propto \exp\left(-\frac{1}{2}\frac{(\theta-\mu_o)^2}{\sigma_o^2}\right)\exp\left(-\sum_{i=1}^{n}\frac{(y_i-\theta)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{\frac{(\theta-\mu_o)^2}{\sigma_o^2}+\sum_{i=1}^{n}\frac{(y_i-\theta)^2}{\sigma^2}\right\}\right) \end{aligned}$$

  La distribución posterior se puede reorganizar y mostrar que,

$$\theta|\boldsymbol{y} \sim N\left(\mu_n, \sigma_n^2\right),$$

  donde,

$$\mu_n = \frac{\frac{1}{\sigma_o^2}\mu_o + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_o^2}+\frac{n}{\sigma^2}} = \frac{\tau_o\mu_o + nr\bar{y}}{\tau_o + nr},$$

  y

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_o^2} + \frac{n}{\sigma^2}$$

Bajo las tres funciones de pérdida el estimador bayesiano para la media será:

$$\hat{\theta} = \mu_n$$

Ejemplo 8.2. Caso Poisson

Sea $y_1, \dots y_n$ una muestra aleatoria de una $Poisson(\lambda)$. Supongamos también que la a priori es una $Gamma(1,1)$. Por lo tanto la a posteriori será $Gamma(1 + \sum_{i=1}^{n} y_i, n+1)$.

El estimador bayesiano para λ

- Bajo la función de pérdida cuadrática es:

$$\hat{\lambda} = \frac{1 + \sum_{i=1}^{n} y_i}{n+1}, \text{ y}$$

- Bajo la función de pérdida escalonada es:

$$\hat{\lambda} = \frac{\alpha^* - 1}{\beta^*} = \frac{\sum_{i=1}^{n} y_i}{n+1} \text{ si } \alpha^* \geq 1,$$

  donde α^* y β^* son los parámetros de la a posteriori.

La siguiente función en R calcula los tres estimadores, bajo el supuesto de una a priori $Gamma(\alpha_0, \beta_0)$:

```
calcula.estimadores.poisson<-function(alfa0,beta0,x,n=lenght(x)) {
alfa1<-alfa0+sum(x)
beta1<-beta0+n
estimador.fpc<-alfa1/beta1
estimador.fpa<-qgamma(0.5,alfa1,beta1)
estimador.fpe<-(alfa1-1)/beta1
list(estimador.fpc=estimador.fpc,
      estimador.fpa=estimador.fpa,
      estimador.fpe=estimador.fpe)
}
```

La utilización será:

```
calcula.estimadores.poisson(1,1,16,n=4)$estimador.fpc
[1] 3.4
calcula.estimadores.poisson(1,1,16,n=4)$estimador.fpa
[1] 3.333571
calcula.estimadores.poisson(1,1,16,n=4)$estimador.fpe
[1] 3.2
```

Las distribuciones a priori Gamma($\alpha_0 = 1$, $\beta_0 = 1$) y a posteriori Gamma($\alpha_1 = 17$, $\beta_1 = 5$), se muestran en la Figura 8.1.

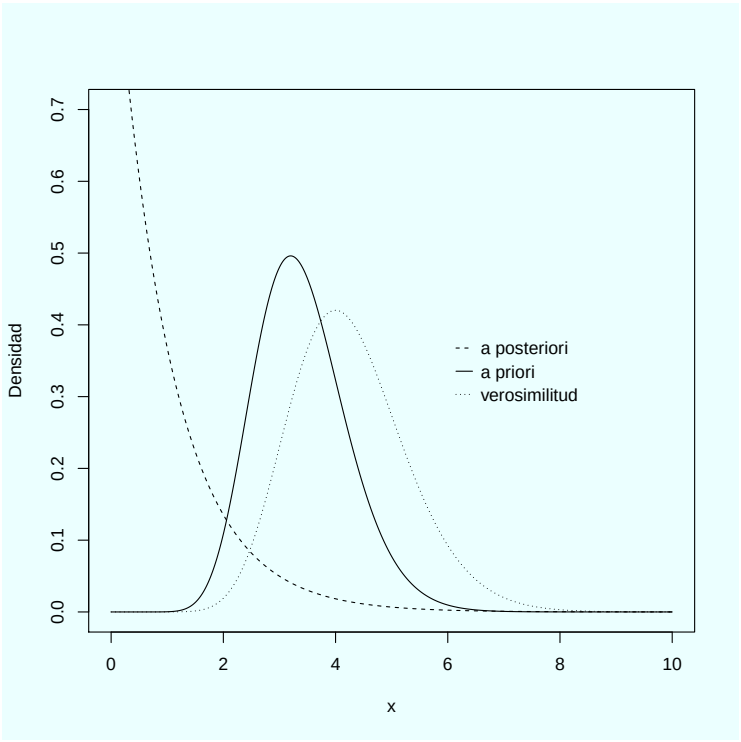

Figura 8.1: *distribuciones a priori Gamma($\alpha_0 = 1$, $\beta_0 = 1$) y a posteriori Gamma($\alpha_1 = 17$, $\beta_1 = 5$) para un problema de conteo Poisson con parámetro λ. En la muestra tomada $n = 4$ y $\sum_{i=1}^{4} y_i = 16$*

Ejemplo 8.3. Goles del equipo visitante. Consideremos el número de goles marcados por el equipo visitante en el torneo profesional de fútbol colombiano. Bajo el supuesto que esta variable se distribuye Poisson con parámetro λ y que tenemos los datos:

	Goles del Visitante					
Torneo	0	1	2	3	4	5
2008-I	61	63	27	9	2	0
2008-II	76	56	23	5	2	0

Si la a priori inicial se escoge no informativa de Laplace entonces:

$$\begin{aligned} \xi_1\left(\lambda | TorneoI - 2008\right) &\propto \lambda^{\sum x_i} \exp\left(-n\lambda\right)\xi_0(\lambda) \\ &\propto \lambda^{(61\times 0+63\times 1+27\times 2+9\times 3+2\times 4)} \exp\left(-162\lambda\right) \\ &\propto \lambda^{(152)} \exp\left(-162\lambda\right) \end{aligned}$$

O sea, ξ_1 es una $Gamma(153, 162)$.

Considerando los datos del segundo torneo la a posteriori es una $Gamma(153 + 125, 162 + 162)$, o sea una $Gamma(278, 324)$.

La media a posteriori es $\frac{278}{324} = 0.8580247$, y la moda a posteriori es $\frac{277}{324} = 0.8549383$.

8.2. Regiones de credibilidad

Los intervalos de confianza clásicos frecuentemente son malinterpretados y los usuarios actúan como si «grado de confianza»fuera sinónimo de uniformidad dentro del intervalo. Por el contrario, los intervalos de credibilidad ofrecen una visión más intuitiva para el investigador y su interpretación se hace realmente práctica.

8.2.1. Región de la densidad posterior más alta (RDPMA)

Si $p(\theta|Y)$ denota la densidad posterior entonces podemos definir un intervalo de credibilidad utilizando la RDPMA.

Definición 8.4. *[13] Una región R en un espacio parametral Θ es llamada la región de la densidad posterior más alta (RDPMA) de contenido α si*

1. $P(\theta \in R|Y) = \alpha$
2. *Para $\theta_1 \in R$ y $\theta_2 \notin R$, se cumple $P(\theta_1 \in R|Y) \geq P(\theta_2 \in R|Y)$*

Para un contenido de probabilidad α, la RDPMA tiene el volumen más pequeño en el espacio parametral. Hyndman [116] presenta diversos gráficos para representar estas regiones y un algoritmo para su determinación.

Ejemplo 8.4. Distribución Exponencial. Elfessi y Reineke [117] construyen intervalos de credibilidad para la media de la distribución exponencial bajo una distribución a priori propia conjugada.

$$\xi(\theta) = \theta^{\alpha-1} \exp\left(-\beta\theta\right),$$

para $\theta > 0$, $-\infty < \alpha < \infty$ y $\beta \geq 0$. Note que esta distribución a priori corresponde al kernel de una distribución gamma cuando $\alpha \geq 0$. La distribución a posteriori es por lo tanto,

$$\xi\left(\theta|x_1, \ldots x_n\right) \propto \theta^{n+\alpha-1} \exp\left(-\theta\left\{\beta + \sum_{i=1}^{n} x_i\right\}\right)$$

Esta distribución posterior es propia cuando $\alpha + n > 0$, y la constante de proporcionalidad es:

$$\frac{(\beta + \sum_{i=1}^{n})^{\alpha+n}}{\Gamma(\alpha+n)}$$

El intervalo de credibilidad de probabilidad $C100\,\%$ es:

$$\left(\frac{\chi^2_{2(\alpha+n),(1-(1-C)/2)}}{2\left(\beta + \sum_{i=1}^{n} x_i\right)}, \frac{\chi^2_{2(\alpha+n),((1-C)/2)}}{2\left(\beta + \sum_{i=1}^{n} x_i\right)}\right)$$

Ejemplo 8.5. Tiempo hasta el primer gol. Del primer torneo de fútbol del 2005 consideramos los tiempos hasta que se marcó el primer gol (en partidos en los cuales se marcó al menos un gol). Si asumimos que el tiempo hasta el primer gol se

distribuye exponencial y que la a priori es una no informativa de Jeffreys, entonces la a posteriori será:

$$gamma\left(n, \sum_{i=1}^{n} x_i\right)$$

```
tiempo<-scan()
9 80 22 46 9 73 91 62 59 6 46
27 19 77 9 29 60 75 75 16 21
40 24 66 83 55 27 50 81 33 43 67

s.x<-sum(tiempo)
s.x
[1] 1480
n<-length(tiempo)
n
[1] 32
qgamma(c(0.025,0.975),n,rate=s.x)
[1] 0.01478917 0.02973110
1/qgamma(c(0.025,0.975),n,rate=s.x)
[1] 67.61703 33.63482
```

Con una credibilidad del 95 %, el tiempo medio para el primer gol en el torneo colombiano de 2005 es marcado entre los 33.6 y 67.6 minutos de juego.

Ejemplo 8.6. Distribución uniforme. Rossman, Short y Parks [118] presentan la construcción de la región de mayor probabilidad para el «parámetro»de la distribución uniforme $U(0, \theta)$. Asumiendo que $X_1, \ldots, X_n$ sea una muestra aleatoria, la estadística clásica nos presenta a $\text{máx}\{X_i\}$ como el estimador de máxima verosimilitud, y $\frac{n+1}{n}\,\text{máx}\{X_i\}$ el estimador de mínima varianza insesgado.

Ahora, si escogemos una distribución a priori impropia o aplanada de la forma $\xi(\theta) = 1$ para $\theta > 0$, la distribución posterior es proporcional a la función de verosimilitud,

$$\xi(\theta|\boldsymbol{X}) \propto \frac{1}{\theta^n} \text{ para } \theta \geq \text{máx}\{X_i\}$$

La constante de proporcionalidad, que vuelve la distribución posterior propia es $(n-1)(\text{máx}\{X_i\})^{n-1}$. Bajo la función de pérdida cuadrática el estimador bayesiano es igual a la media a posteriori

$$E[\theta|\boldsymbol{X}] = \int_{-\infty}^{\infty} \theta \cdot \xi(\theta|\boldsymbol{X})\, d\theta = \frac{n-1}{n-2}\,\text{máx}\{X_i\}$$

Un intervalo de probabilidad del 95 % se halla resolviendo

$$\int_{LI}^{LS} \frac{(n-1)(\text{máx}\{X_i\})^{n-1}}{\theta^n}\, d\theta = 0.95$$

Ejercicio: Distribución potencia

- Una v.a. X con f.d.p. dada por:

$$f(x) = \alpha x^{\alpha-1}, \text{ con } 0 < x < 1 \text{ y } \alpha > 0,$$

se dice que tiene una distribución potencia.

1. Dibujar el gráfico de esta densidad para $\alpha = 1$, $\alpha = 0.5$ y $\alpha = 1.5$. Las tres funciones en el mismo gráfico.
2. Asumir que las observaciones que se van a obtener provienen de esta distribución. Halle la a priori de Jeffreys.
3. Si su muestra es $X_1, X_2, \ldots, X_n$, encuentre la distribución a posteriori bajo la aprioir de Jeffreys.

- Suponga que se observan las proporciones de un cierto compuesto en una muestra de frascos con una droga X sacados al azar y los resultados fueron: 0.03, 0.05, 0.03,0.02,0.03, 0.07,0.03, 0.07,0.09

Bajo el supuesto que el modelo muestral sea el del punto anterior:

1. Grafique la a priori, la a posteriori y la verosimilitud.
2. Encuentre el estimador de máxima verosimilitud bayesiano.
3. Escriba un porgrama en R para construir un intervalo de mayor densidad del 95 %

Ejemplo 8.7. Intervalo para la Poisson. La siguiente función permite construir un intervalo del 95 % de probabilidad de la mayor densidad para el parámetro de la Poisson.

```
intervalo.poisson <-function(a,b){
x1<-1:499/10000; x2<-0.950+x1
dif<-abs(dgamma(qgamma(x1,a,rate=b),a,rate=b)-
         dgamma(qgamma(x2,a,rate=b),a,rate=b))
x3<-qgamma(x1[which.min(dif)],a,rate=b)
x4<-qgamma(x2[which.min(dif)],a,rate=b)
list(x3=x3,x4=x4) }
```

En el ejemplo que teníamos nos da:

```
intervalo.poisson(17,5)$x3
[1] 1.871629
intervalo.poisson(17,5)$x4
[1] 5.045115
```

Mientras que el intervalo tradicional hallado con ambas colas iguales a $\alpha/2$ es:

```
1.980625  y  5.1966,
```

el cual presenta un poco menos precisión que el anterior.

8.2.2. Intervalos frecuentistas tradicionales para la Poisson

Aquí presentamos una de la múltiples dificultades que tiene la aproximación tradicional, en la cual puede existir más de una regla para construir intervalos de confianza y muchas veces sin la suficiente claridad por parte del investigador sobre cúal de ellos usar, debido en parte a la carencia de elementos de juicio que le permita escoger el mejor en una circunstancia particular.

Intervalo basado en transformaciones (M.T.)

Si $\bar{X} \sim N(\lambda, \sigma^2/n)$, entonces $\log(\bar{X}) \sim lognormal(\log(\lambda), 1/n)$, asumiendo que $\lambda > 0$ [119]. El intervalo de confianza está dado por:

$$\left(\bar{X} \frac{1}{\exp\left(z_{\alpha/2}/\sqrt{n}\right)}, \bar{X} \exp\left(z_{\alpha/2}/\sqrt{n}\right) \right)$$

Método basado en el Teorema Central del Límite (T.C.L.)

Si el tamaño muestral es lo suficientemente grande, podemos aplicar el teorema central del límite.

$$\left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}} \right),$$

donde $s^2 = 1/n \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$. Este es el intervalo propuesto en la mayoría de textos básicos en estadística ([120]; [121]; [122]; [50]).

Método basado en la máxima verosimilitud

Se sabe que si $\hat{\theta}$ es el estimador máximo verosímil para θ (puede ser un vector), bajo ciertas condiciones suaves [119], entonces $\hat{\theta} \sim (\theta, I^{-1}(\theta))$, con $I(\theta)$ siendo la matriz de información de Fisher. Entonces, en el caso exponencial,

$$\left(\bar{X} - z_{\alpha/2}\frac{\sqrt{\bar{X}}}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sqrt{\bar{X}}}{\sqrt{n}} \right)$$

Método exacto

Se sabe que $S = \sum_{i=1}^{n} X_i$ se distribuye Poisson con parámetro $n\lambda$. Un intervalo de confianza exacto se obtiene resolviendo:

$$\exp(n\lambda_L) \sum_{i=0}^{s} \frac{(n\lambda_L)^i}{i!} = 1 - \frac{\alpha}{2},$$

y

$$\exp(n\lambda_U) \sum_{i=0}^{s} \frac{(n\lambda_U)^i}{i!} = \frac{\alpha}{2}$$

Intervalos basados en la razón de verosimilitud relativa

Kalbfleish [123] presenta la metodología para construir intervalos de verosimilitud. Si $L(\lambda)$ es la función de verosimilitud, se define la *función de verosimilitud relativa* como:

$$R(\lambda) = \frac{L(\lambda)}{L(\hat{\lambda})}$$

El conjunto de valores de λ para los cuales $R(\lambda) \geq p$ es llamado el *intervalo de 100 %p de verosimilitud* para λ. Los intervalos del 14.7 % y del 3.6 % de verosimilitud corresponden a intervalos de confianza aproximadamente de niveles del 95 % y del 99 %.

Lo que se debe hacer entonces es hallar las raíces que nos dan los límites del intervalo. Para el caso del parámetro de la exponencial, λ, tenemos que un intervalo de confianza del 95 % se halla encontrando el par de raíces tales que,

$$R(\lambda) = \frac{L(\lambda)}{L(\hat{\lambda})} = \left(\frac{\lambda}{\bar{X}}\right)^{n\bar{X}} \geq K(k, \alpha)$$

Esto se resuelve numéricamente.

Bootstrap

El método bootstrap [124] proporciona una manera directa y sencilla para hallar intervalos simultáneos para los parámetros de la distribución multinomial. Para hallarlos se procede así:

1. A partir de la muestra estime el parámetro por máxima verosimilitud.

$$\hat{\lambda} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

2. Genere M muestras de tamaño n de una distribución exponencial con parámetro $\hat{\lambda}$. Para cada muestra estime el parámetro λ, digamos que para la muestra j el estimador es $\hat{\lambda}_j$

3. Para los $\left\{\hat{\lambda}_j\right\}_{j=1}^{M}$, construya un histograma y calcule los percentiles .025/(k-1) y 0.975/(k-1), denotémoslos por $\hat{\lambda}_i^{(0.025)}$ y $\hat{\pi}_i^{(0.975)}$.

Intervalos cuando hay cero eventos

Las tasas bajas de ocurrencia de eventos en ciertos problemas estadísticos son comunes en muchas situaciones, por ejemplo casos de una enfermedad rara. No es raro tener muestras, aun grandes, donde no tengamos eventos, o sea $X_1 = X_2 = \ldots = X_n = 0$.

Presentamos métodos diferentes para estimar el parámetro de la distribución Poisson en el caso donde en una muestra de tamaño n no se observe ninguna ocurrencia del evento de interés. Estos intervalos de confianza tienen aplicaciones en epidemiología, control de calidad, etc.

Para la Poisson tenemos:

$$P\left(x_1 = 0, x_2 = 0, \ldots, x_n = 0|\lambda\right) = e^{-n\lambda}$$

Es posible establecer una cota superior de tal forma que el 95 % de los λ que pueden generar esa muestra particular sean menores o iguales que esa cota. Se halla $e^{-n\lambda} \geq 0.05$. Esto conduce a $\lambda \leq -\log(0.05)/n$. Una interpretación de esta cota es que entre cero y ella se encuentran todas las distribuciones Poisson que con una probabilidad mayor o igual que 0.05 pueden generar la muestra que poseemos.

Si $\hat{\theta}$ es el EMV de θ, entonces,

$$\hat{\theta} \sim AN\left(\theta, \theta/n\right).$$

También lo es $\hat{\theta} + 1/a_n$, donde $\{1/a_n\}$ es una sucesión de constantes que converge a cero y que asumen ciertas condiciones [119], que podemos ajustar de tal forma que se distribuya asintóticamente normal $AN\left(\theta + 1/a_n, \theta/n + 1/a_n\right)$.

Como un caso particular de tales sucesiones tenemos $1/a_n = 1/n^{1+\delta}$, donde $\delta > 0$. En caso de no observar eventos el intervalo será [125].

$$\left(0; 1/n^{1+\delta} + 1.96\sqrt{1/n^{1+\delta}}\right)$$

Si la distribución a priori es una $Gamma(\alpha, \beta)$, entonces para observaciones Poisson, la distribución a posteriori será $Gamma\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$, con media $\mu = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n}$, este parámetro puede ser utilizado como estimador, si escogemos adecuadamente α y β.

El intervalo se construye solucionando la integral

$$\int_{LI}^{LS} \frac{(\beta + n)^{\alpha + \sum_{i=1}^{n} x_i}}{\Gamma(\alpha + \sum_{i=1}^{n} x_i)} \lambda^{\alpha + \sum_{i=1}^{n} x_i - 1} e^{-\lambda(\beta + n)} d\lambda = 0.95$$

Esto nos conduce a un intervalo (LI, LS). $LI = 0$ si el número de casos es cero.

Selección de valores de α y β

- Se seleccionaron valores de α y β tales que la media de la distribución a priori sea pequeña.
- La distribución Gamma es muy sesgada a la derecha (la mayor parte de la densidad se halla a la derecha de la media).

- Se seleccionó una media igual a uno, razón para hacer $\alpha = \beta$, lo que controlamos es la varianza, que nos da una medida del grado de incertidumbre y entre más pequeño sea α, mayor será la varianza.

En el año 1996, en el tramo La Pintada-Primavera (Ruta Nacional 2509), se registraron 152 accidentes, y no hubo muertes. Si estamos interesados en estimar la tasa de muertes por accidentes tenemos la siguiente tabla aplicando la metodología anterior.

Tabla 8.1: *estimación por intervalos de la tasa de muertes por accidentes en el tramo La Pintada-Primavera en el año 1996*

Método	*Intervalo*
Bayesiano	(0,0.0126)
Cota Máxima	(0,0.0197)
Factor de Correción	(0,0.385)

Notamos que los intervalos Bayesiano y de cota máxima, son los que ofrecen mayor precisión.

8.2.3. Intervalos aproximados

Un *intervalo de verosimilitud*, $I(\theta; \beta \,|\mathbf{x})$, se define como:

$$I(\theta; \beta \,|\mathbf{x}) = \left\{\theta;\, l(\theta \,|\mathbf{x}) \geq e^{-\beta} l\left(\hat{\theta} \,|\mathbf{x}\right)\right\},$$

donde β es una constante positiva.

Decimos que una función de verosimilitud tiene la *forma estándar* si, (Hudson, 1971).

1. $\hat{\theta}$ es único,
2. $l(\theta_L) = 0 = l(\theta_R)$ para algún $\theta_L \leq \hat{\theta} \leq \theta_R$
3. $l(\theta)$ es monotónica sobre $\theta_L \leq \theta \leq \hat{\theta}$ y monotónica decreciente sobre $\hat{\theta} \leq \theta \leq \theta_R$.

Los anteriores conceptos pueden generalizarse al caso de la distribución posterior y tendríamos entonces las siguientes definiciones:

Definición 8.5. Intervalo verosímil posterior. *Un intervalo de probabilidad verosimilitud-posterior, $I(\theta; \beta \,|\mathbf{x})$, se define como:*

$$I(\theta; \beta \,|\mathbf{x}) = \left\{\theta;\, \xi(\theta \,|\mathbf{x}) \geq e^{-\beta} \xi\left(\theta^{Moda} \,|\mathbf{x}\right)\right\},$$

donde β es una constante positiva.

Definición 8.6. A posteriori con forma estándar. *Decimos que una densidad posterior tiene la forma estándar si:*

1. θ^{Moda} *es único,*

2. $\xi(\theta_L) = 0 = l(\theta_R)$ *para algún* $\theta_L \leq \theta^{Moda} \leq \theta_R$

3. $\xi(\theta)$ *es monotónica sobre* $\theta_L \leq \theta \leq \theta^{Moda}$ *y monotónica decreciente sobre* $\theta^{Moda} \leq \theta \leq \theta_R$.

Si asumimos una a priori de Laplace, los intervalos de probabilidad a posteriori pueden ser calculados aproximadamente mediante esta metodología.

Si la distribución poblacional es $N(\mu, 1)$ y la a priori $\xi(\mu) \propto k$, entonces un intervalo aproximado con probablidad 0.954 es,

$$\left(\bar{x} - 2\frac{1}{\sqrt{n}};\ \bar{x} + 2\frac{1}{\sqrt{n}}\right)$$

8.3. Pruebas de hipótesis

Ejemplo 8.8. Poderes Sobrenaturales. Bayarri y Berger en la reunión anual que se lleva a cabo en Valencia (España) presentaron el siguiente caso de psicoquinesis: tres investigadores (Schmidt, Jahn y Radin) en 1987 utilizaron un generador cuántico que recibe una fila de partículas y él desvía cada partícula, independientemente de las otras, hacia una luz roja o una luz verde con igual probabilidad. Se le pidió a un sujeto, quien alegaba tener poderes psicoquinéticos, que tratara de influenciar el generador de tal forma que las partículas se fueran para la luz roja. Se generaron 104.490.000 partículas y se contaron 52.263.470 partículas que se fueron hacia la luz roja. Habrá suficiente evidencia que permita decir que el sujeto tiene poderes sicokinéticos?

Podemos pensar en este exprimento así: cada partícula corresponde a un ensayo $Bernoulli(\pi)$, y un éxito será si la partícula se va para la luz roja. Si X denota el número de éxitos, $X \sim Binomial(n, \pi)$. Tenemos $x = 52.263.470$ como la observación real. Se necesita probar:

$$\begin{aligned} H_0 : \pi &= \frac{1}{2} \quad \text{(El sujeto no tiene poderes)}, \\ H_1 : \pi &\neq \frac{1}{2} \quad \text{(El sujeto tiene poderes)} \end{aligned}$$

El $valor - p = P_{H_0}\left(\left|X - \frac{n}{2}\right| \geq \left|x - \frac{n}{2}\right|\right) \approx 0.0003$ nos lleva a concluir que hay una fuerte evidencia contra H_0.

Si pensamos bayesianamente necesitamos una distribución a priori, pero ahora definida sobre las hipótesis en juego:

$$\xi(H_i) = \text{probabilidad a priori de que } H_i \text{ sea cierta, } i = 0, 1.$$

Bajo $H_1 : \pi \neq 1/2$, sea $\xi(\pi)$ la densidad a priori sobre π. El Bayes objetivo selecciona,

$$Pr\left(H_0\right) = Pr\left(H_1\right) = \frac{1}{2},$$

con $\xi(\pi) = 1 \;\; (0 < \pi < 1)$.

La probabilidad posterior de la hipótesis

$$\begin{aligned} Pr\left(H_0|x\right) &= \text{probabilidad de que} H_0 \text{ sea cierta dados los datos } x \\ &= \frac{f\left(x|\pi = 1/2\right) Pr\left(H_0\right)}{Pr\left(H_0\right) f\left(x|\pi = 1/2\right) + Pr\left(H_1\right) \int f\left(x|\pi\right) \xi(\pi)\, d\pi} \end{aligned}$$

Para la a priori objetiva,

$$Pr\left(H_0|x = 52.263.470\right) \approx 0.92$$

La densidad posterior en $H_1 : \pi \neq 1/2$ es:

$$\xi\left(\pi|x, H_1\right) \propto \xi(\pi) f(x|\pi) \propto 1 \times \pi^x (1-\pi)^{n-x},$$

que es una $Beta\left(52.263.470, \;\; 52.226.530\right)$.

«En cualquier etapa de conocimiento es válido preguntar acerca de una hipótesis que ha sido aceptada, '¿cómo lo sabe?' La respuesta usualmente descansará en algunos datos observacionales. Si preguntamos adicionalmente, '¿qué pensaba usted acerca de la hipótesis antes de que obtuviera los datos?' Nos pueden hablar de algunos datos menos convincentes; pero si vamos lo suficientemente atrás siempre llegaremos a una etapa donde la respuesta debe ser: 'yo pensé que valía la pena considerar el asunto, pero no tenía una opinión acerca de si era cierta. ' ¿Cuál es la probabilidad en esta etapa? Ya tenemos la respuesta. Si no hay razón para creer en una hipótesis en lugar de otra, las probabilidades son iguales»[9].

La aproximación bayesiana a las pruebas de hipótesis está basada en el cálculo de de la probabilidad condicional de una hipótesis H_o dada la información disponible, digamos I_o, esto es, $p(H|I_o)$. Cuando la hipótesis nula es $H_o : \boldsymbol{\theta} \in \Theta_o$ y la alternativa $H_1 : \boldsymbol{\theta} \in \Theta_1$, con $\Theta_o \bigcap \Theta_1 = \emptyset$, son formuladas, hay creencias a priori sobre ambas, digamos $\xi(H_o|I_o)$ y $\xi(H_1|I_o)$, con $\xi(H_o|I_o) + \xi(H_1|I_o) = 1$. Por el teorema de la probabilidad total, la distribución a priori de $\boldsymbol{\theta}$ es:

$$\xi(\boldsymbol{\theta}|I_o) = \xi(\boldsymbol{\theta}|H_o, I_o)\xi(H_o|I_o) + \xi(\boldsymbol{\theta}|H_1, I_o)\xi(H_1|I_o),$$

donde $\xi(\boldsymbol{\theta}|H_i, I_o)$, son las densidades a priori de $\boldsymbol{\theta}$, condicionadas en cada hipótesis. La información muestral es utilizada entonces para calcular de los odds a priori:

$$\frac{\xi(H_o|I_o)}{\xi(H_1|I_o)},$$

los odds posteriores en favor de H_o:

$$\frac{\xi(H_o|I_1)}{\xi(H_1|I_1)} = \frac{p(\boldsymbol{y}|H_o)}{p(\boldsymbol{y}|H_1)} \frac{\xi(H_o|I_o)}{\xi(H_1|I_o)},$$

de la cual se deriva la siguiente regla de decisión:

$$\begin{array}{lll} \text{si} & \xi(H_o|I_1) < \xi(H_1|I_1) & \text{Rechace } H_o, \\ \text{si} & \xi(H_o|I_1) > \xi(H_1|I_1) & \text{Acepte } H_o, \\ \text{si} & \xi(H_o|I_1) = \xi(H_1|I_1) & \text{Indecisión acerca de } H_o \end{array}$$

Definición 8.7 (Factor de Bayes)**.** *La razón $p(\boldsymbol{y}|H_o)/p(\boldsymbol{y}|H_1)$ es llamado el factor de Bayes, denotado por BF o $B_{01}(\boldsymbol{y})$.*

Kass [126] presenta una gran discusión sobre las bondades y desventajas de este criterio para la comparación de modelos.

Al querer probar,

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1,$$

y si se considera $f(\boldsymbol{x}|\theta)$ la verosimilitud de $\boldsymbol{x}$ dado θ, entonces, tenemos las siguientes formas del factor de Bayes.

$$\begin{aligned} B_{01}(\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\theta_1)} \qquad \text{(Prueba simple vs. simple).} \\ B_{01}(\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta_0)}{\int_{\Theta_1} f(\boldsymbol{x}|\theta)\xi_1(\theta)d\theta} \qquad \text{(Prueba simple vs. compuesta).} \\ B_{01}(\boldsymbol{x}) &= \frac{\int_{\Theta_0} f(\boldsymbol{x}|\theta_0)\xi_0(\theta)d\theta}{\int_{\Theta_1} f(\boldsymbol{x}|\theta)\xi_1(\theta)d\theta} \qquad \text{(Prueba compuesta vs. compuesta)} \end{aligned}$$

Jeffrey [85], presenta los siguientes criterios sobre el factor de Bayes para decidir cuándo optar por H_0:

Tabla 8.2: *criterios de decisión sobre el factor de Bayes*

1	$<$	B			Hipótesis Nula se sostiene
$10^{-1/2}$	$<$	B	$<$	1	Evidencia contra H_0, pero apenas para mencionar
10^{-1}	$<$	B	$<$	$10^{-1/2}$	Evidencia sustancial contra H_0
$10^{-3/2}$	$<$	B	$<$	10^{-1}	Evidencia fuerte contra H_0
10^{-2}	$<$	B	$<$	$10^{-3/2}$	Evidencia muy fuerte contra H_0
		B	$<$	10^{-2}	Evidencia decisiva contra H_0

Sinharay y Stern [127] discuten la sensibilidad de los factores de Bayes a la selección de la distribuciones aprioris sobre los parámetros de los modelos.

Ejemplo 8.9. Continuación del ejemplo de psicoquinesis. Calculemos el factor de Bayes para la situación del individuo con poderes.

$$B_{01} = \frac{f\left(x \left| \pi = \frac{1}{2}\right.\right)}{\int_0^1 f(x|\pi)\xi(\pi)\, d\pi} \approx 12$$

Note que $\underset{\text{(Odds posterior)}}{\frac{Pr(H_0|x)}{Pr(H_1|x)}} = \underset{\text{(Odds a priori)}}{\frac{Pr(H_0)}{Pr(H_1)}} \times \underset{\text{(Factor de Bayes)}}{B_{01}}$

Una región de credibilidad del 95 % para π bajo el supuesto de H_1 es $C = (0.50008, 0.50027)$. Es decir, se tiene una probabilidad de 0.95 de que la proporción de éxitos en el experimento esté entre estos dos valores.

Cuando las probabilidades a priori son iguales, el factor de Bayes determina la regla de decisión. La evaluación del factor de Bayes involucra el cálculo de:

$$\begin{aligned} p(\boldsymbol{y}|H_0) &= \int_{\Theta_0} p(\boldsymbol{y}|H_0, \boldsymbol{\theta})\xi(\boldsymbol{\theta}|H_0, I_0)\, d\boldsymbol{\theta} \\ p(\boldsymbol{y}|H_1) &= \int_{\Theta_1} p(\boldsymbol{y}|H_1, \boldsymbol{\theta})\xi(\boldsymbol{\theta}|H_1, I_0)\, d\boldsymbol{\theta} \end{aligned}$$

El factor de Bayes proporciona una indicación de cuánto cambian nuestras razones de probabilidad de una situación sin datos, a la luz de los datos, para favorecer un modelo. Puede verse como una medida de la evidencia proporcionada por los datos en favor de un modelo comparado con un competidor. El logaritmo del factor de Bayes ha sido llamado *el peso de la evidencia* proporcionada por los datos [128] [109].

McGee [109] presenta el factor de Bayes relacionándolo con la medida del sonido conocida como decibeles. Él utiliza el logaritmo en base 10 para esto. Así, para comparar la evidencia a favor dada por los datos hacia H_0, se determinaría así:

$$\begin{aligned} 10\log_{10}\left(\frac{Pr\left(H_0|x\right)}{Pr\left(H_0|x\right)}\right) &= 10\log_{10}\left(\frac{Pr\left(H_0\right)}{Pr\left(H_1\right)}\right) + 10\log_{10}\left(B_{01}\right), \\ ev\left(H_0\left|Datos\right.\right) &= ev\left(H_0\right) + 10\log_{10}\left(B_{01}\right), \\ \text{(Evidencia posterior)} &= \text{(Evidencia a priori)} + \text{(Evidencia en datos)} \end{aligned}$$

Por ejemplo, si tenemos dos hipótesis H_1 y H_2 y $\xi\left(H_1\right) = \xi\left(H_2\right) = 0.5$ y además la información muestral corresponde a un experimento Bernoulli donde un éxito favorece H_1, y de 10 ensayos se observan 3 éxitos, entonces la evidencia a priori a favor de H_1 es,

$$10\log_{10}\left(\frac{Pr\left(H_1\right)}{Pr\left(H_2\right)}\right) = 10\log_{10}\left(\frac{0.5}{0.5}\right) = 0$$

Ahora, la evidencia en la muestra sería:

$$10\log_{10}\left(\frac{Pr\left(Resultado\left|H_1\right.\right)}{Pr\left(Resultado\left|H_2\right.\right)}\right) = 10\log_{10}\left(\frac{0.3}{0.7}\right) = -3.679768$$

La evidencia a favor de H_1 se redujo en 3.7 decibeles.

Ejemplo 8.10. La prueba de sabor [129]. Se conduce un experimento para determinar si un individuo tiene poder discriminatorio. El individuo debe identificar correctamente cuál de las dos marcas de un producto ha recibido (obviamente las condiciones experimentales deben ser óptimas). Si θ denota la probabilidad de que seleccione la correcta en el i-ésimo ensayo, entonces la variable Bernoulli x_i denota el resultado del experimento, tomando el valor de 1 si acierta y 0 si falla. Supongamos que en los 6 primeros ensayos los resultados son 1, 1, 1, 1, 1 y 0. Nuestro problema es verificar,

$$H_0 : \theta = \frac{1}{2} \text{ versus } H_1 : \theta > \frac{1}{2}$$

En este caso tenemos una hipótesis simple contra una compuesta donde $\Theta_0 = \frac{1}{2}$ y $\Theta_1 = \left(\frac{1}{2}, 1\right)$. Asumamos una distribución a priori uniforme sobre θ bajo la hipótesis alternativa. Así $\xi_1(\theta) = 2$ si $\frac{1}{2} < \theta < 1$. Ahora el factor de Bayes es:

$$B_{01}(\boldsymbol{x}) = \frac{\left(\frac{1}{2}\right)^6}{\int_{1/2}^1 \theta^5(1-\theta)2d\theta} = \frac{1}{2.86}$$

Esto sugiere que esta persona parece tener algún poder discriminatorio, pero no mucho.

El factor de Bayes puede verse como la versión bayesiana de la prueba clásica de la razón de verosimilitudes [128]. Si se asumen dos hipótesis simples, digamos $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$, el factor de Bayes se reduce a la razón de verosimilitud $f(\boldsymbol{y}|\boldsymbol{\theta}_1)/f(\boldsymbol{y}|\boldsymbol{\theta}_2)$.

Ejemplo 8.11. Sean $y_1, \ldots, y_n|\theta$ variables independientes y distribuidas Poisson con parámetro θ. Así,

$$p\left(y_i|\theta\right) = \frac{\theta^{y_i}\, e^{-\theta}}{y_i!}, \text{ para } \theta > 0, y_i = 0, 1, 2, \ldots$$

Sea $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ dos hipótesis simples, con $\xi\left(H_0|I_0\right) = \xi\left(H_1|I_0\right)$. El Factor Bayes es:

$$\left(\frac{\theta_0}{\theta_1}\right)^{\sum_i y_i} \exp\left(\theta_1 - \theta_0\right),$$

y por tanto, ya que la distribución a priori asigna igual probabilidad a las hipótesis, la regla de decisión será no rechazar H_0 si el Factor de Bayes es mayor que 1.

Paradoja de Lindley

Aitkin [130] presenta y discute la paradoja de Lindley la cual muestra un grave problema que aparece cuando se construye el factor de Bayes con una hipótesis simple y con el uso de distribuciones a priori no informativas. Si consideramos una población normal con varianza conocida σ^2 y una hipótesis tiene la media igual a μ_1 (especificada) y la otra hipótesis no especifica la media, denotada por μ_2. Si se asigna una distribución uniforme para μ_2 pero se acota el espacio parametral ($\mu_2 \in (-C, C)$) tal que:

$$\xi\left(\mu_2\right) = \frac{1}{2C}$$

El factor de Bayes está dado por:

$$BF = \frac{\frac{2C\sqrt{n}}{\sigma}\phi(z)}{\Phi\left(\frac{\bar{y}+C}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\bar{y}-C}{\sigma/\sqrt{n}}\right)}$$

y el denominador tiende a 1 cuando $C \longrightarrow \infty$. El BF crece indefinidamente si $C \longrightarrow \infty$ o si $n \longrightarrow \infty$.

Aitkin [130] comenta que esta paradoja se debe a la presencia de una hipótesis simple y esto se podría resolver re-expresándola o considerando distribuciones a priori informativas, pero esto exigiría un análisis de sensibilidad.

Carlin y Louis [20] presentan algunas propuestas realizadas por otros autores cuando la a priori es impropia al factor de Bayes:

- *Factor de Bayes parcial*: particionar la muestra en dos, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, de tal forma que con una se calculen las distribuciones posteriores bajo las hipótesis y la otra para calcular el Factor de Bayes,

$$BF_{\mathbf{y}_2|\mathbf{y}_1} = \frac{p\left(\mathbf{y}_2 \left| \mathbf{y}_1, H_1\right.\right)}{p\left(\mathbf{y}_2 \left| \mathbf{y}_1, H_2\right.\right)}$$

- *Factor de Bayes intrínseco*: este fue propuesto en 1996 por Berger y Pericchi. Ellos proponen el cálculo el uso de todas las posibles muestras mínimas de entrenamiento que hacen que la a posteriori sea propia. Si el tamaño mínimo es n_1, se obtienen todas las particiones posibles y se calculan los Factores de Bayes Parciales. Luego se calcula la media geométrica de estos.

- *Factor de Bayes fraccional*: este factor fue propuesto por O'Hagan [131]. Se considera una submuestra mínima n_1, y se habla de la fracción $b = n_1/n$. El factor de Bayes fraccional es:

$$BF_{b,\mathbf{y}} = \frac{p\left(\mathbf{y}, b \left| H_1\right.\right)}{p\left(\mathbf{y}, b \left| H_2\right.\right)},$$

y

$$p\left(\mathbf{y}, b \left| H_i\right.\right) = \frac{\int f\left(\mathbf{y} \left| \boldsymbol{\theta}_i, H_i\right.\right) \xi_i\left(\boldsymbol{\theta}_i\right)\, d\boldsymbol{\theta}_i}{\int f\left(\mathbf{y} \left| \boldsymbol{\theta}_i, H_i\right.\right)^b \xi_i\left(\boldsymbol{\theta}_i\right)\, d\boldsymbol{\theta}_i}$$

También sugiere valores para b como $b = \text{máx}\left\{n_1,\, \log(n)\right\}/n$ y $b = \text{máx}\left\{n_1,\, \sqrt{n}\right\}/n$.

Una modificación posible viene de la idea de validación cruzada dejando un punto afuera.

$$BF_i = \frac{\int f\left(y_i \left| \boldsymbol{\theta}_1, H_1\right.\right) \xi_1\left(\boldsymbol{\theta}_1 \left| \mathbf{y}_{(-i)}\right.\right)\, d\boldsymbol{\theta}_1}{\int f\left(y_i \left| \boldsymbol{\theta}_2, H_2\right.\right) \xi_2\left(\boldsymbol{\theta}_2 \left| \mathbf{y}_{(-i)}\right.\right)\, d\boldsymbol{\theta}_2}$$

Finalmente resumimos esto con un promedio.

Ejemplo 8.12. Medición de huevos de gallina: v*eamos el siguiente ejemplo, donde se quiere verificar si es posible considerar un modelo en particular. Para este ejemplo usamos los datos presentados en la tabla 8.3 y el paquete MCMCpack de R [48].*

Consideremos el modelo:

$$Peso = \beta_0 + \beta_1 Largo + \beta_2 Ancho + \epsilon,$$

donde $\epsilon \sim N(0, \sigma^2)$. Queremos verificar si es posible considerar el modelo más simple sin la variable Ancho. O sea, queremos probar $H_0 : \beta_2 = 0$.

```
require(MCMCpack)

eval.promedio.i<-function(res1,X.i,y.i)dnorm(y.i,sum(c(1,X.i)*
  res1[-length(res1)]),sd=sqrt(res1[length(res1)]))

BF.menos.i<-function(i,y,X){
Xi<-X[i,]; yi<-y[i]; y<-y[-i]; X<-X[-i,]
X<-as.matrix(X)
if(ncol(X)==1){
xnam<-"X[,1]"
fmla <-as.formula(paste("y ~ ", xnam)) }
else{
xnam <- paste0("X[,",1:ncol(X),"]")
fmla <- as.formula(paste("y ~ ", paste(xnam,collapse= "+"))) }
res1<-MCMCregress(fmla)
media1<-mean(apply(res1,1,eval.promedio.i,Xi,yi))
return(media1) }
# fin
# BF.menos.i

# Huevos (Luego de leer los datos)
mean(resu1<-sapply(1:length(Peso),BF.menos.i,Peso,cbind(Largo,Ancho)))
[1] 0.1334162
mean(resu2<-sapply(1:length(Peso),BF.menos.i,Peso,matrix(Largo,ncol=1)))
[1] 0.04558463

BF.i<-resu1/resu2

summary(BF.i)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4709  2.1820  3.1150  5.0240  3.9270 25.3400
quantile(BF.i,probs=c(0.025,(1:9)/10,0.975))
      2.5%        10%        20%        30%        40%        50%        60%
 0.6803116  0.9137962  1.7709874  2.3598472  2.7955523  3.1151212  3.3520343
       70%        80%        90%      97.5%
 3.5824575  4.0434896 12.3718778 24.2990179
```

Tabla 8.3: *largo, ancho, peso y volumen para 39 huevos de gallina*

	Largo	Ancho	Peso	Volumen
1	6.70	5.07	98.22	90.00
2	6.58	4.95	87.63	80.00
3	6.81	4.93	94.57	90.00
4	6.82	4.87	90.38	83.00
5	6.70	4.86	89.96	83.00
6	7.19	4.89	96.72	90.00
7	6.76	4.73	85.84	80.00
8	7.01	4.69	87.68	80.00
9	7.31	4.68	90.01	82.00
10	6.21	4.81	80.47	72.00
11	6.53	4.58	76.82	70.00
12	5.87	4.57	67.83	61.00
13	6.18	4.57	71.17	63.00
14	6.19	4.50	70.07	65.00
15	5.75	4.67	69.04	67.00
16	6.04	4.48	67.27	60.00
17	6.06	4.57	70.64	68.00
18	6.11	4.51	68.76	64.00
19	6.81	4.41	72.14	69.00
20	5.89	4.60	67.85	61.00
21	5.38	4.02	44.33	41.00
22	5.37	3.99	44.76	40.00
23	5.21	3.94	42.97	39.00
24	5.44	3.96	44.41	40.00
25	5.52	4.04	50.16	50.00
26	5.12	4.05	43.21	42.00
27	5.19	4.01	45.21	42.00
28	5.27	4.07	46.57	44.00
29	5.59	3.83	40.81	41.00
30	5.35	4.05	46.19	45.00
31	5.41	3.98	43.91	45.00
32	5.41	3.98	43.45	43.00
33	5.78	4.26	54.32	57.00
34	5.65	4.29	53.38	54.00
35	5.62	4.03	50.67	49.00
36	5.75	4.26	54.27	55.00
37	5.71	4.25	54.54	56.00
38	5.87	4.32	56.50	58.00
39	5.65	4.25	53.67	55.00

```
quantile(log10(BF.i),probs=c(0.025,(1:9)/10,0.975))
       2.5%         10%         20%         30%         40%         50%
-0.16865120 -0.03923407  0.24132624  0.37288000  0.44646688  0.49347495
        60%         70%         80%         90%       97.5%
 0.52530830  0.55393899  0.60675568  1.08927641  1.38556819
```

Estos resultados muestran que no hay evidencia en contra de H_0. Ahora, si construimos el modelo clásico, obtenemos:

```
res2<-lm(Peso~Largo+Ancho)
summary(res2)

Call: lm(formula = Peso ~ Largo + Ancho)

Residuals:
   Min     1Q Median     3Q    Max
-3.962 -1.097 -0.163  1.099  3.413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -151.147      4.052  -37.30  < 2e-16 ***
Largo         12.652      1.008   12.55  1.04e-14 ***
Ancho         31.793      1.765   18.01  < 2e-16 ***

Residual standard error: 1.963 on 36 degrees of freedom Multiple
R-squared:  0.9892,    Adjusted R-squared:  0.9886 F-statistic: 1644
on 2 and 36 DF,  p-value: < 2.2e-16
```

El AIC y BIC para el modelo 2 se presenta como sigue (estas medidas son discutidas en la siguiente sección):

```
AIC(res2);BIC(res2)
[1] 168.1631
[1] 174.8174
```

Si eliminamos la variable Ancho obtenemos:

```
res3<-update(res2,.~.-Ancho)
res3

Call: lm(formula = Peso ~ Largo)

Coefficients: (Intercept)        Largo
    -104.68        28.22

summary(res3)

Call: lm(formula = Peso ~ Largo)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.3817  -3.3001  -0.2706   3.3984  13.8029

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -104.681      9.751  -10.73 6.41e-13 ***
Largo         28.224      1.618   17.44  < 2e-16 ***

Residual standard error: 6.127 on 37 degrees of freedom Multiple
R-squared:  0.8915,    Adjusted R-squared:  0.8886 F-statistic:
304.2 on 1 and 37 DF,  p-value: < 2.2e-16
```

El AIC y BIC para el modelo 3 se presenta como sigue:

```
AIC(res3);BIC(res3)
[1] 256.0118
[1] 261.0025
```

Notemos que tanto el AIC como el BIC son inferiores en el modelo 2, esto va de la mano de lo concluido con el factor de Bayes en los modelos bayesianos, donde se concluye que no hay evidencia en contra de H_0.

Ejemplo 8.13. Campeonato colombiano. Suponga que deseamos verificar si la hipótesis que el número promedio de goles del equipo local en el campeonato colombiano es 1.0 o menos es más plausible que si el promedio es mayor que 1.0. Asumamos que el número de goles metidos por el local en el primer tiempo se distribuye Poisson(λ). Las hipótesis serán:

- $H_1 : \lambda \leq 1$
- $H_2 : \lambda > 1$

Por suficiencia $y = \sum_{i=1}^{n} x_i \sim Poisson(n\lambda)$. Suponga que a priori $\xi(H_1) = 0.4$ y $\xi(H_2) = 0.6$.

Bajo H_1 la a priori sobre Θ_1 la escogemos $Beta(\alpha_0, \beta_0)$ y bajo H_2 asumimos una normal truncada con parámetros μ_0 y σ_0^2. El factor de Bayes es:

$$\frac{p(\boldsymbol{y}|H_1)}{p(\boldsymbol{y}|H_2)} = \frac{\int p(\boldsymbol{y}|H_1, \lambda)\xi(\lambda|H_1)\, d\lambda}{\int p(\boldsymbol{y}|H_2, \lambda)\xi(\lambda|H_2)\, d\lambda}$$

Ahora,

$$p(\boldsymbol{y}|H_i) = \int_{\Theta_i} \frac{\lambda^y \exp(-n\lambda)}{y!} \xi(\lambda|H_i)\, d\lambda = E_{\xi_i}[P(Y = y|\lambda)]$$

Para H_1

$$p(\boldsymbol{y}|H_1) = \int_0^1 \frac{\lambda^y \exp(-n\lambda)}{y!} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \lambda^{\alpha_0-1}(1-\lambda)^{\beta_0-1}\, d\lambda$$

Un algoritmo que nos permite estimar este valor sería:

1. Generar $\lambda_1, \lambda_2, \ldots, \lambda_M$ de una $Beta(\alpha_0, \beta_0)$.
2. Calcular $p_i = P(y|n\lambda_i)$, $i = 1, 2, \ldots, M$
3. Calcular

$$\frac{1}{M}\sum_{i=1}^{M} p_i$$

Para H_2

$$p(\boldsymbol{y}|H_2) = \int_1^\infty \frac{\lambda^y \exp(-n\lambda)}{y!} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right) d\lambda$$

Un algoritmo que nos permite estimar este valor sería:

1. Calcule p^* como $P(X > 1)$ donde $X \sim (\mu_0, \sigma_0^2)$
2. Generar $p_1^*, p_2^*, \ldots, p_M^*$ de una $Uniforme(p^*, 1)$.
3. Calcular λ_i tal que,

$$\int_{-\infty}^{\lambda_i} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right) = p_i^*$$

4. Calcular $p_i = P(y|n\lambda_i)$, $i = 1, 2, \ldots, M$
5. Calcular:

$$\frac{1}{M}\sum_{i=1}^{M} p_i$$

```
# Ejemplo de Factor de Bayes # Modelo muestral Poisson(lamb)
# H1:lam<=1 vs H2: lam>1

# a priori bajo H1--> beta(a0,b0)
# a priori bajo H2--> normal truncada(u0,s20)

# Datos observados: Campeonato 2002 I primeras 4 fechas
# Goles marcados por el local el primer tiempo
```

```
x<-c(0,1,0,2,1,0,2,1,1,1,0,1,0,1,0,1,1,0,0,0,3,0,0,0,0,
     1,0,0,2,0,1,0,1,0,1,0)

a0<-1; b0<-1
f.int<-function(la) la^22*exp(-36*la)/48
integrate(f.int,0,1)
# 3.722885e-17 with absolute error < 1.6e-19
f.int2<-function(la)
la^22*exp(-36*la)/48*(dnorm(la,1.5,sd=1)/(1-pnorm(1,1.5,sd=1)))
integrate(f.int2,1,Inf)
# 1.661658e-19 with absolute error < 3.6e-22

3.722885e-17/1.661658e-19
[1] 224.0464

# Cálculo del BF vía simulación
# Valor de numerador
lambdas<-matrix(rbeta(1000000,a0,b0),ncol=1)
prob.pois<-function(lambda,x)
exp(sum(dpois(x,lambda,log=T)))
numerador<-mean(apply(lambdas,1,prob.pois,x))
# Cálculo del denominador
p.1<-pnorm(1,mean=u0,sd=sqrt(s20))
p.s<-runif(1000000,p.1,1)
lambdas<-matrix(qnorm(p.s,mean=u0,sd=sqrt(s20)),ncol=1)
denominador<-mean(apply(lambdas,1,prob.pois,x))
BF<-numerador/denominador
BF
[1] 222.3040
```

El factor de Bayes nos indica que no hay evidencia en contra de H_1. Es decir, No hay evidencia en contra de que el número medio de goles marcados por el local, en el primer tiempo del torneo colombiano 2012-I, sea inferior o igual a un gol.

Ejemplo 8.14. Comparación de dos proporciones. Un problema común en estadística es el de verificar que dos proporciones son iguales ($H_0 : \pi_1 = \pi_2$) contra la alternativa $H_1 : \pi_1 \neq \pi_2$. Bajo el supuesto de H_0 solo tenemos un parámetro que puede tomar un valor en $(0, 1)$ y por lo tanto necesitamos especificar una distribución a priori en esta situación, digamos $\xi_{H_0}(\pi)$ (podemos pensar en una $Beta(\alpha, \beta)$), donde α y β se escogen de tal forma que reflejen el conocimiento a priori (en caso de ignorancia podemos escoger $\alpha = 1$ y $\beta = 1$). Bajo la alternativa H_1 debemos pensar en una distribución conjunta para (π_1, π_2), digamos $\xi_{H_1}(\pi_1, \pi_2)$. Bajo la alternativa una selección obvia es una uniforme en el área $(0, 1) \times (0, 1)$, con $\pi_1 \neq \pi_2$ y esto corresponde al producto de dos uniformes independientes. Además asumamos que la probabilidad a priori de H_0 es 0.5.

Asumamos que nuestros datos son:

	Éxitos	Fracasos	Total
Muestra 1	2	13	15
Muestra 2	14	1	15

El factor de Bayes es 0.0000894 y la probabilidad posterior de la hipótesis nula es 0.0000894.

8.3.1. Comparación de modelos

Si pensamos en términos de modelos, digamos $M_1, \ldots, M_s$, donde asumimos que M_i está parametrizado por $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i$, de dimensión d_i, y con función de densidad de probabilidad de los datos $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ y distribución a priori $\xi(\boldsymbol{\theta}_i)$. Si se tienen las probabilidades a priori para los modelos $p_1, \ldots, p_s$, por el teorema de Bayes tenemos,

$$Pr\left(M_i|\boldsymbol{y}\right) = \frac{p_i m_i(\boldsymbol{y})}{\sum_{j=1}^{s}, p_j m_j(\boldsymbol{y})}$$

donde,

$$m_i(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}_i} f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\xi(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \text{ para } i = 1, \ldots, s,$$

es la distribución marginal de los datos bajo el modelo M_i [128]. La razón de las probabilidades posteriores nos permiten hacer una comparación entre modelos. Para los modelos M_j y M_k se tiene:

$$\frac{Pr\left(M_j|\boldsymbol{y}\right)}{Pr\left(M_k|\boldsymbol{y}\right)} = \frac{p_j}{p_k} B_{jk}(\boldsymbol{y}),$$

donde,

$$B_{jk}(\boldsymbol{y}) = \frac{m_j(\boldsymbol{y})}{m_k(\boldsymbol{y})}$$

es el factor de Bayes para el modelo M_j contra el modelo M_k a partir de los datos $\boldsymbol{y}$.

Densidad predictiva a priori

En la comparación de modelos se puede utilizar la densidad predictiva a priori (PPD), la cual se define como [132].

$$m\left(\mathbf{y}\right) = \int L\left(\theta\,|\mathbf{y}\,\right) \xi\left(\theta\right)\, d\theta$$

También se conoce como log-verosimilitud marginalizada. Este no es un nombre muy adecuado ya que las verosimilitudes son funciones de los parámetros y no de los datos.

Si tenemos I modelos candidatos en una situación particular, tendríamos entonces para el i-ésimo modelo m_i.

El cálculo de la PPD es difícil y algunos autores han sugerido aproximaciones que pueden ser implementadas más fácilmente como el de Lewis y Raftery en 1997. Ellos aproximan el logaritmo de la PPD como:

$$\log\left(\tilde{m}_i\left(\mathbf{y}\right)\right) = \log\left(L_i\left(\tilde{\theta}_i\left|\mathbf{y}\right.\right)\right) + \log\left(\xi\left(\tilde{\theta}_i\right)\right) + \frac{d_i}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\left|\tilde{\mathbf{H}}_i\right|\right),$$

donde,

- d_i es la dimensión de $\boldsymbol{\theta}_i$,
- $\tilde{\theta}_i$ es un valor que maximice la densidad posterior,
- $\tilde{\mathbf{H}}_i$ es la matriz hessiana de la log-posteriori evaluada en $\tilde{\theta}_i$.

Ejemplo 8.15. Geométrica vs. Poisson [129]. Supongamos tenemos una muestra aleatoria $x_1, x_2, \ldots, x_n$ de uno de los dos modelos hipotéticos:

$$\begin{array}{rcll} M_0 & : & f(x|\theta_0) = \theta_0\left(1-\theta_0\right)^x, & x = 0, 1, \ldots \\ M_1 & : & f(x|\theta_1) = e^{-\theta_1}\theta_1^x/x!, & x = 0, 1, \ldots \end{array}$$

Por simplicidad asumamos que θ_0 y θ_1 son conocidos. ¿Cómo nos decidimos entre los dos modelos utilizando la evidencia muestral?

Ya que los parámetros se asumieron conocidos no necesitamos asumir ninguna distribución a priori para ellos. Por lo tanto:

$$f\left(\boldsymbol{x}|M_0\right) = \theta_0^n\left(1-\theta_0\right)^{n\bar{x}}$$

y

$$f\left(\boldsymbol{x}|M_1\right) = \frac{e^{-n\theta_1}\theta_1^{n\bar{x}}}{\prod_{i=1}^n x_i!}$$

ahora, el factor de Bayes es la razón de las dos últimas ecuaciones. Supongamos, $\theta_0 = 1/3$ y $\theta_1 = 2$, o sea que las dos distribuciones tienen la misma media. Si $n = 2$ y $x_1 = x_2 = 0$ entonces $B_{01}(\boldsymbol{x}) = 6.1$, sin embargo, si $n = 2$ y $x_1 = x_2 = 2$ entonces $B_{01}(\boldsymbol{x}) = 0.3$.

Definición 8.8 (Modelos Encajados)**.** *Dos modelos M_k y M_j son encajados (con M_k en M_j), si $\boldsymbol{\theta}_j = (\boldsymbol{\phi}, \boldsymbol{\eta})$ y $\boldsymbol{\theta}_k = \boldsymbol{\phi}$ y $f_k(\boldsymbol{y}|\boldsymbol{\phi}) = f_j(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{\eta}_0)$, donde $\boldsymbol{\eta}_0$ es un valor específico de $\boldsymbol{\eta}$, y $\boldsymbol{\phi}$ es un parámetro común.*

Asumamos que tenemos datos $\boldsymbol{x}$ que surge de uno de los siguientes modelos (hipótesis):

$$\begin{array}{rcl} M_1: & \boldsymbol{X} \text{ tiene densidad } & f_1\left(\boldsymbol{x}\left|\boldsymbol{\theta}_1\right.\right) \\ M_2: & \boldsymbol{X} \text{ tiene densidad } & f_2\left(\boldsymbol{x}\left|\boldsymbol{\theta}_2\right.\right) \\ \vdots & \vdots & \\ M_q: & \boldsymbol{X} \text{ tiene densidad } & f_q\left(\boldsymbol{x}\left|\boldsymbol{\theta}_q\right.\right) \end{array}$$

Le asignamos probabilidades a priori a cada modelo $\xi(M_i)$. Bajo el modelo M_i:

- Densidad a priori de $\boldsymbol{\theta}_i$: $\xi_i(\boldsymbol{\theta}_i)$
- Densidad marginal de $\boldsymbol{X}$:

$$m_i(\boldsymbol{x}) = \int f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\, \xi_i(\boldsymbol{\theta}_i)\; d\boldsymbol{\theta}_i,$$

  que mide qué tan verosímil es $\boldsymbol{x}$ bajo M_i.

- Densidad posterior:

$$\xi_i(\boldsymbol{\theta}_i \,|\boldsymbol{x}) = \frac{f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\, \xi_i(\boldsymbol{\theta}_i)}{m_i(\boldsymbol{x})}$$

- El factor de Bayes de M_j con respecto a M_i:

$$B_{ji} = \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})}$$

- La probabilidad posterior de M_i:

$$\xi(M_i \,|\boldsymbol{x}) = \frac{\xi(M_i)\, m_i(\boldsymbol{x})}{\sum_{j=1}^{q} \xi(M_j)\, m_j(\boldsymbol{x})} = \left[\sum_{j=1}^{q} \frac{\xi(M_j)}{\xi(M_i)} B_{ji}\right]^{-1}$$

  En el caso particular $\xi(M_j) = 1/q$, entonces,

$$\xi(M_i \,|\boldsymbol{x}) = \bar{m}_i(\boldsymbol{x}) = \frac{m_i(\boldsymbol{x})}{\sum_{j=1}^{q} m_j(\boldsymbol{x})} = \frac{1}{\sum_{j=1}^{q} B_{ji}}$$

Ejemplo 8.16. Localización-Escala. Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria con densidad,

$$f(x_i \,|\mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma}\right)$$

Podemos considerar varios modelos:

1. M_N: g es $N(0,1)$
2. M_U: g es $Uniforme(0,1)$
3. M_L: g es Exponencial a la izquierda $\left(\frac{1}{\sigma} e^{x-\mu}\right)$, para $x \leq \mu$
4. M_R: g es Exponencial a la derecha $\left(\frac{1}{\sigma} e^{-(x-\mu)}\right)$, para $x \geq \mu$

Observe que estos modelos no son encajados.

Ejemplo 8.17. Localización-Escala. Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria con densidad,

$$f\left(x_i \,|\mu, \sigma\right) = \frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma}\right)$$

Podemos considerar varios modelos:

1. M_N: g es $N(0,1)$
2. M_U: g es $Uniforme(0,1)$
3. M_L: g es Exponencial a la izquierda $\left(\frac{1}{\sigma} e^{x-\mu}\right)$, para $x \leq \mu$
4. M_R: g es Exponencial a la derecha $\left(\frac{1}{\sigma} e^{-(x-\mu)}\right)$, para $x \geq \mu$

Observe que estos modelos no son encajados.

- Normal:

$$m\left(\boldsymbol{x} \,|M_N\right) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{(2\pi)^{(n-1)/2}\sqrt{n}\left[\sum_i \left(x_i - \bar{x}\right)^2\right]^{(n-1/2)}}$$

- Uniforme:

$$m\left(\boldsymbol{x} \,|M_U\right) = \frac{1}{n(n-1)\left(x_{(n)} - x_{(1)}\right)^{(n-1)}}$$

- Exponencial izquierda:

$$m\left(\boldsymbol{x} \,|M_L\right) = \frac{(n-2)!}{n^n\left(x_{(n)} - \bar{x}\right)^{(n-1)}}$$

- Exponencial derecha:

$$m\left(\boldsymbol{x} \,|M_R\right) = \frac{(n-2)!}{n^n\left(\bar{x} - x_{(1)}\right)^{(n-1)}}$$

8.4. Cálculo del factor de bayes vía MCMC

Hemos visto el cálculo del factor de Bayes mediante el uso de técnicas de simulación. Esto es fácil de realizar cuando la distribución que genera datos es discreta. Un problema que no es tan fácil de resolver es cuando la distribución muestral es continua, ya que si aplicamos directamente la metodología usada, obtendríamos el valor esperado de la densidad, no la probabilidad requerida. Han y Carlin [133] realizan un recuento de los métodos propuestos para el cálculo del factor de Bayes en el caso más general.

8.4.1. Método de Carlin y Chib

Suponga que para el j-ésimo modelo la verosimilitud es:

$$f\left(\boldsymbol{y}\left|\boldsymbol{\theta}_j, M=j\right.\right)$$

y la a priori,

$$\xi\left(\boldsymbol{\theta}_j\left|M=j\right.\right)$$

Bajo estas condiciones tenemos que $\boldsymbol{y}$ es independiente de $\{\boldsymbol{\theta}_{j'\neq j}\}$. El muestreador opera sobre el espacio producto $\mathcal{M}\times\prod_{j\in\mathcal{M}}\boldsymbol{\Theta}_j$. Se requieren distribuciones a priori propias. Se asume independencia a priori entre los $\boldsymbol{\theta}_j$ dado M.

$$\begin{aligned} p\left(y\left|M=j\right.\right) &= \int f\left(\boldsymbol{y}\left|\boldsymbol{\theta}, M=j\right.\right)\xi\left(\boldsymbol{\theta}\left|\boldsymbol{M=j}\right.\right)\,d\boldsymbol{\theta} \\ &= \int f\left(\boldsymbol{y}\left|\boldsymbol{\theta}_j, M=j\right.\right)\xi\left(\boldsymbol{\theta_j}\left|\boldsymbol{M=j}\right.\right)\,d\boldsymbol{\theta}_j \end{aligned}$$

El muestreador de Gibbs, también conocido como dinámica Glauber o algoritmo de la bañera caliente [134], es definido sobre el espacio producto por las distribuciones condicionales completas,

$$\xi\left(\boldsymbol{\theta}_j\left|\boldsymbol{\theta}_{j'\neq j}, M, \boldsymbol{y}\right.\right)\propto\begin{cases} f\left(\boldsymbol{y}\left|\boldsymbol{\theta}_j, M=j\right.\right)\xi\left(\boldsymbol{\theta_j}\left|\boldsymbol{M=j}\right.\right) & \text{si } M=j \\ \xi\left(\boldsymbol{\theta_j}\left|\boldsymbol{M\neq j}\right.\right) & \text{si } M\neq j \end{cases}$$

y

$$\xi\left(M=j\left|\boldsymbol{\theta}, \boldsymbol{y}\right.\right)\propto f\left(\boldsymbol{y}\left|\boldsymbol{\theta}_j, M=j\right.\right)\left\{\prod_{j'\in\mathcal{M}}\xi\left(\boldsymbol{\theta}_{j'}\left|M=j\right.\right)\right\}\pi_j$$

Bajo las condiciones de regularidad corrientes este muestreador de Gibbs produce muestras de la distribución posterior conjunta correcta. La distribución posterior del modelo j puede estimarse como:

$$\hat{\xi}\left(M=j\left|\boldsymbol{y}\right.\right)=\frac{1}{G}\sum_{g=1}^{G}I\left(M^{(g)}=j\right),$$

que puede ser usada para estimar el factor de Bayes como:

$$\hat{B}_{jj'}=\frac{\hat{\xi}\left(M=j\left|\boldsymbol{y}\right.\right)/\hat{\xi}\left(M=j'\left|\boldsymbol{y}\right.\right)}{\xi\left(M=j\right)/\xi\left(M=j'\right)}$$

8.4.2. Método de Dellaportas, Foster y Ntzoufras

Este método es una versión «metropolizada»[1] del algoritmo de Carlin y Chib. El algoritmo es

1. Sea $(j, \boldsymbol{\theta}_j)$, donde $\boldsymbol{\theta}_j$ es de dimensión n_j.
2. Proponga un nuevo modelo j' con probabilidad $h\,(j, j')$.
3. Genere $\boldsymbol{\theta}_{j'}$ de una seudo-a priori $\xi\,(\boldsymbol{\theta}_{j'}\,|M \neq j\,)$ como en el método de Carlin y Chib.
4. Acepte el movimiento propuesto (de j a j') con probabilidad

$$\alpha = \text{mín}\left\{1, \frac{f\,(\boldsymbol{y}\,|\boldsymbol{\theta}_{j'}, M = j')\,\xi\,(\boldsymbol{\theta}_{j'}\,|M = j')\,\pi_{j'} h\,(j', j)}{f\,(\boldsymbol{y}\,|\boldsymbol{\theta}_j, M = j\,)\,\xi\,(\boldsymbol{\theta}_j\,|M = j\,)\,\pi_j h\,(j, j')}\right\}$$

8.5. Otras aproximaciones al factor de Bayes

Han y Carlin [133] presentan otras aproximaciones que han sido propuestas en la literatura para manejar el caso de usar distribuciones a priori no informativas, conocidos como seudo-factores de Bayes, entre ellos:

- El factor de Bayes intrínseco de Berger y Pericchi
- El factor de Bayes fraccionado de O'Hagan

8.6. La aproximación BIC

Esta sección está basada en Raftery [135]. La cantidad básica que subyace en el factor Bayes es la verosimilitud integrada para el modelo, dada por:

$$p(D|M_1) = \int p(D|\boldsymbol{\theta}_1, M_1)\xi(\boldsymbol{\theta}_1|M_1)\,d\boldsymbol{\theta}_1$$

Primero se derivará una aproximación simple para esta cantidad, y se mostrará posteriormente como lleva a aproximar los factores de Bayes al criterio BIC para cualificar modelos. Por simplicidad la ecuación anterior se escribe como,

$$p(D) = \int p(D|\boldsymbol{\theta})\xi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

Consideremos el caso donde D consiste de n observaciones i.i.d. $y_1, \ldots, y_n$, que pueden ser vectores.

[1] Este término se refiere a la similitud respecto al algoritmo de Metropolis-Hastings que se presenta en el próximo capítulo.

Considere la expansión en series de Taylor de $g(\boldsymbol{\theta}) = \log\left(p(D|\boldsymbol{\theta})\xi(\boldsymbol{\theta})\right)$ alrededor de $\bar{\boldsymbol{\theta}}$, el valor de $\boldsymbol{\theta}$ que maximiza $g(\boldsymbol{\theta})$, esto es, la moda posterior. La expansión es:

$$g(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}) + \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T g'(\hat{\boldsymbol{\theta}}) + \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T g''(\hat{\boldsymbol{\theta}})\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) + o\left(||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||^2\right),$$

donde,

$$g'(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_d} \end{bmatrix},$$

y $g''(\boldsymbol{\theta})$ es la matriz hessiana de segundas derivadas parciales.

$$g''(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_d} \end{bmatrix}$$

Ahora, ya que $g'(\hat{\boldsymbol{\theta}}) = 0$, debido a que la moda posterior es obtenida con $g'(\hat{\boldsymbol{\theta}}) = 0$, tenemos,

$$g(\boldsymbol{\theta}) \approx g(\hat{\boldsymbol{\theta}}) + \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T g''(\hat{\boldsymbol{\theta}})\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)$$

Esta aproximación es buena si $\boldsymbol{\theta}$ está cercano a $\hat{\boldsymbol{\theta}}$. Cuando n es grande la verosimilitud $p(D|\boldsymbol{\theta})$ está concentrada alrededor de su máxima y declina rápidamente cuando se aleja de $\hat{\boldsymbol{\theta}}$, así que los únicos valores de $\boldsymbol{\theta}$ que contribuyen a $p(D) = \int p(D|\boldsymbol{\theta})\xi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ son los que están cercanos a $\hat{\boldsymbol{\theta}}$. Se sigue por lo tanto que:

$$p(D) = \int \exp\left(g(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \approx \exp\left(g(\hat{\boldsymbol{\theta}})\right) \int \exp\left(\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T g''(\hat{\boldsymbol{\theta}})\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right) d\boldsymbol{\theta}$$

La integral en la ecuación anterior es proporcional a una densidad normal multivariable, por lo tanto,

$$p(D) \approx \exp\left(g(\hat{\boldsymbol{\theta}})\right)(2\pi)^{d/2}\,|A|^{-1/2},$$

donde $A = -g''\left(\hat{\boldsymbol{\theta}}\right)$. El error en la ecuación anterior es $O\left(n^{-1}\right)$, así:[2]

$$\log\left(p(D)\right) = \log\left(p(D|\hat{\boldsymbol{\theta}})\right) + \log\left(\xi(\hat{\boldsymbol{\theta}})\right) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(|A|\right) + O\left(n^{-1}\right)$$

Ahora, si la muestra es grande, $\hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$, donde $\hat{\boldsymbol{\theta}}$ es el estimador de máxima verosimilitud, y $A \approx nI$, donde I es la matriz de Información de Fisher esperada para una observación. Así $|A| \approx n^d|I|$. Estas dos aproximaciones introducen un error $O\left(n^{-1/2}\right)$ en la ecuación anterior, la cual se convierte en:

[2] La expresión $O\left(n^{-1}\right)$ se refiere a una cota superior asintótica de orden n^{-1}.

$$\log\left(p(D)\right) = \log\left(p(D|\hat{\boldsymbol{\theta}})\right) + \log\left(\xi(\hat{\boldsymbol{\theta}})\right) + \frac{d}{2}\log(2\pi) - \frac{d}{2}\log(n) - \frac{1}{2}\log\left(|I|\right) + O\left(n^{-1/2}\right) \quad (*)$$

Removiendo los términos de orden $O(1)$ o menores queda

$$\log\left(p(D)\right) = \log\left(p(D|\hat{\boldsymbol{\theta}})\right) - \frac{d}{2}\log(n) + O(1)$$

La ecuación anterior nos dice que la verosimilitud log-integrada, $\log(p(D))$, es igual a la log-verosimilitud maximizada, $\log\left(p(D|\hat{\boldsymbol{\theta}})\right)$, menos un factor de corrección.

La ecuación anterior es la aproximación en la cual está basada el BIC, y su error $O(1)$ significa que, en general, el error no se desaparece aún con una cantidad infinita de datos. Esto no es tan malo como parece, ya que los otros términos de la derecha de la ecuación tienden a infinito cuando n lo hace, por lo tanto ellos eventualmente dominarán. Así el error en la ecuación tenderá hacia cero como una proporción del $\log\left(p(D)\right)$, asegurando que el error no afectará la conclusión a la cual se llegue, dado que se tengan suficientes datos.

Suponga que la a priori $\xi\left(\boldsymbol{\theta}\right)$ es normal multivariable con media $\hat{\boldsymbol{\pi}}$ y matriz de covarianzas I^{-1}. Así, hablando aproximadamente, la distribución a priori contiene la misma cantidad de información que una solo observación. Esto parece razonable en una situación en la cual haya poca información a priori. Entonces,

$$\log\left(\xi(\hat{\boldsymbol{\theta}})\right) = -\frac{d}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(|I|\right)$$

y sustituyendo en (*) se llega a:

$$\log\left(p(D)\right) = \log\left(p(D|\hat{\boldsymbol{\theta}})\right) - \frac{d}{2}\log(n) + O(n^{-1/2})$$

Así para la distribución a priori particular seleccionada, el error en la aproximación es $O(n^{-1/2})$ en vez de $O(n^{-1/2})$.

Esta aproximación puede usarse para aproximar el factor de Bayes:

$$B_{12} = \frac{p\left(D|M_2\right)}{p\left(D|M_1\right)}.$$

Esto queda mejor en la escala logarítmica:

$$2\log\left(B_{12}\right) = 2\left(\log\left(p\left(D|\hat{\boldsymbol{\pi}}_2, M_2\right)\right) - \log\left(p\left(D|\hat{\boldsymbol{\pi}}_1, M_1\right)\right)\right) - \left(d_2 - d_1\right)\log(n) + O(n^{-1/2})$$

Si M_1 está encajado en M_2, la ecuación anterior puede re-escribirse

$$2\log\left(B_{12}\right) \approx \chi^2_{21} - \left(d_2 - d_1\right)\log(n)$$

donde χ^2_{21} es el estadístico de la prueba de la razón de verosimilitud corriente para probar M_1 contra M_2, y $d_2 - d_1$ son los grados de libertad asociados con la prueba.

8.7. Verosimilitud cruzada para selección

Rust y Scmittlein [136] proponen un método que ellos llaman de verosimilitud de validación cruzada[3] bayesiana (BCVL) para comparar modelos cuantitativos. Esta metodología puede usarse tanto para modelos encajados o no encajados. Ellos revisan algunas de las propuestas hechas por otros autores tales como la del supermodelo de Atkinson, en la cual, si se tienen dos modelos posibles, digamos $f(x)$ y $g(x)$, se define la siguiente función:

$$h(x) = f(x)^{\lambda} g(x)^{1-\lambda}$$

Note que, cuando $\lambda = 1$, $h(x)$ se convierte en $f(x)$ y cuando es cero, $h(x)$ es $g(x)$. Las inferencias con respecto a λ nos permiten determinar la aceptabilidad de un modelo o el otro. Atkinson también presentó una modificación:

$$h(x) = \lambda f(x) + (1 - \lambda) g(x)$$

Desde el punto de vista bayesiano la primera expresión es más natural.

También es posible realizar una análisis de validación cruzada partiendo la muestra en dos, una para estimar y la otra para validar. Esta técnica es computacionalmente muy costosa. Esta propuesta fue hecha por Mosteller y Tukey [136]. Otro criterio es el de Akaike, el cual es bastante simple:

$$AIC = \log(\text{Máxima Verosimilitud}) - (\text{Número de Parámetros Estimados})$$

Otro procedimiento usado es el criterio de Schwarz:

$$B = \log(\text{Máxima Verosimilitud}) - \frac{1}{2}\log(n)(\text{Número de Parámetros Estimados})$$

Los métodos bayesianos para la comparación de modelos son usados para proporcionar probabilidades posteriores de los modelos que compiten. Estas probabilidades pueden ser usadas para determinar la calidad del modelo. La idea del proceso de validación cruzada es mantener una parte de la muestra y seleccionar el modelo con base en su capacidad predictiva.

El método BCVL para comparación de modelos se puede resumir como:

1. Especifique las formas matemáticas de los modelos que compiten.
2. Seleccione las aprioris para los modelos.

[3] Las tres aproximaciones a la validación cruzada son [137]:

- Partición de los datos, en el cual se separa la estimación del modelo y la validación del mismo
- Reutilización de la muestra tales como el jackknife, que se concentra principalmente en estimación
- El método «simultáneo»que estima y realiza validación cruzada al tiempo de la estimación de los parámetros del modelo sujeto a la restricción de validación cruzada

3. Divida al azar la muestra u obtenga dos muestras diferentes de observaciones empíricas para probar la validez de los modelos.

4. Estime los parámetros de los modelos.

5. Utilizando los parámetros estimados en el paso anterior, obtenga las verosimilitudes para los modelos usando los datos de la segunda muestra.

6. Calcule las probabilidades posteriores para cada uno de los modelos.

Expresado más formalmente: Sean M_j, $j = 1, \ldots, K$ los modelos candidatos, que pueden ser encajados o no, con probabilidades a priori $\xi(M_j)$. Sea $\boldsymbol{\theta}_j$ el vector de parámetros del modelo M_j, definido en el espacio parametral Θ_j. Sean D_1 y D_2 que representan los datos de las dos muestras.

Primero, obtenga las estimaciones de los parámetros del modelo j, digamos θ_j^*, se calculan a partir de la primera muestra:

$$\theta_j^* = \underset{\theta_j \in \Theta_j}{\text{máx}} \, L\left(D_1; M_j\left(\theta_j\right)\right),$$

donde L representa la verosimilitud y $M_j\left(\theta_j\right)$ representa el modelo. Las probabilidades:

$$\begin{aligned} \xi\left(M_j \left| D_2 \right.\right) &= \frac{L\left(D_2 \left| M_j \right.\right) \xi\left(M_j\right)}{L\left(D_2\right)} \\ &\approx \frac{L\left(D_2 \left| M_j\left(\theta_j^*\right)\right.\right) \xi\left(M_j\right)}{\sum_{j=1}^{K} L\left(D_2 \left| M_j\left(\theta_j^*\right)\right.\right) \xi\left(M_j\right)} \end{aligned}$$

En términos de los puntos de la segunda muestra,

$$\begin{aligned} \xi\left(M_j \left| D_2 \right.\right) &= \frac{\prod_{i=1}^{n_2} L\left(x_{i2} \left| M_j\left(\theta_j^*\right)\right.\right) \xi\left(M_j\right)}{\sum_{k=1}^{K} \left[\prod_{i=1}^{n_2} L\left(x_{i2} \left| M_k\left(\theta_k^*\right)\right.\right)\right] \xi\left(M_k\right)} \\ &= \frac{\exp\left(\sum_{i=1}^{n_2} \log\left(L\left(x_{i2} \left| M_j\left(\theta_j^*\right)\right.\right)\right) + \log\left(\xi\left(M_j\right)\right)\right)}{\sum_{k=1}^{K} \exp\left(\sum_{i=1}^{n_2} \log\left(L\left(x_{i2} \left| M_k\left(\theta_k^*\right)\right.\right)\right) + \log\left(\xi\left(M_k\right)\right)\right)} \end{aligned}$$

$$\begin{aligned} = \Bigg(\sum_{k=1}^{K} \exp \Bigg[& \sum_{i=1}^{n_2} \log\left(L\left(x_{i2} \left| M_k\left(\theta_k^*\right)\right.\right)\right) + \log\left(\xi\left(M_k\right)\right) \\ & - \sum_{i=1}^{n_2} \log\left(L\left(x_{i2} \left| M_j\left(\theta_j^*\right)\right.\right)\right) - \log\left(\xi\left(M_j\right)\right) \Bigg] \Bigg)^{-1} \end{aligned}$$

Los pasos anteriores deben repetirse de tal forma que se agoten las particiones del conjunto de datos original en los dos subconjuntos. Si se usa una observación para el segundo subconjunto, entonces los paso anteriores se realizan n veces.

8.7.1. Análisis exploratorio de datos

Gelman [97] discute el uso de conceptos que se desarrollaron en el análisis de datos, conocido como EDA, como ayuda en el análisis de modelos bayesianos. En esto el EDA puede aportar la gran riqueza en técnicas de visualización y construcción de medidas de resumen. Por ejemplo, se pueden considerar medidas como los valores-p que en su forma clásica es:

$$Valor - p\left(y\,|\theta\right) = Pr\left(T\left(y^{rep}\right) > T\left(y\right)|y, \theta\right),$$

a la versión bayesiana, el cual es un promedio sobre la distribución posterior de θ,

$$Valor - p\left(y\right) = Pr\left(T\left(y^{rep}\right) > T\left(y\right)\right)$$

8.8. Estadística bayesiana empírica

La estadística Bayesiana Empírica Paramétrica es un híbrido que asume la existencia de datos previos para estimar los parámetros de $f(\theta)$. El principio básico asume que la distribución a priori es conocida solo parcialmente y usa los datos para estimar la a priori [138]. Este método fue propuesto por Robbins en 1955 [139], también conocida como estadística bayesiana empírica noparamétrica, ya que dejaba inespecificada la distribución a priori.

La estimación bayesiana asume que una sucesión de experimentos similares son llevados a cabo, cada uno teniendo su propio parámetro. Supongamos que tenemos p variables observadas, cada una de ellas de una población normal

$$X_i \sim N\left(\mu,\, \sigma^2\right) \text{ para } i = 1, \ldots, p$$

Si tenemos N poblaciones (la población aquí es de carácter general y podrían ser individuos como ha sido el caso estudiado por Efron sobre bateadores en el béisbol norteamericano) y obtenemos muestras independientes. Por ejemplo, un jugador de un deporte particular tiene sus propios parámetros, tales como tasa de éxito en ciertas ejecuciones, etc. Asumimos además que los parámetros que caracterizan estas poblaciones forman una muestra aleatoria obtenida de alguna distribución a priori. Si la a priori es conocida completamente, el análisis se convierte en un análisis bayesiano estándar. Si la a priori no es conocida, uno puede tener tres posibilidades:

1. Seleccionar una forma paramétrica para la a priori considerando los parámetros de esta distribución desconocidos pero fijos. Esto lleva al modelo paramétrico Bayes empírico. Muchas de las aplicaciones que aparecen en la literatura asumen un modelo normal.

2. Podríamos no tener una forma específica para la a priori. Esto lleva al modelo Bayes empírico no paramétrico de Robbins.

3. Se puede seleccionar una forma a priori paramétrica y sus parámetros a su vez ser considerados cantidades aleatorias con distribuciones hiperaprioris completamente determinadas.

Suponga que se realizan m experimentos donde el resultado del i-ésimo es Y_i, los datos serán $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$, donde $Y_i \sim F(y_i \,|\theta_i)$. La a priori para θ_i la denotamos por $\xi(\theta_i \,|\tau)$. La distribución marginal de $\boldsymbol{Y} \,|\tau$ es:

$$\prod_i \int F(Y_i \,|\theta_i)\, \xi(\theta_i \,|\tau)\, d\theta_i$$

El método bayesiano empírico usa $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ para estimar $\hat{\xi} = \xi(\theta_i \,|\tau)$ como $\xi(\theta_i \,|\hat{\tau})$. El análisis que se realiza es uno bayesiano estándar usando como distribución a priori la distribución a priori estimada.

Basu y Rigdon [140] ilustran el procedimiento Bayes empírico paramétrico en problemas de confiabilidad. Dado que el tipo de variable de interés que surge en esta área es de tipo positivo (duración de un artefacto), ellos utilizan como a priori la familia gamma con parámetros desconocidos. Si consideramos N artefactos que funcionan con vida exponencial, pero cada uno tiene su propia tasa λ_i, $i = 1, \ldots, N$ y además cada aparato tiene n registros de tiempos de funcionamiento hasta fallar (suponga que el plan de remplazo permite n reparaciones). El i-ésimo aparato genera registros $t_{i,1}, t_{i,2}, \ldots, t_{i,n}$. El tiempo de funcionamiento del aparato i, $t_i = t_{i,1} + t_{i,2} + \ldots + t_{i,n}$, será una $Gamma(n, \lambda_i)$ y la f.d.p. conjunta es:

$$p(\boldsymbol{t} \,|\boldsymbol{\lambda}) = \prod_{i=1}^{N} \frac{\lambda_i^n}{\Gamma(n)} t_i^{\alpha-1} \exp(-\lambda_i t_i)$$

$$\xi(\lambda_1, \ldots, \lambda_N \,|\alpha, \theta) = \xi(\boldsymbol{\lambda} \,|\alpha, \theta) \propto \prod_{i=1}^{N} \lambda_i \exp(-\theta \lambda_i)$$

Ahora,

$$\begin{aligned} p(\boldsymbol{t} \,|\alpha, \theta) &= \int p(\boldsymbol{t}, \boldsymbol{\lambda} \,|\alpha, \theta)\; d\boldsymbol{\lambda} \\ &= \int p(\boldsymbol{t} \,|\boldsymbol{\lambda})\, \xi(\boldsymbol{\lambda} \,|\alpha, \theta)\; d\boldsymbol{\lambda} \\ &= \left(\frac{\theta^\alpha}{\Gamma(\alpha)}\right)^N \prod_{i=1}^{N} \int_0^\infty \lambda_i^{n+\alpha-1} \exp(-(\theta + t_i)\,\lambda_i)\; d\lambda_i \end{aligned}$$

Una vez se tienen estimaciones de α y θ, las estimaciones puntuales de las λ_i's pueden estar basadas en la distribución posterior de los λ_i dados los valores estimados por máxima verosimilitud de α y θ.

La distribución posterior de λ_i es una $Gamma\left(n + \hat{\alpha}, \hat{\theta} + t_i\right)$ y el estimador puntual para λ_i será:

$$\hat{\lambda}_i = \frac{n + \hat{\alpha}}{\hat{\theta} + t_i}$$

Intervalos de confianza pueden ser construidos usando los percentiles de la distribución posterior.

Parte II

Estadística Bayesiana Computacional

Capítulo 9

Estadística bayesiana vía simulación

Muchos de los problemas a los que nos enfrentamos a través del enfoque bayesiano requieren de solución a través de métodos numéricos. Esto debido a que los modelos a posteriori resultantes en los análisis, usualmente son de alta complejidad y no tienen solución cerrada. Es por esto, que toma mucha importancia la implementación de métodos numéricos para llegar a la solución de nuestros problemas.

En este capítulo, se presentan diferentes métodos para abordar este tipo de problemas.

Ejemplo 9.1. La necesidad de utilizar métodos numéricos en el análisis bayesiano queda ilustrado con el siguiente ejemplo [141]. Considere la siguiente tabla que presenta información sobre la sobrevivencia en pacientes que sufrieron un ataque al miocardio y que fueron tratados con un medicamento (Blocadren) o un placebo.

Tabla 9.1: *número de pacientes en riesgo en un momento determinado y número de muertes en 8 diferentes momentos*

Tratamiento			
Blocadren		Placebo	
En riesgo	Muertos	En riesgo	Muertos
945	19	939	31
926	10	908	9
916	7	899	4
909	3	895	10
906	5	885	7
901	6	878	10
895	5	869	10
890	2	858	3

La probabilidad de no sobrevivir una semana se estima como el cociente entre el número de pacientes que murieron y el número de pacientes que entraron con infarto de miocardio. Usemos la siguiente notación:

$$
\begin{aligned}
y_{ij} &= \#\text{ de muertes en el } j\text{-ésimo tratamiento en el momento } i \\
n_{ij} &= \#\text{ de pacientes en riesgo en el } j\text{-ésimo tratamiento en el momento } i
\end{aligned}
$$

Entonces podemos pensar en el siguiente modelo:

$$
\begin{aligned}
Y_{ij}|n_{ij}, \pi_{ij} &\sim Binomial\left(\pi_{ij}, n_{ij}\right) \\
\pi_{ij}|\alpha_j, \beta_j &\sim Beta\left(\alpha_j, \beta_j\right)
\end{aligned}
$$

$$
\begin{aligned}
\alpha_j &\sim \sigma\text{Gamma}\left(d_\alpha\right) \\
\beta_j &\sim \sigma\text{Gamma}\left(d_\beta\right)
\end{aligned}
$$

Una parametrización alternativa para α y β es:

$$
\begin{aligned}
\lambda_{1j} &= \frac{\alpha_j}{\alpha_j + \beta_j} \qquad \text{(La media)} \\
\lambda_{2j} &= \alpha_j + \beta_j \qquad \text{(La precisión)} \\
\lambda_{1j} &\sim Beta\left(d_\alpha, d_\beta\right) \\
\lambda_{2j} &\sim \sigma\text{Gamma}\left(d_\alpha + d_\beta\right)
\end{aligned}
$$

Tenemos un modelo, tenemos las distribuciones a priori y para realizar el análisis bayesiano solo necesitamos hallar la distribución posterior conjunta que se halla como:

$$
\xi\left(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{n}, d_\alpha, d_\beta, \sigma\right) =
$$

$$
\frac{\overbrace{\int \cdots \int}^{15} \prod_{j=1}^{2} \prod_{i=1}^{I_j} f\left(y_{ij}|n_{ij}, \pi_{ij}, \alpha_j, \beta_j\right) \xi\left(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}|d_\alpha, d_\beta, \sigma\right) d\boldsymbol{y}}{\underbrace{\int \cdots \int}_{34} \prod_{j=1}^{2} \prod_{i=1}^{I_j} f\left(y_{ij}|n_{ij}, \pi_{ij}, \alpha_j, \beta_j\right) \xi\left(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}|d_\alpha, d_\beta, \sigma\right) \; d\boldsymbol{y}\, d\boldsymbol{\pi}\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}}
$$

Obviamente este es un trabajo que no se puede realizar a mano, lo cual obliga a implementar procedimientos numéricos para su solución.

9.1. MCMC: Monte Carlo por cadenas de Markov

El análisis bayesiano requiere realizar integraciones sobre distribuciones de probabilidad posiblemente de alta dimensión para realizar inferencias acerca de los parámetros de un modelo o realizar predicciones. En el pasado los analistas bayesianos resolvían este problema mediante métodos de integración numérica. Este problema es grave cuando se trata de resolver integraciones en alta dimensión. Algunos métodos

de integración numérica aproximados, como la cuadratura gaussiana o de Laplace, han sido utilizados. La integración Monte Carlo extrae muestras de la distribución de probabilidad de interés y trabaja sobre promedios que aproximen las esperanzas de interés.

Cuando las distribuciones a posteriori son de alta dimensión, las soluciones analíticas o las numéricas comunes no se pueden obtener. Una solución es considerar un procedimiento Monte Carlo iterativo o Monte Carlo por Cadenas de Markov. La metodología MCMC es una herramienta de gran alcance para la modelación estadística y se ha vuelto muy popular en la computación bayesiana en modelos estadísticos de gran complejidad. Se simula una cadena de Markov con distribución estacionaria dada por la distribución a posteriori $\xi\left(\boldsymbol{\theta}|Datos\right)$. Brooks [142] realiza una revisión de la metodología MCMC. Las características de ξ son obtenidas encontrando promedios ergódicos.

$$\hat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} h\left(\boldsymbol{\theta}_r\right)$$

Los métodos MCMC son algoritmos iterativos que se utilizan cuando el muestreo directo de una distribución de interés ξ no es factible.

La aproximación corriente en pregrado a la teoría de cadenas de Markov, que es familiar para todos nosotros, es iniciar con alguna distribución de transición (una matriz de transición en el caso discreto) que modela algún proceso de interés, para determinar las condiciones bajo la cual existe una distribución estacionaria o invariante y entonces identificar la forma de la distribución límite. Los métodos MCMC involucran la solución inversa de este problema ya que la distribución estacionaria es conocida, y es la distribución de transición la que necesita ser determinada, a pesar que el la práctica existen un número infinito de distribuciones de las cuales escoger.

Una cadena de Markov es generada muestreando

$$\boldsymbol{\theta}^{(t+1)} \sim p\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$$

Este p es llamado el *kernel de transición* de la cadena de Markov. Así $\boldsymbol{\theta}^{(t+1)}$ depende solo de $\boldsymbol{\theta}^{(t)}$, y no de $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t-1)}$.

Existen dos problemas mayores que rodean la implementación e inferencias de los métodos MCMC. El primero tiene que ver con la convergencia y el segundo con la dependencia entre las muestras de la distribución posterior.

Las condiciones bajo las cuales una cadena de Markov tiene una única distribución estacionaria son bien conocidas teóricamente:

- Tiene que se aperiódica
- Irreducible
- Positiva recurrente.

La forma en que construimos nuestras cadenas garantiza la existencia de la distribución estacionaria. Sin embargo en la práctica esta convergencia puede ser penosamente lenta y el mayor problema es saber si se ha logrado una convergencia razonable (esto se conoce como un «burn-in»). Por lo tanto las muestras obtenidas hasta el punto de «burn-in»son descartadas.

Un asunto relacionado con la convergencia es la tasa de mezclado. Informalmente, el mezclado es la tasa con la cual la cadena de Markov se mueve a través del soporte de la distribución estacionaria. Así, si una cadena tiene un mezclado lento, puede quedarse en cierta porción del espacio de estados por un período de tiempo muy largo, y a menos que la longitud de la cadena sea ajustada acordemente, las inferencias serán afectadas sin ninguna duda. Lombardi [143] señala, «Uno de los problemas más serios con los algoritmos MCMC es el paradigma 'usted solo ve donde usted ha estado', que es el hecho que la cadena parece haber convergido pero ha fallado explorar completamente el espacio muestral. En lugar de una cadena larga, varias cadenas paralelas empezando desde puntos ampliamente dispersos pueden resolver este problema».

El segundo asunto está relacionado con el hecho que los valores observados, siendo un camino muestral de una cadena de Markov, no son independientes entre sí. Asumiendo que se ha logrado la convergencia, los valores observados formarán una muestra dependiente de la distribución posterior. Esto puede ser molesto para uno pero no es necesariamente malo en MCMC. En la mayoría de los problemas, la estimación típica se obtiene por un promedio sobre las muestras. Aunque las muestras no sean independientes, el teorema ergódico asegura que estos promedios muestrales convergen a las verdaderas esperanzas. Así que la aproximación corriente al problema de dependencia es ignorarla. Pero si uno, por alguna razón, necesita una muestra independiente, puede resolver el problema corriendo varias cadenas de Markov con puntos de comienzo independientes y utilizar el último punto de cada cadena [143].

9.1.1. Glosario de cadenas de Markov

Definición 9.1 (Irreducibilidad)**.** *Una cadena de Markov $X_1, X_2, \ldots$ es irreducible si la cadena puede moverse libremente a través del espacio de estados; esto es, para dos estados cualesquiera x y x', existe un n tal que,*

$$P\left(X_n = x' | X_0 = x\right) > 0.$$

Definición 9.2 (Recurrencia)**.** *Una cadena de Markov es recurrente si el número promedio de visitas a un estado arbitrario es infinito.*

Definición 9.3 (Período)**.** *Un estado x tiene período d si $P\left(X_{n+t} = x | X_t = x\right) = 0$ cuando n no es divisible por d, donde d es el mayor entero con esta propiedad.*

Definición 9.4 (Aperiodicidad)**.** *Si un estado x tiene período $d = 1$ se dice que es aperiódico.*

En una cadena irreducible todos los estados tienen el mismo período. Si ese período es $d = 1$, la cadena de Markov es aperiódica.

Definición 9.5 (Distribución estacionaria)**.** *Sea X_n, $n \geq 1$, una cadena de Markov con espacio de estados x y matriz de transición P. Sea $\pi(i)$, $i \in x$, una distribución de probabilidad, es decir,*

$$\pi(i) = 0, \text{ para todo } i \in x, \quad \sum_{i \in x} \pi(i) = 1$$

Si

$$\sum_i \pi(i) P(i,j) = \pi(j), \quad j \in x,$$

decimos que p es una distribución estacionaria para la cadena X_n.

Definición 9.6 (Cadena de Markov recurrente positiva)**.** *Una cadena de Markov irreducible es recurrente positiva sí y sólo sí tiene una distribución estacionaria.*

Teorema 9.1 (Convergencia a una distribución estacionaria)**.** *Si una cadena de Markov $X_1, X_2, \ldots$ con espacio de estados contable es positiva, recurrente y aperiódica con distribución estacionaria π, entonces desde cualquier estado inicial:*

$$X_n \to X \sim \pi$$

Definición 9.7 (Ergodicidad)**.** *Una cadena de Markov positiva, recurrente y aperiódica es llamada ergódica.*

Teorema 9.2 (Convergencia de Sumas (Teorema Ergódico))**.** *Si una cadena de Markov con espacio de estados contable $X_1, X_2, \ldots$ es ergódica con distribución estacionaria π, entonces desde cualquier estado inicial:*

$$\frac{1}{n} \sum_{i=1}^{n} h\left(X_i\right) \to E_\pi \left[h(X)\right],$$

donde $h : x \to \mathbb{R}$, tal que $\pi|h| = \sum_{i \in x} \pi(i)|h(i)| < \infty$.

Una de las dificultades que surgen en el trabajo bayesiano aparecen cuando tratamos de manipular la distribución a posteriori que usualmente aparece de la siguiente forma:

$$\xi\left(\boldsymbol{\theta}|Datos\right) \propto L\left(\boldsymbol{\theta}|Datos\right) \xi\left(\boldsymbol{\theta}\right),$$

que no es una densidad de probabilidad en sí misma, sino que debe ajustarse por un factor que se calcula como:

$$\int_\Theta L\left(\boldsymbol{\theta}|Datos\right) \xi\left(\boldsymbol{\theta}\right) \, d\boldsymbol{\theta}$$

Solo en problemas muy sencillos es posible evaluar exactamente las expresiones anteriores, lo cual limitaría el uso de los métodos bayesianos, si no fuera por la posibilidad de utilizar métodos computacionales como es el *Método Monte Carlo*. Con esta técnica es posible:

- Generar muestras $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_R$, de una distribución de probabilidad dada, digamos $F\left(\boldsymbol{\theta}\right)$

- Estimar valores esperados de funciones bajo esta distribución, por ejemplo,

$$\Phi = E\left[h\left(\boldsymbol{\theta}\right)\right] = \int h\left(\boldsymbol{\theta}\right)\, dF\left(\boldsymbol{\theta}\right)$$

Denotamos por $f\left(\boldsymbol{\theta}\right)$ la densidad asociada con la distribución y la llamaremos *densidad objetivo*, ella puede ser la distribución a posteriori, que en nuestro caso es el interés y es una distribución condicionada en los datos. La generación de muestras es más importante ya que Φ puede ser estimada como:

$$\hat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} h\left(\boldsymbol{\theta}_r\right)$$

Es claro que si los vectores $\{\boldsymbol{\theta}_r\}_{r=1}^{R}$ corresponden a una muestra de $F\left(\boldsymbol{\theta}\right)$, entonces $E\left[\hat{\Phi}\right] = \Phi$. También, a medida que R se incrementa la varianza de $\hat{\Phi}$ disminuye ya que es σ^2/R, donde σ^2 es la varianza de $h(\boldsymbol{\theta})$.

De lo anterior se desprende una propiedad importante del método Monte Carlo: la exactitud de la estimación Monte Carlo no depende de la dimensionalidad del espacio muestreado. La varianza de $\hat{\Phi}$ es siempre σ^2/R. En teoría, si tenemos una muestra aún pequeña de observaciones independientes podemos obtener una estimación buena de Φ. El problema está en que obtener muestras independientes de F puede no ser una tarea fácil.

Asumamos que la densidad de la cual deseamos obtener muestras es la correspondiente a la distribución a posteriori de un parámetro de un experimento exponencial, digamos λ para el cual la única información a priori que disponíamos era que $\lambda \sim U(0,5)$. Se obtuvieron cinco muestras con resultados $x_1 = 1, x_2 = 1, x_3 = 4, x_4 = 2, x_5 = 3$. Por lo tanto la distribución posterior será

$$\xi(\lambda|Datos) \propto \lambda^5 e^{-11\lambda} I(0,5)$$

Si la constante de normalización fuera difícil de calcular (obviamente en este problema no lo es) entonces no sería fácil muestrear de ξ. Si el problema fuera unidimensional podemos pensar en una discretización y muestrear de esta distribución discreta como se muestra en la Figura 9.1. Cada punto discretizado en esta gráfica tiene una altura igual al valor de la densidad en ese punto, o sea $p_i^* = \lambda_i^5 e^{-11\lambda_i}$. Podemos calcular una constante de normalización Z como:

$$Z = \sum_i p_i^*,$$

y

$$p_i = \frac{p_i^*}{Z},$$

y muestreamos de la distribución de probabilidad $\{p_i\}$. Cuál es el costo de este procedimiento? Para poder calcular Z se requiere visitar cada punto en la discretización. En nuestro caso la dimensión del espacio era uno, pero si el espacio tuviera dimensión 100, el número de puntos a visitar sería 50^{100}. Un número inmenso de visitas.

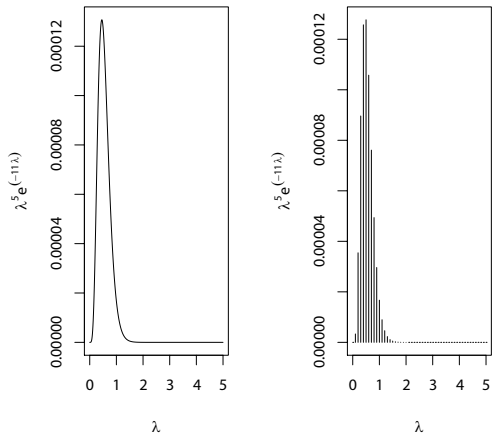

Figura 9.1: *la gráfica izquierda muestra el kernel de la densidad posterior* $\lambda^5 e^{-11\lambda} I(0,5)$. *¿Cómo obtener muestras de esta densidad? La gráfica derecha presenta una discretización del kernel evaluado en 50 puntos equiespaciados en el intervalo* $(0,5)$. *¿Cómo podemos muestrear de esta distribución?*

Ejemplo 9.2. Una proporción. Suponga que estamos interesados en determinar la proporción de estudiantes que sufren gastritis.

- Como a priori supongamos una normal truncada con parámetros $\mu = 0.5$ y $\sigma^2 = 0.2^2$.
- Se saca una muestra al azar de 10 estudiantes y se les evalúa. De éstos solo dos tienen gastritis.

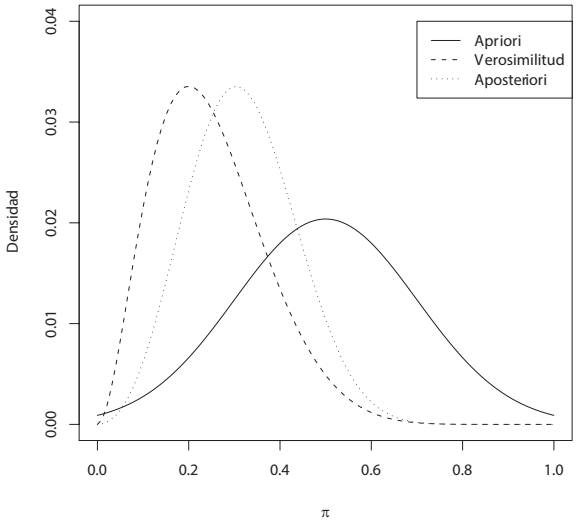

Figura 9.2: *distribución a priori, verosimilitud y a posteriori para el caso de los estudiantes con gastritis*

```
#Generación de muestra de una distribución a posteriori por medio
#del método de discretización

densidad.posteriori<-function(x,media.apriori,dt.a priori,n,
nro.exitos){
#n = tamaño de muestra
#vero = verosimilitud
vero <- x^nro.exitos*(1-x)^(n - nro.exitos)
a priori <- exp(-(x - media.a priori)^2/(2*dt.a priori^2))
a posteriori <- vero*a priori
list(vero = vero, a priori = a priori, a posteriori = a posteriori)
}

#Graficos de la verosimilitud, distribuciones a priori, a posteriori
pis <- seq(0.00001, 0.9999, length = 100)
res <- densidad.posteriori(pis, 0.5, 0.2, 10, 2)
res.a priori<-res$a priori/sum(res$a priori)
plot(pis, res.a priori, type ='l', lty=1,
ylab='',xlab='',,ylim=c(0,0.04))
title(ylab='Densidad',xlab=expression(pi))
res.vero<-res$vero/sum(res$vero)
points(pis, res.vero, type ='l', lty=2)

res.a posteriori<-res$a posteriori/sum(res$a posteriori)

points(pis, res.a posteriori, type ='l', lty=3)
legend(0.7,0.04,c('A priori','Verosimilitud','A
posteriori'),lty=1:3)

resu<-sample(pis,10000,prob=res.a posteriori,replace=T)

hist(resu,main='Distribución Simulada',xlab=expression(pi))
```

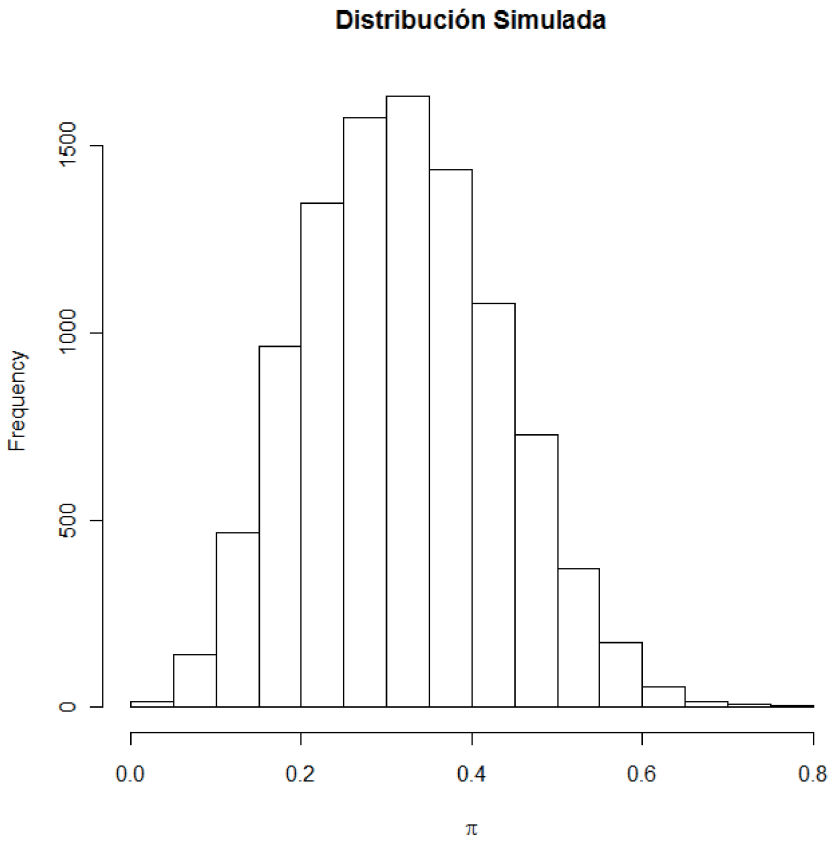

Figura 9.3: *histograma de la distribución a posteriori simulada*

```
mean(resu)
[1] 0.3171478
median(resu)
[1] 0.3131069

quantile(resu, probs=c(0.025, 0.975))
     2.5%     97.5%
0.1111089 0.5555044

require(hdrcde)
hdr(resu)
$hdr
          [,1]      [,2]
99% 0.05800344 0.6060039
95% 0.10100899 0.5371268
50% 0.22220778 0.3912115

$mode
[1] 0.317054

$falpha
       1%        5%       50%
0.1858224 0.6524188 2.6730446
```

Notamos que la proporción promedio a posteriori estimada de estudiantes con gastritis es de 0.31. Además, con una probabilidad de 0.95, dicha proporción está entre 0.101 y 0.537.

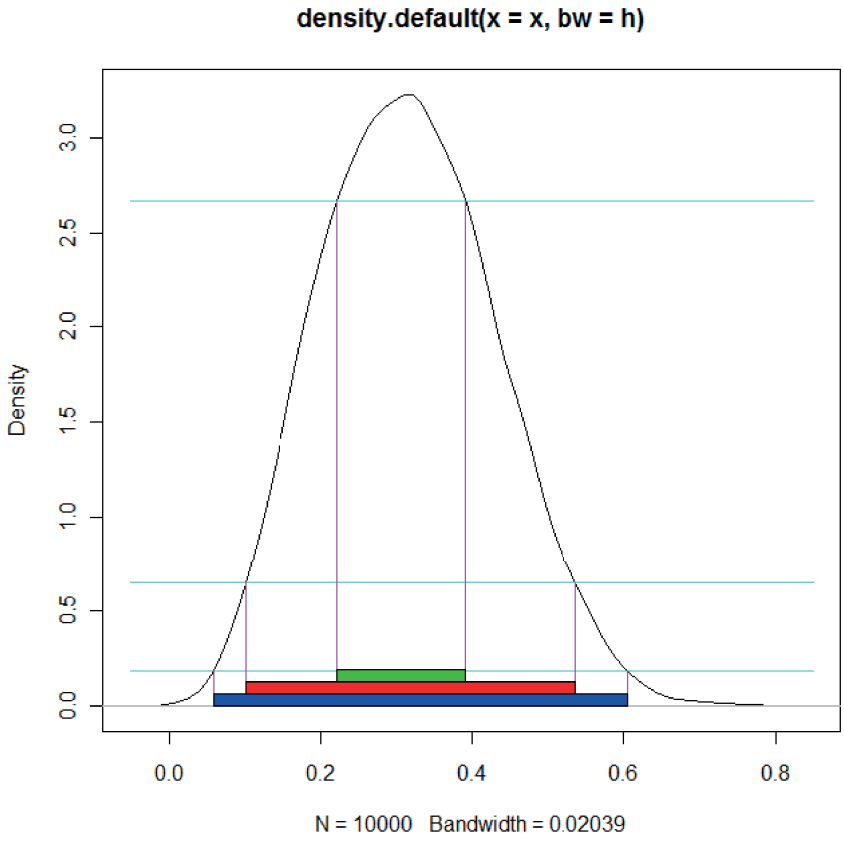

Figura 9.4: *regiones de más alta densidad en el caso de la gastritis*

9.1.2. Muestreo de importancia

Este no es un método para generar muestras. Este es un método para calcular la esperanza de $h(\boldsymbol{\theta})$, [144]. Asumamos que nuestra densidad unidimensional objetivo es $p(\theta)$, y de la cual tenemos su kernel, digamos $p^*(\theta)$ tal que,

$$p(\theta) = \frac{p^*(\theta)}{Z}$$

donde Z es una constante de normalización.

Supongamos que muestrear directamente de $p(\theta)$ es muy complicado. Ahora asumamos que existe una distribución $q(\theta)$ de la cual sabemos es fácil muestrear y que tiene el mismo soporte que p. La densidad q es llamada la *densidad muestreadora*.

En el muestreo de importancia procedemos así:

1. Generamos R muestras $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{R}$ de $q(\theta)$.

2. Calculamos los pesos
$$w_r = \frac{p^*(\theta^{(r)})}{q(\theta^{(r)})}$$

3. Utilizamos los pesos anteriores para ajustar la "importancia" de cada punto en nuestro estimador así:
$$\hat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} w_r h\left(\theta^{(r)}\right)$$

9.1.3. Muestreo por rechazo

Asumamos una densidad unidimensional $p(\theta) = p^*(\theta)/Z$ que suponemos tiene una forma muy complicada para muestrear directamente de ella [144]. Asumamos además que tenemos una distribución que es más simple y de la cual podemos muestrear llamada *densidad propuesta* $q(\theta)$ la cual podemos evaluar hasta un factor multiplicativo Z_q. Además supongamos que conocemos una constante c tal que,

$$cq^*(\theta) > p^*(\theta), \qquad \text{para todo } \theta$$

1. Generamos dos números aleatorios:

   a) El primero, digamos θ, es generado de la densidad propuesta $q(\theta)$. Evaluamos $cq^*(\theta)$.

   b) Generamos un número distribuido uniformemente en el intervalo $[0, cq^*(\theta)]$, digamos u.

2. Evaluamos $p^*(\theta)$. Si $u > p^*(\theta)$ entonces θ es aceptado. En otro caso, es rechazado.

9.1.4. Muestreador de Gibbs

El muestreador de Gibbs es una herramienta de simulación que permite obtener muestras de funciones de distribución conjuntas no normalizadas. El muestreador de Gibbs involucra el muestreo de las distribuciones condicionales completas. Es esencial que el muestreo de las distribuciones condicionales completas sea altamente eficiente desde el punto de vista computacional. El muestreo de rechazo es una técnica posible de muestreo independiente de una densidad general $p(\theta)$ donde la densidad $p(\theta)$ sea analíticamente inmanejable [145].

El muestreo de rechazo requiere una función cobija g de $p(\theta)$ donde $g(\theta) \geq p(\theta)$ para todo θ y un punto muestreado es aceptado con probabilidad $p(\theta)/g(\theta)$.

El muestreo de rechazo adaptativo (ARS), propuesto por Gilks y Wild [146], permite muestrear de densidades condicionales complejas que son log-cóncavas, o sea

$d^2 \ln p(\theta)/d\,\theta^2 < 0$. Ellos mostraron que una función cobija (envelope function) para $\ln p(\theta)$ puede construirse mediante tangentes a $\ln p$ en cada abscisa para un conjunto dado de abscisas. Una cobija se construye entre dos abscisas adyacentes a partir de las tangentes en cada final del intervalo. Secantes son dibujadas a través del $\ln p(\theta)$ en las abscisas adyacentes. La cobija es una función exponencial a tramos, de la cual el muestreo es más fácil.

Para obtener una muestra de la distribución conjunta $p(X_1, \ldots, X_d)$ el Muestreador Gibbs itera sobre este ciclo:

- Muestree $X_1^{(i+1)}$ de $p\left(X_1 \left| X_2^{(i)}, \ldots, X_d^{(i)}\right.\right)$
- Muestree $X_2^{(i+1)}$ de $p\left(X_2 \left| X_1^{(i+1)}, X_3^{(i)} \ldots, X_d^{(i)}\right.\right)$
- $\vdots$
- Muestree $X_d^{(i+1)}$ de $p\left(X_d \left| X_1^{(i+1)}, \ldots, X_{d-1}^{(i+1)}\right.\right)$

Ilustraciones de las bondades de este procedimiento son presentadas en Gelfand, Carlin, Stern y Rubin [147], quienes lo aplican en modelos normales, modelos jerárquicos de crecimiento y otros. Aplicaciones al caso de espacios parametrales restringidos y datos truncados fueron considerados por Gelfand, Carlin, Stern y Rubin [147]. George y McCulloch [148] lo utilizan para el problema de la selección de variables en un modelo. Chan [149] estudia el comportamiento asintótico del muestreador de Gibbs y muestra que bajo condiciones suaves es ergódico geométricamente.

Gelfand [145] señala varios puntos que requieren ser considerados cuando se hace uso de esta herramienta y que son importantes en el trabajo aplicado tales como selección de valores iniciales, transformaciones de parámetros (reparametrizaciones), convergencia del vector completo y convergencia de algunas componentes.

Ejemplo 9.3. Distribución Poisson bivariable. El número de goles que marcan los equipos en un partido de fútbol puede modelarse bastante bien mediante una distribución Poisson bivariable. El número de goles depende de si el equipo es local o visitante y depende también de la calidad del mismo así como de la calidad del visitante. Considere el artículo de Karlis y Ntzoufras [150] donde presenta una distribución Poisson bivariada.

$$\begin{aligned}
P(x,y) &= \exp\left(-\left(\lambda_1+\lambda_2+\lambda_3\right)\right)\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\times \\
&\qquad \sum_{k=0}^{\min(x,y)} \binom{x}{k}\binom{y}{k} k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^2 \\
E(X) &= \lambda_1+\lambda_3 \\
E(Y) &= \lambda_2+\lambda_3 \\
cov(X,Y) &= \lambda_3
\end{aligned}$$

```
# Gibbs para una Poisson Bivariada

logVero.biPoisson<-function(l,X){
l1<-exp(l[1]);l2<-exp(l[2]);l3<-exp(l[3])

sumita<-function(k,xx,yy,l1,l2,l3)
choose(xx,k)*choose(yy,k)*factorial(k)*(l3/(l1*l2))^k

log.densi.un.punto<-function(x,l1,l2,l3){
min.x<-min(x)
xx<-x[1]
yy<-x[2]
if(min.x==0)suma<-1
else{
suma<-sum(sapply(0:min.x,sumita,xx,yy,l1,l2,l3))
#suma<-0
for(k in 0:min.x)
suma<-suma+choose(xx,k)*choose(yy,k)*
 factorial(k)*(l3/(l1*l2))^k
 }#fin else
#print(suma)
log.densi<--l3+dpois(xx,l1,log=T)+dpois(yy,l2,log=T)
+log(suma)

return(log.densi)
} # fin log.densi.en.un.punto

res<-sum(apply(X,1,log.densi.un.punto,l1,l2,l3))
res
}

# Partidos del segundo torneo 2010 hasta la fecha 14
temp<-scan()
2 0 1 1 1 0 2 1 2 0 2 1 1 0 1 1
1 2 1 0 2 0 1 0 1 2 1 2 0 0 1 1 1 2
1 0 3 0 2 1 2 1 1 0 2 0 1 1 3 2 3 1
0 0 1 0 1 1 1 0 4 2 2 2 1 2 0 1 2 2
2 1 0 2 1 0 3 2 0 0 1 3 2 1 4 2 2 0
1 1 2 2 1 0 1 2 3 1 4 1 0 1 1 3 1 0
1 1 2 1 1 1 0 1 3 1 1 1 2 3 2 1 0 2
1 1 1 0 5 3 2 0 1 2 1 2 0 2 0 0 2 1
1 2 2 1 1 1 1 1 0 0 1 1 3 0 0 0 1 0
2 2 1 0 2 2 1 0 2 0 3 3 4 2 2 2 1 0
1 2 2 2 2 1 0 0 0 1 2 0 3 0 2 1 1 1
1 1 1 2 1 4 1 2 2 0 1 1 1 0 1 0 1 0
2 1 2 0 1 1 3 2 6 3 2 1 4 0 2 0 2 0
1 1 1 2 1 2 1 2 1 1 1 1 2 0 0 1 4 1
# La primera columna corresponde a los goles del local y la segunda
# del equipo visitante
```

```
X<-matrix(temp,ncol=2,byrow=T)
mean(X[,1])
[1] 1.536

mean(X[,2])
[1] 1.024

cov(X[,1],X[,2])
[1] 0.2128387

table(X[,1],X[,2])
     0  1  2  3  4
  0  7  5  3  0  0
  1 19 20 15  2  1
  2 13 14  7  1  0
  3  3  3  3  1  0
  4  1  2  3  0  0
  5  0  0  0  1  0
  6  0  0  0  1  0
```

Los resultados más frecuentes son 1-0, 1-1, 1-2, 2-0 y 2-1. Además, vemos que el promedio de goles del local es mayor que el de visitante, 1.54, frente a 1.02 en promedio.

```
# usando el paquete gibbs.met de R, tenemos:
require(gibbs.met)

mc<-gibbs_met(log_f=logVero.biPoisson,no_var=3,
              ini_value=c(1,1,0.2),iters=20000,iters_met=2,
                    stepsizes_met=c(0.2,0.2,0.1), X = X)

plot(exp(mc[,1]),type='l')
title(main=expression(lambda[1]))
plot(exp(mc[,2]),type='l')
title(main=expression(lambda[2]))
plot(exp(mc[,3]),type='l')
title(main=expression(lambda[3]))
media.x<-exp(mc[-c(1:10000),1])+exp(mc[-c(1:10000),3])
plot(density(media.x,from=0),
main='Distribución Marginal de media de X')
media.y<-exp(mc[-c(1:10000),2])+exp(mc[-c(1:10000),3])
plot(density(media.y,from=0),
main='Distribución Marginal de media de Y')
plot(density(media.x-media.y,from=0),
main='Distribución Diferencia de Medias')

quantile(media.x,probs=c(0.025,0.05,1:9/10,0.95,0.975))
    2.5%       5%      10%      20%      30%      40%      50%      60%
1.321991 1.357945 1.395152 1.439441 1.474755 1.505350 1.533398 1.560306
     70%      80%      90%      95%    97.5%
```

```
1.592450 1.627962 1.678656 1.722641 1.764274
quantile(media.y,probs=c(0.025,0.05,1:9/10,0.95,0.975))
    2.5%       5%      10%      20%      30%      40%      50%    60%
0.853732 0.878667 0.908282 0.944443 0.971239 0.996457 1.019010 1.043556
     70%      80%      90%      95%    97.5%
1.067838 1.099439 1.141734 1.176321 1.208515
quantile(media.x-media.y,probs=c(0.025,0.05,1:9/10,0.95,0.975))
    2.5%       5%      10%      20%      30%      40%      50%    60%
0.252218 0.293867 0.342167 0.399882 0.443351 0.478475 0.513161 0.545474
     70%      80%      90%      95%    97.5%
0.580851 0.622917 0.686444 0.734685 0.776613

require(MASS)

# gráfico de la distribución conjunta de la media a posteriori de X y Y.
f1 <- kde2d(media.x, media.y, n = 50)
image(f1)
title(xlab=expression(mu[X]))
title(ylab=expression(mu[Y]))
contour(f1)
persp(f1, phi = 45, theta = 20, d = 2)
```

Vemos que la mediana posterior del número medio de goles de los locales es mayor a la de los visitantes, 1.53 frente a 1.02 en promedio.

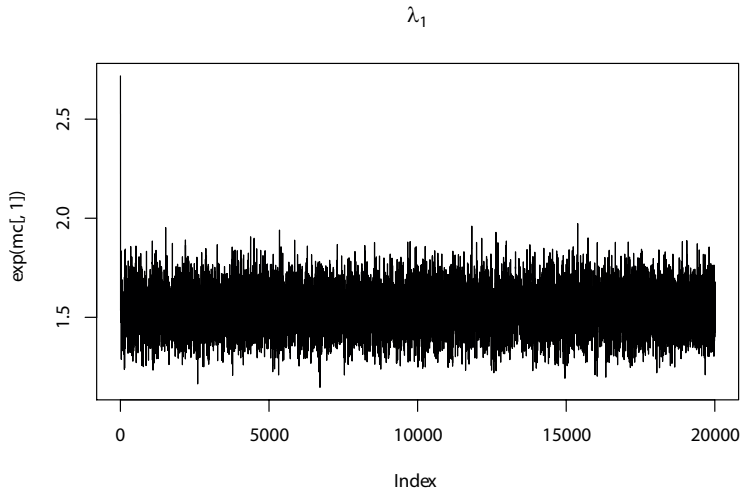

Figura 9.5: *valores tomados por la cadena para el parámetro* λ_1. *Uno de los problemas difíciles es determinar cuántos elementos de la cadena se deben eliminar a su comienzo (burn-in)*

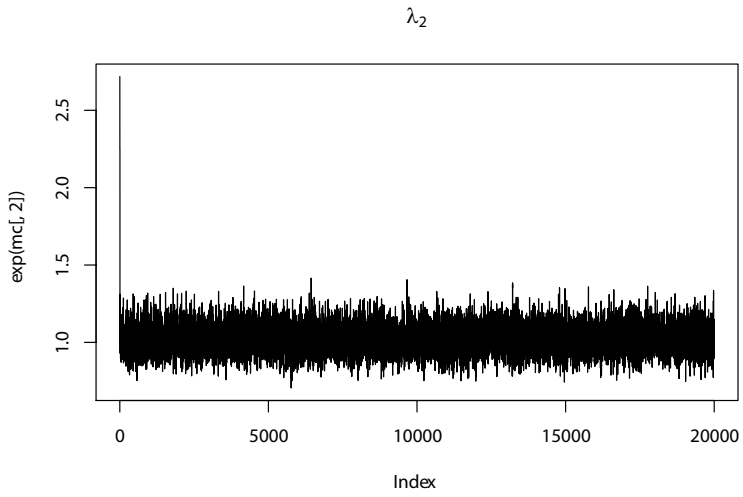

Figura 9.6: *valores tomados por la cadena para el parámetro λ_2. ¿Podemos pensar que empieza a mostrar alguna estabilidad después del punto 15000?*

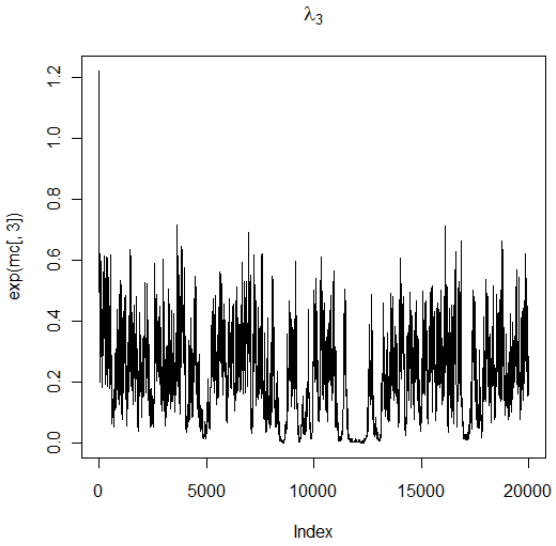

Figura 9.7: *valores tomados por la cadena para el parámetro λ_3. No se aprecia aún convergencia de este parámetro*

En las Figuras 9.8 y 9.9, podemos ver que el número promedio de goles de los locales y de los visitantes se centran alrededor de 1.5 y 1, respectivamente. Entre tanto, la figura 9.10 nos muestra que la diferencia promedio de goles de los locales, respecto a los visitantes, se entra al rededor de 0.5 goles por partido.

La Figura 9.11 muestra la distribución posterior conjunta del número medio de goles de los locales y los visitantes. Aquí se puede apreciar que evidentemente, que el número medio de goles del local tiende a ser mayor que el de visitante en los partidos del segundo torneo del fútbol colombiano en 2010 hasta la fecha 14.

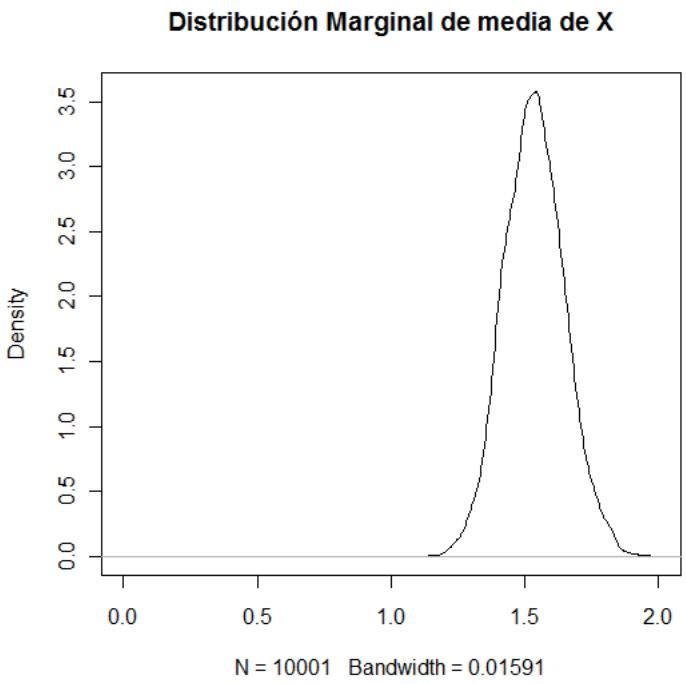

Figura 9.8: *distribución marginal del número promedio de goles de los locales*

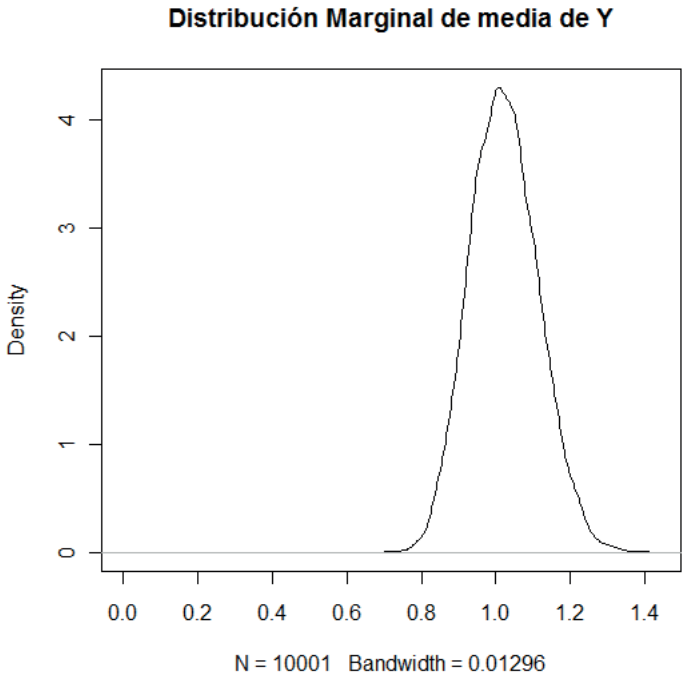

Figura 9.9: *distribución marginal del número promedio de goles de los visitantes*

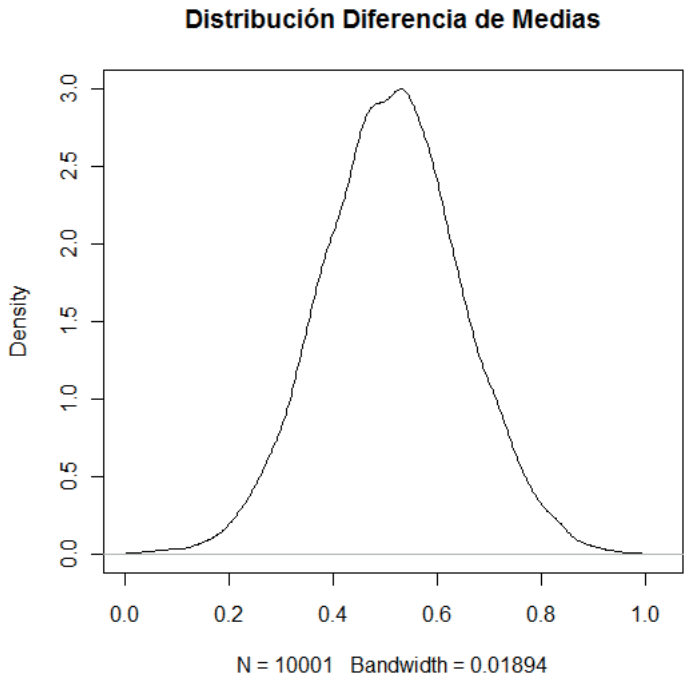

Figura 9.10: *distribución de la diferencia de medias de goles del local versus el visitante*

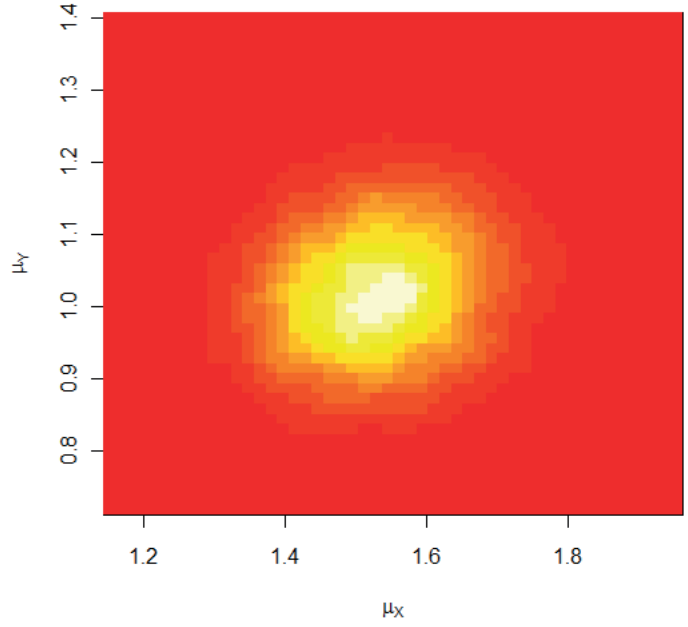

Figura 9.11: *distribución conjunta entre* μ_X *y* μ_Y

Ejemplo 9.4. La distribución Gamma generalizada. Upadhyay, Vasishta y Smith [151] presentan el caso de la distribución gamma generalizada, la cual es reconocida como un modelo flexible para problemas de confiabilidad pero con el cual es difícil de trabajar desde el punto de vista clásico. La f.d.p. de esta distribución es:

$$f\left(x\left|\theta,\beta,\kappa\right.\right)=\frac{\beta}{\Gamma(\kappa)}\frac{x^{(\beta\kappa-1)}}{\theta^{\beta\kappa}}\exp\left(-\left(\frac{x}{\theta}\right)^{\beta}\right),$$

para $x>0$, $\theta>0$, $\beta>0$ y $\kappa>0$.

El parámetro θ es de escala, mientras β y κ determinan la forma de la distribución. Esta familia incluye modelos tales como la gamma de dos parámetros, la Weibull y la exponencial. La lognormal surge cuando se hace tender κ a infinito.

Ya que esta distribución es de uso en confiabilidad, se manejan conceptos como el «tiempo medio hasta que falle (MTF)»y es:

$$MTF=\theta\frac{\Gamma\left(\kappa+\frac{1}{\beta}\right)}{\Gamma(\kappa)}$$

Si $x_1,\ldots,x_n$ es una muestra aleatoria de este modelo (o sea tiempos de falla) y si asumimos aprioris independientes para θ, β y κ

$$\begin{aligned}\xi_1(\theta) &\propto \frac{1}{\theta}\\ \xi_2(\beta) &\propto Gamma\left(a_1,b_1\right)\\ \xi_3(\kappa) &\propto Gamma\left(a_2,b_2\right),\end{aligned}$$

las distribuciones condicionales completas para el muestrador de Gibbs son:

$$\begin{aligned}\xi\left(\theta\left|\beta,\kappa,Datos\right.\right) &\propto \frac{1}{\theta^{(n\beta\kappa+1)}}\exp\left(-\sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^{\beta}\right),\\ \xi\left(\beta\left|\theta,\kappa,Datos\right.\right) &\propto \frac{\beta^{(n+a_1-1)}}{\theta^{(n\beta\kappa)}}\prod_{i=1}^{n}x_i^{\beta\kappa}\exp\left(-\left[\frac{\beta}{b_1}+\sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^{\beta}\right]\right) \quad \text{y}\\ \xi\left(\kappa\left|\theta,\beta,Datos\right.\right) &\propto \frac{1}{\left(\Gamma(\kappa)\right)^{n}}\frac{\kappa^{(a_2-1)}}{\theta^{(n\beta\kappa)}}\prod_{i=1}^{n}x_i^{\beta\kappa}\exp\left(-\frac{\kappa}{b_1}\right)\end{aligned}$$

Ejemplo 9.5. Distribución poli-Weibull. Berger y Sun (1993) discuten la estimación bayesiana de la distribución poli-Weibull. Una variable aleatoria X se dice que sigue esta distribución si su densidad está dada por:

$$f\left(t\left|\beta_j,\theta_j,j=1,\ldots,m\right.\right)=\sum_{j=1}^{m}\frac{\beta_j t^{\beta_j-1}}{\theta_j^{\beta_j}}\exp\left(-\sum_{k=1}^{m}\left(\frac{t}{\theta_k}\right)^{\beta_k}\right),\ \text{para } t>0.$$

Esta distribución surge en el contexto de confiabilidad. Suponga que se tienen m aparatos conectados en serie y no sabemos cuál es el elemento que falla cuando el artículo falla. Por ejemplo las luces de navidad vienen en grupos de m bombillitos y se daña cuando uno de ellos falla, pero usualmente es molesto determinar cuál falló.

Si se prueban r aparatos iguales e independientes con distribución de vida poli-Weibull y se registran $t_1, \ldots, t_n$ tiempos de falla y $t_1^*, \ldots, t_{r-n}^*$ tiempos de funcionamiento de las unidades que no habían fallado aún. La verosimilitud es:

$$L\left(\boldsymbol{\theta}, \boldsymbol{\beta}\right) = \left\{ \prod_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j t_i^{\beta_j - 1}}{\theta_j^{\beta_j}} \right\} \exp\left\{ - \sum_{k=1}^{m} \frac{S\left(\beta_k\right)}{\theta_k^{\beta_k}} \right\},$$

donde

$$S\left(\beta_k\right) = \sum_{i=1}^{n} t_i^{\beta_k} + \sum_{l=1} r - n\left(t_l^*\right)^{\beta_k}$$

Si la distribución a priori se construye de la siguiente forma:

$$\begin{aligned} \xi_1\left(\boldsymbol{\theta}\,|\boldsymbol{\beta}\right) &= \prod_{j=1}^{m} \xi_{1j}\left(\theta_j\,|\beta_j\right) \\ \xi_{1j}\left(\theta_j\,|\beta_j\right) &= \frac{\beta_j b_j^{a_j}}{\Gamma\left(a_j\right)} \theta_j^{-(1+\beta_j a_j)} \exp\left(-\frac{b_j}{\theta_j^{\beta_j}}\right) \\ \xi\left(\boldsymbol{\beta}\right) &= \prod_{j=1}^{m} \xi_{2j}\left(\beta_j\right) I\left(\beta_j > c_j\right) \end{aligned}$$

Cuando hay limitación de datos, las respuestas dependen fuertemente de la selección de ξ_2. En este problema si se escogen aprioris impropias es muy probable terminar con una a posteriori impropia. El algoritmo de Gibbs es complejo, pero aún así permite resolver el problema.

Ejemplo 9.6. Distribución Exponencial generalizada. Kundu y Gupta [152] trabajan la distribución exponencial generalizada de dos parámetros y con la Weibull exponenciada [153] también desarrollan el mismo problema pero hacen referencia del trabajo realizado por los primeros nombrados). Este tipo de distribuciones son usadas en problemas de confiabilidad. En el primer caso la f.d.p. es:

$$f\left(x\,|\alpha\lambda\right) = \alpha\lambda\left(1 - \exp(-\lambda x)\right)^{\alpha-1} \exp(-\lambda x) \text{ para } x > 0,$$

donde $\alpha > 0$ y $\lambda > 0$.

Es común utilizar distribuciones a priori gamma para parámetros positivos. Denotemos por $\xi(\alpha)$ y $\xi(\lambda)$ las a priori de α y λ respectivamente, además asumamos independencia. Entonces,

$$\begin{aligned}\xi(\lambda) &\propto \lambda^{b-1}\exp(-a\lambda),\ \text{para } \lambda>0 \quad \text{y}\\ \xi(\alpha) &\propto \alpha^{d-1}\exp(-c\alpha),\ \text{para } \alpha>0,\end{aligned}$$

donde se asumen los hiperparámetros a, b, c y d conocidos.

Si se tiene una m.a. $x_1, \ldots, x_n$, entonces la verosimilitud es:

$$L(\alpha, \lambda \,|Datos\,) = \alpha^n \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \prod_{i=1}^{n} \left(1-\exp\left(-\lambda x_i\right)\right)^{\alpha-1}$$

La posterior conjunta será por lo tanto,

$$\xi(\alpha, \lambda \,|Datos\,) \propto \alpha^{n+d-1}\lambda^{n+b-1} \exp\left(-\lambda\left(a+\sum_{i=1}^{n} x_i\right)\right) \exp\left(-c\alpha\right) \prod_{i=1}^{n} \left(1-\exp\left(-\lambda x_i\right)\right)^{\alpha-1},$$

y las condicionales completas para el muestreador de Gibbs son:

$$\begin{aligned}\xi(\alpha \,|\lambda, Datos\,) &\propto \alpha^{n+d-1}\exp\left(-c\alpha\right)\prod_{i=1}^{n}\left(1-\exp\left(-\lambda x_i\right)\right)^{\alpha} \quad \text{y}\\ \xi(\lambda \,|\alpha, Datos\,) &\propto \lambda^{n+b-1}\exp\left(-\lambda\left(a+\sum_{i=1}^{n} x_i\right)\right)\exp\left(-c\alpha\right)\prod_{i=1}^{n}\left(1-\exp\left(-\lambda x_i\right)\right)^{\alpha-1}\end{aligned}$$

Ejemplo 9.7. Distribución de Burr La f.d.p. de la distribución Burr generalizada de cuatro parámetros está dada por:

$$f(x \,|\beta, \kappa, \lambda, \mu) = \frac{\beta\kappa}{\lambda}\left(\frac{x-\mu}{\lambda}\right)^{(\beta-1)}\left\{1+\left(\frac{x-\mu}{\lambda}\right)^{\beta}\right\}^{-(\kappa+1)},$$

con $x > \mu$; $\beta, \kappa, \lambda, \mu > 0$. Los parámetros β y κ determinan la forma de la distribución, λ es el parámetro de escala y μ es el de frontera. Si $\mu = 0$ y $\lambda = 1$ se conoce como la distribución Burr Tipo XII. Si μ es conocido o cero, esta es la distribución Burr de tres parámetros.

La función de riesgo (hazard rate, HZR) para el tiempo t está dada por:

$$HZR(t) = \frac{\beta\kappa}{\lambda}\left(\frac{t-\mu}{\lambda}\right)^{(\beta-1)}\left\{1+\left(\frac{t-\mu}{\lambda}\right)^{\beta}\right\}^{-1},$$

la función de confiabilidad,

$$R(t) = \left\{1+\left(\frac{t-\mu}{\lambda}\right)^{\beta}\right\}^{-\kappa},$$

y el tiempo medio hasta fallar (MTF) es:

$$MTF = \mu + \kappa\lambda \frac{\Gamma\left(\beta^{-1}+1\right)\Gamma\left(\kappa+\beta^{-1}\right)}{\Gamma\left(\kappa+1\right)}$$

La función de riesgo es monótona decreciente para $\beta \leq 1$. Tiene curva en forma de bañera invertida si $\beta > 1$.

Upadhyay et al. (2004) utilizan como a priori la siguiente distribución

$$\xi(\beta,\kappa,\lambda,\mu) = \xi_1\left(\kappa\,|\beta\right)\xi_2\left(\beta\right)\xi_3\left(\lambda\right)\xi_4\left(\mu\right),$$

donde,

$$\begin{aligned}
\xi_1\left(\kappa\,|\beta\right) &= \frac{\beta^{(a+1)}}{\Gamma(a+1)b^{(a+1)}}\kappa^a \exp\left[-\left(\frac{\kappa\beta}{b}\right)\right], \\
\xi_2\left(\beta\right) &= \frac{1}{\Gamma(d)c^d}\beta^{(d-1)} \exp\left[-\left(\frac{\beta}{c}\right)\right], \\
\xi_3\left(\lambda\right) &\propto \frac{1}{\lambda} \quad \text{y} \\
\xi_4\left(\mu\right) &\sim Uniforme\left(0, x_1\right),
\end{aligned}$$

donde $a > -1$, $b, c, d > 0$.

Para el muestreador de Gibbs se tienen las siguientes condicionales completas:

$$\begin{aligned}
\xi\left(\beta\,|\kappa,\lambda,\mu,Datos\right) &\propto \beta^{(n+a+d)}\lambda^{-n\beta}\prod_{i=1}^{n}\left(x_i-\mu\right)^{\beta}\left\{1+\left(\frac{x_i-\mu}{\lambda}\right)^{\beta}\right\}^{-(\kappa+1)} \\
&\quad \times \exp\left[-\beta\left(\frac{\kappa}{b}+\frac{1}{c}\right)\right] \\
\xi\left(\kappa\,|\beta,\lambda,\mu,Datos\right) &\propto \kappa^{(n+a)}\prod_{i=1}^{n}\left\{1+\left(\frac{x_i-\mu}{\lambda}\right)^{\beta}\right\}^{-\kappa}\exp\left[-\left(\frac{\kappa\beta}{b}\right)\right] \\
\xi\left(\lambda\,|\beta,\kappa,\mu,Datos\right) &\propto \lambda^{(n\beta+1)}\prod_{i=1}^{n}\left\{1+\left(\frac{x_i-\mu}{\lambda}\right)^{\beta}\right\}^{-(\kappa+1)} \\
\xi\left(\mu\,|\beta,\kappa,\lambda,Datos\right) &\propto \prod_{i=1}^{n}\left(x_i-\mu\right)^{\beta-1}\left\{1+\left(\frac{x_i-\mu}{\lambda}\right)^{\beta}\right\}^{-(\kappa+1)}
\end{aligned}$$

Problemas con el muestreador de Gibbs

Natarajan y McCulloch [154] han señalado que el uso del muestreador de Gibbs con aprioiris propias y difusas pueden llevar a estimaciones inexactas de la distribución posterior y esto lo han ilustrado con modelos jerárquicos.

Justel y Peña [155] han mostrado que el algoritmo de Gibbs no converge fácilmente cuando existen problemas de enmascaramiento fuerte, debido a la presencia de outliers con alto nivel de apalancamiento (leverage), lo cual puede hacer que la estructura de correlación de los parámetros sea muy alta. Las iteraciones en el proceso pueden estabilizarse luego de miles de iteraciones alrededor de valores límites erróneos.

9.1.5. Algoritmo Metropolis-Hastings

El muestreo de importancia y el muestreo de rechazo trabajan bien si la densidad propuesta $q(\theta)$ es similar a $p(\theta)$. En problemas complejos puede ser difícil crear una única $q(\theta)$ que tenga esta propiedad. La construcción de una cadena de Markov no es difícil. Primero describimos el algoritmo de Metropolis-Hastings. Este algoritmo es una generalización de Hastings [156] del método propuesto por Metropolis, Rosenbluth, Teller y Teller [157]. El algoritmo Metropolis utiliza una densidad propuesta q que depende del estado actual de $\theta^{(t)}$. La densidad $q\left(\theta'|\theta^{(t)}\right)$ puede ser tan simple como una normal localizada en $\theta^{(t)}$ y no es necesario que se parezca a $p(\theta)$.

Hitchcock [158] presenta la historia del desarrollo del algoritmo Metropolis-Hastings, la cual revela el poco interés que los estadísticos prestaron a esta metodología durante cuatro décadas. La propuesta inicial fue desarrollada en el laboratorio Los Álamos, el cual contaba con el primer computador, llamado MANIAC (Mathematical Analyzer, Numerical Integrator and Computer). A Metropolis se le conoce como la persona que bautizó los métodos desarrollados por Ulam y von Neuman como Métodos Monte Carlo.

El algoritmo se resume así:

1. Comience en cualquier lugar, y digamos que estamos en $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$.
2. Genere $\boldsymbol{\theta}^*$ de $q\left(\boldsymbol{\theta}^*|\theta\right)$. $\boldsymbol{\theta}^*$ es llamado un *punto candidato* y q es llamada una *distribución propuesta*.
3. Calcule
$$\alpha\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right) = \text{mín}\left\{1, \frac{\xi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^*\right)}{\xi\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}\right)}\right\}$$
4. Acepte $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ con probabilidad $\alpha\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right)$.
5. En otro caso $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}$

Note que la densidad objetivo ξ solo entra en al proceso a través del cociente $\frac{\xi(\boldsymbol{\theta}^*)}{\xi(\boldsymbol{\theta})}$ y por lo tanto no hay necesidad de conocer la constante de normalización para implementar el algoritmo.

Casos especiales:

1. $q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}\right)$: Algoritmo Metropolis.

2. $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*)$: Muestreador independiente.

3. $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = \prod_{i=1}^{k} \xi(\theta_i|\boldsymbol{\theta}^* < i, \boldsymbol{\theta}_{>i}) \Rightarrow \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = 1$: Muestreador de Gibbs.

9.1.6. El algoritmo Metropolis

Aquí la distribución propuesta es simétrica, esto es,

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}),$$

como en el caso de una Normal centrada en el punto actual, entonces el factor:

$$\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} = 1,$$

y el algoritmo Metropolis simplemente se limita a comparar el valor de la densidad objetivo en los dos puntos.

Ejemplo 9.8. Modelo de regresión simple. Asumamos

$$Y_i \sim N\left(\beta_1 X_{i1} + \beta_2 X_{i2}, \sigma^2\right)$$

La formulación bayesiana del modelo consiste en:

1. La función de verosimilitud $f(\boldsymbol{y}|\beta_1, \beta_2, \sigma^2)$
2. La distribución a priori $\xi(\beta_1, \beta_2, \sigma^2)$

Estamos interesados en estimar las siguientes distribuciones posteriores:

- La distribución posterior conjunta

$$\xi\left(\beta_1, \beta_2, \sigma^2|\boldsymbol{y}\right) \propto f\left(\boldsymbol{y}|\beta_1, \beta_2, \sigma^2\right) \times \xi\left(\beta_1, \beta_2, \sigma^2\right)$$

- Distribuciones marginales posteriores $\xi(\beta_1|\boldsymbol{y})$, $\xi(\beta_2|\boldsymbol{y})$ y $\xi(\sigma^2|\boldsymbol{y})$

1. **El Muestreador de Gibbs:** este muestreador genera muestras iterativamente de cada distribución posterior condicional completa.
   - Genere β_1 de $\xi(\beta_1|\beta_2, \sigma, \boldsymbol{y})$
   - Genere β_2 de $\xi(\beta_2|\beta_1, \sigma, \boldsymbol{y})$
   - Genere σ^2 de $\xi(\sigma^2|\beta_1, \beta_2, \boldsymbol{y})$
2. **El Algoritmo Metropolis**
   - Genere un vector de candidatos nuevo $(\beta_1', \beta_2', \sigma^{2\prime})$ de una distribución conocida y fácil de usar

$$q\left(\beta_1, \beta_2, \sigma^2|\beta_1', \beta_2', \sigma^{2\prime}\right)$$

- Acepte los valores propuestos con probabilidad:

$$\alpha = \text{mín}\left\{1,\ \frac{\xi\left(\beta_1', \beta_2', \sigma^{2\prime}|\boldsymbol{y}\right) q\left(\beta_1', \beta_2', \sigma^{2\prime}|\beta_1, \beta_2, \sigma^2\right)}{\xi\left(\beta_1, \beta_2, \sigma^2|\boldsymbol{y}\right) q\left(\beta_1, \beta_2, \sigma^2|\beta_1', \beta_2', \sigma^{2\prime}\right)}\right\}$$

Ejemplo 9.9. Tabla 2×2**.** Se tiene la siguiente tabla 2×2, que presenta información sobre el fumar y desarrollar cáncer pulmonar [74]. La pregunta que se hace un investigador es: ¿existe una diferencia significativa entre los hábitos de los grupos (los que desarrollan cáncer y los que no)? Denotemos por π_L y π_C las probabilidades poblacionales de desarrollar cáncer pulmonar para los casos y controles, respectivamente. Podemos responder a la pregunta mirando la distribución posterior de la diferencia $\pi_L - \pi_C$, pero esta distribución es altamente sesgada. Una solución es utilizar el logaritmo de la razón de odds.

Tabla 9.2: *distribución absoluta del número de fumadores según su condición*

		Condición		
		Cáncer	Control	Total
Fumador	Sí	83	72	155
	No	3	14	17
Total		86	86	192

$$\lambda = \log\left(\frac{\pi_L/(1-\pi_L)}{\pi_C/(1-\pi_C)}\right)$$

Se tiene que $\lambda = 0$ cuando ambas proporciones son iguales. Si $\pi_L > \pi_C$ entonces $\lambda > 0$. El logaritmo de la razón de odds tiene una distribución más simétrica, y se parece a la normal, aún para muestras moderadas. La verosimilitud de los datos está dada por:

$$L\left(\pi_L, \pi_C\right) = \pi_L^{83}\left(1-\pi_L\right)^3 \pi_C^{72}\left(1-\pi_C\right)^{14}, \qquad 0 < \pi_L, \pi_C < 1$$

Utilicemos la simulación para recobrar la distribución "exacta" de λ. Supongamos que la distribución π_L es una $Beta(83, 3)$ y de π_C es una $Beta(72, 14)$, independientes. Ni la suma ni la diferencia de dos Betas tiene una forma estándar, ni la tiene el logaritmo de los odds de dos Beta. Los pasos a seguir son los siguientes:

1. Muestree $\pi_L^{(t)}$ de una $Beta(83, 3)$.
2. Muestree $\pi_C^{(t)}$ de una $Beta(72, 14)$.
3. Calcule:

$$\lambda^{(t)} = \log\left(\frac{\pi_L^{(t)}/(1-\pi_L^{(t)})}{\pi_C^{(t)}/(1-\pi_C^{(t)})}\right)$$

4. Con los $\lambda^{(t)}$ construya un histograma y calcule los estadísticos requeridos de esta distribución.

9.2. Reflexiones acerca el MCMC

Jones y Hobart [159] dicen:

> Cada que una persona emplee el método MCMC, debe preguntarse seriamente lo siguiente:
>
> - (P1) ¿Cuándo debe empezar el muestreo? Esto es, ¿qué tanto debe correr la cadena de Markov para estar lo suficientemente cerca de la distribución estacionaria, o sea, cuántas muestras a quemar?
> - (P2) ¿Cuántas muestras se deben tomar después del quemado? Esto es, cómo sabemos cuándo las estimadas basadas en el resultado de la cadena son lo suficientemente precisas o, puesto de otra forma, cuáles son los errores estándares de las estimadas?
>
> Observe que tratar con (P1) y (P2) es un 'nuevo' problema, que cuando es posible obtener muestras i.i.d. de ξ, (P1) no es un problema y (P2) es fácil de responder.
>
> En la mayoría de la aplicaciones prácticas de MCMC, (P1) y (P2) no son planteadas correctamente. En su lugar, una mezcla de intuición, experiencia y métodos ad hoc son usados para determinar la cantidad de muestras a quemar y la exactitud de las estimaciones resultantes. Uno debe sorprenderse por la calidad de cualquier inferencia subsiguiente.

9.2.1. Problemas con el muestreador de Gibbs

- Determinar el número de iteraciones es un problema difícil de resolver.
- Puede ser extremadamente demandante desde el punto de vista computacional aún para problemas estadísticos a escala pequeña [160].
- Puede ser muy ineficiente cuando la correlación posterior entre los parámetros es alta.
- En modelos jerárquicos tiende a «pegarse».

9.2.2. Ventajas y desventajas dos esquemas de muestreo

El algoritmo Metropolis-Hastings tiene la ventaja de ser fácilmente implementable. Prácticamente no hay restricción en la distribución posterior. Sin embargo, se debe tener cuidado cuando se selecciona la distribución auxiliar para asegurarse que la cadena se mezcle bien. Algunos ajustes son requeridos para la distribución auxiliar.

El muestreador de Gibbs con ARS para la generación de distribuciones condicionales tiene la ventaja de ser más automático y no requiere ajustes extras. Además, puede trabajar con valores truncados o censurados. Tiene la desventaja de generar una sola variable cada vez y por lo tanto, en modelos grandes, la velocidad computacional puede ser lenta. Otro problema es que la correlación serial puede ser alta.

Raftery y Lewis [160] sugieren que el método funciona bien para la mayoría de los problemas con menos de 5000 iteraciones, aunque hay importantes excepciones, como se mencionó en la parte anterior.

9.2.3. Una prueba simple de convergencia

Casella y George [161] presentan una demostración simple de la convergencia del muestreador de Gibbs. Supongamos el caso de una tabla 2×2 bajo un esquema de muestreo multinomial.

		X		
		0	1	Marginal de Y
Y	0	p_1	p_2	$p_1 + p_2$
	1	p_3	p_4	$p_3 + p_4$
Marginal de X		$p_1 + p_3$	$p_2 + p_4$	1

O sea, la distribución de probabilidad conjunta de (X, Y) está dada por:

$$\begin{bmatrix} f_{xy}(0,0) & f_{xy}(1,0) \\ f_{xy}(0,1) & f_{xy}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$

La distribución condicional de $Y|X = x$ es:

$$\boldsymbol{A}_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix},$$

y la distribución condicional de $X|Y = y$ es:

$$\boldsymbol{A}_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix}$$

Las matrices $\boldsymbol{A}_{y|x}$ y $\boldsymbol{A}_{x|y}$ pueden pensarse como las matrices de transición de alcanzar un estado dado otro.

Si solo estamos interesados en generar la distribución marginal de X, entonces empezando en X_0 tenemos que pasar a través de Y_1 para llegar a X_1, ya que el proceso es $X_0 \to Y_1 \to X_1$, y $X_0 \to X_1$ forma una cadena de Markov con probabilidad de transición,

$$P\left(X_1 = x_1 | X_0 = x_0\right) = \sum_y P\left(X_1 = x_1 | X_0 = y\right) P\left(Y_1 = y | X_0 = x_0\right)$$

La matriz de las probabilidades de transición de la sucesión X, digamos $\boldsymbol{A}_{x|x}$, está dada por:

$$\boldsymbol{A}_{x|x} = \boldsymbol{A}_{y|x} \boldsymbol{A}_{x|y}$$

La distribución de probabilidad de cualquier X_k en la secuencia se halla fácilmente. La matriz de transición que produce $P\left(X_k = x_k | X_0 = x_0\right)$ es $\left(\boldsymbol{A}_{x|x}\right)^k$. Además si,

$$f_k = \left[\begin{array}{cc} P(X_k = 0) & P(X_k = 1) \end{array} \right],$$

denota la distribución de probabilidad marginal de X_k, entonces para cualquier k,

$$f_k = f_0 \left(\boldsymbol{A}_{x|x}\right)^k = f_{k-1} \boldsymbol{A}_{x|x}$$

Para cualquier distribución inicial f_0, cuando $k \to \infty$, f_k converge a una única distribución que es un punto estacionario de la ecuación anterior, y satisface:

$$f \boldsymbol{A}_{x|x} = f$$

Así, si la sucesión de Gibbs converge, entonces f debe ser la distribución marginal de X.

Ejemplo 9.10. Pruebas de tamizado. Supongamos que la Secretaría de Salud quiere determinar la prevalencia de un virus particular en la sangre donada en diferentes partes del departamento. Supongamos además que se aplica una prueba tipo ELISA (las siglas en inglés de *Enzime-Linked Inmunosorbent Assay*) para detectar algún tipo particular de virus, por ejemplo el VIH.

Denotemos por D la condición de una unidad particular de sangre y por T el resultado del test aplicado a esa unidad.

$$D = \left\{ \begin{array}{ll} 1 & \text{si la muestra está infectada} \\ 0 & \text{en caso contrario} \end{array} \right.$$

$$T = \left\{ \begin{array}{ll} 1 & \text{si la muestra prueba positivo} \\ 0 & \text{en caso contrario} \end{array} \right.$$

Denotemos por

$$\begin{array}{rcl} \pi & = & P\left(D = 1\right) = prevalencia \quad \text{y} \\ \tau & = & P\left(T = 1\right) \end{array}$$

Hay varios conceptos asociados con este tipo de pruebas y son

Sensibilidad: $\eta = P\left(T = 1 | D = 1\right)$

Especificidad: $\theta = P\left(T = 0 | D = 0\right)$

Valor Predictivo de una Prueba Positiva: $\gamma = P\left(D = 1 | T = 1\right)$

Valor Predictivo de una Prueba Negativa: $\delta = P\left(D = 0 | T = 0\right)$

El interés es determinar π. Esto puede hacerse de varias formas pero el propósito es utilizar el muestreador de Gibbs.

Si conocemos que la distribución conjunta de las variables aleatorias D y T, podemos hallar la prevalencia directamente como la marginal

$$\pi = P(D = 1) = P(D = 1, T = 1) + P(D = 1, T = 0)$$

En su lugar nosotros conocemos las dos distribuciones condicionales $T|D$ y $D|T$, no la conjunta. Para nosotros la distribución condicional de $T|D$ es determinada por η y θ, y la condicional de $D|T$ es determinada por γ y δ.

Aquí están los pasos para proceder con el muestreador de Gibbs para hallar la prevalencia π.

Paso 1: Comience el paso $m = 1$ con un valor arbitrario de D, digamos $D(1) = 1$.

Paso 2a: En el paso $m = 2$, condicionado en el valor $D(1)$ simule si $T(1)$ es 1 o 0. Esto es, simule $T(1) = 1$ con probabilidad η o $T(1) = 0$ con probabilidad $1 - \eta$ (si hubiésemos comenzado con $D(1) = 0$, entonces simularíamos usando $1 - \theta$ o θ).

Paso 2b: Ahora simulamos el valor de $D(2)$ utilizando γ o δ, como sea apropiado. Por ejemplo, si obtuvimos $T(1) = 1$, entonces simularíamos $D(2) = 1$ con probabilidad $\gamma = P\left(D(2) = 1|T(1) = 1\right)$.

Paso 3a: Esta vez, en el paso $m = 3$, simulamos $T(2)$ usando η o θ.

Paso 3b: Ahora simule $D(3)$ utilizando γ o δ, dependiendo del valor de $T(2)$.

Este proceso se estabilizará en el límite. Así obtenemos $D(1), D(2), D(3), \ldots, D(M_1)$ como valores iniciales de «quemado»(se descartan), donde M_1 es un valor «grande»para lograr estabilidad, y de ahí en adelante obtenemos $D(M_1 + 1), D(M_1 + 2), D(M_1 + 3), \ldots, D(M_2)$ de la distribución estable.

Finalmente, estimamos π como la proporción de pasos para los cuales $D(m) = 1$.

Ejemplo 9.11. Distribución ZIP. Asumamos que X es una variable aleatoria discreta con soporte en los enteros no-negativos (una variable de conteo). Un problema que ocurre con cierta frecuencia en la práctica es que $X = 0$ se observa con una frecuencia significativamente mayor (o menor) que la predicha por el modelo asumido. Entonces la variable aleatoria ajustada Y puede ser descrita como:

$$\begin{aligned} P\left(Y = 0\right) &= \omega + (1 - \omega)P\left(X = 0\right) \\ P\left(Y = j\right) &= (1 - \omega)P\left(X = j\right), \quad j = 1, 2, 3, \ldots \end{aligned}$$

Cuando $0 < \omega < 1$ el modelo tiene más ceros. Si $\omega < 0$ el modelo tiene menos ceros.

Un caso de especial importancia es cuando $X \sim Poisson(\lambda)$. La versoimilitud en este caso es:

$$\begin{aligned} L(\omega, \lambda) &= \prod_{i=1}^{n} P\left(Y_i = y_i\right) \\ &= \prod_{i=1}^{n} \left\{ \left(P\left(Y_i = 0\right)\right)^{I(y_i=0)} \left(P\left(Y_i = y_i\right)\right)^{1-I(y_i=0)} \right\} \\ &= \prod_{i=1}^{n} \left\{ \left(\omega + (1-\omega)e^{-\lambda}\right)^{I(y_i=0)} \left((1-\omega)\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}\right)^{1-I(y_i=0)} \right\} \end{aligned}$$

Así, si asumimos una distribución a priori no informativa para ω y para λ, tenemos:

$$\xi(\omega, \lambda) \propto L(\omega, \lambda)$$

Gupta, Gupta y Tripathi [162] hacen referencia a los datos analizados por Leroux y Puterman en 1992 sobre movimientos fetales. Estos datos se recogieron en un estudio sobre respiración y movimiento corporal en fetos de ovejas diseñado para examinar los posibles cambios en el patrón de la actividad fetal durante las dos terceras partes del período de gestación. El número de movimientos efectuados por el feto fue registrado por ultrasonido. Se analizaron los conteos del número de movimientos en una sucesión particular de 240 intervalos de a 5 segundos.

Tabla 9.3: *cantidad de intervalos de a cinco segundos, por número de movimientos fetal de ovejas*

Número de movimientos	0	1	2	3	4	5	6	7
Cantidad de intervalos	182	41	12	2	2	0	0	1

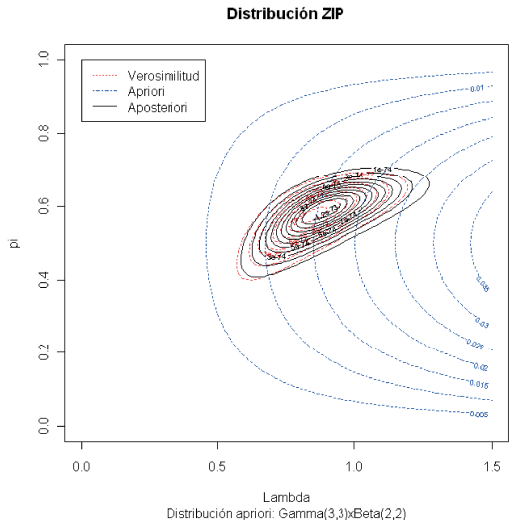

Figura 9.12: *contorno de la función de verosimilitud de la distribución ZIP para el problema de los fetos de ovejas*

El programa completo y descrito en forma detallada sería como el siguiente:

```
# Algoritmo Metropolis
# Función de Verosimilitud
L<-function(omega,lambda,y){
 indicador<-ifelse(y==0,1,0)
 prod1<-prod((omega+(1-omega)*exp(-lambda))^indicador)
 prod2<-prod(((1-omega)*exp(-lambda))^(1-indicador)*lambda^(y*(1-indicador)))
 productoria<-prod1*prod2
 productoria }

# Esta función propone nuevos candidatos
muestreadora<-function(teta.viejo){
omega<-teta.viejo[1]
lambda<-teta.viejo[2]
valor.negativo<-1

# Generamos una normal truncada entre 0 y 1
while(valor.negativo==1){
nuevo1<-rnorm(1,mean=omega)
if(nuevo1>0 & nuevo1<1) valor.negativo<-0
}

valor.negativo<-1
# generamos Una normal truncada mayor que 0
while(valor.negativo==1){
nuevo2<-rnorm(1,mean=lambda)
if(nuevo2>0 ) valor.negativo<-0
}

teta.nuevo<-c(nuevo1,nuevo2)
teta.nuevo
}
```

Nota: Una mejor forma de generar una normal truncada en (a, b)

```
# Genera números aleatorios de una normal truncada
rnormal.trunc<-function(n,media=0,desvi=1,a=0,b=1){
 perce1<-pnorm(a,mean=media,sd=desvi)
 perce2<-pnorm(b,mean=media,sd=desvi)
 res<-qnorm(runif(n,perce1,perce2),mean=media,sd=desvi)
 return(res)
}

# Esta función decide si se acepta el nuevo candidato
qmuestreadora<-function(nuevo,viejo,y){
omega1<-viejo[1]
omega2<-nuevo[1]
lambda2<-nuevo[2]
lambda1<-viejo[2]
```

```
resultado<-(dnorm(omega1,mean=omega2)*dnorm(lambda1,mean=lambda2))/
 (dnorm(omega2,mean=omega1)*dnorm(lambda2,mean=lambda1))
resultado<-resultado*L(omega2,lambda2,y)/L(omega1,lambda1,y)
resultado<-min(1,resultado)
resultado
}

#  DATOS
y<-c(rep(0,182),rep(1,41),rep(2,12),3,3,4,4,7)
# VALOR INICIAL
 viejo<-c(0.05,1)
 Nsim<-10000
 matriz.res<-viejo
 for(i in 1:Nsim){
 nuevo<-muestreadora(viejo)
 prob.acept<-qmuestreadora(nuevo,viejo,y)
 u<-runif(1)
 if(u<prob.acept){
   viejo<-nuevo
   matriz.res<-rbind(matriz.res,nuevo)
   }
 }

 dim(matriz.res)
# con la siguiente instrucción podemos graficar los valores estimados
# para lambda y omega

library(hdrcde)
hdr.boxplot.2d(matriz.res[,2],matriz.res[,1],prob=c(0.001,0.01,0.50,0.80,
0.90,0.95),h = c(5,5),xlab='lambda',ylab='omega' )
points(matriz.res[,2],matriz.res[,1],pch='*')
```

Ahora veamos como estimamos la densidad posterior de λ y ω usando muestreador de Gibbs.

```
# Muestreador de Gibbs con datos del nro. de movimientos fetales
# para encontrar parámetros de una distribución ZIP
muestra.omega <- function(lambda, nro.ceros,n){
rejilla <- seq(0.0001, 0.9999, length=1000)
proba <- (rejilla/(1-rejilla)+exp(-lambda))^nro.ceros*(1-rejilla)^n
proba<-ifelse(is.na(proba),0,proba)
res <- sample(rejilla, 1, prob=proba)
res
}

# ensayo de la funcion 'muestra.omega
# muestra.omega(1,3)
muestra.lambda <- function(omega, Sy, n, n0){
rejilla <- seq(0.000001, Sy, length=100000)
proba <- (omega/(1-omega)+exp(-rejilla))^n0*rejilla^Sy*exp(-rejilla*(n-n0))
```

```
res <- sample(rejilla, 1, prob=proba)
res
}

# Datos: nro. de movimientos fetales en tabla de frecuencias
x <- 0:7
frec <- c(182, 41, 12, 2, 2, 0, 0, 1)
n <- sum(frec)
n0 <- frec[1]
Sy <- sum(x*frec)
teta0 <- c(0.5, 1)
lambda0<-1
omega0<-0.5
resultados <- c(lambda0,omega0)

for(i in 1:2000){
lambda.n<-muestra.lambda(omega0,Sy,n,n0)
resultados<-c(resultados,lambda.n)
omega.0<-muestra.omega(lambda.n,n0,n)
resultados<-c(resultados,omega.0)
}

resultados<-matrix(resultados,ncol=2,byrow=T)
resultados <- resultados[-(1:1001), ]
plot(lambda <- resultados[,1], omega <- resultados[,2],
   ylab='lambda', xlab='omega',pch='*')

library(hdrcde)
hdr.boxplot.2d(lambda,omega,prob=c(0.001,0.01,0.50,0.80,0.90,0.95),
h = c(5,5),xlab='lambda',ylab='omega' )
points(lambda,omega,pch='*')
```

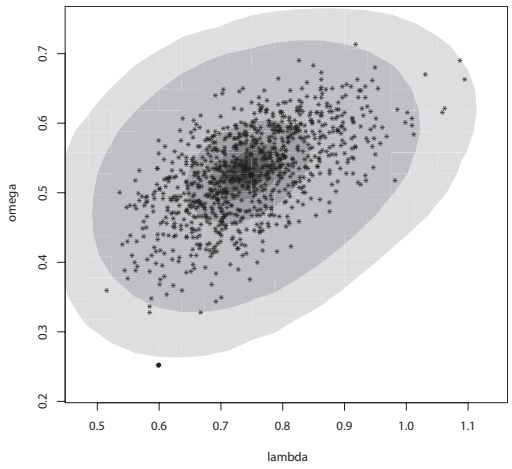

Figura 9.13: *distribución a posteriori conjunta de para λ y ω*

```
hdr.den(omega, prob = c(50, 95, 99),main='Densidad A posteriori de Omega',
 xlab='Omega',ylab='Densidad')

$hdr
         [,1]      [,2]
99% 0.3493410 0.6890070
95% 0.3944155 0.6579649
50% 0.4986371 0.5856638

$mode
[1] 0.5581484

$falpha
       1%        5%       50%
0.2833354 0.9850834 4.8840382

hdr.den(lambda, prob = c(50, 95, 99),main='Densidad A posteriori de lambda',
 xlab='lambda',ylab='Densidad')

$hdr
         [,1]      [,2]
99% 0.5027900 1.0139921
95% 0.5658448 0.9356459
50% 0.6825222 0.8049690

$mode
[1] 0.7425962

$falpha
       1%        5%       50%
0.2070490 0.8193615 3.1736911
```

Podemos ver que los intervalos de probabilidad de alta densidad para las a posteriori de ω y λ muetran que ambos parámetros en promedio se encuentran entre 0 y 1. Lo que indica para ω que hay mucha presencia de ceros en el modelo, y para λ que en promedio el número medio de movimientos del feto por intervalos de cinco segundos es de aproximadamente 0.7 movimientos. Esto se puede apreciar también en las Figuras 9.14 y 9.15.

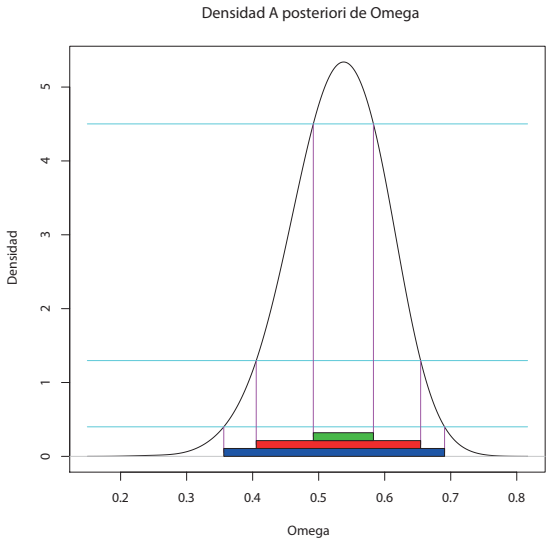

Figura 9.14: *distribución a posteriori de* ω

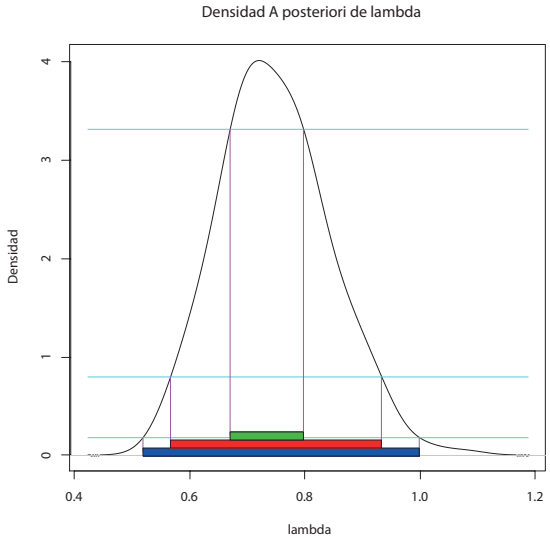

Figura 9.15: *distribución a posteriori de* λ

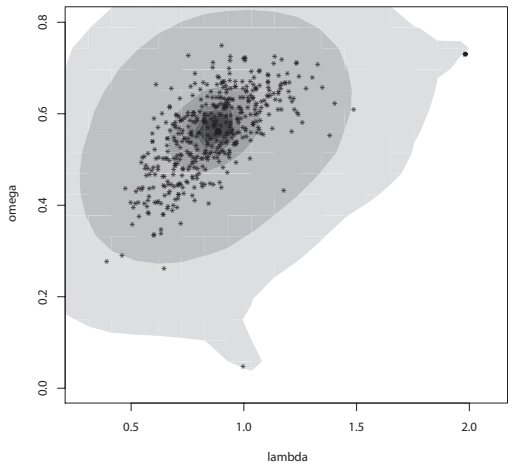

Figura 9.16: *distribución a posteriori conjunta de para λ y ω usando muestreador Metropolis*

9.2.4. Muestreador de Gibbs y problemas con datos censurados

Los datos censurados aparecen con mucha frecuencia en el trabajo estadístico aplicado. Es el dato típico en estudios de sobrevivencia o de confiabilidad. Asumamos que tenemos n observaciones provenientes de una distribución de vida $f(x\,|\theta)$ con r observaciones completas (se observó el tiempo exacto de muerte o falla) $\boldsymbol{x}'(x_1,\ldots,x_r)$. Las demás, $\boldsymbol{x}^*(x_{r+1},\ldots,x_n)$, asumimos, están sujetas a un mecanismo de censura $v_j \leq x_j \leq w_j$, $j = r+1,\ldots,n$. La distribución posterior de θ está dada por:

$$\xi(\theta\,|n,\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}') \propto \xi(\theta)\prod_{i=1}^{r} f(x_i\,|\theta)\prod_{j=r+1}^{n}\int_{v_j}^{w_j} f(x_j\,|\theta)\,dx_j,$$

donde $\boldsymbol{v} = (v_{r+1},\ldots,v_n)$ y $\boldsymbol{w} = (w_{r+1},\ldots,w_n)$.

Las integrales involucradas en la a posteriori anterior usualmente no tienen forma cerrada. Para aplicar el muestreador de Gibbs se puede usar la técnica de «aumentación» donde lo desconocido es $(\theta,\boldsymbol{x}^*)$ [151]. La nuevas condicionales completas tienen las formas:

$$\begin{aligned}
\xi(\theta\,|\boldsymbol{v},\boldsymbol{w},\boldsymbol{x}',\boldsymbol{x}^*) &= \xi(\theta\,|\boldsymbol{x}',\boldsymbol{x}^*)\\
\xi(\boldsymbol{x}^*\,|\boldsymbol{v},\boldsymbol{w},\theta,\boldsymbol{x}') &= f(\boldsymbol{x}^*\,|\boldsymbol{v},\boldsymbol{w},\theta)\\
&= \prod_{j=r+1}^{n}\int_{v_j}^{w_j} f(x_j\,|\theta)\,dx_j
\end{aligned}$$

La primera ecuación es la forma tratable de la posterior que se tendría si no se tuvieran datos censurados, mientras la segunda ecuación es la distribución conjunta de las observaciones censuradas independientes. La generación de variables aleatorias en la segunda ecuación es sencilla, ya que solo es obtener muestras independientes de distribuciones truncadas.

Yu [163] presenta una aplicación de MCMC al caso de datos con doble censura al caso de HIV. Suponga que hay n sujetos y que para el i-ésimo sujeto, el tiempo de origen del evento es denotado por U_i y el tiempo latente de falla V_i^*. El tiempo V_i^* puede estar truncado en C_i así que se observa $V_i = \min\{V_i^*, C_i\}$ y $\delta_i = I(V_i^* \leq C_i)$ en su lugar. Debido a la censura de intervalo, no observamos U_i directamente, sino que U_i cae en el intervalo $[L_i, R_i]$. En adición U_i puede ser truncado a la izquierda por un tiempo de retraso A_i para algunos individuos. El objetivo de análisis es explicar el tiempo $T_i = V_i^* - U_i$ mediante un vector de covariables X_i. Si se considera el modelo de riesgos proporcionales de Cox,

$$\lambda(t\,|X) = \lambda_0(t)\exp(\beta' X),$$

donde β es el vector de coeficientes de regresión y $\lambda_0(t)$ es la función de riesgo base que puede estar sin especificar en el modelo semi-paramétrico de Cox o puede ser una función constante en tramos con $\lambda_0(t) = \lambda_j I(t_{j-1} < t \leq t_j)$ para $j = 1, \ldots, J$.

Otro modelo posible asume la función de sobrevivencia de la forma:

$$S(t\,|X) = \exp(-\pi F(t)),$$

donde $\pi = \exp(\beta' X)$ está relacionado con la tasa de cura y $F(t)$ es una función de distribución.

La estimación del parámetro para $S(t)$ puede ser mejorada relacionando los tiempos de inicio de los eventos a las covariables. Sean $H(u\,|X)$ y $h(u\,|X)$ las funciones de sobrevivencia y de densidad de U, respectivamente. Por ejemplo, para el modelo lognormal,

$$H(u\,|X) = 1 - \Phi\left(\frac{\log(u) - \alpha' X}{\sigma}\right),$$

donde α es el vector de coeficientes de regresión. Si el evento se origina exactamente en $U = u$, la contribución a al verosimilitud es:

$$L_i(\boldsymbol{\theta}\,|u, A_i, X_i, V_i, D_i, \delta_i) = \frac{h(u\,|X_i)}{H(A_i\,|X_i)}\lambda(V_i - u\,|X_i)^{\delta_i}\; S(V_i - u\,|X_i),$$

donde $\boldsymbol{\theta}$ denota el vector de todos los parámetros. En tal caso, H y (β, λ_0) pueden estimarse separadamente factorizando la anterior función de verosimilitud. Ya que U es censurado en intervalo, la función de verosimilitud para los datos observados toma la forma:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\left(\int_{L_i}^{U_i} L_i(\boldsymbol{\theta}\,|u, A_i, X_i, V_i, D_i, \delta_i)\; du\right)$$

La aproximación bayesiana entonces se puede resumir así:

- Se especifican las distribuciones de los tiempos.
- Las aprioris de los parámetros desconocidos son determinadas.

- $\boldsymbol{\alpha} \sim MVN\left(\mathbf{0}, \Sigma_{\alpha}\right)$.
- $\boldsymbol{\beta} \sim MVN\left(\mathbf{0}, \Sigma_{\beta}\right)$.
- $\Sigma_{\beta} \sim Wishart\left(\nu_{\beta}, \Gamma\right)$
- $\sigma^2 \sim Inv - Gamma(a, b)$
- $r \sim Exp(c)$

Los hiperparámetros se seleccionan de tal forma que las a priori sean débilmente informativas.

Ejemplo 9.12. Datos agrupados. En muchas ocasiones los datos muestrales vienen dados en forma de tablas (doble censura). Estimación clásica por máxima verosimilitud para este tipo de datos ha sido estudiada por Tallis [164], entre otros. Consideraremos el caso univariable para las observaciones. Si el espacio muestral S es particionado en $(a_1, a_2], (a_2, a_3], \ldots, (a_{k-1}, a_k)$, donde a_1 puede ser $-\infty$ y a_k puede ser ∞, y asumiendo un modelo muestral $f\left(x \left| \boldsymbol{\theta} \right.\right)$, entonces la probabilidad del i-ésimo intervalo será:

$$\pi_i\left(\boldsymbol{\theta}\right) = \int_{a_{i-1}}^{a_i} f\left(x \left| \boldsymbol{\theta} \right.\right) \, dx,$$

donde $\boldsymbol{\theta}$ pertenece a un subconjunto de R^k. Si tenemos una muestra de n observaciones, denotemos por n_i el número de observaciones que caen en el intervalo $(a_i, a_{i+1}]$, por lo tanto la función de verosimilitud será:

$$L\left(\boldsymbol{\theta} \left| Datos \right.\right) \propto \pi_1\left(\boldsymbol{\theta}\right)^{n_1} \pi_2\left(\boldsymbol{\theta}\right)^{n_2} \ldots \pi_k\left(\boldsymbol{\theta}\right)^{n_k}$$

Si la distribución a priori sobre $\boldsymbol{\theta}$ es denotada por $\xi\left(\boldsymbol{\theta}\right)$, la distribución posterior será:

$$\xi\left(\boldsymbol{\theta} \left| Datos \right.\right) \propto \pi_1\left(\boldsymbol{\theta}\right)^{n_1} \pi_2\left(\boldsymbol{\theta}\right)^{n_2} \ldots \pi_k\left(\boldsymbol{\theta}\right)^{n_k} \xi\left(\boldsymbol{\theta}\right)$$

El algoritmo de Gibbs es el siguiente:

1. Defina un valor inicial $\boldsymbol{\theta}^{(0)}$.
2. (Paso $j+1$). Teniendo $\boldsymbol{\theta}^{(j)}$ calcule
   - (Subpaso 1 del Paso j). Genere
   $$\theta_1^{(j)} \text{ de } \xi\left(\theta_1 \left| \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)} \right.\right)$$
   - (Subpaso 2 del Paso $j+1$). Genere
   $$\theta_2^{(j)} \text{ de } \xi\left(\theta_2 \left| \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)} \right.\right)$$
   - $\vdots$
   - (Subpaso k del Paso $j+1$). Genere
   $$\theta_k^{(j)} \text{ de } \xi\left(\theta_k \left| \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{k-1}^{(j)} \right.\right)$$
3. Repita el paso anterior un gran número de veces hasta que se alcance la distribución estacionaria.

9.3. Cálculo de integrales via simulación

Composición

Supongamos que $f(y|x)$ es una densidad (donde x y y pueden ser vectores). Nuestro objetivo es obtener una muestra aleatoria $y_1, \ldots, y_m$ de:

$$J(y) = \int f(y|x)\, g(x)\, dx$$

El método de composición procede así:

1. Saque $x *^{\sim} g(x)$
2. Saque $x *^{\sim} f(y|x^*)$

Repita los pasos m veces. Los pares $(x_1, y_1), \ldots, (x_m, y_m)$ forman una muestra aleatoria de la densidad conjunta $h(x, y) = f(y|x)\, g(x)$. Las cantidades $y_1, \ldots, y_m$ forman una muestra aleatoria de la marginal $J(y)$.

Capítulo 10

Diagnósticos de los muestreadores MCMC

La utilización de una cadena que no ha convergido aún puede llevarnos a obtener conclusiones con relación a los parámetros de interés o a obtener resultados completamente equivocados con respecto a hipótesis bajo estudio. El asunto de determinar si la cadena ha llegado ya a una etapa estacionaria es un asunto difícil y que solo puede resolverse mediante un desarrollo que dé alguna luz acerca de la estabilidad de la cadena en las últimas iteraciones. Aún así, no podemos nunca estar seguros que hemos llegado a la distribución estacionaria, ya que los resultados teóricos son de carácter asintótico y, sin importar la longitud de la cadena, esta necesariamente es finita.

Recordando que estamos trabajando con una cadena markoviana homogénea, donde el punto de inicio de esta es arbitrario y los valores que toma están correlacionados, además se desea obtener muestras de la distribución estacionaria, los primeros valores generados deben descartarse (burning), denotemos este número por n_B, el cual muchos investigadores los toman como 1000 o 5000. Existen algunas reglas que nos permiten establecer el número a quemar, pero es un tópico que no tiene una única solución. Después de descartar los primeros valores nos queda la muestra definitiva que llamamos muestra a monitorear. Su tamaño lo denotamos por n_M. Cowles y Carlin [165] y Sinharay [166] presentan revisiones extensas de los procedimientos para realizar diagnósticos en MCMC.

Queremos que la estimada $\bar{\theta}$ tenga una alta probabilidad, digamos $1 - \epsilon \times 0.95$, de no estar a más de $d = 0.1$, por ejemplo, del verdadero valor medio $\mu = E(\theta|y)$, o sea,

$$P\left(\left|\bar{\theta} - \mu\right| \leq d\right) = 1 - \epsilon$$

Bajo el supuesto de un $AR_1(\rho)$

$$n_M = \frac{\sigma^2(1+\rho)\left[\Phi^{-1}\left(1-\epsilon/2\right)\right]^2}{d^2(1-\rho)},$$

donde σ es la desviación estándar de un θ_t y Φ es la función de distribución acumulada de una $N(0,1)$.

Como un ejemplo de lo anterior, asumamos que $\hat{\rho} = 0.89$, una cadena que no se mezcla muy bien, $\hat{\sigma} = 3.3$, entonces $n_M \approx 79500$. Si $n_B = 5000$ debemos generar entonces aproximadamente 85000 muestras para un solo parámetro. Si se tienen muchos parámetros a monitorear, como es lo usual en un problema aplicado, puede realmente ser muy restrictivo a nivel de hardware los requerimientos de almacenamiento.

Una de las tareas más difíciles es establecer cuándo podemos decidir que una cadena ha llegado a la distribución límite o de equilibrio, esto puede vislumbrarse a través de pruebas de estacionaridad de los últimos valores generados de la serie, aunque aún teniendo estacionaridad no hay garantía de estar obteniendo valores de la distribución deseada.

Una faceta indeseable en un muestreo MCMC es de no obtener valores bien mezclados, lo cual significa que los valores consecutivos están altamente correlacionados, lo cual podría probarse mediante la correlación de primer orden (correlación serial) de la serie. Una buena cadena tendría un ρ cercano a cero.

10.1. Monitoreo y convergencia de una MCMC

Una cuestión importante, que se relaciona con el monitoreo de una cadena, es sobre cuántos parámetros considerar; realmente la respuesta es chequear todos, ya que si solo se considera un subconjunto de ellos se puede llegar a aceptar una cadena que subconverja (¡que aparentemente converge cuando realmente no!) [166].

10.1.1. Diagnósticos

Existen muchos diagnósticos útiles para analizar los resultados de una cadena, y ya que ninguno de ellos puede garantizar que funcione, Sinharay [167] recomienda que se utilicen varias de la múltiples técnicas disponibles. También es necesario garantizar la convergencia de todos los parámetros involucrados. Entre los diagnósticos tenemos:

Cuatro gráficos MCMC

El monitoreo básico de una cadena se logra mediante gráficos que reflejen el comportamiento secuencial de la misma. Hay cuatro gráficos que se realizan fácilmente y son de gran utilidad como primera aproximación. Se recomienda como paso inicial generar cuatro gráficos para cada parámetro considerado:

1. **Un gráfico de los valores de la cadena en forma de serie temporal**. Estos se pueden presentar como una serie de tiempo en su totalidad, lo cual ayuda a determinar cuántas muestras es necesario quemar antes de recolectar los valores considerados como la muestra de la distribución estacionaria. Algunos programas permiten observar la evolución de la cadena, mediante ventanas con una cierta cantidad de valores generados. Si se tienen varias cadenas, todas las cadenas del mismo parámetro se grafican simultáneamente, de tal forma que uno esperaría que en el momento de lograr convergencia todas las series se entrecrucen.

Otro gráfico que es útil es de medias móviles. Sinharay [166] recomienda graficar medias móviles calculadas a partir de grupos de 50 valores. Si el algoritmo logra convergencia las medias móviles deben ser bastante similares, mostrándose gráficamente como paralelas al eje horizontal. Este gráfico solo mira el comportamiento de la media de los datos y no nos presentan el cuadro completo de la convergencia, ya que, recordemos, la convergencia es a toda una distribución.

2. **Un gráfico de la densidad estimada a partir de estos valores**.
3. **Un gráfico con las autocorrelaciones**. Si este gráfico muestra un decaimiento a cero lento puede ser un indicativo de un mezclado deficiente, lo cual puede sugerir una reparametrización o alguna otra aproximación.
4. **Un gráfico con las autocorrelaciones parciales**.

Prueba de Geweke

Es una prueba de igualdad de medias utilizando el $Z - score$. Si $|Z - score| > 2$ se considera que los niveles son diferentes. Geweke recomendó usar el 5 % para el primer subconjunto y 10 % para el segundo subconjunto. Geweke también implicaba que el procedimiento servía para determinar cuántas observaciones iniciales se descartaban. Obviamente estas dos submuestras deben ser lo suficientemente grande para garantizar la aplicación del teorema central del límite.

Un problema con esta técnica es por ser esencialmente univariable y además solo es útil para una sola cadena MCMC. Otro problema está en que depende en parte de la experiencia del usuario.

Prueba de Heidelberger y Welch

Esta prueba usa el estadístico Cramér-von Mises para estacionalidad. Funciona así: si falla la prueba se descarta el 10 % de las observaciones (las primeras), y así hasta descartar el 50 %.

Prueba de Raftery y Lewis

Este es un diagnóstico de la longitud de la corrida basada en el criterio de la exactitud de la estimación del cuantil q. Pretende usar una cadena de Markov piloto (corta). Se calcula el número de iteraciones requeridas para estimar el cuantil q dentro de una exactitud de $\pm r$ con probabilidad p. Los autores han proporcionado un programa implementando la librería CODA (que está disponible en R), la cual entrega el número de iteraciones a ser realizadas, el número de muestras a ser quemadas y el número de valores k a ser descartados en la cadena de valores aceptados. Sin embargo, algunos autores como MacEachern y Berliner, no favorecen la práctica de descartar resultado intermedios, ya que la calidad de la estimación se degrada [165].

Cowles y Carlin [165] señalan que algunas críticas a este método apuntan a que diferentes puntos de inicio de la cadena pueden resultar en diferentes números de iteraciones y que la información que se obtiene es de carácter univariable.

Prueba de Gelman y Rubin

Esta es una prueba en la que dos o más cadenas paralelas corren con valores iniciales que son sobredispersos (su varianza es superior a la esperada) con respecto a la distribución posterior. Cowles y Carlin [165] recomiendan 10 cadenas cuando se tiene una distribución a posteriori unimodal. La convergencia se diagnostica cuando las cadenas han «olvidado»sus valores iniciales y las salidas de todas las cadenas son indistinguibles. La prueba está basada en una comparación de las varianzas dentro y entre las cadenas y es similar al análisis de varianza clásico. Hay dos formas de estimar la varianza de una distribución estacionaria: la media de la varianza empírica dentro de cada cadena, W, y la varianza empírica de todas las cadenas combinadas, que puede expresarse como:

$$\hat{\sigma}^2 = (n-1)B/n + W/n$$

donde B es la varianza empírica entre las cadenas.

Si las cadenas han convergido, entonces ambas estimadas son insesgadas. De otra manera el primer método subestima la varianza, ya que las cadenas individuales no han tenido tiempo de llegar a la distribución estacionaria y el segundo método sobreestima la varianza, ya que los valores iniciales fueron seleccionados sobredispersos.

El diagnóstico de convergencia está basado en el supuesto que la distribución objetivo es normal. Un intervalo bayesiano de credibilidad puede construirse usando una distribución t con media

$$\hat{\mu} = \text{Media muestral de todas las cadenas combinadas,}$$

y varianza

$$\hat{V} = \hat{\sigma}^2 + B/(mn),$$

donde m es el número de cadenas, y los grados de libertad son estimado por el método de los momentos

$$d = 2\frac{\hat{V}}{Var(\hat{V})}$$

El uso de la distribución t tiene en cuenta el hecho que la media y la varianza de la distribución posterior son estimados.

El diagnóstico de convergencia es:

$$R = \sqrt{(d+3)\hat{V}/((d+1)W)}$$

Valores sustancialmente mayores que 1 indican falta de convergencia.

Una de las críticas que se le hace al proceso es que 10 cadenas que generen 1000 puntos cada una no produce un mejor resultado que una sola cadena que corra y produzca 10000 puntos, ya que uno esperaría que esta última cadena al final estuviera

más cercana a la distribución estacionaria que las 10 cadenas iniciales. Si computacionalmente no es costoso y se tienen los resultados de las cadenas múltiples luego del quemado, se puede implementar un proceso de mezclado de las cadenas de tal forma que se genere una supercadena.

Factor de reducción de escala potencial de Gelman y Rubin (PSRF)

Chequee para cada parámetro si la escala/varianza de su distribución posterior aproximada decrecerá significativamente si las simulaciones continúan indefinidamente.

Factor de reducción de escala potencial multivariable de Brooks y Gelman (MPSRF)

Esta es la versión multivariable de la PSRF. Denote por $\boldsymbol{\theta}$ el vector p-dimensional. Sea $\boldsymbol{\theta}_i^t$ el valor generado en la iteración t de la cadena i del algoritmo MCMC. La matriz de varianzas y covarianzas posterior es estimada por:

$$\hat{\boldsymbol{V}} = \frac{n-1}{n}\boldsymbol{W} + \left(1 + \frac{1}{m}\right)\frac{\boldsymbol{B}}{n},$$

donde n es el número de iteraciones de la cadena,

$$\boldsymbol{W} = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{t=1}^{n}\left(\boldsymbol{\theta}_i^t - \bar{\boldsymbol{\theta}}_{i\cdot}\right)\left(\boldsymbol{\theta}_i^t - \bar{\boldsymbol{\theta}}_{i\cdot}\right)',$$

y

$$\boldsymbol{B} = \frac{n}{m-1}\sum_{i=1}^{m}\left(\bar{\boldsymbol{\theta}}_{i\cdot} - \bar{\boldsymbol{\theta}}_{\cdot\cdot}\right)\left(\bar{\boldsymbol{\theta}}_{i\cdot} - \bar{\boldsymbol{\theta}}_{\cdot\cdot}\right)',$$

denota las estimaciones de la matriz de varianzas y covarianzas dentro y entre del parámetro respectivamente.

Brooks y Gelman en 1998 [166] mostraron que si λ_1 es el mayor valor propio de la matriz $\boldsymbol{W}^{-1}\boldsymbol{B}/n$, entonces:

$$\hat{R}_p = \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\lambda_1 \longrightarrow 1,$$

si las cadenas se mezclan a medida que el tamaño muestral crece. Ellos llaman esta cantidad MPSRF.

Robert

El diagnóstico de convergencia de Robert es un estimador insesgado de $\left\|f^{(n)} - f\right\| + 1$. Requiere correr m cadenas paralelas reversibles del muestreador de Gibbs, todas empezando en el mismo valor inicial $\theta^{(0)}$, y se calcula como:

$$J_n = \frac{1}{m(m-1)} \sum_{l \neq p} \frac{k\left(\theta_l^{(1/2)}, \theta_p^{(2n-1)}\right)}{\theta_p^{(2n-1)}},$$

donde l y p son replicaciones del muestrador, $\theta_l^{(1/2)}$ es el valor obtenido después de la mitad de la primera iteración de la l-ésima cadena, $\theta_p^{(n)}$ es el valor obtenido luego de la n-ésima iteración completa de la p-ésima cadena, y k es el kernel del muestreador reversible. Robert sugiere que se evalúe gráficamente la convergencia monotónica de este diagnóstico.

Una ventaja señalada por Cowles y Carlin [165] es su rigurosidad teórica, pero su problema es la dificultad de su implementación. Otras desventajas fueron señaladas por ellos, y esto ha hecho que este método no se encuentre implementado en el software disponible.

Zellner y Min

Zellner y Min [168] proponen varias pruebas para diagnosticar convergencia en el muestreador de Gibbs.

La prueba se conoce como el criterio de convergencia del muestreador de Gibbs (GSC^2). Se puede aplicar cuando el vector de parámetros del modelo se puede dividir en dos partes, α y β, de tal forma que la distribución posterior conjunta pueda ser escrita analíticamente.

$$\xi\left(\alpha, \beta \left| Datos \right.\right) = c\xi\left(\alpha, \beta \left| I_0 \right.\right) L\left(\alpha, \beta \left| Datos \right.\right),$$

con $(\alpha, \beta) \in \Omega$ y $\xi\left(\alpha, \beta \left| Datos \right.\right) > 0$ in Ω. I_0 representa la información a priori. Los $Datos$ la información muestral.

1. **El criterio de convergencia razón anclada (ARC^2)**. Para cualquier par de puntos (α_1, β_1) y (α_2, β_2), se evalúa el criterio de convergencia de la razón anclada, θ_{12}, está dado por:

   $$\theta_{12} = \frac{\xi\left(\alpha_1, \beta_1 \left| I_0 \right.\right) L\left(\alpha_1, \beta_1 \left| Datos \right.\right)}{\xi\left(\alpha_2, \beta_2 \left| I_0 \right.\right) L\left(\alpha_2, \beta_2 \left| Datos \right.\right)} = \frac{\xi\left(\alpha_1, \beta_1 \left| Datos \right.\right)}{\xi\left(\alpha_2, \beta_2 \left| Datos \right.\right)}$$

   Se usan las estimaciones,

   $$\hat{\xi}_N\left(\alpha_i \left| D \right.\right) = \frac{1}{N} \sum_{j=1}^{N} \xi\left(\alpha_i \left| \beta^{(j)}, D \right.\right),$$

   y

   $$\hat{\xi}_N\left(\beta_i \left| D \right.\right) = \frac{1}{N} \sum_{j=1}^{N} \xi\left(\beta_i \left| \alpha^{(j)}, D \right.\right),$$

para $i = 1, 2$ y donde $j = 1, 2, \ldots, N$ denota la secuencia del muestreador de Gibbs obtenidos de las densidades condicionales, $\xi(\alpha\,|\beta, D)$ y $\xi(\beta\,|\alpha, D)$. De lo anterior se calcula:

$$\hat{\theta}_{12}(N) = \frac{hat\xi_N(\alpha_1\,|D)\,\hat{\xi}_N(\beta_1\,|\alpha_1, D)}{hat\xi_N(\alpha_2\,|D)\,\hat{\xi}_N(\beta_2\,|\alpha_2, D)},$$

y

$$\tilde{\theta}_{12}(N) = \frac{hat\xi_N(\beta_1\,|D)\,\hat{\xi}_N(\alpha_1\,|\beta_1, D)}{hat\xi_N(\beta_2\,|D)\,\hat{\xi}_N(\alpha_2\,|\beta_2, D)}$$

Ellos argumentan que si la corrida es satisfactoria, entonces los valores $\hat{\theta}_{12}(N)$ y $\tilde{\theta}_{12}(N)$ deben estar cerca de θ_{12}.

2. **Criterio de convergencia de la diferencia** DC^2. Tenemos

$$\begin{aligned}\xi(\alpha, \beta\,|D) &= \xi(\alpha\,|D)\,(\beta\,|\alpha, D)\\ &= \xi(\beta\,|D)\,(\alpha\,|\beta, D)\end{aligned}$$

Entonces el criterio de convergencia de la diferencia en cualquier punto en el espacio parametral, digamos (α_i, β_i), es

$$\xi(\alpha_i\,|D)\,(\beta_i\,|\alpha_i, D) - \xi(\beta_i\,|D)\,(\alpha_i\,|\beta_i, D)$$

Si tenemos el resultado del muestreador de Gibbs, entonces,

$$\hat{\eta}_i(N) = \hat{\xi}_N(\alpha_i\,|D)\,(\beta_i\,|\alpha_i, D) - \hat{\xi}_i(\beta_i\,|D)\,(\alpha_i\,|\beta_i, D)$$

Los valores $|\hat{\eta}_i|$ deberían estar muy cercanos a cero.

3. **El criterio de convergencia de la razón** RC^2. De lo expuesto arriba se tiene:

$$\frac{\xi(\alpha_i\,|D)\,(\beta_i\,|\alpha_i, D)}{\xi(\beta_i\,|D)\,(\alpha_i\,|\beta_i, D)} = 1$$

Por lo tanto, si definimos:

$$\hat{\gamma}_i(N) = \frac{\hat{\xi}_N(\alpha_i\,|D)\,(\beta_i\,|\alpha_i, D)}{\hat{\xi}_i(\beta_i\,|D)\,(\alpha_i\,|\beta_i, D)}$$

Esta razón debe estar muy cerca a 1.

Cartas EWMAST

Zhang [169] desarrolló las cartas de control *promedio móvil ponderado exponencialmente para datos estacionarios* (EWMAST) para procesos estacionarios débiles. Si el parámetro que estamos monitoreando ya está en el estado estable, entonces podemos decir que corresponde a un proceso $AR(1)$.

La carta EWMA de X_t se define como:

$$Z_t = (1-\lambda)Z_{t-1} + \lambda X_t, \text{ para } t = 1, 2, \ldots,$$

donde $Z_0 = \mu$, λ es una constante, $(0 < \lambda \leq 1)$, y las X_t's son v.a.'s i.i.d. con media μ y varianza σ_x^2. Tenemos:

$$\sigma_z^2 = var\left[Z_t\right] = \frac{\lambda}{(2-\lambda)}\left(1-(1-\lambda)^{2t}\right)\sigma_x^2.$$

Cuando t es grande, una varianza aproximada es:

$$\sigma_z^2 = var\left[Z_t\right] \approx \frac{\lambda}{(2-\lambda)}\sigma_x^2.$$

Los límites de una carta EWMA son $\mu \pm L\sigma_z$, donde σ_z es la desviación típica de Z_t. Los límites de la carta EWMAST consideran la autocorrelación de los datos.

Si se asume que $\{X_t\}$ es un proceso estacionario discreto con media y función de autocovarianzas constantes. Esto es,

$$\begin{aligned} E\left(X_t\right) &= \mu, \text{ para todo } t \\ R(\tau) &= cov\left(X_t, X_{t+\tau}\right) = E\left[\left(X_t-\mu\right)\left(X_{t+\tau}-\mu\right)\right], \end{aligned}$$

donde $R(\tau)$. la función de autocovarianza solo depende del rezago τ.

Zhang [169] muestra que la función de autocovarianza está dada por:

$$\begin{aligned} cov\left(Z_t, Z_{t+\tau}\right) &= \frac{\lambda}{(2-\lambda)}\sigma_x^2 \\ &\quad \times \left\{ \sum_{k=0}^{t-1} \rho(k+\tau)(1-\lambda)^k \left[1-(1-\lambda)^{2(t-k)}\right] \right. \\ &\quad + \sum_{k=1}^{\tau} \rho(\tau-k)(1-\lambda)^k \left[1-(1-\lambda)^{2t}\right] \\ &\quad \left. + \sum_{k=\tau+1}^{t-1+\tau} \rho(\tau-k)(1-\lambda)^k \left[1-(1-\lambda)^{2(t+\tau-k)}\right] \right\}, \end{aligned}$$

donde $\rho(k)$ es la autocorrelación de X_t de rezago k y

$$\rho(k) = \frac{R(k)}{R(0)} = \frac{R(k)}{\sigma_x^2}.$$

En particular, cuando $\tau = 0$, la varianza de Z_t es:

$$\begin{aligned} \sigma_z^2 &= var(Z_t) = \frac{\lambda}{(2-\lambda)}\sigma_x^2 \\ &\quad \times \left\{ 1-(1-\lambda)^{2t} + 2\sum_{k=1}^{t-1} \rho(k)(1-\lambda)^k \left[1-(1-\lambda)^{2(t-k)}\right] \right\} \end{aligned}$$

Bajo el supuesto de normalidad de las X_t's, Z_t normalmente distribuido con media μ y varianza σ_z^2. La carta EWMAST es construida para monitorear Z_t. La línea central está en μ y los límites están localizados en

$$\mu \pm L\sigma_z,$$

donde σ_z es la desviación típica de Z_t. Asumiendo que no hay cambio en la autocorrelación de $\{X_t\}$, la carta EWMAST monitorea los cambios de señal en la media del proceso.

Tanto μ como σ_z son estimados a partir de registros históricos de X_t cuando el proceso está bajo control, reemplazando mu por la media muestral, σ_x^2 y $\rho(k)$ por sus estimaciones muestrales.

Zhang [169] recomienda seguir los siguientes pasos para la construcción de la carta:

1. Determine un periodo con $N(\geq 100)$ observaciones cuando el proceso esté en una condición estable. Calcule la media muestras del proceso, $\bar{x}$, $\hat{\sigma}_x$ usando la desviación típica del proceso, y las autocorrelaciones muestrales $[\hat{\rho}(k)(k = 1, \ldots, 25)]$ de todas las observaciones en ese punto.

2. Calcule la desviación estándar aproximada EWMA $\hat{\sigma}_z$ con un λ apropiado (usualmente se escoge $\lambda = 0.2$) y $M = 25$.

3. La carta EWMAST es construida monitoreando Z_t, la EWMA. La línea central está en $\bar{x}$ y los límites $L\sigma$ están en $\bar{x} \pm L\hat{\sigma}_z$.

4. Una vez la carta EWMAST da una señal indicando que el proceso está fuera de control, la media del proceso necesita ser re-estimada cuando el proceso esté estable. La varianza y autocorrelaciones del proceso también necesitan ser ajustadas.

Para adaptar esta carta al MCMC procedemos así:

1. Tenemos las muestras $\theta_0, \theta_1, \ldots, \theta_N$. El proceso lo consideramos en reversa, de tal forma que la primera observación sería θ_N, la segunda θ_{N-1}, y así sucesivamente. Consideramos entonces las primeras K observaciones de este proceso, las cuales, si N es lo suficientemente grande, podemos considerar que está en estado estable.

2. Construimos la carta con estas K observaciones.

3. Determinamos a partir de qué observación el proceso se sale de control. De ahí en adelante podemos descartar el resto de observaciones.

10.2. Diagnósticos en CODA

La librería del R CODA posee varios diagnósticos útiles para analizar los resultados de una cadena.

Ejemplo 10.1. Estimación paramétrica del tiempo medio de falla cuando solo es posible observar un punto en el tiempo. Supongamos que en un experimento para determinar la duración de un producto se realiza una prueba para n unidades y que solo es posible observar el resultado en un único punto del tiempo, digamos x_0. Tendremos entonces que n_0 ya fallaron y n_1 quedan aún funcionando. Los datos que obtenemos serán entonces,

$$x_0^-, x_0^-, \ldots, x_0^-, x_0^+, x_0^+, \ldots, x_0^+,$$

donde tenemos n_0 x_0^-'s, aquí la notación indica que ya fallaron, pero no se registró el tiempo exacto, y n_1 x_0^+´s, unidades que no han fallado aún. Asumamos que el tiempo se distribuye exponencial con densidad dada por:

$$f(x;\lambda) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$$

Por lo tanto la verosimilitud será:

$$L(\lambda) = \left(1-\exp\left(-\frac{x_0}{\lambda}\right)\right)^{n_0}\left(\exp\left(-\frac{x_0}{\lambda}\right)\right)^{n_1}$$

Asumamos que la a priori es una distribución no informativa

$$\xi(\lambda) \propto K$$

Por lo tanto, la a posteriori será proporcional a la verosimilitud

$$\xi(\lambda|Datos) \propto \left(1-\exp\left(-\frac{x_0}{\lambda}\right)\right)^{n_0}\left(\exp\left(-\frac{x_0}{\lambda}\right)\right)^{n_1}$$

Suponga que la duración de una resistencia es exponencial con parámetro λ. A priori asumimos una distribución no informativa constante. Se ponen a funcionar 20 resistencias y a las 8 horas se observan. Cuatro de ellas habían fallado y las otras continuaban funcionando.

```
# Muestreador de Metropolis

# Como muestreadora usaremos una gamma.
# Valor inicial
L0<-1 ; res<-L0

for(i in 1:100000){ # genera punto candidato
  Lc<-rgamma(1,L0,scale=1)
  cociente<-4*log(1-exp(-8/Lc))-16*8/Lc-dgamma(Lc,L0,scale=1,log=T)-
           (4*log(1-exp(-8/L0))-16*8/L0-dgamma(L0,Lc,scale=1,log=T))
  cociente<-exp(cociente)
  if(cociente>1){
  L0<-Lc
```

```
  res<-c(res,Lc)
  }
  else{
   if(runif(1)<cociente){
    L0<-Lc
    res<-c(res,Lc)
   }
  }
}
```

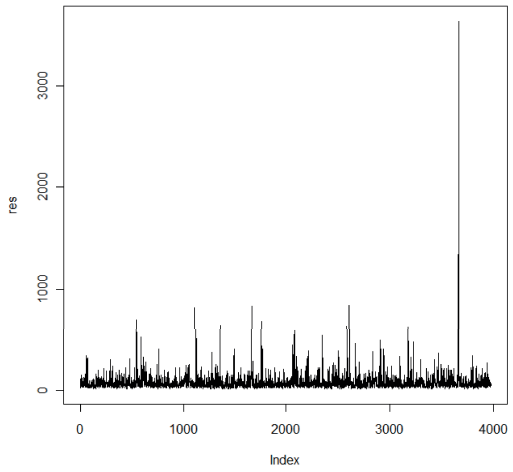

Figura 10.1: *cadena generada para el parámetro λ de la a posteriori*

```
res<-res[-(1:1000)]
plot(res,type='l')
hist(res)
summary(res)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.76   37.49   57.53   80.13   90.26 3631.00

quantile(res,probs=c(0.01,0.05,0.10,0.20,0.25,3:7/10,0.75,0.8,0.95,0.99))
       1%        5%       10%       20%       25%       30%       40%
 16.66239  23.19815  27.36604  34.08504  37.49353  41.20592  48.08089
      50%       60%       70%       75%       80%       95%       99%
 57.53452 67.94006  82.50258  90.25736 100.30699 196.28691 401.89871

acf(res,type='cor')
plot(density(res[res<1000],bw=50,from=0),main='Densidad Posterior')
abline(h=0)
abline(v=0)
```

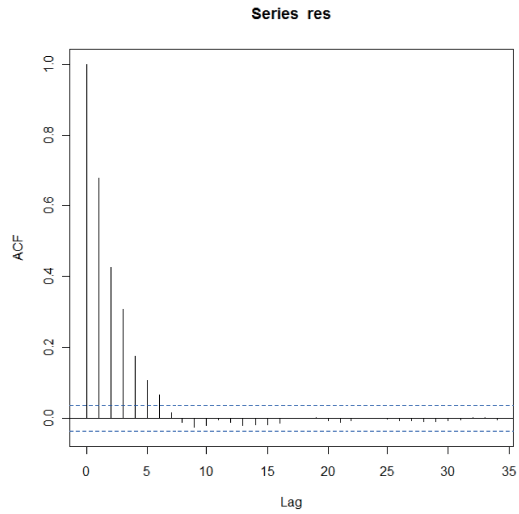

Figura 10.2: *función de autocorrelación para los valores de la cadena generada para el parámetro λ de la a posteriori*

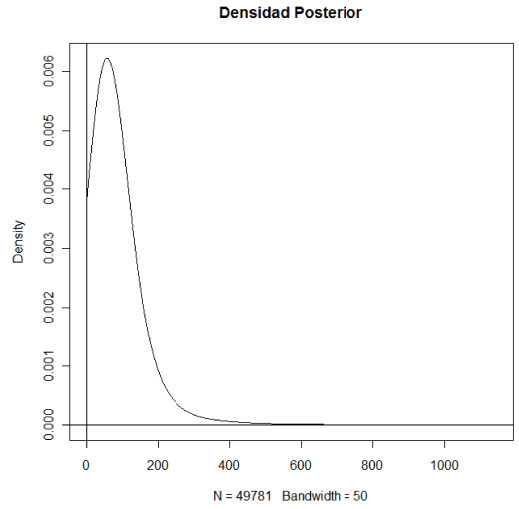

Figura 10.3: *densidad posterior del tiempo medio de duración de las resistencias*

Prueba KPSS [170]

```
library(tseries)
kpss.test(res)

        KPSS Test for Level Stationarity
data:  res KPSS Level = 0.1472, Truncation lag parameter = 12,
p-value = 0.1
Warning message: In kpss.test(res) : p-value greater than printed
p-value
```

El test KPSS muestra que la cadena generada para λ ha alcanzado su convergencia.

Parte III

Aplicaciones

Capítulo 11

Modelos lineales

11.1. La regresión clásica

Un modelo de regresión es un medio formal para expresar los dos ingredientes esenciales de una relación estadística:

1. Una tendencia de la variable dependiente Y que cambia, cuando la variable independiente cambia, en una forma sistemática.

2. Una dispersión de los puntos alrededor de la relación estadística.

Estas características se expresan en un modelo de regresión como:

1. Para cada nivel de X hay una distribución de probabilidad de Y.

2. Las medias de estas distribuciones de probabilidad cambian en una forma sistemática con X.

11.1.1. Regresión simple

El modelo más sencillo, pero el más útil, es el que se conoce como *modelo de regresión simple*. Si tenemos una variable, Y, en cuyo comportamiento estamos interesados cuando la condicionamos en ciertos valores de otra variable, X, el modelo de regresión simple nos dice que la media condicional de Y dado un valor de $X = x$, denotada por $E[Y|X] = \mu_{Y|X}$, es una función lineal de X, o sea,

$$E[Y|X = x] = \mu_{Y|X=x} = \beta_0 + \beta_1 x,$$

donde β_0 y β_1 se conocen como los parámetros del modelo. Estos valores usualmente son desconocidos y el problema es estimarlos a partir de una muestra de individuos de la población.

Sea $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ una muestra aleatoria extraída de la población de referencia. Observe como cada individuo proporciona información simultáneamente sobre X y sobre Y. El individuo i-ésimo puede representarse en términos del modelos así:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Supuestos:

1. $e_i \sim Normal\,(0, \sigma^2)$, varianza constante (homoscedasticidad)
2. $Cov\,(e_i, e_j) = 0$ para todo $i \neq j$

Notación:

$$y_i = \left(Y_i - \bar{Y}\right)$$
$$x_i = \left(X_i - \bar{X}\right)$$

Las minúsculas denotan desviaciones de la media.

11.1.2. Modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e_i,$$

donde:

- Y: Respuesta o variable dependiente
- $X_1, X_2, \ldots, X_k$: k variables explicatorias o independientes (no estocásticas)
- $\beta_0, \beta_1, \ldots, \beta_k$: $k+1$ parámetros (usualmente desconocidos)
- e: Error aleatorio
  1. $E\,(e) = 0$
  2. $Var\,(e) = \sigma_e^2$
  3. Adicionalmente se asume normal

$$Y \sim N\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k, \sigma^2\right)$$
$$E\,[Y|X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

La muestra aleatoria consta de n puntos. El i-ésimo punto se denota como:

$$(X_{i1}, X_{i2}, \ldots, X_{ik}, Y_i)\,, \text{ para } i = 1, 2, \ldots, n$$

Condición,

$$Cov\,(Y_i, Y_j) = 0 \text{ para todo } i \neq j \quad \text{y}$$
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + e_i \text{ para } i = 1, 2, \ldots, n,$$

el modelo aplicado al i-ésimo punto.

Para las n observaciones tenemos:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_k X_{1k} + e_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_k X_{2k} + e_2 \\
&\vdots \quad \vdots \quad \vdots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_k X_{nk} + e_n
\end{aligned}$$

11.1.3. Notación matricial

La notación matricial simplifica todo el trabajo

$$\boldsymbol{Y}_{n\times 1} = X_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{e}_{n\times 1}$$

$$y_i \mid \mathbf{x_i} \sim \mathbf{N}\left(\mathbf{x_i}\beta,\, \sigma^{\mathbf{2}}\right) \text{ ó } \mathbf{y} \mid \mathbf{X} \sim \mathbf{N}\left(\mathbf{X}\beta,\, \sigma^{\mathbf{2}}\mathbf{I}\right)$$

$$\begin{aligned} L\left(\beta, \sigma^{\mathbf{2}} \mid \mathbf{y}\right) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\beta-\hat{\beta}\right)\right), \end{aligned}$$

donde $\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$, el estimador de mínimos cuadrados y utilizamos el hecho que,

$$\begin{aligned} \left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)'\mathbf{X}\left(\beta-\hat{\beta}\right) &= \mathbf{y}'\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\right)\mathbf{X}\left(\beta-\hat{\beta}\right) \\ &= \mathbf{y}'\left(\mathbf{X}-\mathbf{X}\right)\left(\beta-\hat{\beta}\right) = \mathbf{0}. \end{aligned}$$

Por lo tanto, se concluye que $S = \left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)$, $\mathbf{X}'\mathbf{X}$ y $\hat{\beta}$ son estadísticos suficientes para β y σ^2.

11.2. Análisis conjugado

La verosimilitud es de la forma normal-gamma

$$\begin{aligned} L\left(\beta, \sigma^{\mathbf{2}} \mid \mathbf{y}\right) &\propto \left(\sigma^2\right)^{-(n-k-2)/2-1} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right) \\ &\quad \times \left(\sigma^2\right)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\beta-\hat{\beta}\right)\right), \end{aligned}$$

con $\beta \mid \sigma^{\mathbf{2}}$ normal y la distribución marginal de σ^2 es una $Gamma_2$ invertida, denotada por IG_2 con $n-k-2$ grados de libertad. La distribución a priori conjugada también es de la forma normal-gamma.

$$\begin{aligned} \beta \mid \sigma^{\mathbf{2}} &\sim N\left(\beta_{\mathbf{0}}, \sigma^{\mathbf{2}}\mathbf{M_0^{-1}}\right) \\ \sigma^2 &\sim IG_2\left(S_0, v_0\right), \end{aligned}$$

tenemos:

$$\begin{aligned}
\xi\left(\beta,\sigma^{\mathbf{2}}\mid\mathbf{y}\right) &\propto \left(\sigma^2\right)^{-(n-k-2)/2-1}\exp\left(-\frac{S}{2\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\beta-\hat{\beta}\right)\right)\\
&\quad\times\left(\sigma^2\right)^{-v_0/2-1}\exp\left(-\frac{S_0}{2\sigma^2}\right)\left(\sigma^2\right)^{-k/2}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\beta_{\mathbf{0}}\right)'\mathbf{M_0}\left(\beta-\beta_{\mathbf{0}}\right)\right)\\
&= \left(\sigma^2\right)^{-(v_0+n)/2-1}\exp\left(-\frac{S_1}{2\sigma^2}\right)\\
&\quad\times\left(\sigma^2\right)^{-k/2}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\beta_{\mathbf{1}}\right)'\mathbf{M_1}\left(\beta-\beta_{\mathbf{1}}\right)\right),\ \text{donde,}\\
\mathbf{M_1} &= \mathbf{M_0}+\mathbf{X}'\mathbf{X}\\
\beta_{\mathbf{1}} &= \mathbf{M_1^{-1}}\left(\mathbf{M_0}\beta_{\mathbf{0}}+\mathbf{X}'\mathbf{X}\hat{\beta}\right)\\
S_1 &= S_0+S+S_\beta\\
S_\beta &= \left(\beta_{\mathbf{0}}-\hat{\beta}\right)'\left[\mathbf{M_0^{-1}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\left(\beta_0-\hat{\beta}\right)
\end{aligned}$$

El resultado sigue completando el cuadrado para β y reuniendo los otros términos en S_β teniendo en cuenta que:

$$\begin{aligned}
\left(\mathbf{M_0}+\left(\mathbf{X}'\mathbf{X}\right)\right)^{-1} &= \mathbf{M_0^{-1}}-\mathbf{M_0^{-1}}\left[\mathbf{M_0^{-1}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\mathbf{M_0^{-1}}\\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\mathbf{M_0^{-1}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\
\left(\mathbf{X}'\mathbf{X}\right)\left(\mathbf{M_0}+\left(\mathbf{X}'\mathbf{X}\right)\right)^{-1}\mathbf{M_0} &= \left[\mathbf{M_0^{-1}}\left(\mathbf{M_0}+\left(\mathbf{X}'\mathbf{X}\right)\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\\
&= \left[\mathbf{M_0^{-1}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}
\end{aligned}$$

La distribución posterior será entonces,

$$\begin{aligned}
\beta\mid\mathbf{y},\sigma^{\mathbf{2}} &\sim N\left(\beta_{\mathbf{1}},\sigma^{\mathbf{2}}\mathbf{M_1^{-1}}\right)\\
\sigma^2\mid\mathbf{y} &\sim IG_2\left(S_1,v_1\right),
\end{aligned}$$

donde $v_1=v_0+n$.

La distribución marginal posterior para β es una t multivariable[1]. Si integramos para eliminar σ^2 de la distribución conjunta posterior obtenemos el kernel de la marginal posterior como:

$$\xi\left(\beta\mid\mathbf{y}\right)\propto\left[S_1+\left(\beta-\beta_{\mathbf{1}}\right)'\mathbf{M_1}\left(\beta-\beta_{\mathbf{1}}\right)\right]^{-(v1+k)/2}$$

[1]Un vector aleatorio $\boldsymbol{X}$ se dice que tiene una distribución t multivariable con n grados de libertad, vector de localización μ y matriz de precisión $\boldsymbol{T}$, si su densidad es:

$$f\left(\boldsymbol{x}|n,\boldsymbol{\mu},\boldsymbol{T}\right)=c\left(1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{T}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-(n+k)/2},\ \text{donde, } c=\frac{\Gamma\left[(n+k)/2\right]|\boldsymbol{T}|^{1/2}}{\Gamma(n/2)(n\pi)^{(k/2)}},$$

con $2\alpha+n$ grados de libertad, vector de localización $\boldsymbol{\beta}_1$ y matriz de precisión

$$\frac{2\alpha+n}{2\beta_1}\left(\boldsymbol{\tau}+\boldsymbol{X}'\boldsymbol{X}\right)$$

Este es el kernel de la distribución t multidimensional con v_1 grados de libertad y parámetros de escala S_1 y $\mathbf{M_1}$, y denotado por:

$$\beta \mid \mathbf{y} \sim \mathbf{t_k}\left(\beta_\mathbf{1}, \mathbf{S_1}, \mathbf{M_1}, \mathbf{v_1}\right)$$

Resultados similares se obtienen para subconjuntos de parámetros de la regresión. Sea $\beta \sim \mathbf{N}\left(\mathbf{b}, \sigma^\mathbf{2}\mathbf{M^{-1}}\right)$ y $\sigma^2 \sim IG_2\left(S, v\right)$. Asumamos la siguiente partición conformable

$$\begin{aligned} \beta &= \begin{pmatrix} \beta^\mathbf{a} \\ \beta^\mathbf{b} \end{pmatrix} \\ \mathbf{M^{-1}} &= \begin{pmatrix} \mathbf{M^{aa}} & \mathbf{M^{ab}} \\ \mathbf{M^{ba}} & \mathbf{M^{bb}} \end{pmatrix} \end{aligned}$$

Ya que β es normal condicionado en σ^2 tenemos,

$$\begin{aligned} \beta^\mathbf{a} \mid \sigma^\mathbf{2} &\sim N\left(\mathbf{b^a}, \sigma^\mathbf{2}\mathbf{M^{aa}}\right) \\ \beta^\mathbf{a} \mid \beta^\mathbf{b}, \sigma^\mathbf{2} &\sim N\left(\mathbf{b^a} + \mathbf{M^{ab}}\left(\mathbf{M^{bb}}\right)^{-\mathbf{1}}\left(\beta^\mathbf{b} - \mathbf{b^b}\right), \sigma^\mathbf{2}\left(\mathbf{M^{aa}} - \mathbf{M^{ab}}\left(\mathbf{M^{bb}}\right)^{-\mathbf{1}}\mathbf{M^{ba}}\right)\right). \end{aligned}$$

Marginalizando con respecto a σ^2 tenemos,

$$\begin{aligned} \beta^\mathbf{a} &\sim t_{k_a}\left(\mathbf{b^a}, \mathbf{S}, \left(\mathbf{M^{aa}}\right)^{-\mathbf{1}}, \mathbf{v}\right) \\ \beta^\mathbf{a} \mid \beta^\mathbf{b} &\sim t_{k_a}\left(\mathbf{b^a} + \mathbf{M^{ab}}\left(\mathbf{M^{bb}}\right)^{-\mathbf{1}}\left(\beta^\mathbf{b} - \mathbf{b^b}\right), \mathbf{S}, \left(\mathbf{M^{aa}} - \mathbf{M^{ab}}\left(\mathbf{M^{bb}}\right)^{-\mathbf{1}}\mathbf{M^{ba}}\right)^{-\mathbf{1}}, \mathbf{v}\right). \end{aligned}$$

11.2.1. Distribución predictiva

Recordemos que el modelo de interés es $\mathbf{y} = \mathbf{X}\beta + \epsilon$, con $\epsilon \sim \mathbf{N}\left(\mathbf{0}, \sigma^\mathbf{2}\mathbf{I}\right)$.Ya que $\beta \mid \sigma^\mathbf{2} \sim \mathbf{N}\left(\beta_\mathbf{0}, \sigma^\mathbf{2}\mathbf{M_0^{-1}}\right)$, entonces $\mathbf{X}\beta \mid \sigma^\mathbf{2} \sim \mathbf{N}(\mathbf{X}\beta_\mathbf{0}, \sigma^\mathbf{2}\mathbf{X}\mathbf{M_0^{-1}}\mathbf{X}')$. Se sigue que:

$$\mathbf{y} \mid \sigma^\mathbf{2} \sim \mathbf{N}\left(\mathbf{X}\beta_\mathbf{0}, \sigma^\mathbf{2}\left(\mathbf{I} + \mathbf{X}\mathbf{M_0^{-1}}\mathbf{X}'\right)\right),$$

ya que ϵ es independiente de β cuando condicionamos en σ^2. La a priori para σ^2 es $IG_2\left(S_0, v_0\right)$ y marginalizando con respecto a σ^2 produce,

$$\mathbf{y} \sim \mathbf{t_n}\left(\mathbf{X}\beta_\mathbf{0}, \mathbf{S_0}, \left(\mathbf{I} + \mathbf{X}\mathbf{M_0^{-1}}\mathbf{X}'\right)^{-\mathbf{1}}, \mathbf{v_0}\right)$$

La densidad predictiva para un vector $\mathbf{y}^*$ de m componentes condicionado a un conjunto de valores para las variables explicatorias $\mathbf{X}^*$ es:

$$\mathbf{y}^* \mid \mathbf{X}^* \sim \mathbf{t_m}\left(\mathbf{X}^*\beta_\mathbf{1}, \mathbf{S_1}, \left(\mathbf{I} + \mathbf{X}^*\mathbf{M_1^{-1}}\mathbf{X}^{*\prime}\right)^{-\mathbf{1}}, \mathbf{v_1}\right)$$

g-a priori de Zellner

La aproximación g-a priori de Zellner no requiere especificar el conocimiento previo acerca de los parámetros del modelo. La aproximación g-a priori asume que las covarianzas a priori para β son iguales a las obtenidas mediante los datos muestrales. O sea, la matriz de precisión a priori está dada por:

$$\frac{g}{\sigma^2}\mathbf{X}'\mathbf{X}$$

La media a posteriori será:

$$\beta^{(1)} = \frac{\hat{\beta} + g\beta^{(0)}}{1+g},$$

donde $\hat{\beta}$ es el estimador de máxima verosimilitud y $\beta^{(0)}$ es la media a priori. El tamaño de g corresponde al peso que se le asigna a la a priori.

11.2.2. Elicitación en el modelo lineal

Propuesta de Garthwait y Dickey

Garthwaite y Dickey [171] presentan un procedimiento para elicitar los parámetros en el modelo lineal normal. Para esto asumen la elicitación los parámetros de la a priori conjugada. Ellos asumen que σ^2 se distribuye proporcional a una inversa χ^2 con n grados de libertad multiplicada por wn. Además, $\boldsymbol{\beta}$, condicionada en σ^2, se distribuye $N\left(\boldsymbol{b}, \frac{\sigma^2}{w}\boldsymbol{U}\right)$. Los parámetros de la a priori w, n, $\boldsymbol{b}$ y $\boldsymbol{U}$ son los hiperparámetros a ser elicitados. Los autores trabajan dentro del contexto de la experimentación química y se le pide al experto un «diseño»(valores de las variables independientes) donde se observará la respuesta. Ellos presumen que este diseño tiene la menor varianza predictiva y lo llaman CMV (Constrained minimum variance). Garthwaite y Dickey [172] presentan un proceso de elicitación para el caso de la selección de las variables a ser incluidas en el modelo.

Si $\boldsymbol{y} = (y_1, \ldots, y_m)'$ se distribuye $N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 w\boldsymbol{I})$, donde $\boldsymbol{X}_{m\times r}$ es la matriz de diseño e $\boldsymbol{I}_{m\times m}$ es una matriz identidad. Sea $\boldsymbol{t}_n$ que denota la distribución t multivariable estándar con n grados de libertad[2]. Si se define $\boldsymbol{z} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{t}_n$, donde $\boldsymbol{a}$ y $\boldsymbol{B}$ son un vector y una matriz de constantes conocidas, se habla de una t generalizada con n grados de libertad y centro $C(\boldsymbol{z}) = \boldsymbol{a}$ y dispersión $S(\boldsymbol{z}) = \boldsymbol{B}\boldsymbol{B}'$. La matriz de varianzas y covarianzas de $\boldsymbol{z}$ es $[n/(n-2)S(\boldsymbol{z})]$. Las distribuciones de $\boldsymbol{y}|\boldsymbol{X}$ y de $\boldsymbol{\beta}$ son t's multivariables con centros y dispersiones dadas por $C(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{b}$, $S(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{U}\boldsymbol{X} + w^2\boldsymbol{I}$, $C(\boldsymbol{\beta}) = \boldsymbol{b}$ y $S(\boldsymbol{\beta}) = \boldsymbol{U}$.

El procedimiento propuesto requiere que al experto se le pidan los resultados de dos experimentos imaginarios bajo los mismos puntos de diseño. Si Z_1 representa la diferencia entre los dos resultados, esta diferencia sigue una distribución t con n

[2]La distribución t multivariable con n grados de libertad surge como el cociente de una distribución normal multivariable estándar y la raíz cuadrada de una distribución chi cuadrado con n grados de libertad multiplicada por los grados de libertad.

grados de libertad, centro 0 y dispersión $2w$. Al experto se le pregunta acerca de la diferencia Z_1 entre los dos experimentos. Se le pregunta además sobre la mediana de la diferencia absoluta $|Z_i|$ y se denota por k_1, este es el rango semi-intercuartílico de Z_1. Al experto se le pide que en los mismos puntos de diseño imagine otros dos experimentos. Se elicta el valor k_2 (equivalente a k_1). Se usan estos valores (k_1 y k_2) para determinar w y n. Ahora, $Z_1 \sim t_n(0, 2w)$ y $Z_2|Z_1 = z_1 \sim t_{n+1}\left\{0, \left(z_1^2 + 2nw\right) / (n+1)\right\}$.

Si q_n denota el rango semi-intercuartílico de una distribución t con dispersión uno y con n grados de libertad, w y n pueden determinarse de las siguientes ecuaciones:

$$\begin{aligned} k_1 &= \sqrt{2w} q_n \\ k_2 &= \sqrt{\frac{\left(z_1^2 + 2nw\right)}{(n+1)}} q_{n+1} \end{aligned}$$

Otra propuesta

Si se define la matriz de diseño en los puntos que garanticen cubrir la región de interés, digamos $\boldsymbol{X}_{N\times p}$. Podemos, bajo el modelo homoscedástico, pedirle al experto que determine los valores de la respuesta plausibles para cada punto definido en la matriz, entregando el valor medio y valores máximos y mínimos de la respuesta. Con estos valores, en cada punto de la matriz de diseño se estiman los parámetros de la distribución normal de la respuesta. Para cada punto generamos una muestra al azar de n_0 respuestas. Así obtenemos $n_0 N$ puntos. Con estos puntos estimamos el modelo. Aplicando los resultados corrientes obtenemos la distribución de $\boldsymbol{\beta}$ y σ^2.

11.2.3. Inferencias

Intervalos de probabilidad

Regiones de alta probabilidad para conjuntos de parámetros se encuentran directamente de la distribución marginal posterior. Para un solo parámetro tenemos:

$$\beta_i \mid \mathbf{y} \sim \mathbf{t}\left(\beta_{\mathbf{i}}^{\mathbf{1}}, \mathbf{S_1}, \left(\mathbf{M_1^{ii}}\right)^{-\mathbf{1}}, \mathbf{v_1}\right),$$

donde M_1^{ii} es el elemento i, i de $\mathbf{M_1^{-1}}$. La transformación,

$$\frac{\beta_i - \beta_i^1}{\sqrt{M_1^{ii} S_1 / v_1}},$$

tiene una distribución t estándar y una región de más alta probabilidad $1 - \alpha$ está dada por:

$$\left(\beta_i^1 - t_{\alpha/2, v_1}\sqrt{M_1^{ii} S_1 / v_1},\ \beta_i^1 + t_{\alpha/2, v_1}\sqrt{M_1^{ii} S_1 / v_1}\right)$$

Para conjuntos de parámetros notemos que si $\mathbf{x} \sim \mathbf{t_m}\left(\mu, \mathbf{S}, \boldsymbol{M}, \mathbf{v}\right)$ entonces,

$$\frac{(\mathbf{x} - \mu)' \mathbf{M} (\mathbf{x} - \mu) / \mathbf{m}}{S/v} \sim F_{(m,n)}$$

Una región de más alta probabilidad para β está dada por:

$$\left\{\beta:\ \frac{(\beta-\beta_{\mathbf{1}})'\,\mathbf{M_1}\,(\beta-\beta_{\mathbf{1}})\,/\mathbf{k}}{\mathbf{S_1/v_1}} \leq \mathbf{F_{(1-\alpha,k,v_1)}}\right\}$$

11.2.4. Pruebas de hipótesis

Las pruebas de hipótesis puntuales son fáciles de implementar utilizando la técnica de la región de más alta probabilidad y verificando que la hipótesis está contenida en una región apropiada de más alta probabilidad.

Si la hipótesis tiene la forma de q restricciones lineales $\mathbf{B}\beta = \mathbf{r}$, tenemos que la distribución posterior bajo la hipótesis es:

$$\mathbf{R}\beta \sim \mathbf{t_q}\left(\mathbf{R}\beta_{\mathbf{1}}, \mathbf{S_1}, \left(\mathbf{RM_1^{-1}R}\right)^{\mathbf{-1}}, \mathbf{v_1}\right),$$

y por lo tanto,

$$\frac{(\delta - \mathbf{R}\beta_{\mathbf{1}} + \mathbf{r})'\left(\mathbf{RM_1^{-1}R}\right)^{-1}(\delta - \mathbf{R}\beta_{\mathbf{1}} + \mathbf{r})\,/q}{S_1/v_1} \sim F_{(q,v_1)},$$

para $\delta = \mathbf{R}\beta_{\mathbf{1}} - \mathbf{r}$. La hipótesis $\delta = \mathbf{0}$ está contenida en la región de más alta probabilidad si

$$\frac{(\mathbf{R}\beta_{\mathbf{1}} - \mathbf{r})'\left(\mathbf{RM_1^{-1}R}\right)^{-1}(\mathbf{R}\beta_{\mathbf{1}} - \mathbf{r})\,/q}{S_1/v_1} < F_{(q,v_1)}$$

Para el cálculo de las pruebas bayesianas y los factores de Bayes supongamos que deseamos probar $H_1 : \mathbf{R}\beta = \mathbf{r}$ y $H_2 : \mathbf{R}\beta \neq \mathbf{r}$. H_1 implica exactamente q restricciones sobre los parámetros que pueden ser sustituidos en el modelo, lo cual produce

$$\mathbf{y}^* = \mathbf{X}^*\beta^* + \epsilon,$$

donde β es un vector con $k - q$ componentes. Especificando una a priori para β^* y σ^2 bajo H_1, digamos $\beta^* \mid \sigma^{\mathbf{2}} \sim \mathbf{N}\left(\beta_{\mathbf{0}}^*, \sigma^{\mathbf{2}}\mathbf{M_0^{*-1}}\right)$, $\sigma^2 \sim IG_1\left(S_0, v_0\right)$ obtenemos la verosimilitud marginal bajo H_1 como:

$$m\left(\mathbf{y}^* \mid \mathbf{H_1}\right) = t_n\left(\mathbf{X}^*\beta_{\mathbf{0}}^*, \mathbf{S_0}, \left(\mathbf{I} + \mathbf{X}^*\mathbf{M_0^{*-1}}\mathbf{X}^{*\prime}\right)^{\mathbf{-1}}, \mathbf{v_0}\right)$$

Bajo H_2, especificamos una a priori $\beta \mid \sigma^{\mathbf{2}} \sim \mathbf{N}\left(\beta_{\mathbf{0}}, \sigma^{\mathbf{2}}\mathbf{M_0^{-1}}\right)$, $\sigma^2 \sim IG_1\left(S_0, v_0\right)$ y el análisis es igual al anterior. El factor de Bayes será entonces,

$$B_{12} = \frac{t_n\left(\mathbf{X}^*\beta_{\mathbf{0}}^*, \mathbf{S_0}, \left(\mathbf{I} + \mathbf{X}^*\mathbf{M_0^{*-1}}\mathbf{X}^{*\prime}\right)^{\mathbf{-1}}, \mathbf{v_0}\right)}{t_n\left(\mathbf{X}\beta_{\mathbf{0}}, \mathbf{S_0}, \left(\mathbf{I} + \mathbf{XM_0^{-1}X'}\right)^{\mathbf{-1}}, \mathbf{v_0}\right)}$$

Ejemplo 11.1. Precios de Oferta de Vehículos. Consideremos los datos referentes a los precios de oferta de carros Chevrolet Sprint aparecidos en el periódico *El Colombiano*, el 14 de abril de 2002 en la sección de Avisos Clasificados.

Tabla 11.1: *precio de oferta en millones de pesos colombianos de vehículos Chevrolet Sprint según el año*

Año	Precio (en millones)
87	7.0
88	8.0
92	10.4
94	12.5

Si asumimos que el modelo $Precio = \beta_0 + \beta_1 Ao$ nos puede representar de una manera adecuada la relación entre el precio de oferta del vehículo y el año del mismo. Además asumimos que una observación particular tiene una diferencia con el modelo teórico que se distribuye normal con media cero y varianza σ^2. La pendiente β_1 nos indica la diferencia promedio en el precio de dos carros Sprint de años consecutivos. Podemos entonces utilizar un programa estadístico que ajuste el modelo (aún hasta calculadoras de bolsillo ajustan este tipo de modelos). Los resultados son:

$$\begin{array}{rcl} \text{Precio Estimado} = 9.475 & + & 0.74275\text{Año} \\ \text{Error Estándar } 0.17326 & & 0.06055 \end{array}$$

Desviación Típica del Modelo: 0.3465 con 2 grados de libertad; R-Cuadrado: 0.9869.

Obviamente el modelo ajusta bien, pero es claro que tenemos muy pocos datos. El intervalo de confianza del 95 % para la pendiente es (0.4822244, 1.003276), que es bastante amplio. Un problema con esta aproximación es la interpretación frecuentista que hay que darle al intervalo y que se basa en el supuesto de la extracción de infinitas muestras de tamaño 4 de la misma población.

Ahora bien, si asumimos los precios de oferta del mismo tipo de carro que aparecieron en *El Colombiano*, el 16 de diciembre de 2001, en el cual aparecieron los siguientes datos:

Tabla 11.2: *precio de oferta en millones de pesos colombianos de vehículos Chevrolet Sprint según el año*

Año	Precio (en millones)
88	7.8
90	8.8
95	11.8
95	12.3
94	12.0
95	8.8

Si asumimos que β_0 se distribuye normalmente con media 10.86 y precisión de 28.08382 y β_1 se distribuye normalmente con media 0.6522 y precisión 225.2477. Para la varianza del modelo asumimos un modelo poco informativo Gamma(0.001,0.001). Los valores anteriores se construyeron asumiendo inicialmente distribuciones poco informativas y actualizándolos con la información previa, excepto el de la varianza, ya que este nos refleja el nivel de credibilidad en las predicciones de esta actualización, que puede no ser muy alto. Dadas esta nuevas condiciones para nuestro problema, o sea información previa disponible y cuantificada en términos de distribuciones, procedemos a mezclarla, utilizando el teorema de Bayes, para obtener nuestra distribución actualizada o a posteriori. Esta última produce los resultados siguientes:

Tabla 11.3: *resumen estadístico de las a posteriori de los parámetros del modelo*

Parámetro	media	sd	2.5 %	97.5 %
β_0	10.76	0.1925	10.38	11.14
β_1	0.6581	0.06308	0.5339	0.7838
τ	0.5937	0.4655	0.06085	1.825

El intervalo de credibilidad (en la estadística clásica lo llamamos de confianza) para la pendiente del 95 % de probabilidad es (0.5339 , 0.7838), el cual nos dice que el más probable valor para la diferencia promedio en el precio de oferta de dos carros Sprint de años consecutivos está entre \$534.000.00 y \$784.000.00. Este intervalo es mucho más preciso que el intervalo hallado por el método clásico que era \$482.200.00 y \$1.003.000.00.

Ejemplo 11.2. Propiedad raíz en Medellín. El mercado de propiedad raíz es uno de los más importantes y refleja la situación económica de una región. En este caso vamos a considerar el mercado de apartamentos usados en el sector de El Poblado de Medellín-Colombia. Seleccionamos este sector básicamente por las siguientes razones:

1. Es un sector de la ciudad de Medellín con una gran dinámica en el mercado del usado.

2. El nivel de estratificación socioeconómica es muy homogénea.

3. Los apartamentos son relativamente nuevos, en el sentido que la antigüedad de la mayoría no supera los veinte años.

Uno puede considerar muchos factores que expliquen el precio de oferta de un apartamento usado, por ejemplo:

- Antigüedad del inmueble
- Metros cuadrados construidos

- Calidad de la construcción
- Otros

Nosotros consideramos la información disponible para construir un modelo que explique el precio de oferta y básicamente se limita a los metros cuadrados construidos del apartamento.

Tabla 11.4: *precio de 30 apartamentos en millones de pesos colombianos, por metro cuadrado. Fuente: El Colombiano, Avisos Clasificados, Sept. 22 del 2002*

Apto. No.	Metros^2	Precio (en millones)
1	113.00	92.00
2	140.00	130.00
3	140.00	125.00
4	110.00	90.00
5	69.00	65.00
6	152.00	130.00
7	105.00	110.00
8	144.00	120.00
9	103.00	89.00
10	107.00	145.00
11	112.00	85.00
12	103.00	89.00
13	120.00	105.00
14	86.00	75.00
15	143.00	112.00
16	115.00	112.00
17	136.50	125.00
18	168.50	145.00
19	217.00	205.00
20	132.80	115.00
21	120.00	105.00
22	108.00	89.00
23	220.00	150.00
24	110.00	89.00
25	228.00	108.00
26	83.00	66.00
27	78.00	64.00
28	150.00	135.00
29	135.00	125.00
30	90.00	65.00

El código en OpenBUGS es:

```
model { for( i in 1 : N ){ Precio[i ] ~ dnorm(mu[i],tau)
mu[i] <-alpha + beta * (metros[i] - mean(metros[]))
        }

tau ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau)
alpha ~ dnorm(0.0,1.0E-6)
error ~ dnorm(0,tau)
beta ~ dnorm(0.0,1.0E-6)
Precio175<-alpha+beta*(175-mean(metros[]))
Precio175indi<-Precio175+error

for(i in 1:N){
PrecioIndi[i]<-alpha+beta*(metros[i]-mean(metros[]))+error } }

list(N=25, Precio=c(92,130,125,90,65,
                    130,110,120,89,145,
                    85,89,105,75,112,
                    112,125, 145,205,115,
                    105,89,150,89,108,
                    66,64,135,125,65),
           metros=c(113,140,140,110,69,
                    152,105,144,103,107,
                    112,103,120,86,143,
                    115,136.5,168.5,217,132.8,
                    120,108,220,110,228,
                    83,78,150,135,90))

list(tau=1,beta=0,alpha=0,error=0)
```

Los resultados del anterior programa se muestran en la Tabla 11.5.

Tabla 11.5: *resumen estadístico del modelo ajustado*

	Nodo	Media	sd	2.50 %	median	97.5 %
1	alpha	108.70	3.79	101.20	108.70	116.20
2	beta	0.61	0.10	0.41	0.61	0.80
3	Precio175	137.20	5.90	125.40	137.20	148.80
4	Precio175indi	137.20	21.38	94.90	137.20	179.60

Una de las utilidades de OpenBUGS, es que podemos usarlo desde R e implementar toda la herramienta gráfica de R a las salidas del modelo. Para nuestro ejemplo, el código en R es el siguiente:

```
modelo<- function(){ for( i in 1 : N ){ Precio[i ] ~ dnorm(mu[i],tau)
mu[i] <-alpha + beta * (metros[i] - mean(metros[]))
```

```
        }

tau ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau)
alpha ~ dnorm(0.0,1.0E-6)
error ~ dnorm(0,tau)
beta ~ dnorm(0.0,1.0E-6)
Precio175<-alpha+beta*(175-mean(metros[]))
Precio175indi<-Precio175+error

for(i in 1:N){
PrecioIndi[i]<-alpha+beta*(metros[i]-mean(metros[]))+error } }

library(R2OpenBUGS)
# escribimos el modelo en una ubicación temporal
modelo.archivo <- file.path(tempdir(),"modelo.txt")
write.model(model, modelo.archivo)

datos <- list(N=25, Precio=c(92,130,125,90,65,
                  130,110,120,89,145,
                  85,89,105,75,112,
                  112,125, 145,205,115,
                  105,89,150,89,108,
                  66,64,135,125,65),
          metros=c(113,140,140,110,69,
                  152,105,144,103,107,
                  112,103,120,86,143,
                  115,136.5,168.5,217,132.8,
                  120,108,220,110,228,
                  83,78,150,135,90))

param <- c("alpha","beta","Precio175","Precio175indi")

iniciales <- function(){ list(tau=1,beta=0,alpha=0,error=0)}

resul <- bugs(datos, iniciales, param, modelo.archivo,n.chains = 1,
n.iter=50000, n.burnin=20000)

attach.all(resul$sims.list)
par(mfrow=c(2,2))
plot(density(alpha),main="Densidad alpha")
plot(density(beta),main="Densidad beta")
plot(density(Precio175),main="Densidad Precio175")
plot(density(Precio175indi),main="Densidad Precio175indi")

print(resul, digits=2)

resul2 <- bugs(datos, iniciales, param, modelo.archivo, codaPkg=TRUE,
n.iter=50000, n.burnin=20000)
```

```
resul.coda <- read.bugs(resul2)
library(coda)
acfplot(resul.coda)
# gráfico de Gelman-Rubin-Brooks para chequear convergencia
# con el factor de convergencia shrink
gelman.plot(resul.coda)
```

Los resultados se muestran en la Tabla 11.6.

Tabla 11.6: *resumen estadístico del modelo ajustado usando el método clásico*

	Nodo	Media	sd	2.50 %	median	97.5 %
1	alpha	109.98	4.30	101.50	110.00	118.50
2	beta	0.54	0.11	0.32	0.54	0.75
3	Precio175	135.34	6.33	122.80	135.40	147.80
4	Precio175indi	135.37	22.30	91.19	135.40	179.50

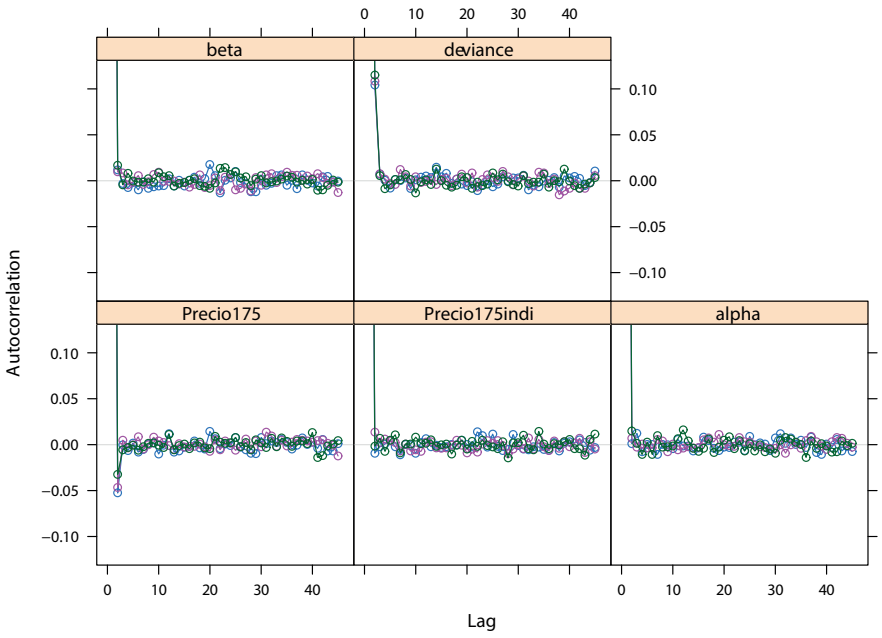

Figura 11.1: *chequeo de convergencia con acf*

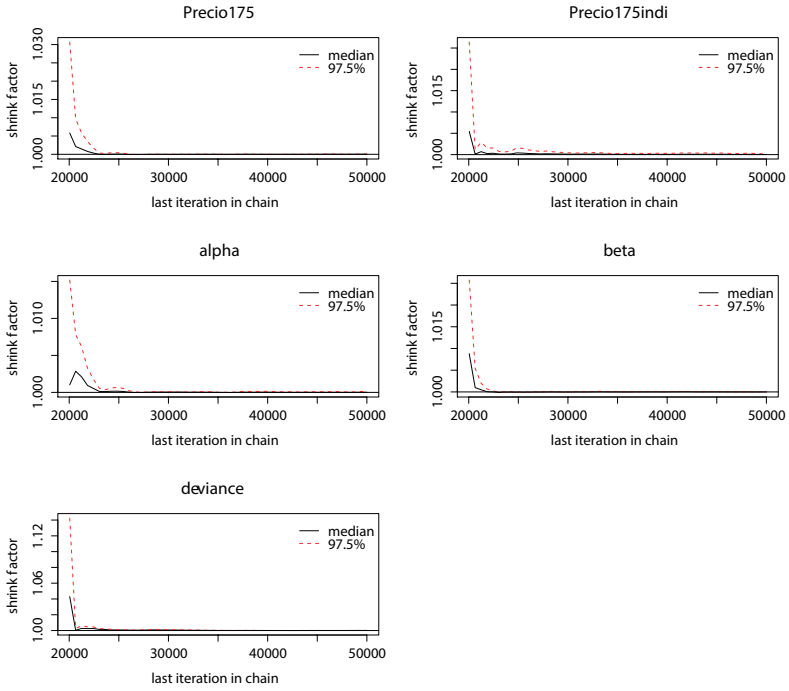

Figura 11.2: *chequeo de convergencia con factor shrink*

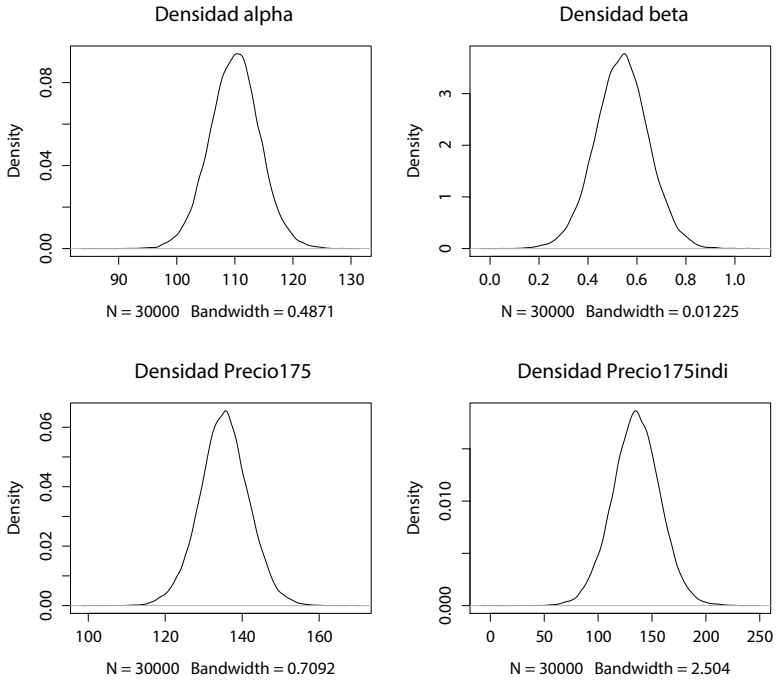

Figura 11.3: *densidades*

Ajustamos en R el modelo $Precio = \alpha + \beta\,(Metros^2 - Media(Metros^2))$ por el método clásico (esto es, no bayesiano) y obtuvimos:

	Parámetro	Estimación	error	Prueba t	Valor-p
1	$\hat{\alpha}$	108.66667	3.61327	30.074	$< 2e-16$
2	$\hat{\beta}$	0.60893	0.09303	6.546	$4.27e-07$

Tabla 11.7: *resumen estadístico de las estimaciones por apartamento*

Obs.	Media	sd	EMC	2.50 %	Mediana	97.50 %
1	99.45	20.96	0.25	58.01	99.35	141.00
2	115.90	20.96	0.25	74.58	115.80	156.90
3	115.90	20.96	0.25	74.58	115.80	156.90
4	97.63	20.98	0.25	56.52	97.54	139.00
5	72.72	21.64	0.26	30.34	72.53	114.60
6	123.10	21.06	0.26	81.64	123.10	164.70
7	94.59	21.02	0.25	53.26	94.54	136.00
8	118.30	20.99	0.25	76.88	118.20	159.40
9	93.38	21.04	0.25	52.12	93.30	134.70
10	95.81	21.00	0.25	54.43	95.74	137.20
11	98.84	20.97	0.25	57.52	98.76	140.30
12	93.38	21.04	0.25	52.12	93.30	134.70
13	103.70	20.93	0.25	62.52	103.60	144.90
14	83.05	21.28	0.25	41.36	82.92	124.90
15	117.70	20.98	0.25	76.35	117.60	158.70
16	100.70	20.95	0.25	59.33	100.50	142.10
17	113.70	20.94	0.25	72.73	113.60	154.80
18	133.20	21.31	0.26	90.95	133.30	174.80
19	162.60	22.68	0.29	117.50	162.50	208.50
20	111.50	20.93	0.25	70.56	111.30	152.70
21	103.70	20.93	0.25	62.52	103.60	144.90
22	96.41	20.99	0.25	55.03	96.37	137.80
23	164.50	22.79	0.29	119.20	164.40	210.60
24	97.63	20.98	0.25	56.52	97.54	139.00
25	169.30	23.11	0.30	123.40	169.20	216.30
26	81.23	21.33	0.25	39.49	81.10	123.00
27	78.19	21.44	0.25	36.10	78.07	119.70
28	121.90	21.04	0.26	80.46	121.90	163.30
29	112.80	20.94	0.25	71.76	112.70	154.00
30	85.48	21.21	0.25	43.99	85.38	127.10

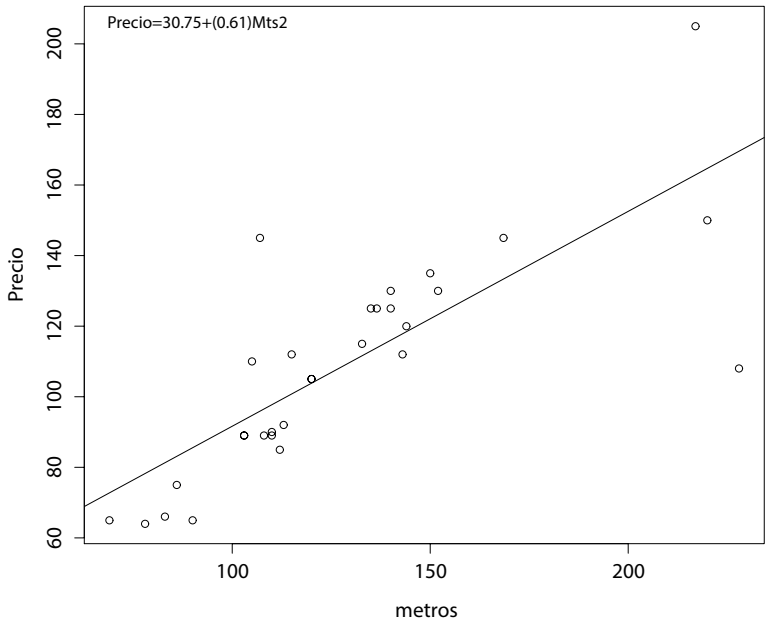

Figura 11.4: *metros cuadrados de un apartamento versus el Precio de oferta (en millones) en el sector de El Poblado. El modelo estimado es* $\widehat{Precio} = 30.7475 + 0.6089 Metros^2$*, el cual indica que el metro cuadrado para los usados en este sector es aproximadamente de $600.000*

El modelo clásico nos da,

$$\widehat{Precio} = 30.7475 + 0.6089 Metros^2$$

$$\boldsymbol{X'X} = \begin{bmatrix} 30.0 & 3838.8 \\ 3838.8 & 536471.3 \end{bmatrix}$$

11.3. Estrategias en modelación

Jeffreys y Berger [173] discuten el principio que es ahora popular entre los modeladores conocido como *la cuchilla de Ockham*, y que dice *Pluritas non est ponenda sine necessitate*, que traduce «La pluralidad no se debe imponer sin necesidad». Aunque el principio es relativamente vago, varias interpretaciones se le han dado tales como:

- «Las entidades no deben ser multiplicadas sin necesidad»
- «Es vano hacer con más lo que se puede hacer con menos»
- «Una explicación de los hechos no debe ser más complicada de lo necesario»
- «Entre hipótesis que compiten, favorezca la más simple»

Este ha sido un principio heurístico, pero ellos argumentan que puede ser justificado y aceptado bajo la escuela bayesiana. Loredo [174] habla de la Cuchilla de Occam Automatizada. Para probabilidades predictivas se prefieren modelos simples.

El Factor de Occam:

$$\begin{aligned} P(D|M_i) &= \int \xi\left(\theta_i|M\right) L(\theta_i)\, d\theta_i \\ &\approx xi\left(\hat{\theta}_i|M\right) L(\hat{\theta}_i)\, \delta\theta_i \\ &\approx L(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\ &\approx \text{Máxima Verosimilitud} \times \text{Factor de Occam} \end{aligned}$$

Los modelos con más parámetros usualmente hacen que los datos produzcan un mejor ajuste. El Factor de Occam penaliza los modelos por el «volumen»del espacio parametral desperdiciado.

No existen reglas rígidas que se deban seguir en el proceso de modelación, más bien lo que se presenta a continuación nace de la experiencia:

- Comenzar con modelos pequeños y simples que han sido utilizados por otras personas y para los cuales los análisis han sido realizados
- Desarrollar modelos más complejos paso a paso
- Cuando realice simulaciones chequee las respuestas finales comenzado desde diferentes puntos iniciales y diferentes semillas para los generadores de números aleatorios

11.4. Librería MCMCpack

Esta librería del R contiene un conjunto de funciones que permiten ajustar una amplia variedad de modelos bayesianos. Un problema es que solo permite ajustar modelos conjugados, lo que en algunas aplicaciones puede ser restrictivo.

`MCMCregress()`

Esta función genera muestras de la distribución posterior del modelo lineal con errores normales usando el muestreador de Gibbs, usando una distribución a priori normal multivariable del vector β, y una Gamma inversa para la varianza condicional.

```
MCMCregress(formula, data = parent.frame(), burnin = 1000,
          mcmc = 10000,thin = 1, verbose = 0, seed = NA,
          beta.start = NA, b0 = 0, B0 = 0, c0 = 0.001,
          d0 = 0.001,marginal.likelihood = c("none",
          "Laplace", "Chib95"), ...)
```

Ejemplo 11.3. Modelación del Precio del Twingo. Para ilustrar el uso de la función `MCMCregress()` vamos a modelar el precio de oferta de carros Renault Twingo considerando el año del vehículo. Estos datos aparecieron en la sección de Avisos Clasificados del El Colombiano, 30 de marzo de 2008.

La lectura de datos es:

```
# Ajuste del modelo para el precio de Twingo # Año (101=2001) y
# Precio (en millones)
datos<-scan()
101 15.2  103 16.9  106 21.4  96 12.3  96 13.0  105 19.9
107 24.5 101 16.5  105 18.9  106 20.5  105 18.7 106 19.0
101 13.8 105 19.0 105 20.0  106 21.5  102 15.5  102 17.5
99 11.0  97 12.5  107 22.5 106 21.5

datos<-matrix(datos,ncol=2,byrow=T)

Precio<-datos[,2]
Año<-datos[,1]
plot(Año,Precio)
```

Ajustamos el modelo clásico usando la función `lm()`:

```
res.lm<-lm(Precio~Año)
abline(res.lm)
lines(smooth.spline(Año,Precio),col='red')
title(main='Precio de Oferta de Twingos vs. Año',
      sub='El Colombiano, Marzo 30 del 2008')

summary(res.lm)

Call: lm(formula = Precio ~ Año)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8500 -0.8438  0.2416  0.7916  2.8387

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -82.81515    8.82414  -9.385  9.1e-09 ***
Año           0.97642    0.08559  11.409  3.3e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.383 on 20 degrees of freedom Multiple
R-squared: 0.8668,     Adjusted R-squared: 0.8601 F-statistic: 130.2
on 1 and 20 DF,  p-value: 3.303e-10
```

Usando la función `MCMCregress()` y bajo el supuesto de a priori no informativas obtenemos:

```
require(MCMCpack)
Loading required package: MCMCpack
Loading required package: coda
Loading required package: MASS
```

```
## Markov Chain Monte Carlo Package (MCMCpack)
## Copyright (C) 2003-2016 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
##
## Support provided by the U.S. National Science Foundation
## (Grants SES-0350646 and SES-0350613)
##
res.bay<-MCMCregress(Precio~Año)
summary(res.bay)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

               Mean      SD  Naive SE Time-series SE
(Intercept) -82.6888 9.35477 0.0935477      0.0917923
Año           0.9752 0.09075 0.0009075      0.0008893
sigma2        2.1353 0.76615 0.0076615      0.0083963

2. Quantiles for each variable:

                2.5%      25%      50%     75%   97.5%
(Intercept) -101.0450 -88.6263 -82.6179 -76.743 -64.137
Año            0.7943   0.9178   0.9744   1.033   1.154
sigma2         1.1229   1.6050   1.9824   2.482   4.019

res.bay[1:15,]
      (Intercept)       Año   sigma2
 [1,]  -100.75066 1.1479148 1.862622
 [2,]   -84.14091 0.9853853 1.790405
 [3,]   -93.42793 1.0783263 2.214343
 [4,]   -87.98343 1.0297587 1.857563
 [5,]   -60.41204 0.7586823 3.487881
 [6,]   -88.45529 1.0324655 2.841447
 [7,]   -70.67004 0.8607582 1.724942
 [8,]   -78.15466 0.9292954 1.773086
 [9,]   -83.50116 0.9834269 1.864375
[10,]   -97.87236 1.1166745 2.987112
[11,]   -87.44617 1.0212019 1.897684
[12,]   -70.76430 0.8667845 3.726868
[13,]   -89.75502 1.0418003 2.435048
[14,]   -81.57685 0.9694492 2.832919
[15,]   -89.03969 1.0361920 2.303039
```

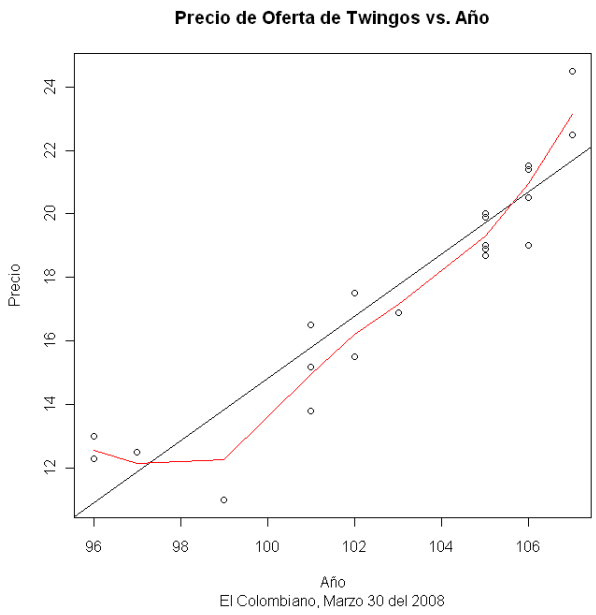

Figura 11.5: *diagrama de dispersión y modelo ajustado del precio de oferta del vehículo según el año*

Ejemplo 11.4. Carros Sprint: incorporando información previa. Estamos interesados en modelar el precio de oferta del Sprint. En El Colombiano, el 10 de octubre de 2010, apareció la siguiente información:

```
# Precio Carros Sprint # Año Precio (en millones) # Oct. 10 2010

datos<-scan()
2003 11 1991 6.2 1991 5.7 1992 7.5 1995 8.3 1996 6.5
1994 8.3 1993 7.6 1993 7.6

datos<-matrix(datos,ncol=2,byrow=T)
año1<-datos[,1]
precio1<-datos[,2]
plot(año1,precio1)
```

Si tenemos información previa (y si asumimos que el proceso se ha mantenido estable) podemos construir la a priori a partir de ella.

```
# Precio Carros Sprint # Año Precio (en millones) # junio 21 2009

datos<-scan()
1988 6.0 1993 6.8 1996 10.0 1996 9.8 1999 10.2 1987
6.0 1993 8.0 1994 7.5 1994 8.8

datos<-matrix(datos,ncol=2,byrow=T)
año2<-datos[,1]
precio2<-datos[,2]
# actualiza valores de precios-> precios corrientes
```

```
# IPC mensual desde julio 2009 hasta sept 2010 (DANE)
IPC<-c(-0.04,0.04,-0.11,-0.13,0.07,0.08,
2.0,0.69,0.83,0.25,0.46,0.10,0.11,-0.04, 0.11,-0.14)

IPC.acum<-sum(IPC); precio2<-precio2*(1+IPC.acum/100)

require(MCMCpack)
res.bay<-MCMCregress(precio2~año2)
summary(res.bay)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
                Mean        SD  Naive SE Time-series SE
(Intercept) -823.2670 164.00857 1.6400857       1.590848
año2           0.4173   0.08228 0.0008228       0.000798
sigma2         0.7868   0.60859 0.0060859       0.007932

2. Quantiles for each variable:
                  2.5%       25%       50%       75%     97.5%
(Intercept) -1145.1565 -922.3210 -823.3099 -725.8357 -490.0331
año2            0.2501    0.3684    0.4173    0.4670    0.5786
sigma2          0.2480    0.4327    0.6184    0.9237    2.3171

res.bay[1:10,]
      (Intercept)      año2    sigma2
 [1,]   -848.2590 0.4297464 0.4481977
 [2,]   -977.8810 0.4947555 0.2565116
 [3,]   -859.3918 0.4352304 0.7775673
 [4,]   -911.6296 0.4616164 0.3732648
 [5,]  -1184.8577 0.5983969 0.9113289
 [6,]   -703.8946 0.3573862 2.4943903
 [7,]   -873.8297 0.4425935 0.2302472
 [8,]   -852.9071 0.4321825 0.8055657
 [9,]   -640.3850 0.3255312 0.5367643
[10,]   -750.3874 0.3807708 0.5411309

library(MASS)
fitdistr(1/res.bay[,3],'gamma')
     shape         rate
  3.51567358   1.98056756
 (0.04754901) (0.02879434)

b0<-c(mean(res.bay[,1]),mean(res.bay[,2]))
b0
```

```
[1] -823.2670280    0.4172582

B0<-solve(cov(res.bay[,1:2]))
B0
            (Intercept)        año2
(Intercept)    11.24989    22424.72
año2        22424.72321 44699971.20

B0[1,2]<-B0[2,1]

res.bay2<-MCMCregress(precio1~año1,b0=b0,B0=B0,
                   c0=3.51567358,d0=1/1.98056756 )
summary(res.bay2)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean        SD  Naive SE Time-series SE
(Intercept) -726.8450 118.92973 1.1892973      1.1882190
año1           0.3686   0.05966 0.0005966      0.0005962
sigma2         0.9444   0.55779 0.0055779      0.0075168

2. Quantiles for each variable:

                2.5%       25%       50%       75%     97.5%
(Intercept) -963.9608 -805.0377 -724.6296 -647.8292 -495.7323
año1           0.2528    0.3289    0.3675    0.4078    0.4876
sigma2         0.3377    0.5818    0.8017    1.1464    2.4118
```

Si usáramos el modelo clásico para los datos tendríamos

```
summary(lm(precio1~año1))

Call: lm(formula = precio1 ~ año1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7684 -0.2823  0.3888  0.4032  0.7460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -704.70446  166.43371  -4.234  0.00387 **
año1           0.35720    0.08346   4.280  0.00366 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8735 on 7 degrees of freedom Multiple
R-squared: 0.7235,     Adjusted R-squared: 0.684 F-statistic: 18.32
on 1 and 7 DF,  p-value: 0.003655
```

11.5. Detección de outliers

Peña y Guttman [175] presentan varias aproximaciones al tema de la detección de valores extraños (outliers) en modelos lineales.

Sea $H = X\left(X'X\right)^{-1}X'$ la matriz hat. Denote por I el conjunto de k enteros diferentes tomados del conjunto $\{1,\ldots,n\}$. El vector y puede descomponerse como $y' = \left(y'_I, y'_{(I)}\right)$, donde (I) significa 'conjunto eliminado I'. Similarmente la matriz X puede ser particionada como $X' = \left(X'_I, X'_{(I)}\right)$. Siguiendo esta notación denotamos a $\hat{\beta}_{(I)}$ y $s^2_{(I)}$ para los estimadores de β y σ^2 basados en $X_{(I)}$ y $y_{(I)}$.

Hay dos modelos alternos:

$$\begin{aligned} y_I &= X_I\beta + a + \epsilon_I \\ y_{(I)} &= X_{(I)}\beta + \epsilon_{(I)}, \end{aligned}$$

donde a es un vector de k componentes de constantes que ajustan la media y $\epsilon_I \sim N\left(0, \sigma^2 I_k\right)$ y es independiente de $\epsilon_{(I)} \sim N\left(0, \sigma^2 I_{n-k}\right)$.

La idea es usar la densidad predictiva:

$$p\left(y_I|y_{(I)}\right) = \int f\left(y_I|\theta\right)\xi\left(\theta|y_{(I)}\right)d\theta$$

Para el modelo lineal con el supuesto de normalidad presentado arriba tenemos:

$$p\left(y_I|y_{(I)}\right) = K\left(s^2_{(I)}\right)^{-k/2}\left|I - H_I\right|^{1/2}\left(1 + Q_I\right)^{-(n-p)/2},$$

donde,

$$K = \frac{\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k \Gamma\left(\frac{n-p-k}{2}\right)(n-p-k)^{k/2}},$$

y

$$Q_I = \frac{\left(y_I - X_I\hat{\beta}_{(I)}\right)'\left(I - H_I\right)\left(y_I - X_I\hat{\beta}_{(I)}\right)}{(n-p-k)s^2_{(I)}}$$

Capítulo 12

Modelo lineal generalizado

El modelo lineal clásico ha sido utilizado extensivamente y con mucho éxito en múltiples situaciones. En el análisis de regresión estamos interesados en predecir la media de una variable, llamada la respuesta, basados en un conjunto de variables, llamadas los predictores. La regresión clásica asume que la respuesta es continua y distribuida normalmente.

El modelo lineal clásico cae en una clase mayor de modelos que se conoce como *modelo lineal generalizado, M.L.G.,* la cual tiene tres componentes básicas:

1. Un conjunto de variables aleatorias independientes que pertenecen a la familia exponencial.
2. Una matriz de diseño y un vector de parámetros.
3. Una función link (enlace, conexión) que relaciona las medias del modelo lineal.

Dentro de la clase de modelos lineales generalizados tenemos el modelo lineal clásico, el modelo loglineal, la regresión Poisson, la regresión logística, etc.

En el modelo lineal generalizado clásico observamos respuestas Y_i y covariables k-dimensionales $\boldsymbol{x}_i$, donde las respuestas condicionales $(Y_i|\theta_i, \phi)$ se asume son variables aleatorias independientes con una densidad que pertenece a la familia exponencial de un parámetro.

$$f\left(y_i|\theta_i, \phi\right) = \exp\left[\frac{y_i\theta_i - \mu\left(\theta_i\right)}{a(\phi)} + c\left(y_i, \phi\right)\right] \qquad i = 1, \ldots, n$$

El modelo clásico asume que la media $E\left(Y_i\right) = \mu'\left(\theta_i\right)$ está relacionada al intercepto β_0 y al vector de parámetros de las covariables $\boldsymbol{\beta}$ a través de una función de encadenamiento monótona y diferenciable, y el espacio parametral no es vacío.

El modelo lineal generalizado aparece imponiendo una a priori jerárquica sobre los parámetros $(\beta_0, \boldsymbol{\beta})$. Una selección particularmente conveniente es usar aprioris normales con aprioris conjugadas para los hiperparámetros [176], así:

$$
\begin{aligned}
(\beta_0|b_0,\sigma_0) &\sim N(b_0,\sigma_0)\\
(\boldsymbol{\beta}|\boldsymbol{b},\boldsymbol{W}) &\sim N_k(\boldsymbol{b},\boldsymbol{W})\\
(b_o|B_0) &\sim N(0,B_0)\\
(\boldsymbol{b}|B) &\sim N(\boldsymbol{0},B\boldsymbol{I})\\
\left(\sigma_0^{-1}|s_1,s_2\right) &\sim gamma(s_1,s_2)\\
\left(\boldsymbol{W}^{-1}|\boldsymbol{V},v\right) &\sim Wishart\left(\boldsymbol{V}^{-1},v\right)
\end{aligned}
$$

12.1. Modelo logístico

Supongamos que observamos proporciones como respuesta $y_1,\ldots,y_N$ de poblaciones binomiales con proporciones $\pi_1,\ldots,\pi_N$ y sus correspondientes tamaños muestrales $n_1,\ldots,n_N$. Asociado con la i-ésima observación hay un vector de covariables $\boldsymbol{x}_i$ y la proporción π_i es encadenada a las covariables $\boldsymbol{x}_i$ por medio del modelo logístico,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\boldsymbol{x}_i^T\boldsymbol{\beta}$$

La verosimilitud del vector de regresión $\boldsymbol{\beta}$ está dada por:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{N}\pi_i^{n_iy_i}\left(1-\pi_i\right)^{n_i(1-y_i)},\ \text{donde,}$$

$$\pi_i=\frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}$$

Si $\xi(\boldsymbol{\beta})$ es la densidad a priori para $\boldsymbol{\beta}$, entonces la densidad posterior para $\boldsymbol{\beta}$ es proporcional a

$$\xi\left(\boldsymbol{\beta}|\boldsymbol{y}\right)\propto\xi(\boldsymbol{\beta})L(\boldsymbol{\beta})$$

Una ventaja grande de la aproximación bayesiana es que prácticamente se elimina el problema de separación[1]. Este problema aparece con cierta frecuencia en los ajustes clásicos de la regresión logística y no tiene soluciones satisfactorias a no ser que sea incrementar el número de observaciones.

[1]El conjunto de datos está *completamente separado* si existe un vector $\theta\in R^p$ tal que

$$x_i^T\theta>0 \text{ si } y_i=1$$

$$x_i^T\theta<0 \text{ si } y_i=0,$$

para $i=1,\ldots,n$. El conjunto de datos está *cuasicompletamente separado* si existe un vector $\theta\in R^p$ tal que

$$x_i^T\theta\geq 0 \text{ si } y_i=1$$

$$x_i^T\theta\leq 0 \text{ si } y_i=0,$$

para $i=1,\ldots,n$ y si existe $j\in 1,\ldots,n$ tal que $x_j^T\theta=0$. Un conjunto de datos se dice que se *traslapa* si no está completamente separado ni cuasicompletamente separado. El estimador de máxima verosimilitud de θ existe si y solo si el conjunto de datos se traslapa. Los autores miden el traslapado. Ellos definen $n_{notraslapado}$ el menor número de observaciones que necesitan removerse para que los estimadores de máxima verosimilitud no existan. Observe que $n_{notraslapado}\leq n_{completo}$.

12.1.1. Selección de la distribución a priori

Para este caso es difícil asignar una distribución a priori directamente al vector de parámetros de la regresión $\boldsymbol{\beta}$ ya que está relacionado de una forma no lineal a las probabilidades $\{\pi_i\}$. Puede ser más fácil especificar indirectamente una a priori para $\boldsymbol{\beta}$ haciendo suposiciones sobre el valor promedio del valor de la proporción $E(\pi)$ para valores seleccionados de las covariables. Si el rango de la matriz de covariables es k, entonces uno considera las proporciones $\pi_1, \ldots, \pi_k$ para k conjuntos diferentes de la covariable x. Las medias condicionales a priori (MCA) asumen que $\pi_1, \ldots, \pi_k$ son independientes, con π_i distribuida $Beta\left(w_i m_i, w_i\left(1-m_i\right)\right)$, donde m_i es una adivinanza a priori de π_i y w_i es la precisión de esta adivinanza. La distribución sobre $\pi_1, \ldots, \pi_k$ es proporcional a

$$\xi\left(\pi_1, \ldots, \pi_k\right) \propto \prod_{i=1}^{k} \pi_i^{w_i m_i - 1}\left(1-\pi_i\right)^{w_i(1-m_i)-1}$$

Para el linkeo logístico, esta a priori sobre $\{\pi_i\}$ es equivalente a una a priori sobre $\boldsymbol{\beta}$ que es de la misma forma que la verosimilitud con «observaciones a priori»$\{(m_i, w_i, \boldsymbol{x}_i)\}$. Esta es llamada una a priori de datos aumentados (ADA). Es fácil actualizar la densidad a posteriori de $\boldsymbol{\beta}$ utilizando esta forma de distribución a priori. La densidad posterior es proporcional a

$$\xi\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \propto \prod_{i=1}^{N} \pi_i^{n_i y_i}\left(1-\pi_i\right)^{n_i(1-y_i)} \prod_{i=1}^{k} \pi_i^{w_i m_i - 1}\left(1-\pi_i\right)^{w_i(1-m_i)-1}$$

En otras palabras, la distribución a posteriori de $\boldsymbol{\beta}$ es equivalente a la verosimilitud de los datos observados $\{(y_i, n_i, \boldsymbol{x}_i)\}$ aumentados con los «datos a priori»$\{(m_i, w_i, \boldsymbol{x}_i)\}$.

Bedrick, Christensen y Johnson [177] presentan un resumen de diversos procedimientos para la elicitación de la a priori en el caso de la regresión logística. Entre ellas se encuentra la elicitación de la probabilidad de éxito a diferentes niveles de las covariables. Al-Awadhi y Garthwaite [178] presentan una metodología de elicitación del modelo logístico en el área de ecología.

Ejemplo 12.1. Modelo logístico cuadrático. Dellaportas y Smith [179] presentan este ejemplo que considera un modelo logístico cuadrático. Los datos hacen referencia a la retinopatía, una enfermedad de los ojos, y el tiempo que un paciente ha tenido diabetes. La siguiente tabla presenta información sobre pacientes que sufrían de este padecimiento en dos muestras (una pasada y otra actual).

Tabla 12.1: *número de pacientes con diabetes y retinopatía en dos diferentes estados del tiempo*

Duración de la Diabetes	Retinopatía			
	Datos Previos		Datos Actuales	
z	Si	No	Si	No
0-2 (1)	17	215	46	290
3-5 (4)	26	218	52	211
6-8 (7)	39	137	44	134
9-11 (10)	27	62	54	91
12-14 (13)	35	36	38	53
15-17 (16)	37	16	39	42
18-20 (19)	26	13	23	23
21+ (24)	23	15	52	32

El modelo considerado fue:

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = \beta_1 + \beta_2 z_j + \beta_3 z_j^2 = \eta_j$$

Un análisis que se realizó tomó como información a priori la generada por los estimadores de máxima verosimilitud a partir de los datos previos:

$$\boldsymbol{\beta}_o = \begin{pmatrix} -3.17 \\ +0.33 \\ -0.007 \end{pmatrix} \quad \text{y}$$

$$\boldsymbol{D}_o = 10^{-4} \begin{pmatrix} 638.0 & & \\ -111.0 & 24.1 & \\ 3.9 & -0.9 & 0.04 \end{pmatrix},$$

y se consideró como la distribución a priori de $\boldsymbol{\beta}$ la normal trivariable $N(\boldsymbol{\beta}_o, \boldsymbol{D}_o)$. Por lo tanto con los datos presentes la distribución a posteriori de $\boldsymbol{\beta}$ será proporcional a:

$$\xi(\boldsymbol{\beta}|Datos) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_o)' \boldsymbol{D}_o^{-1}(\boldsymbol{\beta}_o) \sum_{j=1}^{8} \{x_{1j}\log(\eta_j) - (x_{1j}+x_{2j})\log(1+e^{\eta_j})\}\right),$$

donde x_{1j} y x_{2j} son los números actuales en cada categoría de edad con o sin retinopatía. Para obtener la constante de normalización se necesita una integración numérica tridimensional.

Dellaporta y Smith [179] comentan que Knuiman y Speed optaron por una aproximación normal basados en la moda posterior, una solución de:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}|Datos) = 0,$$

y una medida de dispersión dada por la matriz:

$$\boldsymbol{D}(\boldsymbol{\beta}) = -\left[\frac{\partial^2 \{\log(\boldsymbol{\beta}|Datos)\}}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]^{-1},$$

evaluada en la moda posterior

$$\boldsymbol{\beta} = \begin{pmatrix} -2.37 \\ +0.21 \\ -0.004 \end{pmatrix}$$

$$\boldsymbol{D} = 10^{-4}\begin{pmatrix} 207.0 & & \\ -36.0 & 8.1 & \\ 1.2 & -0.3 & 0.01 \end{pmatrix},$$

y Dellaporta y Smith utilizando el muestrador de Gibbs obtuvieron:

$$\boldsymbol{\beta}^* = \begin{pmatrix} -2.36 \\ +0.21 \\ -0.004 \end{pmatrix}$$

$$\boldsymbol{D}^* = 10^{-4}\begin{pmatrix} 201.0 & & \\ -35.7 & 7.9 & \\ 1.2 & -0.3 & 0.01 \end{pmatrix}$$

Ejemplo 12.2. Niñas Polacas. Consideramos la famosa base sobre la edad de la menarquia de una niñas en Polonia en los años 60 [180]. En esta base se presenta la edad de una niña y si ya ha tenido su primera menstruación. Los datos vienen categorizados.

```
modelo<- function(){
  for( i in 1 : N ){
    r[i]~dbin(p[i],n[i])
    logit(p[i])<-alpha.star+beta*(x[i]-mean(x[]))
    rhat[i]<-n[i]*p[i]}

  alpha <- alpha.star - beta * mean(x[])
  beta ~ dnorm(0.0,0.001)
  alpha.star ~ dnorm(0.0,0.001)
}

library(R2OpenBUGS)
# escribimos el modelo en una ubicación temporal
modelo.archivo <- file.path(tempdir(),"modelo.txt")
write.model(modelo, modelo.archivo)

datos <-list( x = c(10.83,11.08,11.33,11.58,11.83,12.08,
            12.33,12.58,12.83,13.08,13.33,13.58,13.83,14.08,
            14.33,14.58,14.83,15.08,15.33,15.58),
            n =c(120,90,88,105,111,100,93,100,108,99,106, 105,
```

```
                117,98,97,120,102,122,111,94),
           r =c(2,2,5,10,17,16,29,39,51,47,67,81,88,79,90,113,
                95,117,107,92),N =20)
param <- c("alpha","beta")
iniciales <-list(alpha.star=0, beta=0)

resul <- bugs(datos,iniciales,param,modelo.archivo,n.chains = 2,
          n.iter=50000,n.burnin=20000)
print(resul)
```

Tabla 12.2: *resumen estadístico del modelo de edad de menarquia*

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta	1.561	0.05498	5.717E-4	1.458	1.56	1.673	1000	10001
alfa	-20.17	0.7105	0.007317	-21.61	-20.16	-18.83	1000	10001

Procedimiento Clásico en R

```
edad<- c(10.83,11.08,11.33,11.58,11.83,12.08,12.33,12.58,12.83,
13.08,13.33,13.58,13.83,14.08,14.33,14.58,14.83,15.08,15.33,15.58)
exitos<-c(2,2,5,10,17,16,29,39,51,47,67,81,88,79,90,113,95,117,107,92)
n<-c(120,90,88,105,111,100,93,100,108,99,106,105,117,98,97,120,102,
122,111,94)

summary(glm(cbind(exitos,n-exitos)~edad,family='binomial'))

Call: glm(formula = cbind(exitos, n - exitos) ~ edad, family =
"binomial")

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.2267  -0.8613  -0.3124   0.7507   1.2841

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.46917    0.83475  -24.52   <2e-16 ***
edad          1.57545    0.06379   24.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1) Null
deviance: 1278.571  on 19  degrees of freedom Residual deviance:
14.893  on 18  degrees of freedom AIC: 100.35

Number of Fisher Scoring iterations: 3
```

12.1.2. Extensiones del modelo logístico

Sea

$$P = \int_{-\infty}^{z} f_1(x)\,dx$$

La distribución logística tiene densidad dada por:

$$f_1(z) = \frac{e^{-z}}{\left(1+e^{-z}\right)^2},\ z \in \mathbb{R},$$

entonces,

$$P = \int_{-\infty}^{z} f_1(x)\,dx = F_1(z) = \frac{1}{1+e^{-z}},\ z \in \mathbb{R}$$

El complemento de la probabilidad es:

$$Q = 1 - P = \frac{1}{1+e^{z}}$$

En la mayoría de las aplicaciones $z_i = x_i'\beta$, entonces $P_i = F_1\left(x_i'\beta\right)$.

En economía los modelos logísticos y probit son generados a partir de un modelo de utilidad estocástica (random utilty model).

La utilidad por obtener un éxito se modela como:

$$z_i = x_i'\beta + u_i,$$

donde u_i es una variable aleatoria no observable con distribución conocida.

Hay éxito si la utilidad está por encima de un límite (convencionalmente se escoge 0)

$$p_i = Pr\left(x_i'\beta + u_i > 0\right) = Pr\left(u_i > -x_i'\beta\right)$$

Si $u_i \sim F_u$ entonces $p_i = 1 - F_u\left(-x_i'\beta\right)$.

Suponga que $p_i \sim F_p\left(x_i'\beta\right)$. O sea, $F_p\left(x_i'\beta\right) = 1 - F_u\left(-x_i'\beta\right)$. Esto implica que f_p debe ser un reflejo (alrededor de cero) de f_u.

Modelo SCOBIT

El modelo SCOBIT fue propuesto por Nagler en 1994.

$$\begin{aligned} Q_i^* &= Q_i^{\alpha} = \frac{1}{\left(1+\exp\left(x_i'\beta\right)\right)^{\alpha}},\ \text{para } \alpha > 0 \\ P_i^* &= 1 - Q_i^* = 1 - \frac{1}{\left(1+\exp\left(x_i'\beta\right)\right)^{\alpha}} \end{aligned}$$

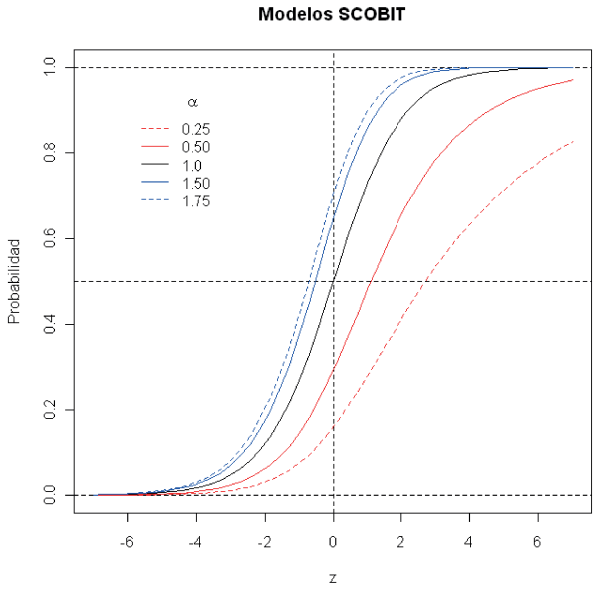

Figura 12.1: *comportamiento del modelo SCOBIT para diferentes valores de* α

$$F^*\left(x_i'\beta\right) = p_i^*$$

Una distribución que tiene una estructura similar es una de las formas de la distribución Burr

$$F_{BURR}(x) = 1 - \left(1 + x^c\right)^{-k} \qquad (x \geq 0)$$

Algunos llaman a F^* la distribución Burr-exponencial o la distribución SCOBIT. Observe que, excepto para $\alpha = 1$, esta distribución no es simétrica alrededor de 0.5.

Continuemos con el ejemplo de las niñas polacas, pero hagamos la estimación usando el modelo SCOBIT. El código es el siguiente:

```
# Niñas polacas

modelo<-function() {
  for( i in 1 : N ) {
  r[i] ~ dbin(p[i],n[i])
  p[i]<-1-exp(-alfa*log(1+exp(alpha.star+beta*(x[i]-mean(x[])))))
  rhat[i]<-(r[i]/n[i] - p[i])/sqrt(p[i]*(1-p[i])/n[i])
  }
  alpha <-alpha.star-beta*mean(x[])
  beta ~ dnorm(0.0,0.001)
  alpha.star ~ dnorm(0.0,0.001)
  alfa~dgamma(1,1) }

library(R2OpenBUGS)
modelo.archivo <- file.path(tempdir(),"modelo.txt")
write.model(modelo, modelo.archivo)
```

```
datos <-list( x = c(10.83,11.08,11.33,11.58,11.83,12.08,
            12.33,12.58,12.83,13.08,13.33,13.58,13.83,14.08,
            14.33,14.58,14.83,15.08,15.33,15.58), n =
        c(120,90,88,105,111,100,93,100,108,99,106,
          105,117,98,97,120,102,122,111,94), r =
        c(2,2,5,10,17,16,29,39,51,47,67, 81,88,79,90,113,95,117,107,92),
        N=20)
param <- c("alpha","beta")
iniciales <-list(alpha.star=0, beta=0,alfa=1)
resul <- bugs(datos,iniciales,param,modelo.archivo,n.iter=50000,
  n.burnin=20000)
print(resul)

          mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
alpha    -23.5 2.3 -28.7 -25.0 -23.3 -21.8 -19.5    1   380
beta       1.9 0.2   1.5   1.7   1.8   2.0   2.3    1   360
deviance  97.4 2.4  94.7  95.7  96.8  98.5 103.6    1  2000
```

12.2. Regresión Poisson

La distribución Poisson juega un papel de fundamental importancia en el trabajo aplicado para modelar problemas de conteo en muchas áreas. Los problemas de regresión donde la variable dependiente es un conteo, ocurre con bastante frecuencia. Como ejemplo tenemos el número de muertos por una cierta enfermedad extraña que puede explicarse por un número grande de factores, como el clima, salubridad, educación, etc. Otro ejemplo es el número de defectos que aparece en cierto rollo de tela, que depende de la longitud del rollo, época de elaboración. Es común asumir una respuesta poissoniana, que perteneciendo a la familia exponencial puede resolverse con la metodología que estamos desarrollando.

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \\ f\left(y_i;\lambda_i\right) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\ &= \exp\left(y_i \log\lambda_i - \lambda_i - \log(y_i!)\right) \\ &= \exp\left(y_i\theta_i - \lambda_i - \log(y_i!)\right), \end{aligned}$$

donde,

$$\theta_i = \log(\lambda_i),$$

el cual es el parámetro natural.

$$\begin{aligned} E[y_i] &= \lambda_i \\ var[y_i] &= \lambda_i, \end{aligned}$$

ya que $\mathrm{g}(\lambda_i) = \theta_i$ cuando g es la función logaritmo. El link canónico es el link log

$$\log(\lambda_i) = \boldsymbol{x_i}'\boldsymbol{\beta},$$

ya que $\lambda_i = \exp(\eta_i)$ se tiene que,

$$\frac{\partial \lambda_i}{\partial \eta_i} = \exp(\eta_i) = \lambda_i,$$

las ecuaciones de verosimilitud,

$$\sum_{i=1}^{n} \frac{(y_i - \lambda_i)}{var(y_i)} x_{ij} \frac{\partial \lambda_i}{\partial \eta_i} = 0 \quad j = 1, .., p,$$

se reduce a

$$\sum_{i=1}^{n} (y_i - \lambda_i) x_{ij} = 0,$$

ya que,

$$w_i = \left(\frac{\partial \lambda_i}{\partial \eta_i}\right)^2 \frac{1}{var(y_i)} = \lambda_i,$$

la matriz de covarianza estimada de $\hat{\boldsymbol{\beta}}$ es $\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}$ donde $\widehat{\boldsymbol{W}}$ es la matriz diagonal con elementos de $\hat{\lambda}$ en la diagonal principal.

Ejemplo 12.3. Una Regresión Poisson. Veamos el siguiente ejemplo para los datos que representan el número de hijos de una pareja y sus años de casado en un estudio de cohorte transversal. El interés es modelar el número de hijos de las parejas según el tiempo de casados usando regresión Poisson.

La Figura 12.2 muestra la relación entre el número de hijos de una pareja y sus años de casados en una muestra de corte transversal. Estos datos presentan un efecto de cohorte, esto es, hay cambios estructurales en estos modelos no observables en los datos, y que se pueden detectar solo en datos que se generan en forma temporal siguiendo cohortes. Estos cambios se originan en cambios de la composición familiar, en cambios económicos, etc. Se observa como la media y la dispersión aumentan a medida que aumenta el número de años.

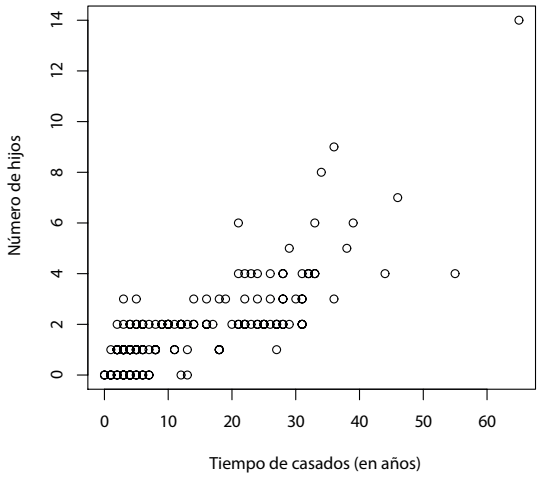

Figura 12.2: *relación entre el número de hijos de una pareja y sus años de casados en una muestra de corte transversal*

```
modelo<-function() { for( i in 1 : N ) {
  NHIJOS[i] ~ dpois(media[i])
  log(media[i])<-alpha.star+beta*(TPOCAS[i]-mean(TPOCAS[]))
}
  alpha <- alpha.star - beta * mean(TPOCAS[])
  beta ~ dnorm(0.0,0.001)
  alpha.star ~ dnorm(0.0,0.001)
}

library(R2OpenBUGS) modelo.archivo <-
file.path(tempdir(),"modelo.txt") write.model(modelo,
modelo.archivo)

datos=list(N=149, TPOCAS=c(28,38,22,1,5,2,3,44,33,10,30,9,21,9,
 5,4,3,2,26,5,4,5,18,5,2,23,3,5,1,21,34,10,3,3,10,31,27,
 24,8,4,12,32,3,6,55,32,65,13,7,31,36,1,6,29,33,18,7,4,
 2,25,20,28,19,6,8,11,2,22,25,26,4,31,28,4,2,24,31,22,27,
 4,11,4,14,29,39,21,2,4,0,3,16,3,14,21,3,18,2,6,11,8,16,4,
 5,10,24,12,12,28,6,25,3,16,1,4,14,33,17,8,3,22,23,6,16,6,
 46,6,8,13,12,24,7,13,26,4,22,31,28,18,27,27,5,28,7,3,12, 5,36,31,0),
 NHIJOS=c(2,5,3,1,1,1,0,4,4,2,3,2,2,2,2,2,1,1,3,2,2,3,1,1,
 1,2,1,0,0,4,8,2,2,1,2,3,1,3,2,1,0,4,0,2,4,4,14,0,2,4,9,0,
 0,5,6,1,0,1,1,2,2,2,3,2,1,2,1,4,2,4,1,2,4,0,0,2,3,2,2,0,1,
 1,2,2,6,2,0,0,0,3,2,1,3,6,1,3,2,1,1,1,3,0,1,2,2,2,2,3,2,2,
 0,2,0,1,2,4,2,1,1,2,4,1,2,0,7,1,1,2,2,4,1,1,2,1,2,2,3,1,2,
 2,0,4,0,1,2,1,3,2,0))

param <- c("alpha","beta") iniciales=list(alpha.star=0, beta=0)
resul <-
bugs(datos,iniciales,param,modelo.archivo,n.chains=2,n.iter=10000,
    n.burnin=1000)
print(resul)
```

Tabla 12.3: *resumen estadístico de los parámetros del modelo a posteriori*

node	mean	sd	MC error	2.5 %	median	97.5 %
beta	0.04284	0.003621	7.762E-5	0.03573	0.04286	0.04996
alpha	-0.1066	0.1046	0.002465	-0.3118	-0.1043	0.09703

El mismo problema con la aproximación clásica:

```
> summary(glm(nrohijos~tpocasados,family='poisson'))

Call: glm(formula = nrohijos ~ tpocasados, family = "poisson")
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.02693  -0.54123  -0.06717   0.43187   2.09419

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.105135   0.102899  -1.022    0.307
tpocasados   0.042891   0.003568  12.020   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 222.330  on 148  degrees of freedom
Residual deviance:  98.788  on 147  degrees of freedom AIC: 436.91

Number of Fisher Scoring iterations: 4
```

Sobredispersión es un fenómeno que ocurre en algunos datos, en especial cuando provienen de binomiales o Poisson. Si la estimación de una medida de dispersión después de ajustar el modelo, como lo es la deviance o el chi cuadrado de Pearson dividido por sus respectivos grados de libertad no está cerca a 1, entonces los datos pueden ser sobredispersos si este cociente es mayor que 1 o subdispersos si es menor que 1.

En el caso de sobredispersión se deben considerar los siguientes cuatro aspectos:

- Pruebas formales para sobredipersión
- Los errores estándares de los coeficientes de la regresión que tomen en cuenta la sobredispersión
- Estadísticos de prueba para la inclusión de variables que tengan en cuenta la sobredispersión
- Modelos más generales con parámetros en la función de varianza

```
# Modelo de regresión Poisson para el número
# de hijos con sobredispersión

modelo<-function(){
  for( i in 1 : N ) {
  NHIJOS[i] ~ dpois(media[i])
  log(media[i]) <- alpha.star + beta * (TPOCAS[i] - mean(TPOCAS[]))
  + tau*TPOCAS[i]
  }
  alpha <- alpha.star-beta*mean(TPOCAS[])
  beta ~ dnorm(0.0,0.001)
  alpha.star ~ dnorm(0.0,0.001)
```

```
  gamma ~ dnorm(0.0,1.0E-6)
  tau ~ dgamma(0.001, 0.001)
  sigma <- 1/sqrt(tau)
}

library(R2OpenBUGS)
modelo.archivo <- file.path(tempdir(),"modelo.txt")
write.model(modelo, modelo.archivo)

datos=list(N=149,
        TPOCAS=c(28,38,22,1,5,2,3,44,33,10,30,9,21,
        9,5,4,3,2,26,5,4,5,18,5,2,23,3,5,1,21,34,10,3,3,10,31,
        27,24,8,4,12,32,3,6,55,32,65,13,7,31,36,1,6,29,33,18,7,
        4,2,25,20,28,19,6,8,11,2,22,25,26,4,31,28,4,2,24,31,22,
        27,4,11,4,14,29,39,21,2,4,0,3,16,3,14,21,3,18,2,6,11,8,
        16,4,5,10,24,12,12,28,6,25,3,16,1,4,14,33,17,8,3,22,23,6,
        16,6,46,6,8,13,12,24,7,13,26,4,22,31,28,18,27,27,5,28,7,3,
        12,5,36,31,0), NHIJOS=c(2,5,3,1,1,1,0,4,4,2,3,2,2,2,2,2,1,1,
        3,2,2,3,1,1,1,2,1,0,0,4,8,2,2,1,2,3,1,3,2,1,0,4,0,2,4,4,14,
        0,2,4,9,0,0,5,6,1,0,1,1,2,2,2,3,2,1,2,1,4,2,4,1,2,4,0,0,2,3,
        2,2,0,1,1,2,2,6,2,0,0,0,3,2,1,3,6,1,3,2,1,1,1,3,0,1,2,2,2,
        2,3,2,2,0,2,0,1,2,4,2,1,1,2,4,1,2,0,7,1,1,2,2,4,1,1,2,1,2,
        2,3,1,2,2,0,4,0,1,2,1,3,2,0))

param <- c("alpha","beta","alpha.star","tau")
iniciales=list(alpha.star=0,beta=0,gamma=0,tau=0.1)
resul <- bugs(datos,iniciales,param,modelo.archivo,n.chains=4,n.iter=10000,
     n.burnin=1000)
print(resul)
```

Tabla 12.4: *resumen estadístico de los parámetros del modelo a posteriori considerando sobredispersión en los datos*

node	mean	sd	MC error	2.5 %	median	97.5 %
beta	0.04282	0.003487	7.991E-5	0.03587	0.04282	0.04955
alpha	-0.1081	0.1015	0.002578	-0.3082	-0.1091	0.09012
alpha.star	0.5482	0.06567	0.001591	0.4201	0.5485	0.6774
tau	1.501E-5	1.492E-4	8.288E-6	3.352E-33	6.125E-17	3.031E-5

Podemos ver que el hecho de considerar sobredispersión en los datos no modifica considerablemente los resultados obtenidos inicialmente (ver Tabla 12.4).

Capítulo 13

Inferencia predictiva

Muchas situaciones aplicadas implican realizar inferencias sobre una observación futura de una variable aleatoria, cuya distribución depende de un número finito de parámetros (desconocidos), esta distribución se conoce como distribución predictiva. Smith [181] argumenta que afirmaciones predictivas acerca de variables aleatorias no observadas tiene más sentido a menudo que la estimación tradicional de parámetros.

13.1. Procedimiento exacto

Asumiendo que $\xi(\theta)$ es la distribución a priori, y que $\xi(\theta|\boldsymbol{x})$ es la posterior, la distribución predictiva bayesiana se calcula como:

$$\begin{aligned} p(z|\boldsymbol{x}) &= \frac{p(z,\boldsymbol{x})}{p(\boldsymbol{x})} \\ &= \frac{\int_\Theta p(z,\boldsymbol{x},\theta)\,d\theta}{\int_\Theta p(\boldsymbol{x},\theta)\,d\theta} \\ &= \frac{\int_\Theta p(z,\boldsymbol{x}|\theta)\xi(\theta)\,d\theta}{\int_\Theta p(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta} \\ &= \frac{\int_\Theta p(z|\theta)p(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta}{\int_\Theta p(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta} \\ &= \int_\Theta p(z|\theta)\left\{\frac{p(\boldsymbol{x}|\theta)\xi(\theta)}{\int_\Theta p(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta}\right\}d\theta \\ &= \int p(z|\theta)\xi(\theta|\boldsymbol{x})\;d\theta \end{aligned}$$

Así

$$\begin{aligned} p(z|\boldsymbol{x}) &= \int p(z|\theta)\xi(\theta|\boldsymbol{x})\;d\theta \\ &= E_{\theta|\boldsymbol{x}}\left[p(z|\theta)\right] \end{aligned}$$

La función $p(z|\theta)$ es la de verosimilitud de θ evaluada en z.

Ejemplo 13.1. Caso Bernoulli. Suponga que $x_1, \ldots, x_n$ es una muestra aleatoria de una $Bernoulli(\pi)$ y suponga que la distribución a priori de π es una $Beta(\alpha, \beta)$. Encontremos la distribución predictiva de una observación futura z.

Tenemos

$$p\left(z|\boldsymbol{x}\right) = \int p(z|\pi)\xi\left(\pi|\boldsymbol{x}\right)\, d\pi$$

Ahora,

$$p(z|\pi) = \pi^z(1-\pi)^{1-z}, \quad z = 0, 1,$$

y

$$\xi(\pi|\boldsymbol{x}) \propto \pi^{\sum x_i+\alpha-1}(1-\pi)^{n-\sum x_i+\beta-1}$$

Ahora, si denotamos por $\alpha^* = \sum x_i + \alpha$ y $\beta^* = n - \sum x_i + \beta$ tenemos que:

$$\begin{aligned} p(z|\boldsymbol{x}) &= \int_0^1 \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\pi^{z+\alpha^*-1}(1-\pi)^{\beta^*+1-z-1}\, d\pi \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\frac{\Gamma(z+\alpha^*)\Gamma(1-z+\beta^*)}{\Gamma(n+\alpha+\beta+1)} \end{aligned}$$

Así,

$$\begin{aligned} P(z=0|\boldsymbol{x}) &= \frac{\Gamma(n+\alpha+\beta)\Gamma(1+\beta^*)}{\Gamma(\beta^*)\Gamma(n+\alpha+\beta+1)} \\ &= \frac{\beta^*}{n+\alpha+\beta} \\ &= \frac{\beta^*}{\alpha^*+\beta^*}, \end{aligned}$$

y

$$P(z=1|\boldsymbol{x}) = \frac{\alpha^*}{\alpha^*+\beta^*}$$

Vale la pena notar que:

$$P(z=1|\boldsymbol{x}) = E(\pi|\boldsymbol{x}),$$

es la media posterior.

Ejemplo 13.2. Caso Poisson. Suponga que $x_1, \ldots, x_n$ es una muestra aleatoria de una $Poisson(\theta)$. Además supongamos que la distribución a priori de θ es una $Gamma(\alpha, \beta)$. Encontremos la distribución predictiva $p(z|\boldsymbol{x})$.

Sabemos que la distribución a posteriori es una $Gamma(\alpha^* = \alpha + \sum x_i, \beta^* = \beta + n)$. Ahora,

$$p(z|\boldsymbol{x}) = \frac{\theta^z e^{-\theta}}{z!}$$

Así,

$$\begin{aligned} p(z|\boldsymbol{x}) &= \int_0^\infty \frac{\theta^z e^{-\theta}}{z!} \frac{(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)} e^{-\beta^*\theta}\, d\theta \\ &= \frac{(\beta^*)^{\alpha^*}}{z!\Gamma(\alpha^*)} \int_0^\infty \theta^{z+\alpha^*-1} e^{-(\beta^*+1)\theta}\, d\theta \\ &= \frac{(\beta^*)^{\alpha^*}}{z!\Gamma(\alpha^*)} \frac{\Gamma(z+\alpha^*)}{(\beta^*+1)^{(z+\alpha^*)}} \\ &= \binom{z+\alpha^*-1}{z} \left(\frac{\beta^*}{\beta^*+1}\right)^{\alpha^*} \left(\frac{1}{\beta^*+1}\right)^z \end{aligned}$$

para $z = 0, 1, 2, \ldots$ Por tanto,

$$z|\boldsymbol{x} \sim Binomial - Negativa\left(\alpha^*, \frac{1}{\beta^*+1}\right)$$

Ejemplo 13.3. Caso Exponencial. Sea $x_1, \ldots, x_n$ una muestra aleatoria de una exponencial con densidad $\theta e^{-\theta x}$, con $x > 0$, $\theta > 0$. Sea Z que denota una observación futura de la misma densidad. Estamos interesados en la probabilidad predictiva que $Z > z$ para algún nivel dado z. Cuando θ es conocido, esto está dado por $\phi = \phi(z|\theta) = e^{-\theta z}$.

Asumiendo que la distribución a priori de θ es $\xi(\theta) \propto \theta^{a-1} e^{-b\theta}$, una a priori Gamma con parámetros (a, b). La distribución a posteriori de θ es también una Gamma con parámetros $(a+n, b+S_n)$, donde $S_n = x_1 + \ldots + x_n$, y la esperanza posterior de ϕ se calcula como:

$$\hat{\phi} = \left(\frac{b+S_n}{b+S_n+z}\right)^{a+n}$$

Cuando $a = b = 0$ se tiene una distribución a priori Jeffreys y la esperanza se reduce a

$$\hat{\phi} = \left(\frac{S_n}{S_n+z}\right)^n$$

Ejemplo 13.4. Distribución Multinomial. En el caso de la distribución multinomial tenemos, bajo una a priori Dirichlet, la a posteriori es también Dirichlet con parámetros $n_i + \alpha_i$, para $i = 1, \ldots, k$. Bajo la distribución a priori de Jeffreys, que corresponde a una Dirichlet con $\alpha_i = 1/2$ para todo $i = 1, \ldots, k$, la distribución predictiva es:

$$p\left(X_i = i \,|\boldsymbol{N}\right) = \frac{n_i + \frac{1}{2}}{\sum_{j=1}^{k} n_j + \frac{k}{2}},$$

y, bajo a a priori uniforme,

$$p\left(X_i = i \,|\boldsymbol{N}\right) = \frac{n_i + 1}{\sum_{j=1}^{k} n_j + k}$$

13.2. Distribución predictiva vía MCMC

- A veces es difícil resolver la integral para calcular la distribución predictiva

$$\begin{aligned} p\left(z|\boldsymbol{x}\right) &= \int p(z|\theta)\xi\left(\theta|\boldsymbol{x}\right)\, d\theta \\ &= E_{\theta|\boldsymbol{x}}\left[p(z|\theta)\right] \end{aligned}$$

- Una solución es usar MCMC.

13.2.1. Algoritmo

- *(Paso 1)* Genere una muestra de tamaño M, luego de haber quemado n_B muestras de $\xi\left(\theta\,|Datos\right)$, puede usar un thin (botar valores intermedios si es necesario para controlar la autocorrelación). Esta muestra la denotamos por comodidad como:

$$\theta_1, \theta_2, \ldots, \theta_M$$

- La distribución predictiva $p\left(z\,|Datos\right)$ podemos aproximarla así:

$$\begin{aligned} p\left(z|\boldsymbol{x}\right) &= \int p(z|\theta)\xi\left(\theta|\boldsymbol{x}\right)\, d\theta \\ &= E_{\theta|\boldsymbol{x}}\left[p(z|\theta)\right] \approx \frac{1}{M}\sum_{i=1}^{M} p\left(z\,|\theta_i, Datos\right) \end{aligned}$$

- *(Paso 2)* Sacamos al azar un número en $\{1, 2, \ldots, M\}$ con probabilidad $1/M$, digamos m.

- *(Paso 3)* De $p(z|\theta_m, Datos)$ sacamos un número al azar, digamos z.

- *(Paso 4)* Repetimos los pasos 2 y 3 una gran cantidad de veces, digamos K. Al final obtenemos un conjunto de valores

$$z_1, z_2, \ldots z_K$$

- *(Paso 5)* Construimos un estimador de la densidad $p(z|Datos)$. Si z es discreta simplemente calculamos la densidad aproximada como:

$$p(z = j|Datos) \approx \frac{\#\left\{x_k = j\right\}}{K}$$

Ejemplo 13.5. Distribución Discreta. Asumamos:

- $X \sim Poisson(\lambda)$
- $\xi(\lambda)$ es $U(0,3)$
- $x_1, x_2, \ldots, x_n$ es una m.a. de la distribución $Poisson(\lambda)$
- La distribución a posteriori es:

$$\xi\left(\lambda \,|Datos\right) = \frac{\lambda^{\sum_{i=1}^n x_i} \exp(-n\lambda)}{\prod_{i=1}^n x_i!},$$

  para $0 < \lambda < 3$.

- La distribución predictiva de z dado los Datos es:

$$p(z\,|Datos) = \int_0^3 \frac{\lambda^z \exp(-\lambda)}{z!} \frac{\lambda^{\sum_{i=1}^n x_i} \exp(-n\lambda)}{\prod_{i=1}^n x_i!}\, d\lambda$$

$$p(z\,|Datos) = \frac{1}{z! \prod_{i=1}^n x_i!} \int_0^3 \exp(-\lambda(n+1))\lambda^{z+\sum_{i=1}^n x_i}\, d\lambda$$

$$\begin{aligned} p(z\,|Datos) &= \frac{1}{z! \prod_{i=1}^n x_i!} \frac{\Gamma\left(z + \sum_{i=1}^n x_i + 1\right)}{(n+1)^{z+\sum_{i=1}^n x_i + 1}} \\ &\quad \times \int_0^3 \frac{(n+1)^{z+\sum_{i=1}^n x_i + 1}}{\Gamma\left(z + \sum_{i=1}^n x_i + 1\right)} \exp(-\lambda(n+1))\lambda^{z+\sum_{i=1}^n x_i}\, d\lambda \end{aligned}$$

Esta última integral corresponde a la función de distribución acumulada de una gamma con parámetros $z + \sum_{i=1}^n x_i + 1$ y $n+1$ evaluada en 3.

Si observamos del proceso 0,0,2,1,2,0,0,2,2,1,1,1,3,4,4,3. Tenemos

```
# Cálculo de la distribución predictiva
# Distr. muestral: Poisson

# A priori: U(0,3)
Datos<-c(0,0,2,1,2,0,0,2,2,1,1,1,3,4,4,3)
p.pred<-function(z,x){
n<-length(x); S.x<-sum(x)
P.x<-prod(factorial(x))
a<-z+S.x+1; b<-n+1
res<-gamma(a)/(factorial(z)*b^a*P.x)*pgamma(3,a,rate=b)
return(res)
}

temp<-p.pred(0:20,Datos)
prob.poste<-temp/sum(temp)
plot(0:20,prob.poste,type='h')
prob.poste
```

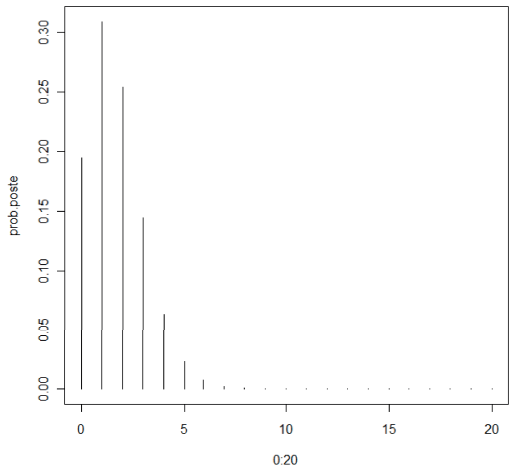

Figura 13.1: *distribución de probabilidad a posteriori de los datos*

Ejemplo 13.6. Distribución continua.

- Suponga $X \sim Gamma(\alpha, \beta)$
- Distribución a priori

$$\xi(\alpha, \beta) \propto 1$$

- Distribución posterior

$$\xi\left(\alpha, \beta \,|Datos\right) \propto \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\beta \sum_{i=1}^{n} x_i\right)$$

```
# Distribución predictiva para una va continua Gamma
tiempos<-c(1.2,0.5,1.6,2.0,2.1,2.0)
prod.tiempos<-prod(tiempos)
sum.tiempos<-sum(tiempos)
n<-length(tiempos)
u<-mean(tiempos)
v<-var(tiempos)
a<-u^2/v
b<-u/v

a
[1] 6.347701
b
[1] 4.051724

veros<-function(a,b,datos){
 res<-1
 for(i in 1:length(datos))
 res<-res*dgamma(datos[i],a,rate=b)
 return(res)
```

```
}

a1<-seq(0.01,16.0,length=50); b1<-seq(0.01,10.0,length=50)
z<-outer(a1, b1, FUN="veros", tiempos)
contour(a1,b1,z,ylab=expression(beta),xlab=expression(alpha))

dist.a.con<-function(a,b,produ,n) exp(n*a*log(b)
                      -n*lgamma(a)+a*log(produ))

# dist.b.con es una gamma(n*a+1,sum.tiempos) # Proceso de muestreo

a.viejo<-a ;b.viejo<-b
result<-c(a,b)
resulta<-matrix(NA,ncol=2,nrow=10000)

for(i in 1:nrow(resulta)){
pesos<-dist.a.con(a1,b.viejo,prod.tiempos,n)
a.nuevo<-sample(a1,1,prob=pesos)
b.nuevo<-rgamma(1,n*a.nuevo+1,sum.tiempos)
resulta[i,]<-c(a.nuevo,b.nuevo)
b.viejo<-b.nuevo }

points(resulta,col='grey')
par(mfrow=c(2,1))
plot(resulta[,1],type='l',ylab=expression(alpha))
plot(resulta[,2],type='l',ylab=expression(beta))
par(mfrow=c(1,1))

# Función que genera muestra de la predictiva

genera.muestra.predictiva<-function(a)rgamma(1,a[1],rate=a[2])
z<-apply(resulta,1,genera.muestra.predictiva)
plot(density(z,from=0),main='Distribución Predictiva')
```

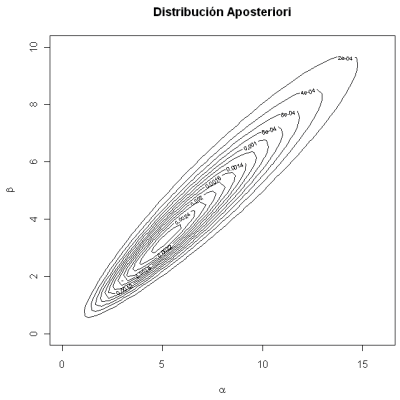

Figura 13.2: *distribución a posteriori para α y β*

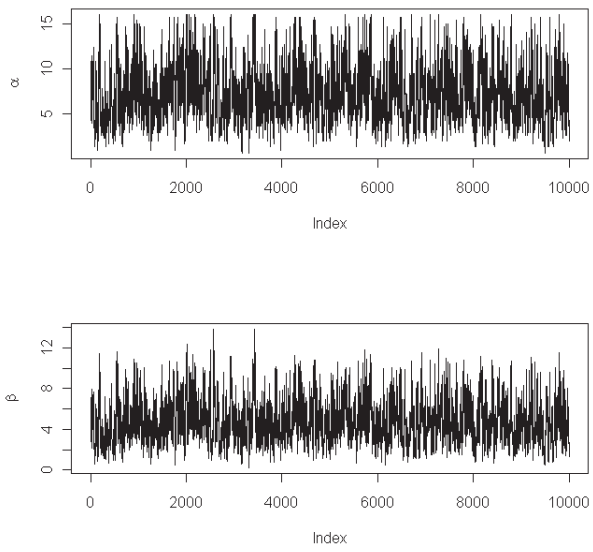

Figura 13.3: *cadenas generadas para α y β*

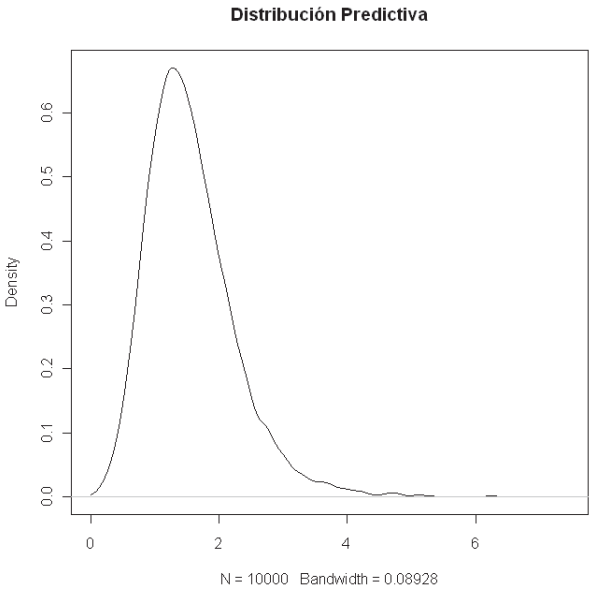

Figura 13.4: *distribución predictiva para una variable aleatoria gamma*

Capítulo 14

Software para estadística bayesiana

14.1. Estadística bayesiana en R

R ha llegado a ser un estándar en el trabajo estadístico aplicado. Su facilidad de extensión por parte de los usuarios lo hace ideal para crear funciones y librerías que ejecuten tareas específicas. Entre las librerías para trabajo bayesiano tenemos:

Librería	Descripción
MCMCpack	Paquete para realizar modelación bayesiana tradicional.
mcmc	Simulación vía MCMC para funciones del usuario.

14.1.1. Librería MCMCpack

Una librería que es muy fácil de utilizar y que posee una amplia base de funciones para ajustar muchos modelos tradicionales.

MCMCregress

Esta función permite ajustar un modelo lineal bayesiano conjugado. Por defecto trabaja con distribuciones a priori no informativas. Esta función asume que la distribución generadora de datos es normal y la distribución de los parámetros es Normal-Gamma.

MCMClogistic

Esta función ajusta modelos de regresión logísticos. Asume que la variable respuesta es 0 o 1. A diferencia de la función $glm()$, no permite considerar respuestas en forma de tabla con columnas fracasos-éxitos.

14.2. Tutorial sobre OpenBUGS

Uno de los inconvenientes que han tenido los métodos bayesianos para ser utilizados en la práctica ha sido la carencia de software especializado. Ninguno de los

grandes paquetes en estadística, SAS, SPSS, etc., tienen módulos para hacer estadística bayesiana.

Existe un programa de acceso gratuito al público que permite utilizar simulación estadística basada en cadenas de Markov en una forma simple y efectiva para gran variedad de modelos llamado *BUGS*, que es un acrónimo de *Bayesian analysis Using the Gibbs Sampler* (Muestreador Gibbs, que lo veremos en un capítulo posterior). Este programa está disponible en:

`http://www.openbugs.net/w/FrontPage`

Hay dos versiones de BUGS, WinBUGS y OpenBUGS. La segunda, trabaja bajo Windows con una interfaz muy similar a la de WinBUGS y con una interfaz de texto plano bajo Linux. Además, puede ser llamado desde R para ejecutar el modelo y analizar las cadenas generadas.

Existen otros programas que permiten resolver problemas bayesianos como el BACC, First Bayes, JAGS (Just Another Gibbs Sampler), etc. El *R* trae algunas librerías con soluciones a ciertos problemas específicos, por ejemplo la MCMCPack y CODA.

14.3. ¿Qué se espera de un software para estadística bayesiana?

Koop [182] señala algunos requisitos claves que todo software bayesiano debería cumplir:

1. Debe ser computacionalmente eficiente
2. Debe estar bien documentado
3. El grupo de soporte debe ser amplio y reconocido
4. Debe proporcionar simuladores posteriores para la clase de modelos que los investigadores quieran usar
5. Para los modelos no incluidos, debe ser fácil la inclusión de los simuladores posteriores que se necesitan por parte del usuario
6. Debe tener una base amplia de funciones $g(\theta)$
7. Debe proporcionar medidas del error en la aproximación para las estimadas de $E\left(g(\theta)|Y\right)$ y las verosimilitudes marginales
8. Debe permitir al usuario graficar la a posteriori y la a priori
9. Debe permitirle al usuario realizar un análisis de sensibilidad a priori de una manera fácil
10. Todo lo anterior debe poderse llevar a cabo de una manera simple, transparente y conveniente para el usuario

14.3.1. Utilización de WinBUGS y OpenBUGS

La utilización por primera vez del programa puede ser una experiencia extraña, ya que el programa no funciona en una forma lineal, sino que requiere múltiples pasos que pueden parecer repetitivos, pero que en realidad no lo son.

En *WinBUGS* y *OpenBUGS* el símbolo $\sim$ significa «distribuido como»y se utiliza para:

- Especificar la distribución de los datos
- Especificar la distribución a priori

Los valores a la izquierda de $\sim$ son llamados «estocásticos».

La flecha (conformada por dos símbolos) a la izquierda `<-` se utiliza como el igual. Por ejemplo `var <- 1/precision` . Los valores a la izquierda de `<-` son llamados «lógicos».

Los pasos en el programa para correr un modelo son:

1. Los comandos anteriores los escribimos en una ventana que abrimos seleccionando *File* y luego *New*. Si usted ya tiene algún archivo con un programa creado y salvado con anterioridad en formato *.odc* puede abrirlo para trabajar con él.

   Si seleccionamos *New* el programa muestra una ventana en blanco en la cual podemos escribir los comandos apropiados, como los que se encuentran enseguida. Con el cursor seleccionamos toda la parte correspondiente al modelo y seleccionamos *Edit* y luego *Copy*.

```
model {
 # likelihood
   n.ij ~ dbin(n, P.ij)
 # prior
   P.ij ~ dbeta(nu, omega)
} list(k = 201, n = 372, nu = 1, omega = 1)
```

   Ejemplo con la longitud máxima del pie de estudiantes universitarios:

```
model
        {
            for(i in 1:N){
                Y[i] ~ dnorm(mu[i], tau)
                mu[i] <- beta
            }
            sigma <- 1/sqrt(tau)
            beta ~ dnorm(0, 1.0E-6)
            tau ~ dgamma(1.0E-3, 1.0E-3)
        }
```

El programa $WinBUGS$ permite utilizar un lenguaje conciso para expresar un modelo: β y τ son expresados con distribuciones a priori propias pero lo más mínimo informativas que se pueda, mientras que la expresión lógica sigma permite que la desviación estándar sea estimada.

2. Primero seleccionamos el menú *Model*.

3. Abrimos la herramienta *Specification*. Aquí nos aparece una ventana con varias opciones.

4. Señalamos la palabra *check model* en el comienzo de la descripción del modelo. Necesitamos chequear que la descripción del modelo define completamente un modelo de probabilidad. Si el modelo fue especificado correctamente aparece el mensaje *model is syntactically correct* en la parte inferior izquierda de la ventana principal. Sino, nos aparece el tipo de error que tenemos en el modelo.

5. Luego señalamos los datos (los cuales deben estar en un formato especial, estilo $S-Plus$) y los copiamos con *Edit* y luego *Copy*.

6. Nuevamente nos vamos a la ventana *Specification Tool* y seleccionamos *load data*. Si los datos están conformes al modelo, aparece un mensaje en la parte inferior izquierda de la ventana principal donde se informa que los datos fueron cargados. (Estos datos pueden estar copiados en la misma ventana en la cual escribimos nuestro modelo. Lo que hacemos es señalarlos y copiarlos y luego oprimimos el cuadro *load data*).

```
list(Y = c(24.2,25.4,25.0,25.9,25.5,24.4), N = 6)
```

7. El siguiente paso se ejecuta en la ventana *Specification Tool* y seleccionamos *compile*.

8. A continuación en la ventana *Specification Tool* seleccionamos *load inits*. Los valores iniciales para el proceso iterativo (Estos valores iniciales pueden estar copiados también en la misma ventana en la cual escribimos nuestro modelo y los datos. Lo que hacemos es señalarlos y copiarlos y luego oprimimos el cuadro *load inits*).

```
list( beta = 0, tau = 1)
```

   Otra opción nos permite que el programa genere automáticamente valores iniciales, esto lo hace generando números aleatorios de la distribución a priori. El programa permite correr más de una cadena simultáneamente, para lo cual se necesita especificar más de un conjunto de valores iniciales.

9. Del menú *model* seleccione *Update...* y del menú *Inference* seleccione *Samples*. Ahora usted tiene dos nuevas ventanas, una con el nombre *Update Tool* y la otra con el nombre *Sample Monitor Tool*.

La ventana *Update Tool* nos permite generar muestras. En MCMC usualmente hay que dejar correr el muestreador durante algún tiempo (quizá 1000 iteraciones) para asegurarnos de que el proceso está estable antes de guardar valores.

Después de una corrida inicial nos ubicamos en la ventana *Sample Monitor Tool*. Para empezar escribimos los nombres de los nodos (parámetros) que queremos estudiar. Escribimos en la parte de *node* `beta` y seleccionamos luego *set*. Procedemos igual con `tau`.

10. De la ventana *Update Tool* seleccionamos la opción *update*. Esto lo podemos realizar tantas veces como sea necesario para que el proceso converja.

11. De la ventana *Sample Monitor Tool* seleccionamos ya lo que sea de nuestro interés. Por ejemplo, seleccionamos un nodo, digamos `beta` y luego *stats*, nos aparece una nueva ventana con algunos resultados de interés acerca de este parámetro. Lo mismo hacemos para `tau`.

node	mean	sd	MC error	2.5 %	median	97.5 %	start	sample
beta	25.06	0.3443	0.006615	24.34	25.06	25.75	1	3000
sigma	0.778	0.3284	0.0068	0.4095	0.698	1.589	1	3000

Ejemplo 14.1. El caso normal con varianza conocida. Supongamos que tenemos una muestra aleatoria de una normal y asumimos que su varianza es conocida e igual a 1.

```
model normalIID \{ for(in 1:10) \{y[i] ~ dnorm(mu,1) \}

# Distribucion a priori
   mu~dnorm(0,1)
\}

list(y=c(1.84,-0.23,1.12,0.35,-0.24,
            -0.89,1.65,-1.01,2.01,1.12))
```

14.3.2. Algunos de los comandos de WinBUGS y OpenBUGS

Model

Specification. Este comando activa una ventana llamada **Specification Tool**, que nos permite definir completamente el modelo.

check model: chequea el modelo.

load data: carga los datos.

compile: compila el modelo.

load inits: carga los valores iniciales.

gen inits: genera los valores iniciales en caso de que el usuario no los especifica con anterioridad.

num of chains: especifica el número de cadenas a generar.

for chain: establece el salto por cadenas usadas.

Update. Este comando se activa una vez el modelo ha sido compilado e inicializado. El produce la ventana **Update Tool** con los siguientes comandos:

updates: número de actualizaciones MCMC a ser llevadas a cabo.

refresh: el número de actualizaciones entre re-actualizaciones de la pantalla.

thin: las muestras de cada k-ésima iteración será guardada, donde k es el valor de **thin**. Hacer $k > 1$ puede ayudar a reducir la autocorrelación en la muestra.

update: cliquear para comenzar a actualizar el modelo.

over relax: esta selección permite trabajar con una versión más relajada del MCMC.

adapting: esta selección permite un proceso de adaptación inicial para un mejor ajuste de los parámetros. Toda la información generada en este proceso es descartada.

Inference

La opción **Inference** tiene varias opciones, pero la más importante es **Samples**.

Samples...: bajo este comando aparece una ventana con título **Sample Monitor Tool**. Contiene los siguientes campos:

node: se especifica el parámetro o variable de interés para el análisis.

chains: se pueden seleccionar las cadenas con las que se construirán los estadísticos.

to: opera junto con el comando anterior.

beg: cuando se utiliza una submuestra para el análisis este comando nos indica desde dónde empezamos a utilizar los valores originales. Marca el comienzo de la submuestra.

end: marca el final de la submuestra que se inició con el comando anterior.

thin: las muestras de cada k-ésima iteración será utilizada para los estadísticos a producir, donde k es el valor de **thin**.

percentiles:

clear: remueve cualquier valor guardado de las variables.

set: debe utilizarse para empezar a guardar los valores para una variable.

trace: presenta una gráfica del valor de la variable contra el número de la iteración. La traza es dinámica y se está reactualizando.

history: grafica la traza completa para la variable.

density: presenta un gráfico de densidad para la variable si es continua, o un histograma si es discreta,

stats: produce un resumen estadístico para la variable.

coda: produce una representación ASCII del proceso para ser reanalizada con *CODA*.

quantiles:

GR diag: calcula el estadístico para convergencia de Gelman-Rubin.

autoC: grafica la función de autocorrelación de variable hasta un rezago de 50.

Fit...: Fit Tool

Referencias

[1] R. Schlaifer. *Probability and Statistics for Business Decisions: An Introduction to Managerial Economics Under Uncertainty*. McGraw-Hill, New York, 1959.

[2] H. Chernoff and L. E. Moses. *Teoría y Cálculo Elemental de las Decisiones*. Compañía Editorial Continental, S.A., México, D.F., 1959.

[3] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Harvard University Press, Boston, 1964.

[4] J. W. Pratt, H. Raiffa, and R. Schlaifer. *Introduction to Statistical Decision Theory*. McGraw-Hill, New York, 1964.

[5] B. W. Morgan. *An Introduction to Bayesian Statistical Decision Processess*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1968.

[6] H. Raiffa. *Decision Analysis: Introductory Lectures on Choice Under Uncertainty*. Addison-Wesley: Reading, Masschusetts, 1970.

[7] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw Hill, New York, 1970.

[8] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. World Scientific Publishing Co., New York, 2 edition, 1985.

[9] H. Jeffreys. *Theory of Probability*. Clarendon Press, Londres, 3 edition, 1961.

[10] L. J. Savage. *Subjective Probability and Statistical Practice. En The Foundations of Statistical Inference, PUBLISHER = G.A. Barnard y D. R. Cox. John Wiley & Sons, YEAR = 1962, edition = , ADDRESS = Londres.*

[11] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, New York, 1971.

[12] R. L. Winkler. *An Introduction to Bayesian Inference and Decision*. Holt, Rinehart and Winston,Inc., New York, 1972.

[13] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York, 1973.

[14] S. J. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Dover: Mineola, New York, 2 edition, 1982.

[15] S. J. Press. *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, New York, 1989.

[16] P. M. Lee. *Bayesian Statistics: An Introduction*. Arnold, Londres, 2 edition, 1997.

[17] T. Leonard and J. S. Hsu. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researches*. Cambridge University Press, Cambridge, UK, 1999.

[18] C. P. Robert. *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer-Verlag, New York, 1994.

[19] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, 2 edition, 2001.

[20] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2 edition, 2000.

[21] P. Congdon. *Bayesian Statistical Modelling*. John Wiley & Sons: Chichester, UK, 2001.

[22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2 edition, 2004.

[23] W. M. Bolstad. *Introduction to Bayesian Statistics*. John Wiley & Sons: Hoboken, New Jersey, 2004.

[24] J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Mtehods*. Springer, New York, 2006.

[25] J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 2000.

[26] M. West and J. harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 1989.

[27] G. Koop. *Bayesian Econometrics*. Wiley: West Sussex, England, 2003.

[28] G. G. Woodworth. *Biostatistics: A Bayesian Introduction*. John Wiley & Sons, Hoboken, New Jersey, 2004.

[29] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, Hoboken, 2004.

[30] A. B. Lawson. *Disease Mapping with WinBUGS and MLwiN*. John Wiley & Sons, Chichester, UK, 2003.

[31] J. Gill. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC, Boca Raton, 2002.

[32] S. M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York, 2007.

[33] H. F. Martz and R. A. Waller. *Bayesian Reliability Analysis*. Wiley, New York, 1982.

[34] M. S. Hamada, A. G. Wilson, C. S. Reese, and H. F. Martz. *Bayesian Reliability*. Springer, New York, 2008.

[35] P. E. Rossi, G. M. Allenby, and McCulloch. *Bayesian Statistics and Marketing*. John Wiley & Sons, Chichester, UK, 2005.

[36] K. Yuen. *Bayesian Methods for Structural Dynamics and Civil Engineering*. John Wiley & Sons, Singapore, 2010.

[37] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

[38] B. A. Berg. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis With Web-Based Fortran Code*. World Scientific Publishing Co., New Jersey, 2004.

[39] J. Albert. *Bayesian Computation With R*. Springer, Nueva York, 2007.

[40] J. Marin and C. P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York, 2007.

[41] I. Ntzoufras. *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, Hoboken, New Jersey, 2009.

[42] C. P. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Springer, New York, 2010.

[43] F. J. Samaniego. *A Comparison of the Bayesian and Frequentist Approaches to Estimations*. Springer, New York, 2010.

[44] D. Qin. Bayesian econometrics: The first twenty years. *Econometric Theory*, 12(3):500–516, 1996.

[45] D. Poole, A. Mackworth, and R. Goebel. *Computational Intelligence: A Logical Approach*. Oxford University Press, New York, 1998.

[46] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, Austria, 2016.

[47] A. D. Martin and K. M. Quinn. Applied bayesian inference in r using mcmcpack. *R News*, 6(1):2–7, 2006.

[48] A. D. Martin, K. M. Quinn, and J. H. park. Mcmcpack: Markov chain monte caro in r. *Journal of Statistical Software*, 42(9):`htt://www.jstatsoft.org/`, 2011.

[49] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Collier MacMillan International, New York, 4 edition, 1978.

[50] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill Kogasakua, Ltd, Tokyo, 3 edition, 1986.

[51] S. Greenland. Putting background information about relative risks into conjugate prior distributions. *Biometrics*, 57(3):663–670, 2001.

[52] Z. Dienes. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3):274–290, 2011.

[53] J. Albert. *MATLAB as an Enviroment for Bayesian Computation*. Dept. of Math. and Statistics. Bowling Green State University, 1997.

[54] B. Western and S. Jackman. Bayesian inference for comparative research. *The American Political Science Review*, 88(2):412–423, 1994.

[55] H. E. Kyburg Jr. Probability and decision. *Philosophy of Science*, 33(3):250–261, 1966.

[56] G. K. Chacko. *Decision-Making under Uncertainty*. Praeger Pub., New York, 1991.

[57] J. Franklin. *The Science of Conjecture: Evidence and Probability Before Pascal*. The Johns Hopkins University Press, Baltimore, Maryland, 2001.

[58] D. S. Sivia. *Data Analysis: A Bayesian Tuorial*. Oxford University Press, Oxford, 1996.

[59] T. S. Wallsten and D. V. Budescu. Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29(2):151–173, 1983.

[60] B. M. Ayyub. *t Opinions for Uncertainty and Risks. Elicitation of Exper*. CRC Press, Boca Raton, 2001.

[61] J. O. Berger. *Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. Technical Report*. Duke University, 1999.

[62] T. A. Loredo. *From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics*. Fougeré. In Maximum Entropy Bayesian Methods. Kluwer Acadaemic Publishers, Dordrecht, 1990.

[63] A. G. Sawyer and J. P. Peter. The significance of statistical significance tests in marketing research. *Journal of Marketing Research*, 20(2):122–133, 1983.

[64] S. Labovitz. The nonutility of significance tests: The significance of tests of significance reconsidered. *The Pacific Sociological Review*, 13(3):141–148, 1970.

[65] B. Lecoutre, M. Lecoutre, and J. Poitevineau. Uses, abuses and misuses of significance tests in the scientific community: Won't the bayesian choice be unavoidable? *International Statistical Review / Revue Internationale de Statistique*, 69(3):399–417, 2001.

[66] F. E. Harrel Jr. *An Introduction to Bayesian Methods with Clinical Applications*. Dept. of Health Evaluation Sciences. School of Medicine, University of Virginia, Charlottesville, 1998.

[67] J. O. Berger and R. L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics: Hayward, California, 2 edition, 1998.

[68] Y. Pawitan. A reminder of the fallibility of the wald statistic: Likelihood explanation. *The American Statistician*, 54(1):54–56, 2000.

[69] L. J. Savage. *The Foundations of the Statistical Inference*. Methuen & Co., Londres, 1962.

[70] R. M. Royall. The effect of sample size on the meaning of significance tests. *The American Statistician*, 40(4):313–315, 1986.

[71] F. M. Rosekrans. Statistical significance and reporting test results. *Journal of Marketing Research*, 6(4):451–455, 1969.

[72] R. L. Winkler. Why bayesian analysis has´nt caught on in healthcare decision making. *International Journal of Technology in Health Care*, 17(1):56–76, 2001.

[73] M. J. Bayarri, M. H. DeGroot, and J. B. Kadane. *Statistical Decision Theory and Realted Topics IV, Gupta, S. S. and Berger, J. O., Eds.* Springer-Verlag, New York, 1988.

[74] S. Jackman. *Bayesian Modelin in the Social Sciences: an Introduction to Markov-Chain Monte Carlo. Technical Report*. Dept. of Political Science, Stanford University, 1999.

[75] D. Ashby. Bayesian statistics in medicine: a 25 year review. *Statistics in Medicine*, 25:3589–3631, 2006.

[76] J. O. Berger and D. Sun. Bayesian analysis for the poly-weibull distribution. *Journal of the American Statistical Association*, 88(424):1412–1418, 1993.

[77] R. A. Evans. *Bayes*. Theory & Practice: The Theory and Applications of Reliability With Emphasis on Bayesian and Nonparametric Methods, New York, 2 edition, 1987.

[78] I. Horowitz. *Introducción al Análisis Cuantitativo de los Negocios*. Ediciones del Castillo, Madrid, 1968.

[79] D. J. Poirier. Frequentist and subjectivist perspectives on the problemof model building in economics. *The Journal of Economic Perspectives*, 2(1):121–144, 1988.

[80] G. Shafer. Probability judgment in artificial intelligence and expert systems. *Statistical Science*, 2(1):3–16, 1987.

[81] D. R. Cox. Some remarks on model criticism. *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS. edited by Adams, N. et al., Imperial College Press: London*, pages 13–21, 2004.

[82] G. D'Agostini. *Role and Meaning of Subjective Probability: Some Comments on Common Misconceptions*. XX International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Gif sur Yvette, Francia, 2000.

[83] F. J. Anscombe and J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.

[84] R. M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, Oxford, 1991.

[85] R. Jeffrey. *Probability and the Art of Judgement*. Cambridge University Press, New York, 1992.

[86] J. B. Kadane and R. L. Winkler. Separating probability elicitation from utilities. *Journal of the American Statistical Association*, 83(402):357–363, 1988.

[87] A. P. Dawid. Probability, causality and the empirical world: A bayes-de finetti-popper-borel sythesis. *Statistical Science*, 19(1):44–57, 2004.

[88] J. O. Berger, B. Liseo, and R. L. Wolpert. *Integrated Likelihood Methods for Eliminating Nuisance Parameters*. Purdue Univ. Dept. of Statistics Technical Report No. 96-7C, 1998.

[89] A. W. F. Edwards. *Likelihood: Expanded Edition*. The Johns Hopkins University Press, Baltimore, 1992.

[90] B. de Finetti. Sulla proprietá conglomerativa delle probabilitá subordinate. *Atti R. Ist. Lomb. Scienze Lettere*, 63:418–418, 1930.

[91] D. Draper, J. S. Hodges, C. L. Mallows, and D. Pregibon. Exchangeability and data analysis. *Journal of the Royal Statistical Society. Series A*, 156(1):9–37, 1993.

[92] J. M. Dickey. Approximate posterior distributions. *Journal of the American Statistical Association*, 71(355):680–689, 1976.

[93] D. Fink. *A Compdium of Conjugate Priors*. Technical Report. Dept. of Biology. Montana State University, Bozeman, MT, 59717 edition, 1997.

[94] M. Ramoni and P. Sebastiani. *Bayesian Methods for Intelligent Data Analysis. KMi Technical Report*. KMi-TR-67, 1998.

[95] D. Draper. *Bayesian Hierarchical Modeling*. Tutorial 1: ISBA, Crete, 2000.

[96] G. C. G. Wei and M. A. Tanner. Posterior computations for censored regression data. *Journal of the American Statistical Association*, 85(411):829–839, 1990.

[97] A. Gelman. A bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, 71(2):369–382, 2003.

[98] E. Damsleth. Conjugate classes for gamma distributions. *Scandinavian Journal of Statistics*, 2(2):80–84, 1975.

[99] G. Meeden. *An Elicitation Procedure Using Piecewise Conjugate Priors.* Bayesian Analysis in Statistics and Econometrics. Goel, P. K. y Iyengar, N. S., Eds. Springer-Verlag, New York, 1992.

[100] J. B. Kadane, M. J. Schervish, and T. Seidenfeld. *Rethinking the Foundations of Statistics.* Cambridge University Press, New York, 1999.

[101] R. L. Winkler. The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical Association*, 62(319):776–800, 1967a.

[102] R. Yang and J. O. Berger. *A Catalog of Noninformative Priors. Technical Report.* Duke University, 1998.

[103] S. K. Bhattacharya. Bayesian approach to life testing and reliability estimation. *Journal of the American Statistical Association*, 62(317):48–62, 1967.

[104] J. Aldrich. The statistical education of harold jeffreys. *International Statistical Review*, 73(2):289–307, 2005.

[105] R. E. Kass and L. Wasserman. *Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography. Techinical Report.* Carnegie Mellon University, 1994.

[106] J. G. Ibrahim and P. W. Laud. On bayesian analysis of generalized linear models using jeffreys' prior. *J. Amer. Statist. Assoc.*, 86:981–986, 199i.

[107] F. Tuyl, R. Gerlach, and K. Mengersen. A comparison of bayes-laplace, jeffreys, and other priors: The case of zero events. *The American Statistician*, 62(1):40–44, 2008.

[108] M. Papathomas and R. J. Hocking. Bayesian updating for binary variables: An application in the uk water industry. *The Statistician*, 52(4):483–499, 2003.

[109] V. E. McGee. *Principles of Statistics: Traditional and Bayesian.* Meredith Co., New York, 1971.

[110] M. Zhu and A. Y. Lu. The counter-intuitive non-informative prior for the bernoulli family. *Journal of Statistics Education*, 12(2), 2004.

[111] J. M. Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society, Series B*, 41:113–147, 1979.

[112] J. M. Bernardo. Reference analysis. *Handbook of Statistics. Dey, D. K. y Rao, C. R.*, 25:17–90, 2005.

[113] B. Clarke and D. Sun. Reference priors under the chi-squared distance. *Sankhya: The Indian Journal of Statistics, Series A*, 59(2):215–231, 1997.

[114] B. Liseo. The elimination of nuisance parameters. *Handbook of Statistics. Dey, D. K. y Rao, C. R.*, pages 193–219, 2005.

[115] J. Albert. Nuisance parameters and the use of exploratory graphical methods in a bayesian analysis. *The American Statistician*, 43(4):191–196, 1989.

[116] R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

[117] A. Elfessi and D. M. Reineke. A bayesian look at classical estimation: The exponential distribution. *Journal of Statistics Education*, 9(1), 2001.

[118] A. J. Rossman, T.H. Short, and M. T. Parks. Bayes estimators for continuous uniform distribution. *Journal of Statistics Education*, 6(3), 1998.

[119] R. J. Serfling. *Aproximation Theorems of Mathematical Statistics*. John Wiley Sons, New York, 1980.

[120] G. Canavos. *Probabilidad y Estadística: Aplicaciones y Métodos*. McGraw Hill, Madrid, 5 edition, 1998.

[121] T. H. Wonnacott and R. J. Wonnacott. *Fundamentos de Estadística para Administración y Economía*. Limusa, México, 1979.

[122] G. G. Roussas. *A First Course in Mathematical Statistics*. Addison-Wesley Publishing Company: Reading, Massachusetts, 1973.

[123] J. G. Kalbfleish. *Probability and Statistical Inference*. Springer-Verlag, New York, 2 edition, 1985.

[124] B. Efron. Computers and theory of statistics: Thinking the unthinkable. *SIAM Review*, 21:460–480, 1979.

[125] T. Wright. When zero defectives appear in a sample: Upper bounds on confidence coefficients of upper bounds. *The American Statitician*, 44(1):40–41, 1990.

[126] R. E. Kass. Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42(5):551–560, 1993.

[127] S. Sinharay and H.S. Stern. On the sensitivity of bayes factors to the prior distributions. *The American Statistician*, 56(3):196–201, 2002.

[128] F. De Santis and F. Spezzaferri. Methods for default and robust bayesian model comparison: the fractional bayes factor approach. *International Statistical Review*, 67(4):267–286, 1999.

[129] S. K. Sahu. *Bayesian Statistics. Lecture Notes, Faculty of Mathematical Studies*. University of Southhampton, 2000.

[130] M. Aitkin. Posterior bayes factors. *Journal of the Royal Statistical Society, Series B (Methodological)*, 53(1):111–142, 1991.

[131] A. O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138, 1995.

[132] E. G. Tsionas. Bayesian inyternational evidence on heavy tails, non-stationary and asymmetry over the business cycle. *International Statistical Review*, 71(1):151–168, 2003.

[133] C. Han and B. P. Carlin. Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

[134] P. Diaconis, K. Khare, and L. Saloff-Coste. Gibbs sampling, conjugate priors and coupling. *Sankhya: The Indian Journal of Statistics, Series A (2008-)*, 72(1):136–169, 2010.

[135] A. E. Raftery. *Bayesian Model Selection in Social Research (with Discussion by Andrew Gelman & Donald B. Rubin, and Robert M. Hauser, and a Rejoinder). Technical Report*. Dept. of Sociology, University of Washington, Washington, 1994.

[136] R. T. Rust and D. C. Schmittlein. A bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. *Marketing Science*, 4(1):20–40, 1985.

[137] B. Cooil, R. S. Winer, and D. L. Rados. Cross-validation for prediction. *Journal of Marketing Research*, 24(3):271–279, 1987.

[138] G. Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[139] R. W. Miller. Parametric empirical bayes tolerance intervals. *Technometrics*, 31(4):449–459, 1989.

[140] A. P. Basu and S. E. Rigdon. Examples of parametric empirical bayes methods forthe estimation of failure processes for repairable systems. *Actuarial Research Clearing House*, 2:179–195, 1985.

[141] M. Woodward. *Epidemiology. Study Design and Data Analysis*. Chapman & Hall/CRC., Florida, 1999.

[142] S. P. Brooks. Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):60–100, 1998.

[143] M. J. Lombardi. Bayesian inference for α-stable distributions: A random walk mcmc approach. *Computational Statistics & data Analysis*, 51(5):2688–2700, 2007.

[144] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, 2000.

[145] A. E. Gelfand. Gibbs sampling. *Journal of the American Statistical Association*, 95(452):1300–1304, 2000.

[146] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Appl. Statist.*, 41(2):337–348, 1992.

[147] A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985, 1990.

[148] E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[149] K. S. Chan. Asymptotic behavior of the gibbs sampler. *Journal of the American Statistical Association*, 88(421):320–326, 1993.

[150] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *The Statistician*, 52(3):381–393, 2003.

[151] S. K. Upadhyay, N. Vasishta, and A. F. M. Smith. Bayes inference in life testing and reliability via markov chain monte carlo simulation. *Sankhya: The Indian Journal of Statistics, Series A*, 62(2):203–222, 2000.

[152] D. Kundu and R. D. Gupta. Generalized exponential distribution: Bayesian estimations. *Computational Statistics & Data Analysis*, 52(4):1873–1883, 2008.

[153] M. Z. Raqab and M. Madi. Bayesian inference for the generalized exponential distribution. *Journal of Statistical Computation and Simulation*, 75(10):841–852, 2005.

[154] R. Natarajan and C. E. McCulloch. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7(3):267–277, 1998.

[155] A. Justel and D. Peña. Gibbs sampling will fail in outlier problems with strong masking. *Journal of Computational and Graphical Statistics*, 5(2):176–189, 1996.

[156] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika Trust*, 57(1):97–109, 1970.

[157] N. Metropolis, A. W. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[158] D. B. Hitchcock. A history of the metropolis-hastings algorithm. *The American Statistician*, 57(4):254–257, 2003.

[159] G. L. Jones and J. P. Hobert. Honest exploration of intractable probability distributions via markov chain monte carlo. *Statistical Science*, 16(4):312–334, 2001.

[160] A. E. Raftery and S. Lewis. *How Many Iterations in the Gibbs Sampler? Technical Report*. Dept. of Statistics, University of Washington, Washington, 1991.

[161] G. Casella and I. E. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[162] P. L. Gupta, R. C. Gupta, and R. C. Tripathi. Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis*, 23:207–218, 1996.

[163] B. Yu. A bayesian mcmc approach to survival analysis with doubly-censored data. *Computational Statistics & data Analysis*, 54(8):1921–1929, 2010.

[164] G. M. Tallis. Approximate maximum likelihood estimates from grouped data. *Technometrics*, 9(4):599–606, 1967.

[165] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[166] S. Sinharay. *Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review*. Research Report RR-03-07. ETS. Princeton, NJ 08541, 2003.

[167] S. Sinharay. Experiences with markov chain monte carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, 29(4):461–488, 2004.

[168] A. Zellner and C. Min. Gibbs sampler convergence criteria. *Journal of the American Statistical Association*, 90(431):921–927, 1995.

[169] N. F. Zhang. A statistical control chart for stationary process data. *Technometrics*, 40(1):24–38, 1998.

[170] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 1992.

[171] P. H. Garthwaite and J. M. Dickey. Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society. Series B*, 50(3):462–474, 1988.

[172] P. H. Garthwaite and J. M. Dickey. Elicitation of prior distributions for variable-selection problems in regression. *The Annals of Statistics*, 20(4):1697–1719, 1992.

[173] W. H. Jefferys and J. O. Berger. Ockham's razor and bayesian analysis. *American Scientist*, 80:64–72, 1992.

[174] T. A. Loredo. *Bayesian Inference: A Practical Primer.* Dept. of Astronomy,Cornell University, 199*.

[175] D. Peña and I. Guttman. Comparing probabilistic methods for outlier detection in linear models. *Biometrika*, 80(3):603–610, 1993.

[176] H. Ishwaran. *Applications of Hybrid Monte Carlo to Bayesian generalized Linear Models: Quasicomplete Separation and Neural Networks. Dept. of Biostatistics and Epidemiology.* The Cleveland Clinic Foundation, 1997.

[177] E. J. Bedrick, R. Christensen, and W. Johnson. Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, 51(3):211–218, 1997.

[178] S. A. Al-Awadhi and P. H. Garthwaite. Quantifying expert opinion for modelling fauna habitat distributions. *Computational Statistics*, 21:121–140, 2006.

[179] P. Dellaportas and F. M. Smith. Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Applied Statistics*, 42(3):443–459, 1993.

[180] H. Milicer and F. Szczotka. Age at menarche in warsaw girls in 1965. *Human Biology*, 38:199–203, 1966.

[181] R. L. Smith. Bayesian and frequentist approaches to parametric predictive inference. *Bayesian Statistics*, 6, 1998.

[182] G. Koop. *Review of: Bayesian Analysis, Computation and Communication Software. Techinical Report.* Dept. of Economics, University of Edinburgh, 1999.

Juan Carlos Correa Morales
Docente Asociado, Escuela de Estadística
Universidad Nacional de Colombia Sede Medellín
Medellín-Colombia
Email: jccorrea@unal.edu.co

Carlos Javier Barrera Causil
Docente Titular, Facultad de Ciencias Exactas y Aplicadas
Instituto Tecnológico Metropolitano -ITM-
Medellín-Colombia
Email: carlosbarrera@itm.edu.co

Introducción a la Estadística Bayesiana

Fuentes tipográficas: Computer Modern para texto corrido, en 12 puntos, para títulos en Computer Modern, en 24 puntos y subtítulos

La relevancia que ha tomado la estadística bayesiana en distintas áreas lleva a escribir este texto, cuyo objetivo es contribuir en el crecimiento de los métodos bayesianos en América Latina e incentivar a los estudiantes a aplicar dichas herramientas en sus investigaciones. Aquí, se presentan los elementos básicos de la estadística bayesiana, estadística bayesiana computacional y aplicaciones. Esta estructura contiene en total 14 capítulos que ilustran al lector en un gran número de procedimientos. El lector puede solicitar al correo electrónico de los autores la información correspondiente de las bases de datos necesarias para implementar paso a paso los códigos de R y OpenBUGS presentados en esta obra.

The new relevance of Bayesian statistics in different fields led to the creation of this text. Its two goals are contributing to the growth of Bayesian methods in Latin America and encouraging students to use such tools in their research projects. It presents basic elements of Bayesian statistics, computational Bayesian statistics and their applications. It is divided into 14 chapters that instruct in a great deal of procedures. The reader may email the authors requesting the corresponding information of the necessary databases to implement the R and OpenBUGS codes herein step by step.

Institución Universitaria
Acreditada en Alta Calidad

ISBN 978-958-5414-24-2