Universidad Tecnológica de la Mixteca

Selección bayesiana de modelos aplicada al ANOVA usando distribuciones a priori intrínsecas

Tesis para obtener el título de:

Licenciado en Matemáticas Aplicadas

Presenta:

Iván Méndez García

Directora de Tesis:

M. C. Norma Edith Alamilla López

Huajuapan de León, Oaxaca. Febrero del 2007

Índice general

Dedicatoria

Esta tesis se la dedico a mis padres, Ricardo y Esperanza, por su comprensión, confianza y apoyo incondicional que me han brindado en todo momento de mi carrera, por todos los sacrificios que hiceron para que pudiera realizar mis estudios de Licenciatura.

A mis hermanos, Oscar y Eder, por escucharme y porque de alguna manera me han apoyado para que siga adelante. Que esta tesis sea un incentivo para que ellos le sigan echando ganas a la escuela y puedan terminar sus respectivas carreras.

Un agradecimiento especial a mi directora de tesis, la M.C. Norma Edith Alamilla López, por encontar un tema a mi medida, por darme el empuje para terminar ésta tesis, por transmitirme esa energía y gusto por la estadística; asi como, por sus consejos, paciencia y amistad que me ha brindado.

A mis maestros, por haberme enseñado a querer las matemáticas, por compartir sus conocimientos conmigo; asi también, por sus buenos consejos y por todo lo bueno que me han enseñado durante toda la carrera.

A mis amigos, Alfredo, Jesus, Kanain, Miguel, Zavaleta, Iván López y Ricardo, pues con ellos comparti buenos y malos momentos durante mi estancia en la universidad; por su compañía, por brindarme su amistad, por escucharme, animarme y aconsejarme en los momentos dificiles.

A mis sinodales, M.C. José del Carmen Jiménez Hernández, M.C. José Margarito Hernández Morales, M.C. Tirso Miguel Angel Ramírez Solano, quienes participaron en la revisión de esta tesis; por sus observaciones y recomendaciones.

A los amigos y familiares que se me olvida nombrar y que de alguna forma me dieron animos para seguir adelante.

Y finalmente a la Universidad Tecnológica de la Mixteca, por brindarme la oportunidad de realizar mis estudios en la Licenciatura en Matemáticas Aplicadas.

Capítulo 1

Introducción

La interpretación de gran parte de las investigaciones en ingeniería y ciencias de la computación depende cada vez más de métodos estadísticos. La estadística se ocupa fundamentalmente del análisis de datos que presentan variabilidad con el fin de comprender el mecanismo que los genera, o para ayudar en un proceso específico de toma de decisiones.

En cualquiera de los casos existe una componente de incertidumbre involucrada, por lo cual el estadístico se ocupa en reducir lo más posible esa componente, así como también en describirla de forma apropiada.

Además, se supone que toda forma de incertidumbre debe describirse por medio de modelos de probabilidad, y más aún la probabilidad es el único lenguaje posible para describir una lógica que trata con todos los niveles de incertidumbre y no solo con los extremos de verdad o falsedad, este enfoque se conoce como estadística bayesiana.

La estadística bayesiana, es un enfoque alternativo a la estadística clásica y frecuentista en la resolución de los problemas típicos estadísticos que son: estimación, contraste de hipótesis y predicción. Además, los métodos bayesianos pueden ser aplicados a problemas que han sido inaccesibles a la teoría frecuentista clásica. Utiliza el teorema de Bayes combinando la información a priori y la de los datos (función de verosimilitud), para obtener la distribución a posteriori. De la distribución a posteriori es de donde se derivan los estimadores y, cuya interpretación, difiere radicalmente de la interpretación proporcionada por la inferencia clásica.

Esta teoría está firmemente basada en fundamentos matemáticos y una metodología coherente con la que es posible incorporar información relevante. La inferencia sobre una cantidad de interés queda descrita mediante modificaciones en la incertidumbre a la luz de la evidencia, y el teorema de Bayes especifíca como estas modificaciones deben llevarse a cabo.

La estadística bayesiana, de acuerdo con la interpretación subjetiva, dice que la probabilidad que un estadístico asigna a uno de los posibles resultados de un proceso, representa su propio juicio sobre la verosimilitud de que se tenga el resultado. Este juicio estará basado en opiniones e información acerca del proceso.

La información que el estadístico tiene sobre la verosimilitud de los distintos eventos relevantes al problema de decisión debe ser cuantificada a través de una medida de probabilidad sobre el espacio muestral Θ.

Un modelo estadístico bayesiano esta formado por un modelo estadístico paramétrico, $f(x|\theta)$, y una distribución a priori sobre los parámetros, $\pi(\theta)$. Lo que se requiere es hacer inferencias sobre el verdadero valor de θ. La inferencia se basa en la distribución de θ condicional sobre x, llamada distribución a posteriori y definida mediante el teorema de Bayes por:

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

El teorema de Bayes actualiza la información sobre θ debida a la información contenida en la observación de x. Teniendo las distribuciones $f(x|\theta)$ y $\pi(\theta)$, además de obtener la distribución a posteriori de θ, también se pueden obtener, la distribución conjunta de (x, θ)

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

y la distribución marginal de x

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

Los problemas específicos de la inferencia estadística son estimación puntual, estimación por intervalos y pruebas de hipótesis. Desde el punto de vista bayesiano se dice que la inferencia sobre θ, debe depender estrictamente de la distribución a posteriori $\pi(\theta|x)$, por lo que esta puede ser usada para describir las propiedades de θ. En el caso de la estimación puntual pueden ser usadas la media, la moda, la varianza, la mediana como estimadores de θ, dependiendo de las condiciones del problema.

En el caso de estimación por intervalos, el conocimiento de la distribución a posteriori también se utiliza para la determinación de regiones de confianza. Similarmente es posible determinar de manera natural la probabilidad de una hipótesis nula H_0, condicionado sobre las observaciones.

La planeación y análisis de un experimento es mucho más fácil para el estadístico si puede asumirse que las observaciones son elegidas de una familia de distribuciones en la cual existe un estadístico suficiente de dimensión fija.

En ciertos problemas, el conocimiento inicial sobre el verdadero valor del parámetro θ puede ser muy débil, o vaga, ésto ha llevado a generar un tipo de distribuciones iniciales llamada *distribuciones a priori no informativas*, las cuales reflejan un estado de ignorancia inicial.

Las ventajas principales del enfoque bayesiano son su generalidad y coherencia, pues todos los problemas de inferencia se resuelven con los principios de cálculo de probabilidades y teoría de decisión. Además, posee una capacidad de incorporar información a priori adicional que se tenga del parámetro, a la muestra.

Esta última también es una desventaja, pues algunos investigadores rechazan que la información inicial se incluya en un proceso de inferencia científica. Pero esta situación se puede evitar estableciendo una distribución a priori no informativa o de referencia, la cual se introduce cuando no se posee mucha información previa acerca del problema.

A un problema específico se le puede asignar cualquier tipo de distribución a priori, ya que finalmente al actualizar la información a priori que se tenga acerca del parámetro, mediante el teorema de Bayes y obtener la distribución a posteriori del parámetro, es con ésta distribución con la que se hacen las inferencias acerca del mismo.

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

Desde el enfoque clásico o frecuentista esta regla de decisión esta basada en las características del estadístico de contraste y en las de su distribución muestral, y la evaluación de las hipótesis se realiza en términos de las probabilidades de dos tipos de error. Estas probabilidades de error representan la "casualidad" de que se observe una muestra para la cual el contraste acepte la hipótesis errónea.

Desde el enfoque bayesiano la tarea de evaluación de las hipótesis resulta conceptualmente más sencilla. La regla de decisión se basa en el cálculo de las probabilidades a posteriori de las hipótesis contrastadas y su evaluación depende de los resultados obtenidos. La ventaja conceptual de este último análisis es que dichas probabilidades a posteriori son las verdaderas probabilidades de las hipótesis que reflejan los datos observados y la distribución a priori.

Se abordará el problema de contraste de hipótesis asumiendo que las varianzas de las distribuciones normales son distintas, pero cuyo cociente es conocido, a este procedimiento se le conoce como análisis de varianza bajo heterocedasticidad; así también se analizará el análisis de varianza bajo homocedasticidad, el cual a diferencia del anterior se supone que las varianzas de las distribuciones normales son iguales. Se han propuesto distintos procedimientos de contraste pero la mayoría de ellos han sido tema de controversia respecto a su validez o utilidad.

Aunque las probabilidades a posteriori de las hipótesis son las principales medidas en los problemas de contraste, se abordará un concepto de gran interés: el factor de Bayes; el cual utilizaremos para la comparación de modelos.

En el contexto de selección de modelos o comparación de modelos, éste se considera como un problema estadístico. Existen diferentes criterios para resolver el problema. En Bernardo y Smith (1994) se plantean tres criterios de éstos.

El primero, el cual es llamado criterio $\mathcal{M}$-cerrado, se cree que uno de los modelos $\{M_i, i \in I\}$, es el verdadero, sin el conocimiento explicito de cuál de ellos es el verdadero modelo. Desde este criterio tiene sentido asignar probabilidades a los modelos, cuyas probabilidades son las que cree el estadístico, por su experiencia que tiene acerca de los mismos.

El segundo criterio es el llamado $\mathcal{M}$-completo, corresponde a una actuación individual, como si $\{M_i, i \in I\}$, fuera un conjunto de modelos específicos, disponibles para comparación de modelos. Desde esta perspectiva no tiene sentido asignar probabilidades a los modelos.

El tercer criterio, llamado $\mathcal{M}$-abierto, también se considera que el conjunto de modelos son simplemente un rango de modelos específicos, disponibles para comparación, tampoco tiene sentido asignarles probabilidades a los modelos, puesto que no se tiene creencia de que el verdadero modelo esté en el conjunto considerado.

En nuestro caso, se trabajará con el criterio $\mathcal{M}$-cerrado. Considerando el criterio $\mathcal{M}$-cerrado, tradicionalmente los procedimientos para la solución de problemas de selección de modelos se basan en los llamados *Factores de Bayes.* El interés ha sido desarrollado cuando se trabaja con distribuciones a priori impropias, ya que éstos no están definidos, con éste tipo de distribución a priori. Hay algunos métodos propuestos por Aikin(1991), O'Hagan (1995), Berger y Pericchi (1996), entre otros, los cuales tratán de resolver éste problema.

Aikin (1991), propone el uso de medias de las distribuciones a posteriori de la verosimilitud bajo cada modelo en lugar de la usual media inicial. Aikin, reitera que el uso de la media final tiene una gran ventaja, incluyendo la reducción de variación en la distribución a priori. Sin embargo éste procedimiento ha sido criticado severamente por algunos estadísticos entre ellos Lindley, quien dice que éste no es recomendable, él propone un ejemplo en donde el uso de los factores de Bayes finales, dan un resultado ilógico e inconsistente, el lector interesado en este caso puede consultar la discusión.

O'Hagan (1995) describe para comparación de modelos una alternativa a los factores de Bayes parciales, los cuales ofrecen una solución del problema eliminando parte de los datos y tomando ésta como una muestra adicional. Esta muestra adicional es usada para obtener una distribución a posteriori informativa que se supone inicial de los parámetros en cada modelo, calculando entonces los

factores de Bayes sobre el resto de los datos.

Una modificación de los factores de Bayes parciales, son los factores de Bayes fraccionales. En general los factores de Bayes fraccionales son preferibles por su gran robustes y su conformidad con el principio de verosimilitud.

Esta alternativa ha causado gran polémica entre los bayesianos, ya que algunos opinan que O'Hagan proporciona una solución elegante al problema mientras que otros opinan que no es alternativa que convenza, de hecho Lindley da un contra ejemplo para la metodología Bayes fraccionales, el lector que quiera ahondar en este tema, lo puede hacer en el artículo de O'Hagan (1995).

Por otro lado Berger y Pericchi (1996), proponen una solución para el problema de comparación de modelos, llamado factores de Bayes intrínsecos, el cual está basado sobre los datos y distribuciones a priori no informativas estándar, y tomando en cuenta distribuciones a priori intrínsecas, éste metodo parece corresponder al método de factores de Bayes, al menos asintóticamente. Puede ser usado para modelos anidados o no anidados y para comparación múltiple de modelos y predicción; esta tesis desarrollará la propuesta de Berguer y Pericchi en el capítulo 4.

El enfoque bayesiano clásico se aproxima al análisis de varianza asumiendo la condición de homocedasticidad y usando distribuciones a priori uniforme convencionales para los parámetros de localización y el logaritmo del parámetro de escala común. El problema ha sido desarrollado como un problema de estimación de parámetros de localización. Moreno, Bertolino y Racugno han argumentado que esto no conduce a una solución bayesiana apropiada.

Se propone una solución basada en la selección bayesiana de modelos. Nuestro desarrollo es en el contexto general de heterocedasticidad , para el cual no existen contrastes frecuentistas. El factor de Bayes involucra el uso de distribuciones a priori intrínsecas en vez de las distribuciones a priori por defecto.

El objetivo central de la presente tesis es revisar y analizar exhaustivamente el método propuesto por Berger y Pericchi (1996), donde se propone una solución para el problema de comparación de modelos; asi tambien se analiza el articulo "Bayesian models Selection Approach to Analysis of Variance under Heterocedasticy" propuesto por Elias Moreno, Francesco Bertolino y Walter Racugno.

La tesis está estructurada de la manera siguiente:

En el segundo capítulo, se presenta una introducción a la teoría de la estadística bayesiana, así como también se definen algunos conceptos básicos del análisis bayesiano, se abordan problemas típicamente estadísticos, como son: estimación, y predicción, los cuales son necesarios para abordar los temas que se presentan a lo largo de la tesis.

En el tercer capítulo se analizan temas como contraste de hipótesis y análisis de varianza, los cuales serán empleados en los siguientes capítulos.

En el cuarto capítulo, se analizarán los "factores de Bayes intrínsecos". Como es bien sabido, en el enfoque bayesiano para problemas de selección de modelos y contraste de hipótesis la herramienta principal son los factores de Bayes. En el caso en que al definir los factores de Bayes, se usa una distribución a priori impropia, éstos quedan bien definidos salvo una constante multiplicativa. Los "factores de Bayes intrínsecos" (FBI), definidos por Berger y Pericchi (1996), es un método interesante para resolver dicho problema, por lo que se hace un análisis exhaustivo del método.

En el capítulo cinco, se hará un análisis de varianza desde el enfoque bayesiano asumiendo homocedasticidad, así como también un análisis de varianza desde el enfoque bayesiano asumiendo heterocedasticidad.

Por último se presentan las conclusiones, así como algunos comentarios. También se incluye al final una sección de la notación y abreviaturas usadas en ésta tesis.

Para las soluciones numéricas de este trabajo se usó el paquete computacional Mathematica.

Capítulo 2

Conceptos básicos

2.1. Probabilidad subjetiva

El concepto clásico de probabilidad envuelve una larga secuencia de repeticiones de una situación dada. Por ejemplo decir que una moneda legal (balanceada) tiene probabilidad $1/2$ de caer cara cuando fue lanzada, quiere decir que, en una serie larga de lanzamientos independientes de la moneda, el resultado cara ocurre aproximadamente la mitad de las veces. Desafortunadamente, este concepto frecuentista no es suficiente cuando se asignan probabilidades alrededor de una varia-ble aleatoria θ. Por ejemplo, consideremos el problema de intentar determinar θ, la proporción de fumadores en alguna ciudad de México. ¿Qué significa que $P(0.3 < \theta < 0.35) = 0.5$? Aquí θ es simplemente algún número que nosotros no conocemos. Claramente este se encuentra en el intervalo $(0.3, 0.35)$ con probabilidad 0.5. Aquí no hay nada aleatorio. Como un segundo ejemplo sea θ que representa la proporción de desempleados para el siguiente año en alguna ciudad de la República Mexicana, en este caso pensamos en θ como una variable aleatoria, mientras el futuro es incierto. Ahora ¿puede $P(0.03 < \theta < 0.04)$ ser interpretada en términos de una secuencia de situaciones idénticas? La teoría de probabilidad subjetiva ha sido creada para hablar acerca de probabilidades cuando el punto de vista frecuentista no es aplicable (algunos argumentan que el concepto frecuentista nunca se aplica, esto es porque es imposible tener una sucesión infinita de repeticiones independientes e identicamente distribuidas ($i.i.d$) de cualquier situación).

La idea principal de la probabilidad subjetiva es permitir que la probabilidad de un evento refleje la creencia personal en el sentido de la ocurrencia de dicho evento. El cálculo de probabilidades frecuentistas es teóricamente directa; uno simplemente determina la frecuencia relativa del evento de interés. Una pro - babilidad subjetiva es típicamente determinada por introspección. El camino mas

simple para calcular probabilidades subjetivas es comparar eventos determinando probabilidades relativas. Es decir, por ejemplo, si queremos encontrar $P(E)$, simplemente comparamos E con E^c (el complemento de E). Si es dos veces mas factible que ocurra E que E^c, entonces claramente $P(E) = \frac{2}{3}$ y $P(E^c) = \frac{1}{3}$.

2.2. Distribuciones a priori no informativas

2.2.1. Introducción

Las distribuciones a priori no informativas reflejan un estado de ignorancia inicial. Por ejemplo, en el contraste de dos hipótesis simples, la distribución a priori que asigna una probabilidad de 1/2 a cada hipótesis, podría considerarse no informativa . Un ejemplo más complejo sería el siguiente:

Ejemplo 2.1 *Supóngase que la variable de interés es una normal con media θ, además que el espacio del parámetro es $\Theta = (-\infty, \infty)$. Si una densidad a priori no informativa es deseada, es razonable dar pesos iguales a todos los posibles valores de θ. Desafortunadamente, si $\pi(\theta) = c > 0$ es elegida, entonces $\pi(\theta)$ tiene masa infinita (esto es $\int \pi(\theta)d\theta = \infty$) y esta no es una densidad propia. Sin embargo, se puede trabajar con tal $\pi(\theta)$, satisfactoriamente. La elección de c no es importante, además típicamente la densidad a priori no informativa para este problema es elegida como $\pi(\theta) = 1$. Frecuentemente la llamamos* la densidad uniforme *sobre R^1, la cual fue introducida por Laplace (1812).*

Como en el ejemplo anterior, con frecuencia ocurre que la distribución a priori no informativa natural es una distribución a priori impropia, es decir, que tiene "masa infinita".Enseguida consideraremos el problema de determinar distribuciones a priori no informativas.

La situación más simple es considerar cuando Θ es un conjunto finito, consistente de n elementos. La distribución a priori no informativa obvia es asignar a cada elemento de Θ una probabilidad de $1/n$. Uno puede generalizar esto a Θ infinito, dando a cada $\theta \in \Theta$ densidades iguales, llegando a la distribución a priori no informativa uniforme $\pi(\theta) \equiv c$. Esta distribución fue introducida por Laplace (1812) y fue criticada duramente debido a una falta de invarianza bajo transformación (ver Jefreys (1983)).

Ejemplo 2.2 *En lugar de considerar θ, supóngase que el problema ha sido parametrizado en términos de $\eta = \exp\{\theta\}$. Ésta es una transformación uno a uno. Pero si $\pi(\theta)$ es la densidad para θ, entonces la correspondiente densidad para η es (nótese que el Jacobiano de la transformación es $d\theta/d\eta = d\log\eta/d\eta = \eta^{-1}$)*

$$\pi^*(\theta) = \eta^{-1}\pi(\log\eta).$$

Por lo tanto, si la distribución a priori no informativa para θ es elegida como una constante, debemos elegir la distribución a priori no informativa para η proporcional a η^{-1} para mantener consistencia. No podemos mantener consistencia y elegir para ambas distribuciones a priori no informativas θ y η, que sean constantes.

2.2.2. Distribuciones a priori en problemas de localización y escala

Los esfuerzos para derivar distribuciones a priori no informativas considerando los problemas de invarianza bajo transformación comenzaron con Jeffreys (1961).

Definición 2.1 *(Parámetros de localización). Supóngase que χ y Θ son subconjuntos de R^p, y que la densidad de X es de la forma $f(x-\theta)$ (esto es, solo depende de $(x-\theta)$). La densidad entonces se dice que es una* densidad de localización, *y θ es llamado un* parámetro de localización *(o en algunas ocasiones un vector de localización cuando $p \geq 2$).*

Ejemplo 2.3 *La distribución normal con media θ y varianza σ^2 ($\mathcal{N}(\theta, \sigma^2)$), con σ^2 fija; la distribución t con alfa grados de libertad, parámetro de localización μ y parámetro de escala σ^2 ($t(\alpha, \mu, \sigma^2)$), con α y σ^2 fijo; y la normal p-variada, con media μ y matriz de covarianza $\sum$, ($\mathcal{N}_p(\theta, \sum)$) con $\sum$ fija, son todas ejemplos de densidades de localización. Además, una muestra de variables aleatorias i.i.d. se dice que tiene una densidad de localización si su densidad común es una densidad de localización.*

Para derivar una distribución a priori no informativa para esta situación, imagine que, en lugar de haber observado X, hemos observado la variable aleatoria $Y = X + c$ ($c \in R^p$). Definiendo $\eta = \theta + c$, es claro que Y tiene densidad $f(y-\eta)$. Si ahora $\chi = \theta = R^p$, entonces el espacio muestral y el espacio paramétrico para los (Y, η) están también en R^p. Los problemas (X, θ) y (Y, η), son idénticos en estructura, y es razonable insistir en que ellos tienen la misma distribución a priori no informativa.

Sea π y π^* que denotan las distribuciones a priori no informativas en los problemas (X, θ) y (Y, η), respectivamente, el mismo argumento implica que π y π^* deben ser iguales, esto es

$$P^{\pi}(\theta \in A) = P^{\pi^*}(\eta \in A) \tag{2.1}$$

para cualquier conjunto A en R^p. Mientras $\eta = \theta + c$, debe ser verdadero (por un simple cambio de variables) que

$$P^{\pi^*}(\eta \in A) = P^{\pi}(\theta + c \in A) = P^{\pi}(\theta \in A - c), \tag{2.2}$$

donde $A - c = \{z - c : z \in A\}$. Combinando (2.1) y (2.2) se tiene que

$$P^{\pi}(\theta \in A) = P^{\pi}(\theta \in A - c). \tag{2.3}$$

Además, este argumento se aplica para $c \in R^p$, así que (2.3) se cumple para todo $c \in R^p$. Cualquier distribución π que satisface esta relación se dice que es una *distribución a priori invariante de localización.* Asumiendo que la distribución a priori tiene una densidad, entonces podemos escribir la ecuación (2.3) como

$$\int_A \pi(\theta)d\theta = \int_{A-c} \pi(\theta)d\theta = \int_A \pi(\theta - c)d\theta.$$

Si esto se cumple para todo A, entonces se puede demostrar que es verdadero

$$\pi(\theta) = \pi(\theta - c)$$

para todo θ. Y al hacer $\theta = c$ se tiene

$$\pi(c) = \pi(0).$$

La conclusión es que π debe ser una función constante . Es conveniente elegir la constante igual a 1, así *la distribución a priori no informativa para un parámetro de localización* es $\pi(\theta) = 1$.

Definición 2.2 *(parámetros de escala). Una (unidimensional) densidad de escala es una densidad de la forma*

$$\left(\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)\right),$$

donde $\sigma > 0$. El parámetro σ es llamado un parámetro de escala.

Ejemplo 2.4 *La $\mathcal{N}(0, \sigma^2)$ y $t(\alpha, 0, \sigma^2)$ (α fijo) son ejemplos de densidades de escala. Además una muestra de variables aleatorias i.i.d. se dicen que tienen una densidad de escala si su densidad común es una densidad de escala.*

Para derivar una distribución a priori no informativa para ésta situación, imagine que, en lugar de haber observado X, hemos observado la variable aleatoria $Y = cX (c > 0)$. Definiendo $\eta = c\sigma$, y efectuando algunos cálculos se muestra que la densidad de Y es $\eta^{-1} f(y/\eta)$. Si ahora $\chi = R^1$ o $\chi = (0, \infty)$, entonces el espacio muestral y parámetrico para el problema (X, σ) es el mismo que para el problema

(Y, η). Los dos problemas son idénticos en estructura, lo cual nuevamente indica que ellos deben tener la misma distribución a priori no informativa.

Sea π y π^* que denotan las distribuciones a priori no informativas en los problemas (X, σ) y (Y, η), respectivamente, esto quiere decir que la igualdad

$$P^{\pi}(\sigma \in A) = P^{\pi^*}(\eta \in A)$$

debe cumplirse para todo $A \subset (0, \infty)$. Mientras que $\eta = c\sigma$, debe ser verdadero que

$$P^{\pi^*}(\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A)$$

donde $c^{-1}A = \{c^{-1}z : z \in A\}$. De lo anterior tenemos que se debe satisfacer

$$P^{\pi}(\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A). \tag{2.4}$$

Esta ecuación se debe cumplir para toda $c > 0$. Cualquier distribución $\pi(\theta)$ para la cual esto es verdadero es llamada escala invariante. Rescribiendo (2.4) (asumiendo densidades) como

$$\int_A \pi(\sigma)d\sigma = \int_{c^{-1}A} \pi(\sigma)d\sigma = \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma.$$

Concluimos que para todo A y toda σ debe ser verdadero que

$$\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma).$$

Eligiendo $\sigma = c$, se tiene que

$$\pi(c) = c^{-1}\pi(1).$$

Tomando $\pi(1) = 1$ por conveniencia y notando que la igualdad debe ser para todo $c > 0$, se sigue que una distribución a priori no informativa razonable para un parámetro de escala es $\pi(\sigma) = \sigma^{-1}$. Observe que ésta además es una a distribución a priori impropia, esto es que $\int_0^{\infty} \sigma^{-1}d\sigma = \infty$.

2.2.3. Distribuciones a priori no informativas en conjuntos generales

Para problemas más generales, se han hecho varias propuestas para determinar distribuciones a priori no informativas . El método usado más ampliamente considerado es el propuesto por Jeffreys (1961), que elige

$$\pi(\theta) = [I(\theta)]^{1/2} \tag{2.5}$$

como una distribución a priori no informativa, donde $I(\theta)$ es la información esperada de Fisher, que está dada por

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

Si $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ es un vector, Jeffreys (1961) sugiere el uso de

$$\pi(\theta) = [\det \mathbf{I}(\boldsymbol{\theta})]^{1/2} \tag{2.6}$$

donde $\mathbf{I}(\boldsymbol{\theta})$ es la $(p \times p)$ matriz de información esperada de Fisher, que bajo condiciones que comúnmente se satisfacen, es la matriz con elementos (i, j)

$$I_{ij}(\boldsymbol{\theta}) = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\theta) \right].$$

Ejemplo 2.5 *(Parámetros de localización - escala). Una densidad de localización - escala es una densidad de la forma $\sigma^{-1} f((x-\theta)/\sigma)$, donde $\theta \in R^1$ y $\sigma > 0$ son los parámetros desconocidos.*

Ejemplo 2.6 *La $\mathcal{N}(\theta, \sigma^2)$ es un ejemplo crucial de densidad de localización - escala. Una muestra de variables aleatorias i.i.d se dice que tienen una densidad de localización - escala si hay una densidad común que es una densidad de localización - escala.*

Trabajando con la distribución normal por simplicidad, y notando que $\boldsymbol{\theta} = (\theta, \sigma)$ en este problema, la matriz de información de Fisher es

$$\begin{aligned} I(\boldsymbol{\theta}) &= -E_\theta \begin{pmatrix} \frac{\partial^2}{\partial \theta^2}\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) & \frac{\partial^2}{\partial \theta \partial \sigma}\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \\ \frac{\partial^2}{\partial \theta \partial \sigma}\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) & \frac{\partial^2}{\partial \sigma^2}\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \end{pmatrix} \\ &= -E_\theta \begin{pmatrix} -1/\sigma^2 & 2(\theta - X)/\sigma^3 \\ 2(\theta - X)/\sigma^3 & -3(X-\theta)^2/\sigma^4 \end{pmatrix} \\ &= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 3/\sigma^2 \end{pmatrix} \end{aligned}$$

por lo tanto

$$\pi(\theta) = \left(\frac{1}{\sigma^2} \cdot \frac{3}{\sigma^2} \right)^{1/2} \propto \frac{1}{\sigma^2}.$$

2.2.4. La distribución marginal

Si X tiene una densidad de probabilidad $f(x|\theta)$, y θ tiene una densidad de probabilidad $\pi(\theta)$, entonces la densidad conjunta de X y θ es

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

Definición 2.3 *La densidad marginal de X es*

$$m(x) = \int_\Theta f(x|\theta)dF^\pi(\theta) = \begin{cases} \int_\Theta f(x|\theta)\pi(\theta)d\theta & \textit{(caso continuo)} \\ \sum_\Theta f(x|\theta)\pi(\theta) & \textit{(caso discreto).} \end{cases} \tag{2.7}$$

La cual se le conoce como densidad marginal a priori.

2.2.5. La distribución a posteriori

El análisis bayesiano se hace combinando la información a priori $\pi(\theta)$ y la información muestral x dando como resultado la así llamada distribución a posteriori de θ dado x, sobre la cual se basan todas las decisiones e inferencias.

La distribución a posteriori de θ dado x, se define como la distribución condicional de θ dada la observación x, y se denota por $\pi(\theta|x)$. Nótese que la densidad conjunta de θ y X es

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

y que la densidad marginal de X es

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta.$$

Además es claro que (con $m(x) \neq 0$),

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

Así como la distribución a priori refleja nuestra creencia acerca de la distribución a priori θ antes de la experimentación, $\pi(\theta|x)$ refleja la creencia acerca de θ después de haber observado la muestra x.

Ejemplo 2.7 *Supóngase que $X \sim \mathcal{N}(\theta, \sigma^2)$, donde θ es desconocida pero σ^2 es*

conocida. Sea $\pi(\theta)$ *una* $\mathcal{N}(\mu, \tau^2)$*, donde* μ *y* τ^2 *son conocidos. Entonces*

$$\begin{aligned} h(x,\theta) &= \pi(\theta) f(x|\theta) = (\sqrt{2\pi}\tau)^{-1} \exp\left\{ -\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} \right] \right\} (\sqrt{2\pi}\sigma)^{-1} \\ &\quad \times \exp\left\{ -\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2} \right] \right\} \\ &= (2\pi\sigma\tau)^{-1} \exp\left\{ -\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2} \right] \right\}. \end{aligned}$$

Para encontrar $m(x)$*, note que definimos*

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

y completando cuadrados obtenemos

$$\begin{aligned} & \frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2} \right] \\ = \; & \frac{1}{2}\left[\frac{\theta^2 - 2\mu\theta + \mu^2}{\tau^2} + \frac{x^2 - 2x\theta + \theta^2}{\sigma^2} \right] \\ = \; & \frac{1}{2}\left[\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)\theta^2 - 2\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right)\theta + \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \right] \\ = \; & \frac{1}{2}\rho\left[\theta^2 - \frac{2}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right)\theta \right] + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ = \; & \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 - \frac{1}{2\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right)^2 + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ = \; & \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 - \frac{\tau^2\sigma^2}{2(\tau^2+\sigma^2)}\left(\frac{\sigma^2\mu + \tau^2 x}{\tau^2\sigma^2} \right)^2 + \frac{1}{2}\left(\frac{\sigma^2\mu^2 + \tau^2 x^2}{\tau^2\sigma^2} \right) \\ = \; & \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 - \frac{\tau^2\sigma^2}{2(\tau^2+\sigma^2)}\left(\frac{\sigma^4\mu^2 + 2\sigma^2\tau^2 x\mu + \tau^4 x^2}{\tau^4\sigma^4} \right) \\ & + \frac{1}{2}\left(\frac{\sigma^2\mu^2 + \tau^2 x^2}{\tau^2\sigma^2} \right) \\ = \; & \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 - \frac{\sigma^4\mu^2 + 2\sigma^2\tau^2 x\mu + \tau^4 x^2}{2(\tau^2+\sigma^2)\tau^2\sigma^2} + \frac{\sigma^2\mu^2 + \tau^2 x^2}{2\tau^2\sigma^2} \\ = \; & \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 \\ & + \frac{(\sigma^2\mu^2 + \tau^2 x^2)(\tau^2+\sigma^2) - (\sigma^4\mu^2 + 2\sigma^2\tau^2 x\mu + \tau^4 x^2)}{2\left(\sigma^2+\tau^2\right)\tau^2\sigma^2} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2 \\
&\quad + \frac{\sigma^2\mu^2\tau^2+\tau^4x^2+\sigma^4\mu^2+\sigma^2\tau^2x^2-\sigma^4\mu^2-2\sigma^2\tau^2x\mu-\tau^4x^2}{2\left(\sigma^2+\tau^2\right)\tau^2\sigma^2} \\
&= \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2 + \frac{\sigma^2\tau^2\left(\mu^2+x^2-2x\mu\right)}{2\left(\sigma^2+\tau^2\right)\tau^2\sigma^2} \\
&= \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2 + \frac{\left(\mu-x\right)^2}{2\left(\sigma^2+\tau^2\right)}.
\end{aligned}$$

Por lo tanto

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

y

$$\begin{aligned}
m(x) &= \int_{-\infty}^{\infty} h(x,\theta)d\theta = \frac{1}{2\pi\tau\sigma} \\
&\quad \times \int \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\exp\left\{-\frac{\left(\mu-x\right)^2}{2\left(\sigma^2+\tau^2\right)}\right\}d\theta \\
&= \frac{1}{2\pi\tau\sigma}\times\exp\left\{-\frac{\left(\mu-x\right)^2}{2\left(\sigma^2+\tau^2\right)}\right\}\times\rho^{-1/2}\sqrt{2\pi} \\
&\quad \times \int \frac{1}{\sqrt{2\pi}\rho^{-1/2}}\exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}d\theta \\
&= \left(2\pi\rho\right)^{-1/2}(\sigma\tau)^{-1}\exp\left\{-\frac{\left(\mu-x\right)^2}{2\left(\sigma^2+\tau^2\right)}\right\},
\end{aligned}$$

de aqui se sigue que

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

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

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

Como un ejemplo concreto tenemos el siguiente:

Ejemplo 2.8 *Considérese la situación en la que a un individuo se le aplica una prueba de inteligencia. Asuma que el resultado de la prueba X es $\mathcal{N}(\theta, 100)$, donde θ es el verdadero nivel de inteligencia (CI) del individuo. Supóngase además que en la población, θ se distribuye como una $\mathcal{N}(100, 225)$. Usando las ecuaciones obtenidas anteriormente, se sigue que marginalmente X se distribuye como una $\mathcal{N}(100, 325)$, mientras la distribución a posteriori de θ dado x es normal con media*

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

y varianza

$$\rho^{-1} = \frac{100(225)}{(100 + 225)} = \frac{900}{13} = 69.23.$$

Si el individuo tiene un resultado de 115 puntos en la prueba de inteligencia, su CI verdadero tiene una distribución a posteriori $\mathcal{N}(110.39, 69.23)$.

2.2.6. Familias conjugadas

En general, $m(x)$ y $\pi(\theta|x)$ no son fáciles de calcular. Si, por ejemplo, X es $\mathcal{N}(\theta, \sigma^2)$ y θ es $\mathcal{C}(\mu, \sigma)$ entonces $\pi(\theta|x)$ puede ser calculada solo numéricamente. Una gran parte de la estadística bayesiana se dedica a encontrar distribuciones a priori intrínsecas para la cual $\pi(\theta|x)$ puede ser calculada fácilmente. Estas son las llamadas distribuciones a priori conjugadas, y están desarrolladas extensivamente en Raiffa y Schlaifer (1961).

Definición 2.4 *Sea F que denota la clase de funciones de densidad $f(x|\theta)$. Una clase $\wp$ de distribuciones a priori se dice que es una familia conjugada para F si $\pi(\theta|x)$ esta en la clase $\wp$ para todo $f \in F$ y $\pi \in \wp$.*

El ejemplo (2.5) muestra que la clase de una a priori normal es una familia conjugada para la clase de densidades normales (muestras). (Si X tiene una densidad normal y θ tiene una distribución a priori normal, entonces la distribución a posteriori de θ dado X es además normal.)

Para una clase dada de densidades F, una familia conjugada puede ser determinada analizando la función de verosimilitud $l_x(\theta) = f(x|\theta)$, y eligiendo como una familia conjugada, la clase de distribuciones con la misma forma funcional de la función de verosimilitud. Las distribuciones a priori resultantes son frecuentemente llamadas *a priori conjugadas naturales* .

Ejemplo 2.9 *Supóngase que* $X = (X_1, X_2, ..., X_n)$ *es una muestra de una distribución de Poisson. Esto es* $X_i \sim \wp(\theta)$, $i = 1, 2, ..., n$ *y*

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

Aquí, F es la clase de todas las densidades de esta forma. Observando que la función de verosimilitud parece una densidad gamma es posible determinar una familia conjugada de distribuciones a priori. Esto es, suponga que $\theta \sim Ga(\alpha, \beta)$ y observe que

$$\begin{aligned} h(x, \theta) &= f(x|\theta)\pi(\theta) = \frac{e^{-n\theta}\theta^{n\bar{x}}}{\prod_{i=1}^{n} [x_i!]} \cdot \frac{\theta^{\alpha-1} e^{-\theta/\beta} I_{(0,\infty)}(\theta)}{\Gamma(\alpha)\beta^{\alpha}} \\ &= \frac{e^{-\theta(n+1/\beta)} \theta^{(n\bar{x}+\alpha-1)} I_{(0,\infty)}(\theta)}{\Gamma(\alpha)\beta^{\alpha} \prod_{i=1}^{n} [x_i!]}. \end{aligned}$$

El factor que envuelve a θ en esta última expresión es claramente reconocible como una distribución $Ga(n\overline{x} + \alpha, [n + 1/\beta]^{-1})$. Esto debe ser entonces $\pi(\theta|x)$. Por lo tanto, la clase de distribuciones gamma es verdaderamente una famila conjugada para F.

En este ejemplo, $m(x)$ puede ser determinado dividiendo $h(x, \theta)$ entre $\pi(\theta|x)$ y cancelando factores

$$m(x) = \frac{h(x, \theta)}{\pi(\theta|x)} = \frac{\left(\Gamma(\alpha)\beta^{\alpha} \prod_{i=1}^{n} [x_i!] \right)}{\{\Gamma(\alpha + n\bar{x})[n + 1/\beta]^{-(\alpha+n\bar{x})}\}^{-1}}.$$

2.3. Estimación

Los problemas de inferencia concernientes a θ pueden ser atacados fácilmente con analisis bayesiano. La idea es que, mientras la distribución a posteriori contiene toda la información promedio acerca de θ, cualquier inferencia concerniente a θ debe depender solamente de características de esta distribución.

El uso inferencial más simple de la distribución a posteriori es hallar una estimación puntual para θ, con una medida de precisión asociada.

Para estimar θ, pueden aplicarse varias técnicas clásicas a la distribución a posteriori. La técnica clásica más común es la estimación de máxima verosimilitud, la cual elige, como estimador de θ, al valor $\hat{\theta}$ que maximiza la función de verosimilitud $l(\theta) = f(x|\theta)$. La definición bayesiana del estimador de θ se define como sigue.

Definición 2.5 *El estimador generalizado de máxima verosimilitud de θ es definido como la moda más grande, $\hat{\theta}$, de $\pi(\theta|x)$ (es decir el valor $\hat{\theta}$ el cual maximiza $\pi(\theta|x)$ es considerado como función de θ).*

Obviamente $\hat{\theta}$ tiene la interpretación de ser "el más probable" valor de θ, dada la distribución a priori y la muestra x.

Ejemplo 2.10 *Cuando f y π son distribuciones normales, la densidad a posteriori es una $\mathcal{N}(\mu(x), \rho^{-1})$. De aqui que el valor estimado de máxima verosimilitud de θ es*

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

2.4. Error de estimación

Cuando presentamos una estimación estadística, es necesario indicar la precisión de la estimación. La medida bayesiana de la precisión de un estimador (en una dimensión) es el error cuadrático medio a posteriori de la estimación, el cual se define a continuación.

Definición 2.6 *El error cuadrático medio (ECM) de un estimador $\hat{\delta}$ del parámetro θ es la función definida por*

$$E_{\pi(\theta|x)}\left[(\hat{\delta} - \theta)^2\right].$$

Para propósitos de cálculo, nótese que

$$\begin{aligned}
& E_{\pi(\theta|x)}\left[(\hat{\delta} - \theta)^2\right] \\
= \; & E_{\pi(\theta|x)}\left[(\hat{\delta} - \mu_{\pi(\theta|x)} + \mu_{\pi(\theta|x)} - \theta)^2\right] \\
= \; & E_{\pi(\theta|x)}\left[(\hat{\delta} - \mu_{\pi(\theta|x)})^2\right] + 2E_{\pi(\theta|x)}\left[(\hat{\delta} - \mu_{\pi(\theta|x)})(\mu_{\pi(\theta|x)} - \theta)\right] \\
& + E_{\pi(\theta|x)}\left[(\mu_{\pi(\theta|x)} - \theta)^2\right] \\
= \; & (\hat{\delta} - \mu_{\pi(\theta|x)})^2 + 2(\mu_{\pi(\theta|x)} - \hat{\delta})(E[\theta] - \mu_{\pi(\theta|x)}) + Var_{\pi(\theta|x)} \qquad (2.8) \\
= \; & (\hat{\delta} - \mu_{\pi(\theta|x)})^2 + Var_{\pi(\theta|x)}. \qquad (2.9)
\end{aligned}$$

Observe de (2.9) que la media a posteriori, $\mu_{\pi(\theta|x)}$, *minimiza* $E_{\pi(\theta|x)}\left[(\hat{\delta}-\theta)^2\right]$ *(sobre todo* $\hat{\delta}$*), y de aqui* $\hat{\delta}$ *es el estimador con menor error estándar. También tenemos que* $\sqrt{E_{\pi(\theta|x)}\left[(\hat{\delta}-\theta)^2\right]}$ *es el error estándar.*

Ejemplo 2.11 *(continuación del ejemplo 2.8). Es claro que la varianza a posteriori*

$$Var_{\pi(\theta|x)} = \rho^{-1} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Esto es, en el ejemplo de la prueba de inteligencia, el individuo con un resultado de x=115 reportará un CI estimado de $\mu_{\pi(\theta|x)}(115) = 110.39$, *con un error estándar asociado de* $\sqrt{Var_{\pi(\theta|x)}(115)} = \sqrt{69.23} = 8.32$.

El estimador clásico de θ *para el problema normal general es justamente* $\delta = x$, *el cual usando (2.9) se tiene*

$$\begin{aligned} E_{\pi(\theta|x)}\left[(\hat{\delta}-\theta)^2\right] &= Var_{\pi(\theta|x)} + (\hat{\delta}-\mu_{\pi(\theta|x)})^2 \\ &= Var_{\pi(\theta|x)} + \left(\frac{\sigma^2\mu}{\sigma^2+\tau^2} + \frac{\tau^2 x}{\sigma^2+\tau^2} - x\right)^2 \\ &= Var_{\pi(\theta|x)} + \frac{\sigma^4}{\sigma^2+\tau^2}(\mu-x)^2. \end{aligned}$$

Nótese que, en el ejemplo de CI, el estimador clásico $\hat{\delta} = x = 115$ *debe tener un error estándar (con respecto a* $\pi(\theta|x)$*) de*

$$\sqrt{Var_{\pi(\theta|x)}(115) + (\hat{\delta}-\mu_{\pi(\theta|x)})^2} = [69.23 + (110.39-115)^2]^{1/2} = \sqrt{90.48} = 9.49.$$

2.5. Estimación multivariada

La estimación bayesiana de un vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ es directo. La estimación de máxima verosimilitud generalizada (la moda a posteriori) es frecuentemente un estimador razonable, aunque se encuentren dificultades en la existencia y unicidad en el caso multivariado. La media a posteriori

$$\mu_{\pi(\theta|x)} = (\mu^1_{\pi(\theta_1|x)}, \mu^2_{\pi(\theta_2|x)}, ..., \mu^p_{\pi(\theta_p|x)})^t = E_{\pi(\theta|x)}[\theta]$$

es un atractivo estimador bayesiano, y su precisión puede ser descrito por la matriz de covarianza a posteriori

$$V_{\pi(\theta|x)} = E_{\pi(\theta|x)}[(\theta-\mu_{\pi(\theta|x)})(\theta-\mu_{\pi(\theta|x)})^t]. \tag{2.10}$$

(Por ejemplo, el error estándar estimado de $\mu^i_{\pi(\theta_i|x)}$ de θ_i, sería $\sqrt{V^{ii}_{\pi(\theta_i|x)}}$, donde $V^{ii}_{\pi(\theta_i|x)}$ es el (i,i) elemento de $V_{\pi(\theta|x)}$).

El análogo de (2.8), para una estimación general $\hat{\delta}$ de θ es

$$\begin{aligned} E\left[(\hat{\delta}-\theta)^2\right] &= E\left[(\hat{\delta}-\theta)(\hat{\delta}-\theta)^t\right] \\ &= V_{\pi(\theta|x)} + (\mu_{\pi(\theta|x)} - \hat{\delta})(\mu_{\pi(\theta|x)} - \hat{\delta})^t. \end{aligned} \tag{2.11}$$

Nuevamente, es claro que la media "minimiza" $E\left[(\hat{\delta}-\theta)^2\right]$.

Otra forma de inferir es mediante un intervalo de confianza para θ. El análogo bayesiano de un intervalo de confianza es llamado conjunto creíble.

2.6. Conjuntos creíbles

Definición 2.7 *Un conjunto creíble de* $100(1-\alpha)\,\%$ *para* θ *es un subconjunto* C *de* θ *tal que*

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

Puesto que la distribución a posteriori es una distribución de probabilidad de θ, uno puede hablar del significado (usando subjetividad) de la probabilidad de que θ esté en C. Esto está en contraste con los procedimientos de confianza clásicos que son interpretados en términos de probabilidad de cobertura (la probabilidad de que la variable aleatoria X sea tal que el conjunto de confianza $C(X)$ contenga a θ).

Al elegir un conjunto creíble para θ, se desea intentar minimizar su tamaño. Para hacer esto, se pueden incluir en el conjunto sólo aquellos puntos con la densidad a posteriori más grande, es decir, los valores "más probables" de θ.

Definición 2.8 *El conjunto creíble de máxima densidad a posteriori (MDP) a un nivel de confianza de* $100(1-\alpha)\,\%$, *es el subconjunto* C *de* θ, *de la forma*

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

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

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

Ejemplo 2.12 *Sabemos que la densidad a posteriori de* θ *dado* x *es* $\mathcal{N}(\mu(x), \rho^{-1}) = \mathcal{N}(110.39, 69.23)$, *que es unimodal y simétrica respecto a* $\mu(x)$. *El conjunto creíble de máxima densidad a un nivel de confianza* α *bilateral de* $100(1-\alpha)\,\%$ *está dado por*

$$C = (\mu(x) - z_{\frac{\alpha}{2}}\rho^{-\frac{1}{2}}, \mu(x) + z_{\frac{\alpha}{2}}\rho^{-\frac{1}{2}}),$$

donde z_α *es el* $\alpha - percentil$ *de la distribución* $\mathcal{N}(0,1)$. *Si queremos un nivel de significación del* $95\,\%$, $\alpha = 0.05$, *entonces* $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

$$\begin{aligned} C &= \left(110.39 - 1.96(\sqrt{69.23}), 110.39 + 1.96(\sqrt{69.23})\right) \\ &= (94.0819, 126.6981)\,. \end{aligned}$$

Los conjuntos creíbles bayesianos son usualmente mucho más fáciles de calcular que los intervalos de confianza clásicos, principalmente en situaciones donde un estadístico suficiente no existe.

Cuando θ es multivariado, el uso de la aproximación normal para la distribución a posteriori es valioso, pues los cálculos se vuelven difíciles de otra manera.

2.7. Inferencia predictiva

Ahora se intenta predecir una variable aleatoria $Z \sim g(z|\theta)$ basado en las observaciones de $X \sim f(x|\theta)$. Supongamos que X y Z son independientes y que g es una función de densidad.

La idea de la inferencia predictiva bayesiana es que, dado que $\pi(\theta|x)$ es la distribución de θ (creencia a posteriori), entonces $g(z|\theta)\pi(\theta|x)$ es la distribución conjunta de z y θ dado x, e integrando sobre θ se obtiene la distribución de z dado x.

Definición 2.9 *La densidad predictiva de* Z *dado* x, *cuando la distribución a priori para* θ *es* π, *es definida por*

$$p(z|x) = \int_\Theta g(z|\theta) dF^{\pi(\theta|x)}(\theta).$$

Ejemplo 2.13 *Considere el modelo de regresión lineal*

$$Z = \theta_1 + \theta_2 Y + \varepsilon, \tag{2.12}$$

donde $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ *(σ^2 conocida). Los datos* $((z_1, y_1), ..., (z_n, y_n))$ *son independientes de la regresión. Un estadístico suficiente para* $\boldsymbol{\theta} = (\theta_1, \theta_2)$ *es el estimador*

de mínimos cuadrados $\mathbf{X} = (X_1, X_2)$, *donde*

$$\begin{aligned} X_2 &= \sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})/SSY, \qquad X_1 = \bar{Z} - X_2\bar{Y}, \\ SSY &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad y \;\; \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \end{aligned}$$

por lo tanto, $X \sim \mathcal{N}_2(\theta, \sigma^2 \sum)$, *donde*

$$\sum = \frac{1}{SSY}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i^2 & -\bar{y} \\ -\bar{y} & 1 \end{pmatrix}.$$

Si la a priori no informativa $\pi(\theta) = 1$ *se usa para* θ, *la distribución a posteriori,* $\pi(\theta|x)$, *es* $\mathcal{N}_2(\theta, \sigma^2 \sum)$.

Supóngase que ahora se desea predecir el futuro de Z *correspondiente a un* y *dado. Claramente* $g(z|\theta)$ *es entonces* $\mathcal{N}(\theta_1 + \theta_2 y, \sigma^2)$, y *la densidad conjunta de* (z, θ) *dado* x *es normal. La distribución predictiva* $p(z|x)$, *es la distribución marginal de* Z *de la distribución a posteriori normal conjunta, esta marginal debe ser una distribución normal, y los cálculos nos dan una media de* $(x_1 + x_2 y)$ y *varianza* $V = \sigma^2(1 + n^{-1} + (\bar{y} - y)^2/SSY)$.

Esta distribución predictiva puede usarse para cualquier decisión de inferencia bayesiana. Por ejemplo, un conjunto creíble al $100(1-\alpha)\,\%$ *de* Z *debe ser*

$$\left((x_1 + x_2 y) + z(\frac{\alpha}{2})\sqrt{V}, \qquad (x_1 + x_2 y) - z(\frac{\alpha}{2})\sqrt{V}\right).$$

Capítulo 3

Contraste de hipótesis y análisis de varianza

3.1. Contraste de hipótesis

En el contraste de hipótesis clásico se especifica la hipótesis nula $H_0 : \theta \in \Theta_0$ y la hipótesis alternativa $H_1 : \theta \in \Theta_1$. Un procedimiento de contraste es evaluar en término de probabilidades los errores tipo I y tipo II. Estas probabilidades de los errores representan la posibilidad de que una muestra observada para dicho procedimiento de contraste pueda ser rechazada o no rechazada.

En análisis bayesiano la decisión entre H_0 y H_1 es conceptualmente más fácil. Esencialmente se calcula la probabilidad a posteriori $\alpha_0 = P(\Theta_0|x)$ y $\alpha_1 = P(\Theta_1|x)$ y se decide entre H_0 y H_1 respectivamente. La ventaja conceptual es que α_0 y α_1 son las probabilidades (subjetivas) actuales de las hipótesis en base a los datos y a la opinión a priori.

Definición 3.1 *La proporción α_0/α_1 es llamada la razón de probabilidad a posteriori de H_0 contra H_1, y π_0/π_1 es llamado la razón de probabilidad a priori (donde π_0 y π_1 son las probabilidades a priori de Θ_0 y Θ_1). La cantidad*

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

es llamado el factor de Bayes en favor de Θ_0.

El interés en el factor de Bayes es que algunas veces puede ser interpretado como "la razón para H_0 contra H_1 dado los datos". Esta interpretación es claramente válida cuando las hipótesis son simples, esto es, cuando $\Theta_0 = \{\theta_0\}$ y

$\Theta_1 = \{\theta_1\}$, entonces

$$\begin{aligned} \alpha_0 &= \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}, \qquad \alpha_1 = \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}, \\ \frac{\alpha_0}{\alpha_1} &= \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)} \qquad \text{y} \qquad B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{f(x|\theta_0)}{f(x|\theta_1)}. \end{aligned}$$

En otras palabras, B es entonces justamente la razón de verosimilitud de H_0 contra H_1, que es comúnmente vista como la probabilidad de H_0 contra H_1 que esta dada por los datos.

En general, B dependerá de la distribución a priori de entrada. Para explorar esta dependencia, es conveniente escribir la distribución a priori como

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{si } \theta \in \Theta_0, \\ \\ \pi_1 g_1(\theta) & \text{si } \theta \in \Theta_1, \end{cases} \tag{3.1}$$

donde g_0 y g_1 son densidades a priori (propias) las cuales describen, como la masa a priori se extiende sobre las dos hipótesis. (Recordemos que π_0 y π_1 son las probabilidades a priori de Θ_0 y Θ_1). Con esta representación se puede escribir

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

por lo tanto

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

el cual es la razón de verosimilitud de Θ_0 contra Θ_1. Por lo cual g_0 y g_1 no pueden ser interpretadas como una medida del soporte relativo para la hipótesis proporcionando solamente los datos. Algunas veces, B puede ser relativamente insesgado para elecciones razonables de g_0 y g_1, en tal caso la interpretación es razonable. La ventaja operacional principal de tener tal factor de Bayes estable es que un reporte científico puede proporcionar este factor de Bayes, y cualquier lector puede determinar su probabilidad a posteriori personal, simplemente multiplicando el factor de Bayes proporcionado por la probabilidad personal a priori.

Ejemplo 3.1 *(continuación del ejemplo 2.6) Supóngase que el individuo al cual se le hace la prueba de inteligencia, es clasificado en razón de su CI, que puede ser menor o igual que el CI medio (menor que 100) o mayor que el CI medio (más grande que 100). Formalmente, se tomará una decisión para contrastar* $H_0 : \theta \leq 100$ *contra* $H_1 : \theta > 100$. *Recordando que la distribución a posteriori de* θ *es*

$\mathcal{N}(110.39, 69.23)$, *estandarizando los resultados y usando tablas de la distribución normal estándar, obtenemos que*

$$\alpha_0 = P(\theta \leq 100|x) = 0.106, \qquad \alpha_1 = P(\theta > 100|x) = 0.894,$$

y por lo tanto la razón de probabilidad a posteriori es $\alpha_0/\alpha_1 = 8.44$. *Además* $\pi_0 = P^{\pi}(\theta \leq 100) = \frac{1}{2} = \pi_1$ *y la razón de probabilidad a priori es 1. (Note que una razón a priori de probabilidades de 1 indica que* H_0 *y* H_1 *son vistas como igualmente probables inicialmente). El factor de Bayes es entonces* $B = \alpha_0\pi_1/(\alpha_1\pi_0) = 8.44$.

3.1.1. Contraste de hipótesis unilateral

Un contraste de hipótesis unilateral ocurre cuando $\Theta \subset R^1$, éste contraste no tiene una característica especial. El interés esta en que es una de las pocas situaciones de contraste clásico para el cual el uso del valor-p tiene una justificación bayesiana. Considérese el siguiente ejemplo.

Ejemplo 3.2 *Cuando* $X \sim \mathcal{N}(\theta, \sigma^2)$ *y* θ *tiene la distribución a priori no informativa* $\pi(\theta) = 1$, *se puede ver que* $\pi(\theta|x)$ *es* $N(x, \sigma^2)$. *Consideremos ahora la situación de contraste* $H_0 : \theta \leq \theta_0$ *contra* $H_1 : \theta > \theta_0$. *Entonces*

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

donde, nuevamente Φ *es la función de distribución acumulada de una normal estándar. El valor-p clásico contra* H_0 *es la probabilidad cuando* $\theta = \theta_0$, *de observar un* X *"más extremo" que el dato* x *actual. Aquí el valor-p puede ser*

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

Por la simetría de la distribución normal, se sigue que α_0 *es igual que el valor-p contra* H_0.

3.1.2. Contraste de hipótesis nula puntual

Es muy común en la estadística clásica plantear un contraste de la forma $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$. Tal contraste de hipótesis nula puntual es interesante, particularmente porque el enfoque bayesiano contiene algunas características originales, pero principalmente porque las respuestas bayesianas difieren radicalmente de las respuestas clásicas. Antes de discutir este tema, se deben hacer algunos comentarios acerca de este tipo de hipótesis. Primero, los contraste de

hipótesis nula puntual son comúnmente realizados en situaciones inapropiadas. En realidad nunca se dá el caso que se considere la posibilidad que $\theta = \theta_0$ exactamente. Más razonable sería la hipótesis nula que $\theta \in \theta_0 = (\theta_0 - b, \theta_0 + b)$, donde $b > 0$ es alguna constante elegida tal que todo $\theta \in \theta_0$ pueda considerarse "indistinguible" de θ_0.

Dado que uno debe contrastar $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, necesitamos saber ¿cuándo es adecuada la apoximación de H_0 por $H_0 : \theta = \theta_0$? Desde la perspectiva bayesiana, la respuesta es "la aproximación es razonable si la probabilidad a posteriori de H_0 está cerca de la igualdad en ambos contrastes". Una condición fuerte es que la función de probabilidad sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$.

Ejemplo 3.3 *Sea $X_1, ..., X_n$ una muestra aleatoria con una distribución $\mathcal{N}(\theta, \sigma^2)$ donde σ^2 es conocida. La función de verosimilitud observada es entonces proporcional a una densidad $\mathcal{N}(\bar{x}, \sigma^2/n)$ de θ. Esto puede ser constante en $(\theta_0 - b, \theta_0 + b)$ cuando b es pequeña comparada con $\sigma/\sqrt{n}$. Por ejemplo, en un caso interesante donde es una prueba clásica $z = \sqrt{n}\,|\bar{x} - \theta_0|\ /\sigma$ más grande que 1, la función de verosimilitud puede variar no más del 5 % en $(\theta_0 - b, \theta_0 + b)$ si*

$$b \leq (0.024) z^{-1} \sigma / \sqrt{n}.$$

Cuando $z = 2$, $\sigma = 1$, y $n = 25$, esto impone la cota $b \leq 0.024$. Note que el límite en b, depende de $|\bar{x} - \theta_0|$, así como de $\sigma/\sqrt{n}$.

Para realizar un contraste bayesiano la hipótesis nula puntual $H_0 : \theta = \theta_0$, no se puede usar una distribución a priori continua. Pues tal distribución a priori y la distribución a posteriori darán una probabilidad de cero. Una aproximación razonable es dar a θ_0 una probabilidad positiva π_0 mientras que a $\theta \neq \theta_0$ la densidad $\pi_1 g_1(\theta)$, donde $\pi_1 = 1 - \pi_0$ y g_1 es propia. Uno puede pensar a π_0 como la masa que se le asignaría a la hipótesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$.

La densidad marginal de X es

$$m(x) = \int f(x|\theta) dF^{\pi}(\theta) = f(x|\theta_0)\pi_0 + (1 - \pi_0) m_1(x),$$

donde

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

es la densidad marginal de X con respecto a g_1. Por lo tanto la probabilidad a

posteriori de que $\theta = \theta_0$ es

$$\begin{aligned} \pi(\theta_0|x) &= \frac{f(x|\theta_0)\pi_0}{m(x)} \\ &= \frac{f(x|\theta_0)\pi_0}{f(x|\theta_0)\pi_0 + (1-\pi_0)m_1(x)} \\ &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}. \end{aligned} \tag{3.2}$$

Note que esto es $\alpha_{0,}$ y que $\alpha_1 = 1 - \alpha_0$ es por lo tanto la probabilidad a posteriori de H_1. Así la razón a posteriori de probabilidades es (recordar que $\pi_1 = 1 - \pi_o$)

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

y el factor H_0 contra H_1 es

$$B = f(x|\theta_0)/m_1(x). \tag{3.3}$$

3.1.3. Contraste de hipótesis múltiple

La característica interesante del contraste de hipótesis múltiple es que no es más difícil que un contraste de dos hipótesis, desde una perspectiva bayesiana. Uno simplemente calcula la probabilidad a posteriori de cada una de las hipótesis.

Ejemplo 3.4 *(continuación del ejemplo 2.6). Considerando que el individuo al cual se le hace la prueba de inteligencia es clasificado de acuerdo a su CI, en por debajo del promedio CI (menos que 90), el promedio CI (90 a 110) o por encima del CI promedio (más que 110). Llamando a estas tres regiones como* $\Theta_{1,}$ $\Theta_{2,}$ *y* Θ_3, *respectivamente, y recordando que la distribución a posteriori es N(110.39,69.23), estandarizando los resultados y usando tablas de la distribución normal estándar, obtenemos que para* $x = 115$*:* $P(\Theta_1|x = 115) = 0.007$, $P(\Theta_2|x = 115) = 0.473$ *y* $P(\Theta_3|x = 115) = 0.520$.

Lo anterior nos indica que el nivel de inteligencia (CI) del individuo en cuestión, está por encima del CI medio.

3.2. Análisis de varianza

En múltiples ocasiones el analista o investigador se enfrenta al problema de determinar si dos o más grupos son iguales, si dos o más cursos de acción arrojan resultados similares o si dos o más conjuntos de observaciones son parecidos.

Pensemos por ejemplo en el caso de determinar si dos niveles de renta producen consumos iguales o diferentes de un determinado producto, si las notas de dos grupos en una asignatura son similares o si tres muestras de análisis químico de una sustancia son iguales.

Una aproximación simple sería comparar las medias de estos grupos y ver si las medias aritméticas de la variable estudiada son parecidas o diferentes. Pero tal aproximación no es válida ya que la dispersión de las observaciones influirá en la posibilidad de comparar los promedios o medias de cada grupo. La dispersión deberá de tenerse en cuenta para realizar una comparación de medias o de grupos y esto es lo que se pretende con el análisis de varianza.

3.2.1. Análisis de varianza unidireccional

Suponga que interesa comparar las medias de k poblaciones. La situación experimental podría ser cualquiera de las siguientes:

1. Se tienen k poblaciones, cada una identificada por alguna característica común que se estudiará en el experimento. Se seleccionan muestras aleatorias independientes de tamaño $n_1, n_2, ..., n_k$, respectivamente, de cada una de las k poblaciones. Las diferencias observadas en las respuestas medidas se atribuyen a diferencias básicas entre las k poblaciones.

2. Se tiene un conjunto de N unidades experimentales homogéneas y se pretende estudiar los efectos de k tratamientos distintos. Estas unidades se dividen aleatoriamente en k subgrupos de tamaños $n_1, n_2, ..., n_k$, y cada subgrupo recibe un tratamiento experimental diferente. Los k subgrupos son conceptuados como muestras aleatorias independientes de tamaños $n_1, n_2, ..., n_k$, extraídas de las k poblaciones.

Aunque las situaciones experimentales descritas son distintas, tienen en común que cada una lleva a muestras aleatorias independientes extraídas de poblaciones con medias $\mu_1, \mu_2, ..., \mu_k$. Interesa probar la hipótesis nula de que las medias poblacionales son iguales, es decir:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_k \\ H_1 &: \mu_i \neq \mu_j \quad \text{para algunas } i \text{ y } j. \end{aligned}$$

El modelo aquí desarrollado se llama modelo de *clasificación unidireccional de efectos fijos.* La expresión "clasificación unidireccional" se refiere a que solo un factor o atributo se estudia en el experimento. El factor estudiado tiene k niveles distintos.

Ejemplo 3.5 *Se diseña un estudio para investigar el contenido de azufre de las cinco vetas de carbón principales de cierta región geográfica. Se obtienen muestras*

en puntos seleccionados aleatoriamente de cada veta y la respuesta medida es el porcentaje de azufre de cada muestra. Se pretende detectar las diferencias que pudiera haber en el contenido promedio de azufre de las cinco vetas. Cada una constituye una población. Se pretende comparar las medias poblacionales al poner a prueba:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ H_1 &: \mu_i \neq \mu_j \quad \textit{para algunas } i \textit{ y } j \end{aligned}$$

con base en muestras independientes extraídas de esas poblaciones. El factor de estudio es la veta de carbón y se investiga en cinco niveles. Estos no se seleccionan al azar, en su lugar, se eligen intencionadamente para estudiar las cinco vetas principales de la región. Se trata de un diseño de efectos fijos. El estudio es un ejemplo de la primera situación experimental antes descrita.

Sea Y_{ij} la $j-$*ésima* respuesta para el $i-$*ésimo* tratamiento o nivel del factor con $i = 1, 2, ..., k$ y $j = 1, 2, ..., n$. En este contexto, n_i es el tamaño de la muestra extraída de la $i-$*ésima* población. El número total de observaciones de las k muestras combinadas es $N = n_1 + n_2 + ... + n_k$. Los datos recopilados en un experimento de un solo factor y algunas estadísticas muestrales importantes se expresan de manera conveniente como se muestra en la siguiente tabla. El signo + que esta junto al subíndice en la notación siguiente indica que la suma se realiza con dicho subíndice.

Tabla 3.1.
Distribución de los datos de clasificación unidireccional

	Tratamiento o nivel del factor					
	1	2	3	...	k	
	Y_{11}	Y_{21}	Y_{31}		Y_{k1}	
	Y_{12}	Y_{22}	Y_{32}		Y_{k2}	
	Y_{13}	Y_{23}	Y_{33}		Y_{k3}	
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	Y_{1n_1}	Y_{2n2}	Y_{3n3}		Y_{kn3}	
Total	T_{1+}	T_{2+}	T_{3+}	...	T_{k+}	T_{++}
Media muestral	$\bar{Y}_{1+}$	$\bar{Y}_{2+}$	$\bar{Y}_{3+}$	...	$\bar{Y}_{k+}$	$\bar{Y}_{++}$

donde

$$\begin{aligned} T_{i+} &= \text{Total de las } i-\textit{ésimas} \text{ respuestas al tratamiento} = \sum_{j=1}^{n_i} Y_{ij} \\ \bar{Y}_{i+} &= \text{Media muestral del } i-\textit{ésimo} \text{ tratamiento} = T_{i+}/n_i \end{aligned}$$

$$T_{++} = \text{Total de todas las respuestas} = \sum_{i=1}^{k} T_{i+} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

$$\bar{Y}_{++} = \text{Media muestral de todas las respuestas} = T_{++}/N.$$

Ejemplo 3.6 *Se obtienen los datos y estadísticas de resumen del ejemplo sobre el contenido de azufre de las cinco vetas de carbón principales de una región geográfica. Notar que* $n_1 = 7$, $n_2 = 8$, $n_3 = 9$, $n_4 = 8$ *y* $n_5 = 9$, $T_{1+} = 1.51 + 1.92 + 1.08 + 2.04 + 2.14 + 1.76 + 1.17 = 11.62$, $\bar{Y}_{1+} = 11.62/7 = 1.66$, *de manera análoga se calcula* $T_{2+}, T_{3+}, T_{4+}, T_{5+}$, $\bar{Y}_{2+}, \bar{Y}_{3+}, \bar{Y}_{4+}$, *y* $\bar{Y}_{5+}$.

Factor (veta de carbón)					
1	*2*	*3*	*4*	*5*	
1.51	*1.69*	*1.56*	*1.30*	*0.73*	
1.92	*0.64*	*1.22*	*0.75*	*0.80*	
1.08	*0.90*	*1.32*	*1.26*	*0.90*	
2.04	*1.41*	*1.39*	*0.69*	*1.24*	
2.14	*1.01*	*1.33*	*0.62*	*0.82*	
1.76	*0.84*	*1.54*	*0.90*	*0.72*	
1.17	*1.28*	*1.04*	*1.20*	*0.57*	
	1.59	*2.25*	*0.32*	*1.18*	
		1.49		*0.54*	
T_{1+}*=11.62*	T_{2+}*=9.36*	T_{3+}*=13.14*	T_{4+}*=7.04*	T_{5+}*=8.8*	T_{++}*=49.96*
$\bar{Y}_{1+}$*=1.66*	$\bar{Y}_{2+}$*=1.17*	$\bar{Y}_{3+}$*=1.46*	$\bar{Y}_{4+}$*=0.88*	$\bar{Y}_{5+}$*=0.88*	$\bar{Y}_{++}$*=1.189*

El modelo

Se requiere crear un modelo estadístico para ver la forma de como probar la hipótesis nula de medias de tratamiento iguales. Por principio de cuentas, advierta que cada respuesta puede expresarse como:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

donde μ_i denota la media teórica de la $i - ésima$ población y ϵ_{ij}, la diferencia aleatoria entre la $j - ésima$ observación tomada de la $i - ésima$ población y la media de esa población. En otras palabras, $\epsilon_{ij} = Y_{ij} - \mu_i$. Una forma alterna de escribir este modelo es hacer que $d_i = \mu_i - \mu$, donde:

$$\mu = \sum_{i=1}^{k} n_i \mu_i / N.$$

En sentido práctico, μ representa un efecto medio global, calculado al combinar las k medias poblacionales. Note que si los tamaños muestrales son iguales, entonces μ es simplemente el promedio de las k medias poblacionales. Puesto que d_i es la diferencia entre la media global μ y la media de la $i-ésima$ población, d_i mide el efecto del $i-ésimo$ tratamiento. Advierta que:

$$\sum_{i=1}^{k} n_i d_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = \sum_{i=1}^{k} n_i\mu_i - N\mu = 0,$$

por sustitución, el modelo de clasificación unidireccional con efectos fijos puede expresarse en cualquiera de las tres formas siguientes:

Modelo de efectos fijos de clasificación unidireccional
$Y_{ij} = \mu_i + \varepsilon_{ij}$
$Y_{ij} = \mu + (\mu_i - \mu) + (Y_{ij} - \mu_i)$
$Y_{ij} = \mu + d_i + \varepsilon_{ij}.$

Estos modelos expresan matemáticamente la idea de que cada respuesta puede dividirse en tres componentes reconocibles, como sigue:

$$Y_{ij} = \mu + (\mu_i - \mu) + (Y_{ij} - \mu_i) \quad \text{ó} \quad Y_{ij} = \mu + d_i + \epsilon_{ij}$$

donde

Y_{ij}	Es la respuesta de la $j-ésima$ unidad experimental al $i-ésimo$ tratamiento
μ	Es la respuesta media global
$(\mu_i - \mu$ o $d_i)$	Es la desviación respecto de la media global, debido a que la unidad recibió el $i-ésimo$ tratamiento
$(Y_{ij} - \mu_i$ o $\epsilon_{ij})$	Es la desviación aleatoria respecto de la $i-ésima$ media poblacional debido a influencias aleatorias.

La hipótesis nula de medias de tratamiento iguales puede expresarse de manera alternativa al tomar nota de que si $\mu_1 = \mu_2 = ... = \mu_k$, entonces:

$$\mu = \sum_{i=1}^{n_i} n_i\mu_i/N = N\mu_i/N = \mu_i \qquad \text{para cada } i = 1, ..., k$$

y $d_i = \mu_i - \mu = 0$ para cada i. Ello implica que verificar

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

es equivalente a probar:

$$H_0 : d_1 = d_2 = ... = d_k = 0.$$

3.2.2. Prueba de H_0

La obtención de un estadístico de prueba requiere ciertos supuestos acerca de las diferencias aleatorias de ϵ_{ij}. En particular, se supone que esas diferencias son variables aleatorias independientes con distribución $\mathcal{N}(0, \sigma^2)$. En términos más comprensibles se supone que:

1. Las k muestras son independientes y extraídas de k poblaciones específicas, con medias desconocidas $\mu_1, \mu_2, ..., \mu_k$.

2. Cada una de las k poblaciones tiene distribución normal.

3. Cada una de las k poblaciones tiene la misma varianza, σ^2.

Se definió ya el análisis de varianza, procedimiento en el que la variación total de una respuesta media se subdivide en componentes atribuibles a fuentes reconocibles. Puesto que $\mu, \mu_1, \mu_2, ..., \mu_k$ son medias poblacionales teóricas, dividir de manera práctica una observación requiere sustituir las medias teóricas con sus estimadores insesgados $\bar{Y}_{++}, \bar{Y}_{1+}, ..., \bar{Y}_{k+}$, respectivamente. Efectuar tal remplazo lleva a la identidad siguiente:

$$Y_{ij} = \bar{Y}_{++} + (\bar{Y}_{i+} - \bar{Y}_{++}) + (Y_{ij} - \bar{Y}_{i+}).$$

Note que $\bar{Y}_{++}$ es un estimador de μ, el efecto medio agrupado global; $\bar{Y}_{i+} - \bar{Y}_{++}$ es un estimador de $d_i = \mu_i - \mu$, el efecto del $i - ésimo$ tratamiento, y $Y_{ij} - \bar{Y}_{i+}$ es un estimador de $\epsilon_{ij} = Y_{ij} - \mu_i$, el error aleatorio. El término $Y_{ij} - \bar{Y}_{i+}$ suele denominarse *residuo*. Esta identidad es equivalente a:

$$Y_{ij} - \bar{Y}_{++} = (\bar{Y}_{i+} - \bar{Y}_{++}) + (Y_{ij} - \bar{Y}_{i+}).$$

Si se eleva al cuadrado y se suma cada miembro de la identidad respecto de todos los valores posibles de i y j, se tiene

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{++})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\bar{Y}_{i+} - \bar{Y}_{++}) + (Y_{ij} - \bar{Y}_{i+})\right]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i+} - \bar{Y}_{++})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i+} - \bar{Y}_{++})(Y_{ij} - \bar{Y}_{i+}) \\
&\quad + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i+})^2
\end{aligned}$$

pero

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i+} - \bar{Y}_{++})(Y_{ij} - \bar{Y}_{i+}) = 0$$

puesto que:

$$\begin{aligned}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i+}) &= \sum_{j=1}^{n_i}Y_{ij}-n_i\bar{Y}_{i+}=n_i\left(\frac{\sum_{j=1}^{n_i}Y_{ij}}{n_i}\right)-n_i\bar{Y}_{i+}\\ &= n_i\bar{Y}_{i+}-n_i\bar{Y}_{i+}=0\end{aligned}$$

así

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{++})^2=\sum_{i=1}^{k}n_i(\bar{Y}_{i+}-\bar{Y}_{++})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i+})^2$$

se tiene la llamada identidad de la suma de cuadrados del análisis de varianza de clasificación unidireccional. Cada uno de los componentes de esta identidad puede interpretarse de manera que tenga sentido. En particular,

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{++})^2 &= \text{Medida de la variabilidad de los datos}\\ &= \text{Suma de cuadrados total } (SC_{Tot})\\ \sum_{i=1}^{k}n_i(\bar{Y}_{i+}-\bar{Y}_{++})^2 &= \text{Medida de la variabilidad de los datos atribuida al}\\ &\quad \text{hecho de que se usan diferentes niveles de factores}\\ &\quad \text{o tratamientos}\\ &= \text{Suma de cuadrados de los tratamientos } (SC_{Tr})\\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i+})^2 &= \text{Medida de la variabilidad de los datos atribuida a la}\\ &\quad \text{fluctuación aleatoria entre los sujetos de un mismo}\\ &\quad \text{nivel de factor}\\ &= \text{Residuo o suma de cuadrados de error } (SCE).\end{aligned}$$

De manera simbólica, la identidad de la suma de cuadrados puede escribirse como:

$$SC_{Tot}=SC_{Tr}+SCE.$$

En caso de existir diferencias entre las medias poblacionales, se espera que en gran parte de la variación en las respuestas se deba al hecho de que se usan tratamientos distintos. En otras palabras, se espera que SC_{Tr} sea grande en comparación con SCE. En los procedimientos del análisis de varianza, se usa esta idea para probar

la hipótesis nula de medias de tratamientos iguales mediante la comparación de la variación de los tratamientos (SC_{Tr}) contra la variación del tratamiento (SCE) con una razón F apropiada.

A fin de considerar una razón F apropiada, deben considerarse los valores esperados de las estadísticas SC_{Tr} y SCE. Ello requiere suponer en el modelo que los errores aleatorios ϵ_{ij} son variables aleatorias normalmente distribuidas e independientes, cada una con media 0 y varianza σ^2. Por principio de cuentas, tome nota de que para cada i:

$$\begin{aligned} \bar{Y}_{i+} &= \sum_{j=1}^{n_i}(\mu + d_i + \epsilon_{ij})/n_i \\ &= \frac{n_i\mu + n_i d_i + \sum_{j=1}^{n_i}\epsilon_{ij}}{n_i} \\ &= \mu + d_i + \bar{\epsilon}_{i+}. \end{aligned}$$

Además, puesto que $\sum_{i=1}^{k} n_i d_i = 0$, se tiene:

$$\begin{aligned} \bar{Y}_{++} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}/N = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu + d_i + \epsilon_{ij})/N \\ &= \frac{N\mu + \sum_{i=1}^{k} n_i d_i + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\epsilon_{ij}}{N} = \mu + \bar{\epsilon}_{++}. \end{aligned}$$

Luego de sustituir, es posible rescribir SC_{Tr} como se muestra:

$$\begin{aligned} SC_{Tr} &= \sum_{i=1}^{k} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2 = \sum_{i=1}^{k} n_i\left[(\mu + d_i + \bar{\epsilon}_{i+}) - (\mu + \bar{\epsilon}_{++})\right]^2 \\ &= \sum_{i=1}^{k} n_i(d_i + \bar{\epsilon}_{i+} - \bar{\epsilon}_{++})^2 = \sum_{i=1}^{k} n_i d_i^2 + 2\sum_{i=1}^{k} n_i d_i \bar{\epsilon}_{i+} + \sum_{i=1}^{k} n_i \bar{\epsilon}_{i+}^2 - N\bar{\epsilon}_{++}^2. \end{aligned}$$

Al tomar el valor esperado de cada término se tiene:

$$E[SC_{Tr}] = \sum_{i=1}^{k} n_i d_i^2 + 2\sum_{i=1}^{k} n_i d_i E[\bar{\epsilon}_{i+}] + \sum_{i=1}^{k} n_i E[\bar{\epsilon}_{i+}^2] - NE[\bar{\epsilon}_{++}^2]$$

pero $E[\bar{\epsilon}_{i+}^2] = \sigma^2/n_i$ pues

$$\begin{aligned} E[\bar{\epsilon}_{i+}^2] &= E\left[\left(\sum_{j=1}^{n_i} \epsilon_{ij}/n_i\right)^2\right] = E\left[\left(\sum_{j=1}^{n_i} \epsilon_{ij}\right)^2\right]/n_i^2 \\ &= \frac{1}{n_i^2} E[\epsilon_{i1}^2 + \epsilon_{i2}^2 + ... + \epsilon_{in_i}^2 \\ &\quad + \epsilon_{i1}\epsilon_{i2} + \epsilon_{i1}\epsilon_{i3} + ... + \epsilon_{i1}\epsilon_{in_i} + \epsilon_{i2}\epsilon_{i1} + \epsilon_{i2}\epsilon_{i3} + ... + \epsilon_{i2}\epsilon_{in_i} \\ &\quad + ... + \epsilon_{in_i}\epsilon_{i1} + ... + \epsilon_{in_i}\epsilon_{in_i-1}] \\ &= \frac{1}{n_i^2}[\sigma^2 + \sigma^2 + ... + \sigma^2] = \frac{n_i\sigma^2}{n_i^2} = \frac{\sigma^2}{n_i}. \end{aligned}$$

Puesto que $E[\epsilon_{ij}^2] = E[(Y_{ij} - \mu_i)^2] = \sigma^2$ y $E[\epsilon_{ij}] = E[(Y_{ij} - \mu_i)] = E[Y_{ij}] - \mu_i = \mu_i - \mu_i = 0$.

Un argumento similar muestra que $E[\bar{\epsilon}_{++}^2] = \frac{\sigma^2}{N}$. Es fácil ver que $E[\bar{\epsilon}_{i+}] = 0$. Así

$$\begin{aligned} E[SC_{Tr}] &= \sum_{i=1}^{k} n_i d_i^2 + \sum_{i=1}^{k} n_i \frac{\sigma^2}{n_i} - N\frac{\sigma^2}{N} \\ &= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i d_i^2. \end{aligned}$$

Luego de dividir SC_{Tr} entre $k-1$, se obtiene el estadístico llamada cuadrado medio del tratamiento, que se denota CM_{Tr}. Dicho de otra manera:

$$CM_{Tr} = SC_{Tr}/(k-1).$$

Se puede apreciar que $E[CM_{Tr}] = \sigma^2 + \sum_{i=1}^{k} n_i d_i^2/(k-1)$. Recuerde que la suma de los cuadrados de los residuos ayuda a estimar σ^2 en el contexto de regresión. Lo mismo es válido aquí. A fin de obtener un estimador insesgado de σ^2, se divide la suma de cuadrados de los residuos SCE entre $N-k$. Es un estimador denominado *cuadrado medio del error* y se denota con CM_E. En otras palabras:

$$CM_E = SCE/(N-k).$$

¿Cómo usar CM_{Tr} y CM_E para probar H_0? A fin de responder a esta pregunta basta tomar en cuenta que si H_0 es verdadera, entonces $d_1 = d_2 = ... = d_k = 0$ y, por ende, $\sum_{i=1}^{k} n_i d_i^2/(k-1) = 0$. En caso de que H_0 no sea verdadera, entonces este término es positivo. Así pues, en el primero de estos dos casos, se esperaría que CM_{Tr} y CM_E tengan valores cercanos, ya que con ambas se estima σ^2,

mientras que en el segundo cabría esperar que la primera sea un tanto mayor que la segunda. Para probar la hipótesis nula $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ contra la hipótesis alternativa $H_1 : \mu_i \neq \mu_j$ para algunas i y j se usa

$$F_{k-1,N-k} = CM_{Tr}/CM_E,$$

como estadístico de prueba lógico. Si H_0 es verdadera, se espera que su valor sea cercano a 1, y en cualquier otro caso, que sea mayor que 1. Esta razón puede usarse como estadística de prueba, puesto que si la hipótesis nula es verdadera, se sabe que tiene distribución F, con $k-1$ y $N-k$ grados de libertad. La prueba siempre es de cola derecha, con rechazo de H_0 en el caso de valores de la variable aleatoria $F_{k-1,N-k}$ que parecen demasiado grandes para haber ocurrido al azar. Es posible que los valores de $F = CM_{Tr}/CM_E$ sean menores que la unidad ya que F es una variable aleatoria. Dicho resultado puede ocurrir únicamente por azar o causa de que el modelo lineal supuesto sea incorrecto. Aunque en la práctica es usual que el análisis de varianza se efectue mediante computadora, se cuenta con algunos atajos de cálculo.

Teorema 3.1

$$\begin{aligned} SC_{Tot} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{++}^2}{N} \\ SC_{Tr} &= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - \frac{T_{++}^2}{N} \\ SCE &= SC_{Tot} - SC_{Tr}. \end{aligned}$$

Demostración *Por definición se tiene que*

$$\begin{aligned} SC_{Tot} &= SC_{Tr} + SCE = \sum_{i=1}^{k} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i+})^2 \\ &= \sum_{i=1}^{k} n_i(\bar{Y}_{i+}^2 - 2\bar{Y}_{i+}\bar{Y}_{++} + \bar{Y}_{++}^2) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}^2 - 2Y_{ij}\bar{Y}_{i+} + \bar{Y}_{i+}^2) \\ &= \sum_{i=1}^{k} n_i \left(\frac{T_{i+}^2}{n_i^2}\right) - 2\bar{Y}_{++}\sum_{i=1}^{k} n_i\bar{Y}_{i+} + N\bar{Y}_{++}^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 \\ &\quad -2\sum_{i=1}^{k} T_{i+}\bar{Y}_{i+} + \sum_{i=1}^{k} n_i\bar{Y}_{i+}^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - 2\bar{Y}_{++} \sum_{i=1}^{k} n_i \bar{Y}_{i+} + N\bar{Y}_{++}^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 \\
&\quad -2\sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} + \sum_{i=1}^{k} n_i \bar{Y}_{i+}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\bar{Y}_{++} \sum_{i=1}^{k} n_i \frac{T_{i+}}{n_i} + N\bar{Y}_{++}^2 - \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} + \sum_{i=1}^{k} n_i \frac{T_{i+}^2}{n_i^2} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\bar{Y}_{++} \sum_{i=1}^{k} T_{i+} + N\bar{Y}_{++}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\bar{Y}_{++} T_{++} + N\bar{Y}_{++}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\frac{T_{++}}{N} T_{++} + N\frac{T_{++}^2}{N^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\frac{T_{++}^2}{N} + \frac{T_{++}^2}{N} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{++}^2}{N},
\end{aligned}
$$

por lo que queda demostrada la ecuación 1. Para la ecuación 2 tenemos que

$$
\begin{aligned}
SC_{Tr} &= \sum_{i=1}^{k} n_i (\bar{Y}_{i+} - \bar{Y}_{++})^2 = \sum_{i=1}^{k} n_i (\bar{Y}_{i+}^2 - 2\bar{Y}_{i+}\bar{Y}_{++} + \bar{Y}_{++}^2) \\
&= \sum_{i=1}^{k} n_i \left(\frac{T_{i+}^2}{n_i^2}\right) - 2\bar{Y}_{++} \sum_{i=1}^{k} n_i \bar{Y}_{i+} + N\bar{Y}_{++}^2 \\
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - 2\bar{Y}_{++} \sum_{i=1}^{k} n_i \frac{T_{i+}}{n_i} + N\bar{Y}_{++}^2 \\
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - 2\frac{T_{++}}{N} \sum_{i=1}^{k} {}_i T_{i+} + N\left(\frac{T_{++}^2}{N^2}\right) \\
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - 2\frac{T_{++}}{N} T_{++} + \frac{T_{++}^2}{N} \\
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - 2\frac{T_{++}^2}{N} + \frac{T_{++}^2}{N} \\
&= \sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - \frac{T_{++}^2}{N}
\end{aligned}
$$

Por lo tanto se cumple la ecuación 2. La ecuacion 3 es solo un despeje de $SC_{Tot} = SC_{Tr} + SCE$.■

Las ideas teóricas subyacentes al procedimiento de análisis de varianza en el caso del modelo de efectos fijos de clasificación unidireccional se resume en la siguiente tabla.

Tabla 3.2.
ANOVA para clasificación unidireccional con efectos fijos

F. de var.	$G.l$	$S.C$	$C.M$	CME	F
Trat. o nivel	$k-1$	$\sum_{i=1}^{k} \frac{T_{i+}^2}{n_i} - \frac{T_{++}^2}{N}$	$\frac{SC_{Tr}}{k-1}$	$\sigma^2 + \sum_{i=1}^{k} \frac{n_i d_i^2}{k-1}$	$\frac{CM_{Tr}}{CM_E}$
Error o residuo	$N-k$	SCE	$\frac{SCE}{N-k}$	σ^2	
Total	$N-1$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{++}^2}{N}$			

donde $G.l$ son los grados de libertad, $S.C$ la suma de cuadrados, $C.M$ cuadrado medio y CME cuadrado medio esperado.

El uso de la razon F se ilustra con la continuación del análisis de los datos del carbón iniciado en el ejemplo 3.5.

Ejemplo 3.7 *Se ponen a prueba*

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

o

$$H_0 : d_1 = d_2 = d_3 = d_4 = d_5 = 0$$

con base en los datos previos. Recuerde que μ_i, i=1,2,3,4,5, denota el contenido medio de azufre de las cinco vetas de carbón principales de una región geográfica. Se tienen las estadísticas de resumen siguientes:

$$\begin{array}{llll} T_{1+} = 11.62 & T_{4+} = 7.04 & n_1 = 7 & n_4 = 8 \\ T_{2+} = 9.36 & T_{5+} = 8.8 & n_2 = 9 & n_5 = 10 \\ T_{3+} = 13.14 & T_{++} = 49.96 & n_3 = 9 & N = 42 \end{array}$$

La única estadística adicional necesaria es $\sum_{i=1}^{5} \sum_{j=1}^{n_i} Y_{ij}^2$. En relación con los datos del ejemplo 3.5, esta estadística asume el valor 67.81. Luego de sustituir en

las fórmulas de cálculo, se obtiene:

$$\begin{aligned}
SC_{Tot} &= \sum_{i=1}^{5}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{++}^2}{N} = 67.861 - \frac{(49.96)^2}{42} = 8.432 \\
SC_{Tr} &= \sum_{i=1}^{5} \frac{T_{i+}^2}{n_i} - \frac{T_{++}^2}{N} \\
&= \frac{(11.62)^2}{7} + \frac{(9.36)^2}{8} + \frac{(13.14)^2}{9} + \frac{(7.04)^2}{8} + \frac{(8.8)^2}{10} - \frac{(49.96)^2}{42} \\
&= 3.935 \\
SCE &= SC_{Tot} - SC_{Tr} = 8.432 - 3.935 = 4.497 \\
CM_{Tr} &= \frac{SC_{Tr}}{k-1} = \frac{3.935}{4} = 0.984 \\
CM_E &= \frac{SCE}{N-k} = \frac{4.497}{37} = 0.122.
\end{aligned}$$

El valor observado de la estadística de prueba $F_{k-1,N-k} = F_{4,37}$ *es:*

$$F_{4,37} = \frac{CM_{Tr}}{CM_E} = \frac{0.984}{0.122} = 8.066.$$

Puesto que $f_{0,05}(4,37) \doteq 2.626$, *es posible rechazar* H_0 *con* $p < 0.05$. *Se tienen evidencias estadísticas que al menos dos vetas de carbón difieren en la media de contenido de azufre. La tabla ANOVA de estos datos es la siguiente:*

Tabla 3.3.
ANOVA de los datos de vetas de carbón

Fuente de variación	*G.l*	*S.C*	*CM*	*F*
Tratamientos	*4*	*3.935*	*0.984*	*8.066*
Error	*37*	*4.497*	*0.122*	
Total	*41*	*8.432*		

Recuerde que se parte del supuesto de que cada una de las k muestras independientes se extrae de poblaciones de distribución normal, con varianzas iguales σ^2.

3.2.3. Comparación de varianzas

Como se mencionó, la prueba F de verificación de la igualdad de medias es sensible a la violación del supuesto de varianzas iguales. Ello reviste validez

particular cuando los tamaños muestrales difieren mucho. Antes de emprender el análisis de varianza, es necesario probar las hipótesis:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \\ H_1 &: \sigma_i^2 \neq \sigma_j^2 \quad \text{para algunas } i \text{ y } j. \end{aligned}$$

Si se rechaza H_0, se debe usar un análisis no paramétrico o realizar una transformación de los datos con la esperanza de estabilizar las varianzas. La prueba más usada para poner a prueba la hipótesis nula de varianzas iguales es la prueba de Bartlett. La prueba de Bartlett se inicia con el cálculo de las varianzas muestrales $S_1^2, S_2^2, ..., S_k^2$ de cada una de las k muestras. También se determina el cuadrado medio del error, la estimacion ponderada de σ^2 bajo el supuesto de que la hipótesis nula es verdadera. En este contexto, resulta conveniente el cálculo directo de CM_E a partir de las varianzas muestrales individuales, mediante la ecuación:

$$CM_E = S_p^2 = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{N - k}.$$

Luego se forma la estadística Q, definida por:

$$Q = (N - k) \log_{10} S_p^2 - \sum_{i=1}^{k} (n_i - 1) \log S_i^2.$$

El valor esperado de ésta estadística es grande cuando las varianzas muestrales S_i^2, $i = 1, 2, ..., k$ son muy distintas, y es cercano a 0 si dichas varianzas también guardan cercanía en sus valores. Para probar la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ contra la hipótesis alternativa $H_1 : \sigma_i^2 \neq \sigma_j^2$ para algunas i y j, se usa el estadístico de Bartlett definido por:

$$B = 2.3026Q/h$$

donde

$$h = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right).$$

A continuación se mostrará el uso de esta prueba.

Ejemplo 3.8 *Nuevamente regresemos a los datos de las vetas de carbón del ejemplo 3.5. Deben calcularse las varianzas muestrales y sus logaritmos para cada uno de los cinco niveles del factor. Los resultados de esos cálculos se resumen a continuación:*

veta de carbón	***varianza muestral*** (S_i^2)	$\mathbf{log_{10}S_i^2}$	***Tamaño muestral*** (n_i)
1	*0.175*	*-0.757*	*7*
2	*0.144*	*-0.842*	*8*
3	*0.115*	*-0.939*	*9*
4	*0.123*	*-0.910*	*8*
5	*0.074*	*-1.131*	*10*

La estimación agrupada de σ^2 *es:*

$$\begin{aligned} CM_E &= S_p^2 = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{N - k} \\ &= \frac{6(0.175) + 7(0.144) + 8(0.115) + 7(0.123) + 9(0.074)}{42 - 5} \\ &= 0.122, \end{aligned}$$

por sustitución, se tiene:

$$\begin{aligned} Q &= (N - k)\log_{10} S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log S_i^2 \\ &= 37\log_{10} 0.122 \\ &\quad -[6(-0.757) + 7(-0.842) + 8(-0.939) + 7(-0.910) + 9(-1.131)] \\ &= 0.692, \end{aligned}$$

y

$$\begin{aligned} h &= 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right) \\ &= 1 + \frac{1}{3(4)}\left[\frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{7} + \frac{1}{9} - \frac{1}{37}\right] \\ &= 1.055. \end{aligned}$$

El valor observado de la estadística de Bartlett es:

$$\begin{aligned} B &= 2.3026Q/h \\ &= 2.3026(0.692)/1.055 \\ &= 1.510. \end{aligned}$$

Basado en la distribución ji − cuadrada, $X_{k-1}^2 = X_4^2$*, el valor-p se ubica entre 0.75 y 0.9. Puesto que es un valor grande resulta imposible rechazar:*

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2.$$

3.3. Análisis de varianza bidireccional

Se recurre al procedimiento llamado uso de bloques cuando se pretende comparar las medias de k poblaciones en presencia de una variable extraña. Un bloque es un conjunto de k unidades experimentales tan parecidas como sea posible en lo relativo a la variable extraña. Cada tratamiento asigna aleatoriamente a una unidad de cada bloque. Puesto que el efecto de la variable extraña se controla con el emparejamiento de unidades experimentales similares, toda diferencia en la respuesta se atribuye a los efectos del tratamiento.

El diseño experimental analizado en esta sección se llama diseño de bloques completos aleatorizados con efectos fijos. En dicha expresión, bloques se refiere a que las unidades experimentales se emparejan en lo concerniente a una variable extraña; aleatorizados, a que los tratamientos se asignan de manera aleatoria al interior de los bloques, y completo, a que cada tratamiento se usa exactamente una vez en cada bloque. La expresión "efectos fijos" es aplicable a los bloques y tratamientos, es decir, se supone que ni los bloques ni los tratamientos se seleccionan al azar. Todas las inferencias elaboradas se aplican únicamente a los k tratamientos y b bloques usados realmente.

La hipótesis que interesa es la de media de tratamientos iguales, dada por:

$$H_0 : \mu_{1+} = \mu_{2+} = ... = \mu_{k+}$$

donde μ_{k+} denota la media del $i - ésimo$ tratamiento.

En cuanto a la notación, Y_{ij} denota la respuesta del $i-ésimo$ tratamiento en el $j-ésimo$ bloque, con $i = 1, 2, ..., k$ y $j = 1, 2, ..., b$. Advierta que b indica el número de bloques usado en el experimento y el número de observaciones por tratamiento; k denota el número de tratamientos que se estudia y el de observaciones por bloque, y $N = kb$, el número total de respuestas. Los datos recopilados en un experimento de bloques completos aleatorizados y algunas estadísticas muestrales importantes se presentan de manera conveniente en la siguiente tabla:

Tabla 3.4.
Disposición de datos de bloques completos aleatorizados

	Tratamiento						
Bloque	1	2	3	$\cdots$	k	Bloq. tot.	Med. del bloque
1	Y_{11}	Y_{21}	Y_{31}		Y_{k1}	T_{+1}	$\bar{Y}_{+1}$
2	Y_{12}	Y_{22}	Y_{32}		Y_{k2}	T_{+2}	$\bar{Y}_{+2}$
3	Y_{13}	Y_{23}	Y_{33}		Y_{k3}	T_{+3}	$\bar{Y}_{+3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
b	Y_{1b}	Y_{2b}	Y_{3b}		Y_{kb}	T_{+b}	$\bar{Y}_{+b}$
Tot. del trat.	T_{1+}	T_{2+}	T_{3+}	$\cdots$	T_{k+}	T_{++}	
Med. del trat.	$\bar{Y}_{1+}$	$\bar{Y}_{2+}$	$\bar{Y}_{3+}$	$\cdots$	$\bar{Y}_{k+}$	$\bar{Y}_{++}$	

Note que:

$$\begin{aligned}
T_{i+} &= \text{total de todas las respuestas al } i-\acute{e}simo \text{ tratamiento} = \sum_{j=1}^{b} Y_{ij} \\
\bar{Y}_{i+} &= \text{media muestral del } i-\acute{e}simo \text{ tratamiento} = T_{i+}/b \\
T_{+j} &= \text{total de respuestas del } j-\acute{e}simo \text{ bloque} = \sum_{i=1}^{k} Y_{ij} \\
\bar{Y}_{+j} &= \text{media muestral del } j-\acute{e}simo \text{ bloque} = T_{+j}/k \\
T_{++} &= \text{total de todas las respuestas } = \sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij} = \sum_{i=1}^{k} T_{i+} = \sum_{j=1}^{b} T_{+j} \\
\bar{Y}_{++} &= \text{media de todas las repuestas } = T_{++}/N.
\end{aligned}$$

3.3.1. El modelo

En la redacción del modelo del diseño de bloques completos aleatorizados con efectos fijos, se requiere la notación siguiente:

$$\begin{aligned}
\mu_{ij} &= \text{media del } i-\acute{e}simo \text{ tratamiento al } j-\acute{e}simo \text{ bloque} \\
\mu_{i+} &= \text{media del } i-\acute{e}simo \text{ tratamiento} = \sum_{j=1}^{b} \mu_{ij}/b \\
\mu_{+j} &= \text{media del } j-\acute{e}simo \text{ bloque} = \sum_{i=1}^{k} \mu_{ij}/k \\
\mu &= \text{media global} = \sum_{i=1}^{k}\sum_{j=1}^{b} \mu_{ij}/kb \\
\tau_i &= \mu_{i+} - \mu = \text{efecto debido al hecho de que la unidad experimental recibió el } i-\acute{e}simo \text{ tratamiento} \\
\beta_j &= \mu_{+j} - \mu = \text{efecto debido al hecho de que la unidad experimental está en el } j-\acute{e}simo \text{ bloque} \\
\epsilon_{ij} &= Y_{ij} - \mu_{ij} = \text{error residual o aleatorio.}
\end{aligned}$$

Ahora, es posible expresar el modelo como sigue:

Modelo para el diseño de bloques completos aleatorizados

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}.$$

Este modelo expresa simbolicamente el concepto de que cada observación se puede dividir en cuatro componentes reconocibles: la media de efecto global μ, el efecto de tratamiento τ_i, el efecto de bloque β_j y una desviación aleatoria ϵ_{ij} que se atribuye a fuentes inexplicadas. Se tienen los siguientes supuestos del modelo:

1. Las $k \cdot b$ observaciones constituyen muestras aleatorias independientes, cada uno de tamaño 1, de $k \cdot b$ poblaciones con medias desconocidas μ_{ij}.

2. Cada una de las $k \cdot b$ poblaciones tienen distribución normal.

3. Cada una de las $k \cdot b$ poblaciones tienen la misma varianza, σ^2.

4. Los efectos de bloque y tratamiento son aditivos, es decir, no existe interacción de los bloques con los tratamientos.

Los supuestos 1-3 son idénticos a los del modelo de clasificación unidireccional, salvo que se consideran $k \cdot b$, no k, poblaciones. Afirmar que los efectos de bloque y tratamiento son aditivos significa que los tratamientos tienen comportamiento constante de un bloque a otro, y los bloques lo tienen de un tratamiento a otro.

En lo matemático, esa naturaleza aditiva significa que:

$$\begin{aligned} \mu_{ij} &= \mu + \tau_i + \beta_j \\ &= \mu + (\mu_{i+} - \mu) + (\mu_{+j} - \mu). \end{aligned}$$

Al sustituir, puede reescribirse el modelo teórico como sigue:

$$\begin{aligned} \mu_{ij} - \mu &= \tau_i + \beta_j \\ &= (\mu_{i+} - \mu) + (\mu_{+j} - \mu) + \left\{\mu_{ij} - \left[\mu + (\mu_{i+} - \mu) + (\mu_{+j} - \mu)\right]\right\}. \end{aligned}$$

El remplazo de los parámetros con sus estimadores insesgados respectivos lleva a:

$$Y_{ij} - \bar{Y}_{++} = (\bar{Y}_{i+} - \bar{Y}_{++}) + (\bar{Y}_{+j} - \bar{Y}_{++}) + \left\{Y_{ij} - \left[\bar{Y}_{++} + (\bar{Y}_{i+} - \bar{Y}_{++}) + (\bar{Y}_{+j} - \bar{Y}_{++})\right]\right\}.$$

Si cada miembro de esta idéntidad se eleva al cuadrado, se suma respecto de todos los valores posibles de i y j, y se simplifica, resulta la idéntidad siguiente de suma de cuadrados del diseño de bloques completamente aleatorizados:

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{++})^2 &= \sum_{i=1}^{k} b(\bar{Y}_{i+} - \bar{Y}_{++})^2 + \sum_{j=1}^{b} k(\bar{Y}_{+j} - \bar{Y}_{++})^2 \\ &\quad + \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{i+} - \bar{Y}_{+j} - \bar{Y}_{++})^2. \end{aligned}$$

La interpretación práctica de cada componente es similar a la del modelo de

clasificación unidireccional. En particular:

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_{++})^2 &= \text{medida de la variabilidad total de los datos} \\
&= \text{suma de cuadrados total } (SC_{Tot}) \\
\sum_{i=1}^{k} b(\bar{Y}_{i+}-\bar{Y}_{++})^2 &= \text{medida de la variabilidad total de los datos} \\
&\quad \text{atribuibles al uso de tratamientos distintos} \\
&= \text{suma de cuadrados de los tratamientos } (SC_{Tr}) \\
\sum_{j=1}^{b} k(\bar{Y}_{+j}-\bar{Y}_{++})^2 &= \text{medida de la variabilidad total de los datos} \\
&\quad \text{atribuibles al uso de bloques distintos} \\
&= \text{suma de cuadrados de los bloques } (SCB) \\
\sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_{i+}-\bar{Y}_{+j}-\bar{Y}_{++})^2 &= \text{medida de la variabilidad de los datos debida} \\
&\quad \text{a factores aleatorios} \\
&= \text{residuo o suma de cuadrados del error } (SCE).
\end{aligned}$$

En forma simbólica, la identidad de la suma de cuadrados es:

Identidad conceptual de suma de cuadrados de bloques aleatorizados

$$SC_{Tot} = SC_{Tr} + SCB + SCE.$$

La hipótesis de medias de tratamientos iguales puede expresarse con base en los efectos de tratamientos τ_i. A fin de ver cómo se logra, note que si $\mu_{1+} = \mu_{2+} = ... = \mu_{k+}$, entonces

$$\begin{aligned}
\mu &= \sum_{i=1}^{k}\sum_{j=1}^{b}\mu_{ij}/kb \\
&= \sum_{i=1}^{k}\mu_{i+}/k \\
&= \mu_{i+} \quad \text{para cada } i = 1, 2, ..., k.
\end{aligned}$$

Por definición, $\tau_i = \mu_{i+} - \mu$. Por ende, si las medias de tratamiento son iguales, su valor común es μ y cada efecto de tratamiento tiene valor 0. La hipótesis nula del experimento:

$$H_0 : \mu_{1+} = \mu_{2+} = ... = \mu_{k+}$$

es equivalente a:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0.$$

Al igual que el caso del modelo de clasificación unidireccional, esta forma de H_0 es útil si se pretende considerar el diseño de bloques completos aleatorizados como un modelo lineal general y analizar los datos mediante técnicas de regresión.

3.3.2. Prueba de H_0

La prueba de esta hipótesis se obtiene de manera similar a la utilizada en el diseño de clasificación unidireccional. El uso de los supuestos del modelo y las reglas de la esperanza permite demostrar que el cuadrado de medias esperadas de los tratamientos esta dada por:

$$\begin{aligned} E[CM_{Tr}] &= E[SC_{Tr}/(k-1)] \\ &= \sigma^2 + \frac{b\sum_{i=1}^{k} \tau_i^2}{(k-1)}. \end{aligned}$$

Definir el cuadrado medio del error requiere notar en primer término que los grados de libertad relacionados con esta estadística corresponden a lo usual, a saber:

$$kb - 1 - [(k-1) + (b-1)] = (k-1)(b-1).$$

Al igual que antes, el cuadrado medio del error es un estimador insesgado de σ^2. En otras palabras :

$$E[CM_E] = E[SC_E/(k-1)(b-1)] = \sigma^2.$$

A fin de probar la hipótesis nula siguiente:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0 \quad \text{ó} \quad H_0 : \mu_{1+} = \mu_{2+} = ... = \mu_{k+}$$

se usa la razón F:

$$F = \frac{CM_{Tr}}{CM_E},$$

donde $CM_{Tr} = \frac{SC_{Tr}}{(k-1)}$. La hipótesis H_0 se debería de rechazar si $\frac{CM_{Tr}}{CM_E} > c$, donde c es una constante apropiada cuyo valor se puede determinar para cualquier nivel de significación a partir de una tabla de la distribución F con $(k-1)$ y $(k-1)(b-1)$ grados de libertad. La prueba sirve para rechazar H_0 si el valor observado del estadístico de prueba es excesivamente grande para haber ocurrido al azar.

Análogamente, supóngase ahora que se van a contrastar las siguientes hipótesis:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_b = 0 \\ H_1 &: \text{La hipótesis } H_0 \text{ no es cierta.} \end{aligned}$$

Cuando la hipótesis nula H_0 es cierta el estadístico de prueba tendrá una distribución F con $(b-1)$ y $(k-1)(b-1)$ grados de libertad:

$$F = \frac{CM_B}{CM_E},$$

donde $CM_B = \frac{SCB}{(b-1)}$. La hipótesis H_0 se debería de rechazar si $\frac{CM_B}{CM_E} > c$, donde c es una constante apropiada cuyo valor se puede determinar para cualquier nivel de significación a partir de una tabla de la distribución F con $(b-1)$ y $(k-1)(b-1)$ grados de libertad. En la tabla 5.5 se resumen las ideas desarrolladas en esta sección y se incluyen algunas fórmulas de cálculo de SC_{Tr} y SCB.

Tabla 3.5.
ANOVA para diseño de bloques completos aleatorizados de efectos fijos

F. de var.	G.l.	S.C	C.M	CME	F
Trat.	$k-1$	$\sum_{i=1}^{k} \frac{T_{i+}^2}{b} - \frac{T_{++}^2}{kb}$	$\frac{SC_{Tr}}{(k-1)}$	$\sigma^2 + b\sum_{i=1}^{k} \frac{\tau_i^2}{(k-1)}$	$\frac{CM_{Tr}}{CM_E}$
Bloque	$b-1$	$\sum_{j=1}^{b} \frac{T_{+j}^2}{k} - \frac{T_{++}^2}{kb}$	$\frac{SCB}{(b-1)}$	$\sigma^2 + k\sum_{j=1}^{b} \frac{\beta_j^2}{(b-1)}$	$\frac{CM_B}{CM_E}$
Error	$(k-1)(b-1)$	sustracción	$\frac{SCE}{(k-1)(b-1)}$		
Total	$kb-1$	$\sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}^2 - \frac{T_{++}^2}{N}$			

A manera de ilustración, se analizan los datos del siguiente ejemplo:

Ejemplo 3.9 *Los funcionarios de un sistema de transporte pequeño, con apenas cinco autobuses, necesitan evaluar el desgaste de cuatro tipos de neumáticos. Cada uno de los autobuses tiene ruta distinta, de modo que las condiciones de terreno y de conducción difieren de un vehiculo a otro. Es apropiado un diseño de bloques bidireccional para controlar el efecto de esta variable extraña. Cada autobús constituye un bloque y cada tipo de neumático, un tratamiento. En otras palabras se intenta probar :*

$$H_0 : \mu_{1+} = \mu_{2+} = \mu_{3+} = \mu_{4+}.$$

Se tienen calculadas las estadísticas de resumen siguientes:

$$\begin{array}{lll} T_{1+} = 61.8 & T_{+1} = 58.8 & T_{++} = 355 \\ T_{2+} = 100 & T_{+2} = 78.0 & \\ T_{3+} = 119.3 & T_{+3} = 80.1 & \\ T_{4+} = 73.9 & T_{+4} = 64.7 & \\ & T_{+5} = 73.4 & \end{array}$$

Tabla 3.6.
ANOVA de datos de desgaste de neumáticos

F. de variación	*G.l*	*S.C*	*C.M*	*F*
Tratamientos	*3*	*401.338*	*133.779*	*61.340*
Bloque	*4*	*81.525*	*20.381*	*9.345**
Error	*12*	*26.167*	*2.181*	
Total	*19*	*509.030*		

En relación con los datos que se tienen $\sum_{i=1}^{4}\sum_{j=1}^{5} Y_{ij}^2 = 6810.28$. El uso de las fórmulas de cálculo de la tabla (3.5) debe permitir que el lector verifique muchas de las cifras de la tabla ANOVA mostradas en la tabla anterior (3.6). Puesto que $f_{0.05}(3, 12) = 3.49$ y $61.34 > 3.49$, es posible rechazar:

$$H_0 : \mu_{1+} = \mu_{2+} = \mu_{3+} = \mu_{4+}$$

con $p < 0.05$. Se tiene evidencia estadística concluyente de las diferencias en la media de desgaste de la superficie de rodamiento de los cuatro tipos de neumáticos. El valor con asterisco en la tabla anterior es el observado de la razón CM_B/CM_E, usada para evaluar la efectivdad de la utilización de los bloques. Puesto que $f_{0.05}(4, 12) = 3.26$ y $9.345 > 3.26$, es posible rechazar la hipótesis nula $H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0$.

Capítulo 4

Factores de Bayes intrínsecos

4.1. Factores de Bayes intrínsecos para selección de modelos

¿Es necesario otro criterio para la selección de modelos? Obviamente pensamos, ¿para qué? Primero, sentimos que el modelo de selección debe tener una base bayesiana. La selección de modelos de métodos bayesianos y contraste de hipótesis son particularmente necesarios por las siguientes razones:

a) Medidas basadas en cálculos frecuentistas, tal como valor-p son las más grandes dificultades y las más engañosas.

b) Análisis de modelos no anidados y/o contraste de hipótesis es muy dificultoso en un marco frecuentista.

c) Métodos no bayesianos tienen dificultad incorporada. Si dos modelos explican datos igualmente de bien, entonces el modelo más simple será preferido, los factores de Bayes hacen esto automáticamente.

d) La predicción es frecuentemente la meta real.

Una premisa básica de nuestra motivación es que uno necesita métodos automáticos de selección de modelos. La comunidad bayesiana continua un debate sobre que métodos deben ser usados, objetivos o subjetivos; muchos bayesianos aceptan que ambos métodos pueden ser usados. El argumento a favor de métodos automáticos de selección de modelos es particularmente evidente, porque frecuentemente uno tiene inicialmente una gran variedad de modelos, y una especificación cuidadosa de distribuciones a priori para todos los parámetros de todos los modelos es típicamente no factible. Sin embargo, estos procedimientos tienen la limitación de que si se trabaja con distribuciones a priori impropias, entonces los factores de Bayes quedan definidos salvo una constante multiplicativa.

A continuación analizaremos una técnica para el desarrollo de factores de

Bayes por defecto, los cuales llamaremos Factores de Bayes intrínsecos (FBI).

4.1.1. Preliminares

Consideremos los modelos $M_1, M_2, \ldots, M_n$ bajo la consideración que los datos tienen densidad $f_i(x|\theta_i)$ bajo el modeloM_i. Los vectores de parámetros θ_i son desconocidos y de dimensión k_i.

En el procedimiento bayesiano de selección de modelos, bajo el enfoque M-cerrado se procede seleccionando distribuciones a priori $\pi_i(\theta_i)$, para los parámetros de cada modelo, junto con probabilidades a priori p_i de que cada modelo sea verdadero. La probabilidad a posteriori que M_i es verdadero es entonces:

$$P(M_i|x) = \left(\sum_{j=1}^{n} \frac{p_j}{p_i} \cdot B_{ji} \right)^{-1}, \tag{4.1}$$

donde B_{ji} es el factor de Bayes de M_j a M_i, definido por:

$$B_{ji} = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}, \tag{4.2}$$

aquí $m_j(x)$ es la marginal o densidad predictiva de X bajo el modelo M_j.

La demostración de la ecuación (4.1) es como sigue

$$\begin{aligned} P(M_i|x) &= \frac{P(x|M_i)P(M_i)}{P(x)} = \frac{P(x|Mi)P(Mi)}{\sum_{j=1}^{n} P(x|M_j)P(M_j)} \\ &= \frac{m_i(x)p_i}{\sum_{j=1}^{n} m_j(x)p_j} = \left(\sum_{j=1}^{n} \frac{m_j(x)p_j}{m_i(x)p_i} \right)^{-1} = \left(\sum_{j=1}^{n} \frac{p_j}{p_i} \cdot B_{ji} \right)^{-1}. \end{aligned}$$

Aunque usamos un lenguaje bayesiano estándar, note que no asumimos estrictamente que uno de los modelos es el verdadero; en particular B_{ji} puede ser visto como el cociente de verosimilitud de M_j con M_i, y por lo tanto puede interpretarse solamente en términos comparativos de soporte de los datos para los dos modelos. Aunque formalmente se discutirá solo el problema de selección de modelos, éste desarrollo puede ser aplicable al contraste de hipótesis.

Para calcular B_{ji} se requiere la especificación de $\pi_i(\theta_i)$ y $\pi_j(\theta_j)$. Frecuentemente en análisis bayesiano, uno puede usar distribuciones a priori no informativas $\pi_i^N(\theta_i)$. Comúnmente se elige la "distribución a priori uniforme", $\pi_i^U(\theta_i) = 1$; la

distribución a priori de Jeffreys, $\pi_i^J(\theta_i) = (\det(I_i(\theta_i)))^{\frac{1}{2}}$, donde $I_i(\theta_i)$ es la matriz de información esperada de Fisher correspondiente al modelo M_i; y la distribución a priori de referencia, $\pi_i^R(\theta_i)$) la cual ha sido probada por Bernardo (1979) y Berger y Bernardo (1992).

Usando cualquiera de las distribuciones a priori antes mencionadas, obtenemos

$$B_{ji}^N = \frac{m_j^N(x)}{m_i^N(x)} = \frac{\int f_j(x|\theta_j)\pi_j^N(\theta_j)d\theta_j}{\int f_i(x|\theta_i)\pi_i^N(\theta_i)d\theta_i}, \tag{4.3}$$

la dificultad con la ecuación (4.3) es que $\pi_i^N(\theta_i)$ son típicamente impropias, y por lo tanto son definidas salvo constantes arbitrarias c_i. Por lo tanto B_{ji} esta definido salvo una constante c_j/c_i la cual es arbitraria.

Una solución común a este problema es usar parte de los datos como una muestra de entrenamiento. Sea $x(l)$ que denota la parte de los datos que será usada. La idea es que $x(l)$ será usada para convertir la $\pi_i^N(\theta_i)$ a una distribución a posteriori propia

$$\pi_i^N(\theta_i|x(l)) = f_i(x(l)|\theta_i)\pi_i^N(\theta_i)/m_i^N(x(l)),$$

donde (abusando un poco de la notación) $f_i(x(l)|\theta_i)$ es la densidad marginal de $X(l)$ bajo M_i, y

$$m_i^N(x(l)) = \int f_i(x(l)|\theta_i)\pi_i^N(\theta_i)d\theta_i. \tag{4.4}$$

La idea es entonces calcular los factores de Bayes con el resto de los datos, usando $\pi_i^N(\theta_i|x(l))$ como distribución a priori, esto puede facilitarse con el siguiente lema.

Lema 4.1 *El factor de Bayes del modelo j al modelo i, condicional a $x(l)$ y asumiendo que $\pi_i^N(\theta_i|x(l))$ es propia, esta dado por:*

$$B_{ji}(x(l)) = B_{ji}^N \cdot B_{ij}^N(x(l)), \tag{4.5}$$

donde

$$B_{ij}^N(x(l)) = m_i^N(x(l))/m_j^N(x(l)).$$

Demostración *Denotando el resto de los datos por $x(n-l)$ el resultado se sigue de la definición.*

$$B_{ji}(x(l)) = \frac{\int f_j(x(n-l)|\theta_j, x(l))\pi_j^N(\theta_j|x(l))d\theta_j}{\int f_i(x(n-l)|\theta_i, x(l))\pi_i^N(\theta_i|x(l))d\theta_i},$$

pero

$$f_j(x(n-l)|\theta_j, x(l)) \cdot f_j(x(l)|\theta_j) = f_j(x|\theta_j),$$

y

$$\pi_j^N(\theta_j|x(l)) = \frac{f_j(x(l)|\theta_j)\pi_j^N(\theta_j)}{m_j^N(x(l))},$$

también

$$f_i(x(n-l)|\theta_i, x(l)) \cdot f_i(x(l)|\theta_i) = f_i(x|\theta_i),$$

y

$$\pi_i^N(\theta_i|x(l)) = \frac{f_i(x(l)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(l))},$$

así sustituyendo en $B_{ji}(x(l))$

$$\begin{aligned} B_{ji}(x(l)) &= \frac{\int \frac{f_j(x|\theta_j)}{f_j(x(l)|\theta_j)} \cdot \frac{f_j(x(l)|\theta_j)\pi_j^N(\theta_j)}{m_j^N(x(l))} d\theta_j}{\int \frac{f_i(x|\theta_i)}{f_i(x(l)|\theta_i)} \cdot \frac{f_i(x(l)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(l))} d\theta_i} = \frac{\int \frac{f_j(x|\theta_j)\pi_j^N(\theta_j)}{m_j^N(x(l))} d\theta_j}{\int \frac{f_i(x|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(l))} d\theta_i} \\ &= \frac{\int f_j(x|\theta_j)\pi_j^N(\theta_j) m_i^N(x(l)) d\theta_j}{\int f_i(x|\theta_i)\pi_i^N(\theta_i) m_j^N(x(l)) d\theta_i} = B_{ji}^N \cdot B_{ij}^N(x(l)). \blacksquare \end{aligned}$$

Claramente, quitando la arbitrariedad en la elección de las constantes multiplicativas de las $\pi_i^N(\theta_i)$; el cociente c_j/c_i que multiplica a B_{ji}^N debe ser cancelado por el cociente c_i/c_j que debe multiplicar a $B_{ji}^N(x(l))$. El uso de muestras aleatorias tiene sentido, sólo si $m_i^N(x(l))$ en (4.4) son finitas. Esto se formaliza en la siguiente definición.

Definición 4.1 *Una muestra de entrenamiento,* $x(l)$*, sera llamada propia si* $0 < m_i^N(x(l)) < \infty$ *para todo modelo* M_i*, y es llamada minimal si es propia y no tiene subconjuntos propios.*

Ejemplo 4.1 *Supóngase que* $X = (X_1, X_2, \ldots, X_n)$ *donde los* X_i *son i.i.d.* $\mathcal{N}(\mu, \sigma_2^2)$ *bajo* M_2*. Bajo* M_1*, los* X_i *son i.i.d* $\mathcal{N}(0, \sigma_1^2)$*. Consideremos las distribuciones a priori no informativas* $\pi_1^N(\sigma_1) = 1/\sigma_1 \quad y \quad \pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$*. Esto es directo para mostrar que* $m_2^N(x_i) = \infty$ *para una simple observación, pero son una muestra de entrenamiento para 2 distintas observaciones propias. Enseguida verificaremos esta afirmación.*

$$m_1^N(x(l)) = \frac{1}{2\pi(x_i^2 + x_j^2)}, \qquad m_2^N = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}. \tag{4.6}$$

De la definición se sigue que

$$m_1^N(x(l)) = \int f_1(x(l)|\theta_1)\pi_1^N(\theta_1)d\theta_1 = \int_0^\infty \frac{1}{2\pi} e^{-\frac{1}{2\sigma_1^2}(x_i^2+x_j^2)} \frac{1}{\sigma_1^3} d\sigma_1,$$

tomando $u = -\frac{1}{2\sigma_1^2}(x_i^2 + x_j^2)$, $\qquad du = \frac{(x_i^2+x_j^2)}{\sigma_1^3} d\sigma_1,$

$$m_1^N(x(l)) = \int_{-\infty}^0 \frac{1}{2\pi} e^u \frac{du}{(x_i^2+x_j^2)} = \frac{1}{2\pi(x_i^2+x_j^2)} e^{-\frac{1}{2\sigma_1^2}(x_i^2+x_j^2)}|_0^\infty = \frac{1}{2\pi(x_i^2+x_j^2)},$$

también

$$\begin{aligned} m_2^N(x(l)) &= \int f_2(x(l)|\theta_2)\pi_2^N(\theta_2)d\theta_2 \\ &= \int_{-\infty}^\infty \int_0^\infty \frac{1}{2\pi\sigma_2^2} e^{-\frac{1}{2\sigma_1^2}\left[(x_i-\mu)^2+(x_j-\mu)^2\right]} d\sigma_2 d\mu \\ &= \int_{-\infty}^\infty \int_0^\infty \frac{1}{2\pi\sigma_2^2} e^{-\frac{1}{2\sigma_1^2}\left\{2\left[\left(\mu-\frac{1}{2}(x_i+x_j)\right)^2 - x_i x_j\right] + \frac{1}{2}(x_i+x_j)^2\right\}} \frac{1}{\sigma_2^2} d\sigma_2 d\mu \\ &= \int_0^\infty \frac{e^{-\frac{1}{4\sigma_1^2}(x_i+x_j)^2} \cdot e^{\frac{1}{\sigma_1^2}x_i x_j}}{2\pi} \frac{1}{\sigma_2^3} \left[\int_{-\infty}^\infty e^{-\frac{1}{\sigma_1^2}\left(\mu-\frac{1}{2}(x_i+x_j)\right)^2} \cdot \frac{1}{\sigma_2} d\mu\right] d\sigma_2, \end{aligned}$$

eligiendo $u = \left(\mu - \frac{1}{2}(x_i+x_j)\right)/\sigma, \qquad du = d\mu/\sigma,$

$$m_2^N(x(l)) = \int_0^\infty \frac{e^{-\frac{1}{4\sigma_1^2}\left[(x_i+x_j)^2 - 4x_i x_j\right]}}{2\pi} \frac{1}{\sigma_2^3} \sqrt{\pi} d\sigma_2,$$

haciendo $u = -\frac{1}{4\sigma_1^2}\left[(x_i+x_j)^2 - 4x_i x_j\right], \qquad du = \frac{\left[(x_i+x_j)^2 - 4x_i x_j\right]}{2\sigma_2^3} d\sigma_2,$

$$\begin{aligned} m_2^N(x(l)) &= \frac{1}{\sqrt{\pi}\left[(x_i+x_j)^2 - 4x_i x_j\right]} \int_{-\infty}^0 e^u du \\ &= \frac{1}{\sqrt{\pi}\left[(x_i+x_j)^2 - 4x_i x_j\right]} = \frac{1}{\sqrt{\pi}\,(x_i-x_j)^2}. \end{aligned}$$

Típicamente, para una muestra de entrenamiento minimal todos los parámetros son identificables. Frecuentemente será una muestra de tamaño $max\{k_i\}$, recordemos que k_i es la dimensión de θ_i. Si los π_i^N son densidades propias entonces la muestra de entrenamiento minimal es el conjunto vacío y $B_{ji}(x(l) = B_{ji}^N$.

4.1.2. Relación con otros métodos bayesianos automáticos

Uso de distribuciones a priori convencionales

Jeffereys (1961) introdujo el uso de distribuciones a priori convencionales para la selección de modelos y contraste de hipótesis. En la situación del ejemplo (4.1), él argumentó el uso de

$$\pi_1(\sigma_1) = \frac{1}{\sigma_1}, \qquad \pi_2(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \frac{1}{\pi\sigma_2(1 + \mu_2/\sigma_2^2)}, \tag{4.7}$$

para la cual utilizó la distribución a priori no informativa estándar para parámetros de escala pero una densidad (propia) de Cauchy$(0, \sigma_2)$ para la a priori condicional de μ dado σ_2. Eligiendo $\pi(\mu|\sigma_2)$ propia, la indeterminación salvo una constante del factor de Bayes es eliminada, al menos en términos de μ. Jeffreys identificó $\sigma_1^2 = \sigma_2^2 = \sigma^2$ en esta situación y por lo tanto no se preocupó acerca de la indeterminación de $\pi(\sigma) = 1/\sigma$, si esta a priori ocurre en ambos modelos, entonces una constante multiplicativa debe ser cancelada.

El argumento de Jeffreys para (4.7) es bastante largo y puede o no ser convincente. Su solución no obstante es bastante razonable, eligiendo la "constante" de la a priori de μ dado σ_2 es de forma natural, y se sabe que las a priori de Cauchy son robustas en varios sentidos. Aunque es fácil objetar teniendo tales "imposiciones" sobre el análisis, es crucial recordar que no hay otra alternativa (excepto subjetividad). Cualquier método alternativo corresponde a la imposición de alguna (propia) a priori, o peor aun, termina por no corresponder a cualquier análisis bayesiano actual. Este problema es suficientemente importante para merecer énfasis.

Principio 1. Los métodos que corresponden a el uso de distribuciones a priori (propias) por defecto plausibles son preferibles a aquellas que no corresponden a cualquier análisis bayesiano actual.

Se intentará mencionar cuales métodos bayesianos por defecto son consistentes y no consistentes con este principio. Algunas propuestas que son consistentes con dicho principio son las siguientes: Albert (1990), George y McCulloch (1993), Madigan Y Raftery (1994), McCulloch y Rosi (1993), Mitchel y Beauchamp (1988), Poirier (1985), Raftery (1993), Stewart (1987), Verdinelli y Wasserman (1995) y Sellner y Siow (1980). La limitación de estas aproximaciones es que ellas fueron construidas para problemas específicos y nuestra meta es construir un método completamente general y automático, pero que sea consistente con el principio anterior.

Métodos asintóticos y criterio de información de Bayes

El método asintótico de Laplace (Haughton 1988, Kass y Raftery 1995) es producido como una aproximación a B_{ji}.

$$B_{ji}^{L} = \frac{f_j(x|\hat{\theta}_j)(\det \hat{I}_j)^{-1/2}}{f_i(x|\hat{\theta}_i)(\det \hat{I}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2}\pi_j(\hat{\theta}_j)}{(2\pi)^{k_i/2}\pi_i(\hat{\theta}_i)}, \tag{4.8}$$

donde $\hat{I}_i$, y $\hat{\theta}_i$, son la matriz de observación y el estimador de máxima verosimilitud (EMV) bajo el modelo M_i. Cuando el tamaño de la muestra tiende a infinito, el primer factor de B_{ji}^{L} típicamente tiende a cero o a infinito, mientras el segundo factor esta acotado. El criterio de información de Bayes (CIB) de Schwarz (1978) surgió reemplazando el segundo factor por una constante apropiada. Kass y Waserman (1995) argumentaron que, para modelos anidados, el CIB corresponde a un análisis bayesiano.

La expresión asintótica (4.8) tiene mucha utilidad teórica. Esta puede ser usada para ayudar a desarrollar distribuciones a priori propias por defecto y comparar criterios en la selección de modelos.

Como un comentario final, existen muchos problemas de selección de modelos para los cuales el resultado asintótico difiere de (4.8) (ver Haughton y Dudley 1992). Es interesante ver que los factores de Bayes intrínsecos (FBI) son compatibles bayesianamente.

Distribuciones a priori no informativas convencionales

En la sección (4.1.1) se observó que el problema con las distribuciones a priori no informativas, π_i^N es que ellas están definidas salvo constantes multiplicativas arbitrarias c_i y que tales constantes deben ser factores multiplicativas de los factores de Bayes. Se han hecho esfuerzos para especificar convencionalmente las constantes c_i. Para instanciarlas, Smith y Spiegelhater (1980) y Spiegelhalter y Smith (1982) proponen elegir los c_i de tal forma que $B_{ji}(x(l))$ sea igual a 1, cuando $x(l)$ se elige como la muestra de entrenamiento minimal, que favorece mas al modelo simple.

Este método se acerca a satisfacer el principio 1. El fallo es que tiene una tendencia en favor del modelo más complejo. Esta tendencia surge de la especificación que $B_{ji}(x(l))$ va a ser 1, aunque la muestra de entrenamiento sea escogida para favorecer al máximo al modelo más simple.

Métodos de entrenamiento y métodos de verosimilitud parcial

La idea de muestras de entrenamiento, ha sido usada muchas veces informalmente. Desarrollos más formales de esta idea pueden ser encontrados en trabajos

de Atkinson (1978), Geisser y Eddy (1979), Gelfand, Dey y Chang (1992), Lempers(1971), San Martíni y Spezzaferri (1984), y Spiegelhalter y Smith (1982), aunque no todos estos trabajos utilizan la idea con factores de Bayes ordinarios. Otras referencias y el comportamiento asintótico general de los modelos de muestra de entrenamiento han sido probados por Gelfand y Dey (1994).

Se puede mostrar que si el tamaño de la muestra de entrenamiento se incrementa con la muestra de tamaño n, entonces la muestra del factor de Bayes es no asintoticamente equivalente a (4.8).

El método de Aitkin (1991) puede ser considerado como un método de muestra de entrenamiento, tomando la muestra entera x, como una muestra de entrenamiento para obtener $\pi_i^N(\theta_i|x)$ y entonces usar esta como la distribución a priori en (4.5) para calcular el factor de Bayes. Este doble uso de los datos es por supuesto no consistente con la lógica bayesiana usual, y el método viola el criterio asintótico de forma severa.

O'Hagan (1995) propuso usar una parte fraccional de la verosimilitud entera, $[f(x|\theta)]^\alpha$, en lugar de una muestra de entrenamiento. Esto tiende a producir una respuesta más estable que la producida por el uso de una muestra de entrenamiento particular, pero falla el criterio asintótico a menos que $\alpha \propto 1/n$ cuando la muestra de tamaño n crece. El tamaño del factor fraccional de Bayes ha sido bastante estudiado, particularmente para las elecciones tales que $\alpha = m_0/n$, donde m_0 es la muestra de entrenamiento minimal. Esta elección puede resultar en los factores de Bayes que corresponden a el uso de distribuciones a priori por defecto, al menos para modelos lineales y ciertas elecciones de distribuciones a priori no informativas, tal justificación debe ser formalmente establecida.

Independientemente del trabajo de Berger y Pericchi, de Vos (1993) ha propuesto un método de muestra de entrenamiento para modelos lineales que es similar al método propuesto por Berger y Pericchi.

4.2. Factores de Bayes para módelos anidados

4.2.1. Modelos anidados

Asumamos que M_1 es anidado en M_2, en el sentido que nosotros podemos escribir $\theta_2 = (\zeta, \eta)$ y que f_1 y f_2 satisfacen

$$f_1(x|\theta_1) = f_2(x|\zeta = \theta_1, \eta = \eta_0) \tag{4.9}$$

donde η_0 es un valor específico de η. Algunas veces simplemente escribimos $\theta_2 = (\theta_1, \eta)$, aunque esto es peligroso (pero común) en la práctica. El peligro es que θ_1 en $f_1(x|\theta_1)$ y θ_2 en $f_2(x|\theta_1, \eta)$ pueden tener interpretaciones muy diferentes,

porque los símbolos son los mismos, es demasiado fácil asignar entonces la misma distribución a priori (especialmente cuando se usan las distribuciones a priori por defecto). Esta es una importante cuestión, porque muchos de los esquemas para el desarrollo de a priori condicionales son basados en una formalización en la identificación de tales parámetros.

El siguiente supuesto es siempre verdadero para modelos anidados.

Supuesto 1. Si M_1 es anidado en M_2, entonces asumimos que como el tamaño de la muestra tiende a infinito ($n \to \infty$),

$$\hat{\theta}_2 \overset{\text{bajo } M_1}{\longrightarrow} \theta_2^* = (\theta_1, \eta_0). \tag{4.10}$$

4.2.2. Los factores de Bayes intrínsecos

Para un conjunto dado de datos x, típicamente existen muchas muestras de entrenamiento como la que se definió en la sección 4.1.1. Sea

$$X_T = \{x(1), x(2), ..., x(L)\} \tag{4.11}$$

que denota el conjunto de todas las muestras minimales $x(l)$. Claramente el factor de Bayes $B_{21}(x(l))$, definida en (4.5), depende de la elección de la muestra de entrenamiento minimal. Para eliminar esta dependencia e incrementar estabilidad, una idea natural es promediar los $B_{21}(x(l))$ sobre todos los $x(l) \in X_T$. Este promedio puede ser hecho aritméticamente o geométricamente, primero el factor de Bayes intrínseco aritmético (FBIA) y enseguida el factor de Bayes intrínseco geométrico (FBIG) quedan definidos de la siguiente manera:

$$B_{21}^{IA} = \frac{1}{L}\sum_{l=1}^{L} B_{21}(x(l)) = B_{21}^{N} \cdot \frac{1}{L}\sum_{l=1}^{L} B_{12}^{N}(x(l)) \tag{4.12}$$

$$B_{21}^{IG} = \left(\prod_{l=1}^{L} B_{21}(x(l))\right)^{1/L} = B_{21}^{N} \cdot \left(\prod_{l=1}^{L} B_{21}^{N}(x(l))\right)^{1/L}, \tag{4.13}$$

donde el $B_{12}^{N}(x(l))$ esta definido en (4.5). Note que $B_{21}^{IG} \leq B_{21}^{IA}$ porque la media geométrica es menor o igual que la media aritmética. Por tanto B_{21}^{IG} favorecerá al modelo anidado más simple que B_{21}^{IA}.

Nota 1: Definimos B_{12}^{IA} como $1/B_{12}^{IA}$, y no por (4.12) con los índices invertidos. La asimetría surge porque M_1 esta anidado en M_2. Para B_{21}^{IG} no hay problema, invirtiendo los índices en (4.13) claramente resulta $1/B_{21}^{IG}$.

$$B_{12}^{IG} = \left(\prod_{l=1}^{L} B_{12}(x(l))\right)^{1/L} = B_{12}^{N} \cdot \left(\prod_{l=1}^{L} B_{21}^{N}(x(l))\right)^{1/L} = \left(\prod_{l=1}^{L} B_{21}(x(l))\right)^{1/L}$$

$$= B_{12}^{N} \cdot \left(\frac{1}{\prod_{l=1}^{L} B_{12}^{N}(x(l))}\right)^{1/L} = B_{12}^{N} \cdot \left(\prod_{l=1}^{L} B_{12}^{N}(x(l))\right)^{-1/L} = 1/B_{21}^{IG}.$$

Nota 2: Si el tamaño de la muestra es muy pequeña, entonces claramente se tendrá problema usando una parte de los datos como muestra de entrenamiento.

Ejemplo 4.2 *Sea* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *una muestra i.i.d. de* $M_1 : \mathcal{N}(0, \sigma_1^2)$ *o* $M_2 : \mathcal{N}(\mu, \sigma_2^2)$. *Usando* $\pi_1^N(\sigma_1) = 1/\sigma_1$ *y* $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$, *y haciendo algunos cálculos se tiene*

$$B_{21}^{N} = \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\bar{x}^2}{s^2}\right)^{n/2}, \tag{4.14}$$

donde

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

Sabemos que

$$B_{21}^{N} = \frac{m_2^N(x)}{m_1^N(x)} = \frac{\int_{\theta_2} f_2(x|\theta_2)\pi_2^N(\theta_2)d\theta_2}{\int_{\theta_1} f_1(x|\theta_1)\pi_1^N(\theta_1)d\theta_1},$$

$$m_2^N(x) = \int_{\theta_2} \frac{1}{(2\pi)^{n/2}\sigma_2^n} e^{-\frac{1}{2\sigma_2}\sum_{i=1}^{n}(x_i-\mu)^2} \cdot \frac{1}{\sigma_2^2} d\theta_2$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}\sigma_2^n} e^{-\frac{1}{2\sigma_2}\sum_{i=1}^{n}(x_i-\mu)^2} \cdot \frac{1}{\sigma_2^2} d\mu d\sigma_2,$$

y si tenemos que

$$\sum_{i=1}^{n} (xi - \mu)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n(\mu - \bar{x})^2,$$

$$m_2^N(x) = \int_0^{\infty}\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}\sigma_2^n} e^{-\frac{1}{2\sigma_2}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n(\mu-\bar{x})^2\right]} \cdot \frac{1}{\sigma_2^2} d\mu d\sigma_2$$

$$= \int_0^{\infty} \frac{1}{(2\pi)^{n/2}\sigma_2^n} e^{-\frac{1}{2\sigma_2}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_2}\left(n(\mu-\bar{x})^2\right)} \frac{1}{\sigma_2^2} d\mu d\sigma_2,$$

tomando $u = \frac{\sqrt{n}(\mu - x)}{\sigma}$ y $du = \frac{\sqrt{n}d\mu}{\sigma}$,

$$m_2^N(x) = \frac{\sqrt{2\pi}}{(2\pi)^{n/2}\sqrt{n}} \int_0^\infty e^{-\frac{1}{2\sigma_2}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right]} \frac{1}{\sigma_2^{n+1}} d\sigma_2,$$

haciendo $y = \sigma_2^{-2}$, $dy = -2\sigma_2^{-3}d\sigma_2$, *se tiene*

$$\begin{aligned} m_2^N(x) &= \frac{\sqrt{2\pi}}{2(2\pi)^{n/2}\sqrt{n}} \int_0^\infty -y^{\frac{n-2}{2}} e^{-y/2\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right]} dy \\ &= \frac{\sqrt{2\pi}}{2(2\pi)^{n/2}\sqrt{n}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{2}\right)^{n/2}}. \end{aligned}$$

Ahora

$$m_1^N(x) = \frac{1}{(2\pi)^{n/2}} \int_{\theta_2} \sigma_1^{-n-1} e^{-\frac{1}{2\sigma_2}\sum_{i=1}^n x_i{}^2} d\sigma_1,$$

eligiendo $y = \sigma_1^{-2}$, $dy = -2\sigma_1^{-3}d\sigma_1$, *y procediendo de forma análoga a la integral anterior tenemos*

$$m_1^N(x) = \frac{\sqrt{2\pi}}{2(2\pi)^{n/2}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{n/2}}.$$

por lo tanto

$$B_{21}^N = \frac{\frac{\sqrt{2\pi}}{2(2\pi)^{n/2}\sqrt{n}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{2}\right)^{n/2}}}{\frac{\sqrt{2\pi}}{2(2\pi)^{n/2}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{n/2}}} = \frac{\sqrt{2\pi}}{\sqrt{n}} \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right)^{n/2},$$

pero $s^2 = \sum_{i=1}^n (xi - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$, *así*

$$B_{21}^N = \sqrt{\frac{2\pi}{n}} \left(\frac{\sum_{i=1}^n x_i^2}{s^2}\right)^{n/2} = \sqrt{\frac{2\pi}{n}} \left(\frac{s^2 + n\bar{x}^2}{s^2}\right)^{n/2} = \sqrt{\frac{2\pi}{n}} \left(1 + \frac{n\bar{x}^2}{s^2}\right)^{n/2}.$$

Usando (4.8) observese que X_T consiste en todos los pares de observaciones diferentes $L = n(n-1)/2$, de esto se sigue que

$$\begin{aligned} B_{21}^{IA} &= B_{21}^N \cdot \frac{1}{L} \sum_{l=1}^L \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi}\left[x_1^2(l) + x_2^2(l)\right]} \\ &= B_{21}^N \cdot \frac{1}{n(n-1)} \sum_{i<j} \frac{(x_i - x_j)^2}{\sqrt{\pi}\left[x_i^2(l) + x_j^2(l)\right]}, \end{aligned} \tag{4.15}$$

y

$$B_{21}^{IG} = B_{21}^{N} \times \prod_{l=1}^{L} \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi}\,[x_1^2(l) + x_2^2(l)]}. \tag{4.16}$$

Nótese que B_{21}^{IA} y B_{21}^{IG} estan definidos esencialmente para cualesquiera modelos anidados inclusive para aquellos no estándares.

Ejemplo 4.3 *Supóngase que* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *es una muestra i.i.d. de* $M_1 : X_1 \sim \mathcal{N}(\theta_1, 1)$ *con* $\theta_1 < 0$ *o* $M_2 : X_2 \sim \mathcal{N}(\theta_2, 1)$ *con* $\theta_2 \in R^1$. *Una vez mas, es importante recordar que* θ_1 *y* θ_2 *podrían ser cantidades distintas a priori, incluso cuando* $\theta_2 < 0$. *Esto puede ser peligroso para formular este problema diciendo* $X_i \sim \mathcal{N}(\theta, 1)$ *con* $M_1 : \theta_1 < 0$ *y* $M_2 : \theta \in R^1$. *El peligro es al usar el mismo símbolo* θ, *apareciendo en* M_1 *y* M_2 *el cual puede causar que en un momento dado uno asume que* $\pi_1(\theta)$ *(bajo* M_1*) sea igual a* $\pi_2(\theta|\theta < 0)$ *bajo* M_2.

La distribución a priori no informativa usual es $\pi_1^N(\theta_1) = 1_{(-\infty,0)}$ y $\pi_2^N(\theta_2) = 1$. De donde se tiene que:

$$B_{21}^{N} = \frac{1}{\Phi(-\sqrt{n}\bar{x})}, \tag{4.17}$$

donde ϕ es la función de distribución acumulada de la normal estándar.

A continuacion se demostrará la ecuación (4.17). Por definición de $B_{21}^{N}(x)$ sabemos lo siguiente:

$$\begin{aligned} B_{21}^{N}(x) &= \frac{m_2^N(x)}{m_1^N(x)} = \frac{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_2)^2} d\theta_2}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_1)^2} d\theta_1} \\ &= \frac{\int_{-\infty}^{0} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_2)^2} d\theta_2 + \int_{0}^{\infty} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_2)^2} d\theta_2}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_1)^2} d\theta_1}, \end{aligned}$$

también tenemos que

$$
\begin{aligned}
& -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_2)^2 \\
= & -\frac{1}{2}\sum_{i=1}^{n}(x_i^2-2x_i\theta_2+\theta_2^2) \\
= & -\frac{1}{2}\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}x_i\theta_2-\frac{1}{2}\sum_{i=1}^{n}\theta_2^2 \\
= & -\frac{1}{2}\sum_{i=1}^{n}x_i^2+n\theta_2\bar{x}-\frac{n\theta_2^2}{2} \\
= & -\left(\frac{n\theta_2^2}{2}-n\theta_2\bar{x}\right)-\frac{1}{2}\sum_{i=1}^{n}x_i^2 \\
= & -\left(\frac{n\theta_2^2}{2}-n\theta_2\bar{x}+\frac{n\bar{x}^2}{2}\right)-\frac{1}{2}\sum_{i=1}^{n}x_i^2+\frac{n\bar{x}^2}{2} \\
= & -\left(\sqrt{\frac{n}{2}}\theta_2-\sqrt{\frac{n}{2}}\bar{x}\right)^2-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)
\end{aligned}
$$

así, si

$$\int_{-\infty}^{0}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_2)^2}d\theta_2=\int_{-\infty}^{0}\frac{1}{(2\pi)^{n/2}}e^{-\left(\sqrt{\frac{n}{2}}\theta_2-\sqrt{\frac{n}{2}}\bar{x}\right)^2-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)}d\theta_2,$$

al hacer $t=\sqrt{2}\left(\sqrt{\frac{n}{2}}\theta_2-\sqrt{\frac{n}{2}}\bar{x}\right)$ $\qquad dt=\sqrt{2}\cdot\sqrt{\frac{n}{2}}d\theta_2$, se obtiene

$$\int_{0}^{\infty}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-\theta_2^2\right)}d\theta_2=\frac{e^{-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)}}{(\sqrt{2\pi})^{n-1}\sqrt{n}}\int_{-\infty}^{-\sqrt{n}\bar{x}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt,$$

y por tanto

$$
\begin{aligned}
B_{21}^{N}(x) &= \frac{\int_{-\infty}^{-\sqrt{n}\bar{x}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt+\int_{-\sqrt{n}\bar{x}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt}{\int_{-\infty}^{-\sqrt{n}\bar{x}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt} \\
&= \frac{\Phi(-\sqrt{n}\bar{x})+(1-\Phi(-\sqrt{n}\bar{x}))}{\Phi(-\sqrt{n}\bar{x})}=\frac{1}{\Phi(-\sqrt{n}\bar{x})}.
\end{aligned}
$$

Luego de (4.12) y (4.13) resulta que

$$B_{21}^{IA} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \frac{1}{n} \sum_{i=1}^{n} \Phi(-x_i) \quad , \tag{4.18}$$

y

$$B_{21}^{IG} = \frac{1}{\Phi(-\sqrt{n}\bar{x})} \cdot \frac{1}{n} \left[\prod_{i=1}^{n} \Phi(-x_i) \right]^{1/n} \quad . \tag{4.19}$$

Esto es un ejemplo no estándar que es difícil de manejar por métodos ordinarios en parte esto se encuentra indicado por el hecho de que la expresión asintótica (4.8) y (por lo tanto el CIB), aquí no es válida. Las correcciones asintóticas han sido probadas por Haughton y Dudley (1992), quienes estudiaron muchos problemas muy generales de este tipo. Para este caso, la expresión análoga a (4.8) es :

$$B_{21} \cong \frac{\pi_2(\bar{x})}{\Phi(-\sqrt{n}\bar{x})\pi_1(\text{mín}\,\{\bar{x}, 0\})} \tag{4.20}$$

la cual es válida si π_2 es continua, y si π_1 es continua y tiene límite finito.

4.2.3. Los factores de Bayes intrínsecos esperados

Para una muestra de tamaño pequeño, el promedio de las muestras de entrenamiento en (4.12) y (4.13) pueden tener varianza grande, lo cual indica una inestabilidad de los factores de Bayes intrínsecos. También, el cálculo puede ser un problema si L es grande. Una atractiva solución fue propuesta por Berger y Pericchi (1996), remplazando los promedios en (4.12) y (4.13) por sus respectivas esperanzas, evaluadas en los estimadores de máxima verosimilitud. Formalmente ellos definieron el factor de Bayes Intrínseco Aritmético Esperado (FBIAE) y el Factor de Bayes Intrínseco Geométrico Esperado (FBIGE) por

$$B_{21}^{IAE} = B_{21}^{N} \cdot \frac{1}{L} \sum_{l=1}^{L} E_{\hat{\theta}_2}^{M_2} \left[B_{12}^{N}(X(l)) \right] \tag{4.21}$$

y

$$B_{21}^{IGE} = B_{21}^{N} \cdot \exp\left\{ \frac{1}{L} \sum_{l=1}^{L} E_{\hat{\theta}_2}^{M_2} \left[\log B_{12}^{N}(X(l)) \right] \right\}, \tag{4.22}$$

donde las esperanzas son bajo el modelo M_2, con θ_2 igual al estimador máximo verosímil $\hat{\theta}_2$. Si las $X(l)$ son intercambiables, como es común, entonces los promedios sobre L salen sobrando. Esto es, (4.21) y (4.22) son justificadas como

aproximación a (4.12) y (4.13) para L grande y bajo M_2 es obvio. Pero estas también son aproximaciones válidas bajo el modelo M_1 si asumimos el *supuesto* 1 en la sección 4.2.1. Es satisfecho por que (bajo el modelo M_1), $\hat{\theta}_2 \cong (\theta_1, \eta_0)$, lo cual junto con (4.9) muestra que las esperanzas en (4.21) y (4.22) son equivalentes.

Ejemplo 4.4 *Supongamos que las X(l) son intercambiables, así por (4.21) se tiene:*

$$B_{21}^{IAE} = B_{21}^{N} \cdot E_{\hat{\theta}_2}^{M_2}\left[\frac{(X_i - X_j)^2}{2\sqrt{\pi}\left(X_i^2 + X_j^2\right)}\right] = B_{21}^{N} \cdot \left(\frac{1 - e^{\left(-n\bar{x}/s^2\right)}}{2\sqrt{\pi}\left[n\bar{x}/s^2\right]}\right) \tag{4.23}$$

(para el cálculo de estas esperanzas ver Berger y Pericchi (1994)). B_{21}^{IGE} puede ser evaluado únicamente como una serie infinita (ver Berger y Pericchi (1994)), pero el cálculo númerico es directo.

Ejemplo 4.5 *(continuación del ejemplo 4.4) Usando (4.18) y (4.21), y considerando que los X(l) son intercambiables se tiene que*

$$B_{21}^{IAE} = \frac{1}{\phi(-\sqrt{n}\bar{x})} \cdot E_{\hat{\theta}_2}^{M_2}\left[\phi\left(-X_i\right)\right] = \frac{1}{\phi(-\sqrt{n}\bar{x})} \cdot \left(-\bar{x}/\sqrt{2}\right) \tag{4.24}$$

aquí $X_i \sim \mathcal{N}(\theta, 1)$ bajo M_2 y $\theta_2 = \bar{x}$. Otra vez B_{21}^{IGE} puede no tener una forma cerrada. Como ocurría con B_{12}^{IA}, Berger y Pericchi definen $B_{12}^{IAE} = 1/B_{21}^{IAE}$. En este caso no hay otra opción porque en muchos problemas $E_{\hat{\theta}_2}^{M_2}\left[B_{21}^{N}(X(l))\right] = \infty$. Esto también explica la definición de $B_{12}^{IAE} = 1/B_{21}^{IAE}$, aunque B_{12}^{IA} podría definirse como en (4.12) con los índices intercambiados, el promedio de $B_{12}^{N}(x(l))$ normalmente diverge cuando $L \to \infty$, resultando un factor de Bayes que viola el principio 1.

4.2.4. Comparaciones

Haremos una pausa para comparar los diferentes factores de Bayes intrínsecos con otros métodos seguros, así como obtener una visión en estos casos simples para ver si nuestra meta es alcanzada. Las comparaciones que hacemos son con la expresión asintótica (4.8), la aproximación de Schwarz y con los métodos de Jeffresys (1961), Smith y Spiegelhalter (1980).

Ejemplo 4.6 *Se puede demostrar que los estimadores de máxima verosimilitud para ambos modelos son: $\hat{\mu} = \bar{x}$, $\hat{\sigma}_1 = (\sum_{i=1}^{n} x_i^2/n)^{1/2} = (\bar{x}^2 + s^2/n)^{1/2}$, $\hat{\sigma}_2 = (s^2/n)^{1/2}$.*

Sabemos que $X_i \sim N(\mu, \sigma_2^2)$ *bajo* M_2.

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}(x_i\ -\mu)^2}, \qquad f(x_1, x_2, ..., x_n) = \frac{1}{(\sqrt{2\pi})^n \sigma_2^n} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n (x_i\ -\mu)^2},$$

asi

$$\begin{aligned} \ln f(x_1, x_2, ..., x_n) &= \ln\left(\frac{1}{(\sqrt{2\pi})^n \sigma_2^n}\right) - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu)^2 \\ &= -\ln\left(\sqrt{2\pi}\sigma_2\right)^n - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu)^2 \\ &= -n\left[\ln\sqrt{2\pi} + \ln\sigma_2\right] - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu)^2, \end{aligned}$$

luego

$$\begin{aligned} \frac{\partial \ln f(x_1, x_2, ..., x_n)}{\partial \mu} &= \frac{2\sum_{i=1}^n (x_i - \mu)}{2\sigma_2^2} = 0 \\ &\Rightarrow \sum_{i=1}^n (x_i - \mu) = 0 \\ &\Rightarrow \sum_{i=1}^n x_i - n\mu = 0 \Rightarrow \hat{\mu} = \bar{x}, \end{aligned}$$

también

$$\begin{aligned} \frac{\partial \ln f(x_1, x_2, ..., x_n)}{\partial \sigma_2} &= -\frac{n}{\sigma_2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_2^3} = 0 \\ &\Rightarrow \sigma_2 = \left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}\right)^{1/2} \Rightarrow \hat{\sigma}_2 = \left(s^2/n\right)^{1/2}. \end{aligned}$$

Se tiene que $X_i \sim N(0, \sigma_1^2)$ *bajo* M_1

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}x_i^2} \qquad f(x_1, x_2, ..., x_n) = \frac{1}{(\sqrt{2\pi})^n \sigma_1^n} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n x_i^2}$$

por lo cual

$$\begin{aligned} \ln f(x_1, x_2, ..., x_n) &= \ln\left(\frac{1}{(\sqrt{2\pi})^n \sigma_1^n}\right) - \frac{1}{2\sigma_1^2}\sum_{i=1}^n x_i^2 \\ &= -\ln\left(\sqrt{2\pi}\sigma_1\right)^n - \frac{1}{2\sigma_1^2}\sum_{i=1}^n x_i^2 \\ &= -n\left[\ln\sqrt{2\pi} + \ln\sigma_1\right] - \frac{1}{2\sigma_1^2}\sum_{i=1}^n x_i^2, \end{aligned}$$

luego

$$\frac{\partial \ln f(x_1, x_2, ..., x_n)}{\partial \sigma_1} = \frac{-n}{\sigma_1} + \frac{\sum_{i=1}^n x_i^2}{\sigma_1^3} = 0 \Rightarrow \hat{\sigma}_1 = \left(\frac{\sum_{i=1}^n x_i^2}{n}\right)^{1/2},$$

pero $s^2 = \sum_{i=1}^n (\bar{x}_i^2 - \mu)^2$, *asi*

$$\begin{aligned}
\hat{\sigma}_1 &= \left(\frac{\sum_{i=1}^n x_i^2}{n}\right)^{1/2} = \left(\frac{\sum_{i=1}^n x_i^2}{n} + (\bar{x} - \mu)^2\right)^2 \\
&= \left(\frac{\sum_{i=1}^n x_i^2}{n} + (\bar{x}^2 - 2\bar{x}\mu + \mu^2)\right)^{1/2} \\
&= \left(\bar{x}^2 + \frac{\sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2)}{n}\right)^{1/2} \\
&= \left(\bar{x}^2 + \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}\right)^{1/2} = \left(\bar{x}^2 + s^2/n\right)^{1/2}.
\end{aligned}$$

A continuación se tienen las siguientes aproximaciones:
Aproximación Asintótica (4.8) dada por:

$$B_{21}^L = B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\mu}, \hat{\sigma}_2)}{\pi_1(\hat{\sigma}_1)}. \tag{4.25}$$

Solución:
De antemano sabemos que

$$B_{ji}^L = \frac{f_j(x|\hat{\theta}_j)(\det \hat{I}_j)^{-1/2}}{f_i(x|\hat{\theta}_i)(\det \hat{I}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2}\, \pi_j(\hat{\theta}_j)}{(2\pi)^{k_i/2}\, \pi_i(\hat{\theta}_i)},$$

donde $\hat{I}_i$ y $\hat{\theta}_i$ son la matriz de información observada y el estimador de máxima verosimilitud.

Bajo M_1, $Xi \sim N(0, \sigma_1^2)$, asi $k_1 = 1$.

$$\begin{aligned}
\hat{I}_1(x) &= -\left[\frac{\partial^2 \log f(x|\sigma_1)}{\partial \sigma_1^2}\right]_{\sigma_1=\hat{\sigma}_1} = -\left[\frac{\partial\left(\frac{-n}{\sigma_1} + \frac{\sum_{i=1}^n x_i^2}{\sigma_1^3}\right)}{\partial \sigma_1}\right]_{\sigma_1=\hat{\sigma}_1} \\
&= -\left[\frac{n}{\sigma_1^2} - \frac{\sum_{i=1}^n x_i^2 \cdot 3\sigma_1^2}{\sigma_1^6}\right]_{\sigma_1=\hat{\sigma}_1} = -\left[\frac{n}{\sigma_1^2} - \frac{3\sum_{i=1}^n x_i^2}{\sigma_1^4}\right]_{\sigma_1=\hat{\sigma}_1},
\end{aligned}$$

como $\hat{\sigma}_1 = (\sum_{i=1}^n x_i^2/n)^{1/2}$,

entonces

$$\begin{aligned}
\hat{I}_1(x) &= -\frac{n}{\frac{\sum_{i=1}^n x_i^2}{n}} + \frac{3\sum_{i=1}^n x_i^2}{\left(\frac{\sum_{i=1}^n x_i^2}{n}\right)^2} \\
&= -\frac{n^2}{\sum_{i=1}^n x_i^2} + \frac{3n^2\sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} \\
&= \frac{2n^2}{\sum_{i=1}^n x_i^2}, \\
(\det \hat{I}_1(x))^{-1/2} &= \frac{1}{(\det \hat{I}_1(x))^{1/2}} = \frac{1}{\left(\frac{2n^2}{\sum_{i=1}^n x_i^2}\right)^{1/2}} \\
&= \left(\frac{\sum_{i=1}^n x_i^2}{2n^2}\right)^{1/2} = \frac{\left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{1/2}}{\sqrt{2}n} \\
&= \frac{\hat{\sigma}_1}{\sqrt{2}n}.
\end{aligned}$$

Bajo M_2 $X_i \sim N(\mu, \sigma_2^2)$

$$\hat{I}_2(x) = -\left[\begin{array}{cc} \frac{\partial^2 \log f(x|\mu,\sigma_2)}{\partial \mu^2} & \frac{\partial^2 \log f(x|\mu,\sigma_2)}{\partial \mu \partial \sigma_2} \\ \frac{\partial^2 \log f(x|\mu,\sigma_2)}{\partial \sigma_2 \partial \mu} & \frac{\partial^2 \log f(x|\mu,\sigma_2)}{\partial \sigma_2^2} \end{array}\right]_{(\mu,\sigma)=(\hat{\mu},\hat{\sigma}_2)}$$

$$\begin{aligned}
ln(f(x_1, x_2, ..., x_n|\mu, \sigma_2) &= \ln\left(\frac{1}{(2\pi)^{n/2}\sigma_2^n}\right) - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu)^2 \\
&= -\left[\ln(2\pi)^{n/2} + \ln(\sigma_2^n)\right] - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu)^2,
\end{aligned}$$

así

$$\begin{aligned}
\frac{\partial \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \mu} &= \frac{2}{2\sigma_2^2}\sum_{i=1}^n (x_i - \mu) = \frac{1}{\sigma_2^2}\sum_{i=1}^n (x_i - \mu), \\
\frac{\partial^2 \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \mu^2} &= \frac{-n}{\sigma_2^2}, \\
\frac{\partial^2 \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \mu \partial \sigma} &= \frac{\partial^2 \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \sigma \partial \mu} = \frac{-2\sum_{i=1}^n (x_i - \mu)}{\sigma_2^3}, \\
\frac{\partial \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \sigma} &= -\frac{n}{\sigma_2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_2^3}, \\
\frac{\partial^2 \ln(f(x_1, x_2, ..., x_n|\mu, \sigma)}{\partial \sigma^2} &= \frac{n}{\sigma_2^2} - \frac{3\sum_{i=1}^n (x_i - \mu)^2}{\sigma_2^4},
\end{aligned}$$

por lo tanto

$$\begin{aligned}\hat{I}_2(x) &= -\left[\begin{array}{cc}\frac{-n}{\sigma_2^2} & \frac{-2\sum_{i=1}^n(x_i-\mu)}{\sigma_2^3}\\ \frac{-2\sum_{i=1}^n(x_i-\mu)}{\sigma_2^3} & \frac{n}{\sigma_2^2}-\frac{3\sum_{i=1}^n(x_i-\mu)^2}{\sigma_2^4}\end{array}\right]_{(\mu,\sigma)=(\hat{\mu},\hat{\sigma}_2)=(\bar{x},(s^2/n)^{1/2})}\\ &= -\left[\begin{array}{cc}\frac{-n}{\sigma_2^2} & \frac{-2\sum_{i=1}^n(x_i-\bar{x})}{\hat{\sigma}_2^3}\\ \frac{-2\sum_{i=1}^n(x_i-\bar{x})}{\hat{\sigma}_2^3} & -\frac{2n}{\hat{\sigma}_2^2}\end{array}\right],\end{aligned}$$

$$\begin{aligned}\det(\hat{I}_2(x)) &= \frac{n}{\sigma_2^2}\cdot\left(\frac{2n}{\sigma_2^2}\right)-\left(\frac{2\sum_{i=1}^n(x_i-\bar{x})}{\sigma_2^3}\right)^2\\ &= \frac{2n^2}{\sigma_2^4}-\frac{4\left(\sum_{i=1}^n(x_i-\bar{x})\right)^2}{\sigma_2^6}\\ &= \frac{2n^2\sigma_2^2-4\left(\sum_{i=1}^n(x_i-\bar{x})\right)^2}{\sigma_2^6}\\ &= \frac{2n^2\sigma_2^2}{\sigma_2^6}=\frac{2n^2}{\hat{\sigma}_2^4},\end{aligned}$$

luego

$$(\det\hat{I}_2(x))^{-1/2}) = \frac{1}{\left(\frac{2n^2}{\hat{\sigma}_2^4}\right)^{1/2}} = \frac{\hat{\sigma}_2^2}{\sqrt{2}n},$$

además

$$B_{ji}^L = \frac{f_j(x|\hat{\theta}_j)(\det\hat{I}_j)^{-1/2}}{f_i(x|\hat{\theta}_i)(\det\hat{I}_i)^{-1/2}}\cdot\frac{(2\pi)^{k_j/2}\,\pi_j(\hat{\theta}_j)}{(2\pi)^{k_i/2}\,\pi_i(\hat{\theta}_i)},$$

$$\begin{aligned}B_{ji}^L &= \frac{\frac{1}{(\sqrt{2\pi})^n\sigma_2^n}e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n(x_i\,-\mu)^2}\cdot\left(\frac{\hat{\sigma}_2^2}{\sqrt{2}n}\right)\cdot(2\pi)^{2/2}\cdot\pi_2(\hat{\theta}_2)}{\frac{1}{(\sqrt{2\pi})^n\sigma_1^n}e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n{x_i}^2}\cdot\left(\frac{\hat{\sigma}_1}{\sqrt{2n}}\right)\cdot(2\pi)^{1/2}\cdot\pi_1(\hat{\theta}_1)}\\ &= \frac{\sigma_1^n e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n(x_i\,-\mu)^2}\cdot\hat{\sigma}_2^2\cdot(2\pi)^{1/2}\cdot\pi_2(\hat{\theta}_2)}{\sigma_2^n e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n{x_i}^2}\cdot\hat{\sigma}_1\cdot\sqrt{n}\cdot\pi_1(\hat{\theta}_1)},\end{aligned}$$

pero

$$\begin{aligned}s^2 &= \sum_{i=1}^n(x_i-\bar{x})^2,\ \hat{\sigma}_1=(\sum_{i=1}^n x_i^2/n)^{1/2}=(\bar{x}^2+s^2/n)^{1/2},\\ \hat{\sigma}_2 &= (s^2/n)^{1/2}\ \text{ y }\ \sum_{i=1}^n x_i^2 = s^2+n\bar{x}^2.\end{aligned}$$

Así

$$\begin{aligned} B_{ji}^{L} &= \frac{\sigma_1^n \cdot \hat{\sigma}_2^2 \cdot e^{-\frac{n}{2}} \cdot (2\pi)^{1/2} \cdot \pi_2(\hat{\theta}_2)}{\sigma_2^n \cdot \hat{\sigma}_1 \cdot e^{-n/2}\sqrt{n} \cdot \pi_1(\hat{\theta}_1)} = \sqrt{\frac{2\pi}{n}} \cdot \frac{\sigma_1^n \cdot \hat{\sigma}_2^2 \cdot \pi_2(\hat{\theta}_2)}{\sigma_2^n \cdot \hat{\sigma}_1 \cdot \pi_1(\hat{\theta}_1)} \\ &= \sqrt{\frac{2\pi}{n}} \cdot \frac{(\bar{x}^2 + s^2/n)^{n/2} \cdot \hat{\sigma}_2^2 \cdot \pi_2(\hat{\theta}_2)}{(s^2/n)^{n/2} \cdot \hat{\sigma}_1 \cdot \pi_1(\hat{\theta}_1)} = \sqrt{\frac{2\pi}{n}} \cdot (\frac{n\bar{x}^2 + s^2}{s^2})^{n/2} \frac{\cdot \hat{\sigma}_2^2 \cdot \pi_2(\hat{\theta}_2)}{\cdot \hat{\sigma}_1 \cdot \pi_1(\hat{\theta}_1)} \\ &= B_{21}^{N} \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\theta}_2)}{\pi_1(\hat{\theta}_1)}. \end{aligned}$$

La aproximación de Schwarz esta dada por

$$B_{21}^{s} = B_{21}^{N}/\sqrt{2\pi}. \tag{4.26}$$

Mientras que la aproximación de Jeffreys, para la distribución a priori en (4.7) es:

$$B_{21}^{J} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \frac{1}{\hat{\sigma}_2 \pi \left[1 + \hat{\mu}^2/\hat{\sigma}_2^2\right]}.$$

El factor de Bayes de Smith y Spiegelhalter (1980) es

$$B_{21}^{ss} = B_{21}^{N}/\sqrt{\pi}.$$

Es usual escribir $\pi_2(\mu, \sigma_2)$ como $\pi_2(\mu, \sigma_2) = \pi_2(\mu|\sigma_2)\pi_2(\sigma_2)$.

Si ahora elegimos las distribuciones a priori no informativas como $\pi_1(\sigma_1) = 1/\sigma_1$ y $\pi_2(\sigma_2) = 1/\sigma_2$, entonces de (4.21) tenemos

$$B_{21}^{L} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \pi_2(\hat{\mu}/\hat{\sigma}_2). \tag{4.27}$$

Notese que B_{21}^{J} es de esta forma, con $\pi_2(\mu/\sigma_2)$ una distribución de Cauchy $(0, \sigma_2)$. Además que es consistente con (4.7). Así, reescribiendo (4.23)

$$B_{21}^{IAE} = B_{21}^{N} \cdot \hat{\sigma}_2 \cdot \left(\frac{1 - e^{\{-\hat{\mu}^2/\hat{\sigma}_2^2\}}}{2\sqrt{\pi}\left[\hat{\mu}^2/\hat{\sigma}_2^2\right]} \right). \tag{4.28}$$

Recordemos que una de las metas de Berger y Pericchi (1996) era desarrollar un método automático que reproduzca un factor de Bayes verdadero. Resulta que B_{21}^{IAE} se comporta bien ya que

$$\pi_2^*(\mu|\sigma_2) = \frac{1 - e^{\{-\hat{\mu}^2/\hat{\sigma}_2^2\}}}{2\sqrt{\pi}\left[\hat{\mu}^2/\hat{\sigma}_2^2\right]}, \tag{4.29}$$

es una distribución a priori propia (integrando sobre μ) y, además es equivalente a la elección de Jeffreys de $\pi_2(\mu|\sigma_2)$ como Cauchy $(0, \sigma_2)$; de hecho, las densidades de las dos distribuciones a priori nunca difieren en mas de 15 %.

Esta interesante propiedad de B_{21}^{IAE} es desafortunadamente no compartida con B_{21}^{IGE}. Este no corresponde a un factor de Bayes de una disitribución a priori propia, difiriendo por una constante multiplicativa.

Comparando B_{21}^{J} ó B_{21}^{IAE} con B_{21}^{SS}, vemos que el último es más grande por un factor de alrededor de $\sqrt{\pi}(1 + \hat{\mu}^2/\hat{\sigma}^2)$, el cual es siempre más grande que 1. Este fenómeno es generalmente verdadero y proporciona soporte para la aceptación de que B_{21}^{SS} esta cargado hacia el modelo más complejo.

Similarmente B_{21}^{S} es más grande por un factor de $\sqrt{\pi}(1 + \hat{\mu}^2/\hat{\sigma}_2^2)$, y este es basado hacia el modelo más complejo. Si uno elige $\pi_2(\mu|\sigma_2)$ a ser $\mathcal{N}(0, \sigma_2^2)$, entonces $B_{21}^{L} = B_{21}^{S} \cdot \exp\{-x^2/2\hat{\sigma}_2^2\}$. Si M_1 fuera el verdadero modelo y n fuese grande, entonces $\bar{x} \cong 0$ y $B_{21}^{L} \cong B_{21}^{S}$. Esta es la base para el argumento de Kass y Wasserman (1994) esto es que B_{21}^{S}es aproximadamente un factor de Bayes cuando el modelo simple es el verdadero, en una situación de modelos anidados.

La historia para B_{21}^{IA} y B_{21}^{IG} es similar. Remarcando, el promedio en B_{21}^{IA} nuevamente corresponde a una distribución a priori propia. El correspondiente término de (4.13) es cualitativamente similar, pero nuevamente no es propia.

4.2.5. Distribuciones a priori intrínsecas

En el ejemplo 1 en sección 4.2.4 vimos que B_{21}^{IA} y B_{21}^{IAE} eran aproximadamente iguales a los factores de Bayes para la distribución a priori (condicional) propia. Tal distribución a priori si existe es llamada *una distribución a priori intrínseca.* Vimos los efectos que los FBI tienden a corresponder a los factores de Bayes verdaderos con respecto a distribuciones a priori intrínsecas para ser su justificación más fuerte. Por lo tanto la determinación de las distribuciones a priori intrínsecas es teóricamente interesante, además proviene del mejor comportamiento de los FBI.

También hay un beneficio potencial en determinar las distribuciones a priori intrínsecas. Un beneficio obvio es que las distribuciones a priori intrínsecas pueden ser usadas en lugar de las π_i^N, para calcular el factor de Bayes actual. Esto eliminaría la necesidad de calcular muestras de entrenamiento y eliminaría preocupaciones acerca de la estabilidad de los factores de Bayes intrínsecos. De hecho, uno puede ver alternativamente el procedimiento de los factores de Bayes, como un método para aplicar a "muestras de entrenamiento imaginarias", para determinar distribuciones a priori convencionales actuales para ser usadas en la selección de modelos y contraste de hipótesis.

Como con las distribuciones a priori de referencia, nuestra definición de dis-

tribuciones a priori intrínsecas será centrada alrededor de una muestra de entrenamiento imaginaria. Frecuentemente, se pueden utilizar argumentos asintóticos para verificar la siguiente aproximación, que Berger y Pericchi establecieron como condición.

Condicion A. Suponga que el tamaño de muestra tiende a infinito, para distribuciones a priori en una clase apropiada, Γ, (4.2) puede aproximarse por

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_i(\hat{\theta}_i)\pi_j^N(\hat{\theta}_j)} \cdot (1 + o(1)), \tag{4.30}$$

donde $\hat{\theta}_j$ y $\hat{\theta}_i$ son los estimadores de máxima verosimilitud bajo los modelos M_i y M_j. (aquí $o(1) \to 0$ en probabilidad bajo M_i y M_j).

Esta condición puede verse fácilmente para comprender la situación asintótica estándar (4.8), pero también para comprender la situación no estándar.

Para definir las distribuciones a priori intrínseca, empezaremos igualando la ecuación (4.30) con (4.12) o (4.13) produciendo

$$\frac{\pi_j(\hat{\theta}_j)\pi_i^N(\hat{\theta}_i)}{\pi_i(\hat{\theta}_i)\pi_j^N(\hat{\theta}_j)}(1 + o(1)) = \tilde{B}_{ij}^N, \tag{4.31}$$

definiendo $\tilde{B}_{ij}^N$ como la media geométrica o la media aritmética de $B_{ij}^N(x(l))$. Enseguida necesitamos asumir algo acerca del comportamiento de los límites de las cantidades en (4.31). La siguiente condición es típicamente satisfecha.

Condición B. Cuando el tamaño de muestra tiende a infinito, se tiene lo si-guiente

i) Bajo M_j, $\hat{\theta}_j \longrightarrow \theta_j$, $\hat{\theta}_i \longrightarrow \Psi_i(\theta_j)$, y $\tilde{B}_{ij}^N \longrightarrow \tilde{B}_j^*(\theta_j)$.

i) Bajo M_i, $\hat{\theta}_i \longrightarrow \theta_i$, $\hat{\theta}_j \longrightarrow \Psi_j(\theta_i)$, y $\tilde{B}_{ij}^N \longrightarrow \tilde{B}_i^*(\theta_i)$.

Donde, $\Psi_i(\theta_j)$ denota el límite del estimador máximo verosímil $\hat{\theta}_i$ bajo el modelo M_j en el punto θ_j.

Típicamente, para $k = i$ o $k = j$

$$\tilde{B}_k^*(\theta_k) = \begin{cases} \lim\limits_{L\to\infty} E_{\theta_k}^{M_k}\left[\frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(X(l))\right] & \text{caso aritmético,} \\ \\ \lim\limits_{L\to\infty} \exp\left\{E_{\theta_k}^{M_k}\left[\frac{1}{L}\sum_{l=1}^{L} \log B_{ij}^N(X(l))\right]\right\} & \text{caso geométrico.} \end{cases} \tag{4.32}$$

Si los $X(l)$ son intercambiables, entonces los límites y los promedios sobre L pueden eliminarse.

Usando la condición B y tomando límites en (4.31), primero bajo M_j y después bajo M_i, resultan las siguientes 2 ecuaciones que definen las distribuciones a priori intrínsecas (π_j^I, π_i^I).

$$\frac{\pi_j^I(\theta_j)\pi_i^N(\Psi_i(\theta_j))}{\pi_j^N(\theta_j)\pi_i^I(\Psi_i(\theta_j))} = B_j^*(\theta_j), \tag{4.33}$$

$$\frac{\pi_i^N(\theta_i)\pi_j^I(\Psi_j(\theta_i))}{\pi_i^I(\theta_j)\pi_j^N(\Psi_j(\theta_i))} = B_i^*(\theta_i). \tag{4.34}$$

La motivación es otra vez, que las distribuciones a priori que satisfagan las ecuaciones (4.33) y (4.34), tendrán respuestas que serán asintóticamente equivalentes a aquellas obtenidas usando los FBI.

Notemos que las soluciones no son necesariamente únicas, ni necesariamente propias.

En el escenario de los modelos anidados de la sección 4.2.1 y bajo el supuesto 1, las soluciones a (4.33) y (4.34), están dadas por

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1), \qquad \pi_2^I(\theta_2) = \pi_2^N(\theta_2) \cdot B_2^*(\theta_2). \tag{4.35}$$

Típicamente, hay otras soluciones, quizás incluso son soluciones que son distribuciones propias, pero las soluciones en (4.35) son las más simples.

Ejemplo 4.7 *Para B_{21}^{IA} se sigue de (4.23) y (4.32) que $B_2^*(\theta_2) = \sigma_2 \cdot \pi_2^*(\mu|\sigma_2)$ donde $\pi_2^*(\mu|\sigma_2)$ fue definido en (4.29). Por lo tanto las distribuciones a priori intrínseca son*

$$\pi_1^I(\sigma_1) = \pi_1^N(\sigma_1) = 1/\sigma_1,$$

$$\pi_2^I(\mu, \sigma_2) = \pi_2^N(\sigma_2)B_2^*(\mu, \sigma_2). = \frac{1}{\sigma_2} \cdot \pi_2^*(\mu|\sigma_2). \tag{4.36}$$

B_{21}^{IA} se comporta (asintóticamente) como un factor de Bayes verdadero que usa las distribuciones a priori no informativas de referencia para σ_1 y σ_2, y la distribución propia $\pi_2^*(\mu|\sigma_2)$, para la distribución propia condicional de μ dado σ_2. Además de la propiedad de $\pi_2^*(\mu|\sigma_2)$, es notable que la distribución a priori intrínseca para σ_2 es la a priori de referencia $1/\sigma_2$, y no la distribución a priori de Jeffreys (formal) $1/\sigma_2^2$ usada para deducir B_{21}^{IA}.

Berger y Pericchi observaron este último comportamiento en otros ejemplos; los FBI parecen intentar convertir la original π_i^N en distribuciones a priori de referencia para parámetros comunes o modelos similares.

Ejemplo 4.8 *Para B_{21}^{IA}, vemos de (4.18), (4.25), y (4.32) que $B_2^*(\theta_2) = \Phi(-\theta_2/\sqrt{2})$. Por lo tanto las distribuciones a priori intrínsecas (4.35) son:*

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1) = 1, \qquad \pi_2^I(\theta_2) = 1 \cdot \ \Phi(-\theta_2/\sqrt{2}) \ . \tag{4.37}$$

Dos características de estas distribuciones a priori intrínsecas son particularmente interesantes. Primero, en $(-\infty, 0)$, π_1^I y π_2^I no son proporcionales. Recordemos que se escribieron los modelos como $M_1 : \theta_1 < 0$, $M_2 : \theta_2 \in R^1$, a diferencia de $M_1 : \theta_1 < 0$, $M_2 : \theta_2 \in R^1$, para enfatizar que los θ_i pueden tener diferentes interpretaciones bajo cada modelo y distribuciones a priori diferentes, incluso en su dominio común. Esta posibilidad parece haberse realizado. Note, sin embargo que sobre $(-\infty, 0)$, π_1^I y π_2^I difieren substancialmente cerca de cero. La segunda característica interesante de la distribución a priori intrínseca es que $\int_0^\infty \pi_2^I(\theta_2)d\theta_2 = \int_0^\infty \Phi(-\theta_2/\sqrt{2})\ d\theta_2 = 1/\sqrt{\pi}$, por lo tanto $\pi_2^I(\theta_2|\{\theta_2 > 0\})$ es propia.

El comportamiento de las distribuciones a priori intrínsecas, observadas en los ejemplos anteriores es típicamente para modelos anidados. Parámetros "comunes" (o al menos parámetros que pueden ser identificados en el sentido de la ecuación (4.9)) típicamente tienen distribuciones a priori intrínsecas que son distribuciones a priori estándares no informativas o ligeras variantes, mientras que paramétros que están solo en modelos más complejos (o que tienen dominios extendidos en los modelos más complejos) tienen distribuciones a priori intrínsecas propias (condicionales).

Teorema 4.1 *Para los FBIA suponga que (4.32) se cumple y que $\pi_1^N(\theta_1)$ es propia. Entonces $\pi_2^I(\theta_2)$, definida en (4.35) es también propia.*

Demostración *Puesto que se asume que el límite existe en (4.32), es cierto que*

$$\begin{aligned}
\int \pi_2^I(\theta_2)d\theta_2 &= \int \pi_2^N(\theta_2) \left(\lim_{L\to\infty} E_{\theta_2}^{M_2} \left[\frac{1}{L} \sum_{l=1}^{L} B_{12}^N(X(l)) \right] \right) d\theta_2 \\
&= \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \int \int \pi_2^N(\theta_2) f_2(x(l)|\theta_2) B_{12}^N(x(l)) d(x(l) d\theta_2 \\
&= \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \int m_2^N(x(l)).\left[m_1^N(x(l))/m_2^N(x(l)) \right] d(x(l))
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \int m_1^N(x(l)) d(x(l) \\
&= \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} (1) = \lim_{L\to\infty} \frac{L}{L} = 1. \blacksquare
\end{aligned} \tag{4.38}
$$

La última fila se sigue del efecto que $\pi_1^N(\theta_1)$ es propia obteniendo m_1^N también lo es. La gran generalidad de este teorema es atenuada por el requerimiento de que $\pi_1^N(\theta_1)$ es propia. Finalmente nótese que no existe un teorema análogo para B_{21}^{GI}.

Capítulo 5

Selección bayesiana de modelos aplicada al ANOVA

Sean $\mathcal{N}(x_1|\mu_1, \sigma_1^2), ..., \mathcal{N}(x_k|\mu_k, \sigma_k^2)$ las distribuciones normales con medias $\mu_1, ...\mu_k$ y varianzas $\sigma_1^2, ..., \sigma_k^2$ desconocidas. Los tamaños de las muestras consideradas son n_i, es decir, $x_i = (x_{i1}, x_{i2}, ..., x_{in_i})$, y las medias muestrales y varianzas son $\bar{x}_i$ y s_i^2/n_i, $i = 1, 2, ..., k$, respectivamente.

El análisis clásico de varianza unidireccional consiste en contrastar si las medias μ_i son todas iguales.

Bajo el punto de vista frecuentista este problema posee la dificultad de que un F-contraste exacto (Scheffe 1959) solo existe bajo la condición de que las varianzas son todas iguales (condición de homocedasticidad). En general, en un contexto de heterocedasticidad la razón de verosimilitudes falla (Scheffe 1959, Stuart y Ord 1991), de tal manera que la teoría normal de modelos lineales no puede aplicarse. Las aproximaciones asintóticas entonces son necesarias: Welch (1951) propuso una aproximación basada en un intervalo de confianza asintótico. Una aproximación asintótica alternativa basada en una modificación del estadístico F del análisis de varianza usual fue dado por Brown y Forsythe (1974). Para comparaciones de estas aproximaciones, ver Tan y Tabatabai (1986) y Brown y Forsythe (1974).

Bajo el punto de vista bayesiano la solución al análisis de varianza bajo homocedasticidad (Lindley 1970; Box y Tiao 1973), es basado en la distribución a posteriori de los parámetros $(\lambda_1, \lambda_2, ..., \lambda_k)$, donde $\lambda_i = \mu_i - \mu$, $i = 1, ..., k$. Para un $\alpha \in (0, 1)$ dado la región más grande que contiene una probabilidad igual a $1 - \alpha$ es calculada. Entonces, cualquier punto de esta región es aceptado como verosímil: en particular, la regla es aplicada a la hipótesis nula $H_0 : \lambda_1 = \lambda_2 = ... = \lambda_k$. Esto no puede ser considerado como una solución bayesiana: la hipótesis nula, que no es tomada en cuenta en la formulación del problema, tiene una probabilidad a posteriori igual a cero. Esta dificultad proviene como consecuencia de que el

problema se ha enfocado como uno de estimación cuando realmente es un problema de contraste. Para distinguir entre problemas de estimación y problemas de contraste es recomendable ver Jeffreys (1961, pp. 245-249).

La propuesta aquí es formular el análisis clásico de varianza como un problema de selección de modelos, en el cual la condición de homocedasticidad no se impone, y para el cual Moreno, Bertolino y Racugno (2001) proponen un procedimiento bayesiano por defecto.

Consideremos los modelos anidados

$$M_1 : \ f_1(z(\theta_1)) = \mathcal{N}(x_1|\mu, \tau_1^2) \cdots \mathcal{N}(x_k|\mu, \tau_k^2), \tag{5.1}$$

y

$$M_2 : \ f_2(z(\theta_2)) = \mathcal{N}(x_1|\mu_1, \sigma_1^2) \cdots \mathcal{N}(x_k|\mu_k, \sigma_k^2), \tag{5.2}$$

donde $z = (x_1, ..., x_k)$, $\theta_1 = (\mu, \tau_1, ..., \tau_k)$ y $\theta_2 = (\mu_1, ..., \mu_k, \sigma_1, ..., \sigma_k)$. Considerando que μ y τ_i son distribuciones a priori independientes, las distribuciones a priori convencionales para μ y $log\ \tau_i$ son distribuciones uniformes (Jeffreys 1961. p. 138), por tanto

$$\pi_1^N(\theta_1) = \frac{c_1}{\prod_{i=1}^{k} \tau_i}. \tag{5.3}$$

Similarmente, asumamos que μ y σ_1 son también distribuciones a priori independientes, la distribución a priori convencional es

$$\pi_2^N(\theta_2) = \frac{c_2}{\prod_{i=1}^{k} \sigma_i}. \tag{5.4}$$

En (5.3) y (5.4), c_1 y c_2 son constantes positivas que no pueden ser especificadas ya que $\pi_1^N(\theta_1)$ y $\pi_2^N(\theta_2)$ no son funciones integrables. Sin la suposición de independencia, la regla de Jeffreys daría lugar a distribuciones a priori ligeramente diferentes. Una discusión interesante sobre la selección de distribuciones a priori por defecto esta dado en Kass y Wasserman (1996).

Los modelos a comparar son $M_1 : \left\{ f_1(z|\theta_1), \pi_1^N(\theta_1) \right\}$ y $M_2 : \left\{ f_2(z|\theta_2), \pi_2^N(\theta_2) \right\}$. Dada una distribución a priori P, sobre $\{M_1, M_2\}$, la probabilidad a posteriori de M_1 es:

$$P(M_1|z) = \left(1 + B_{21}^N(z) \frac{P(M_2)}{P(M_1)} \right)^{-1},$$

donde $z = (x_1, x_2, ..., x_k)$ y el factor

$$B_{ji}^N = \frac{m_2^N(x)}{m_1^N(x)} = \frac{\int f_2(z|\theta_2)\pi_2^N(\theta_2)d\theta_2}{\int f_1(z|\theta_1)\pi_1^N(\theta_1)d\theta_1}, \tag{5.5}$$

es el conocido como el factor Bayes de M_2 frente a M_1. Este factor encierra toda la información que tienen los datos alrededor de la distribución de probabilidad a posteriori de M_1.

Desafortunadamente, el factor de Bayes en (5.5) está definido salvo una constante multiplicativa c_2/c_1. Para evitar esta dificultad se han propuesto algunas alternativas tales como: la aproximación asintótica de Schwarz (1978), llamado también "Criterio de Información Bayesiana" (CIB); el factor de Bayes fraccional (FBF) y el factor de Bayes intrínseco (FBI) propuesto por O'Hagan (1994) y Berger y Pericchi (1996), respectivamente. Críticas y comparaciones de estos métodos relevantes son incluidos en Berger y Pericchi (1998), De Santis y Spezzaferri (1999), Moreno (1997), Moreno, Bertolino y Racugno (1998) y O'Hagan (1995,1997). Debemos enfatizar que aunque el factor de Bayes fraccional o el factor de Bayes intrínseco no son factores de Bayes verdaderos, son asintóticamente consistentes. Esto permite, bajo ligeras condiciones, hallar distribuciones a priori intrínsecas y fraccionales, para obtener un factor de Bayes verdadero. Motivaciones y justificaciones para el uso de distribuciones a priori intrínsecas en problemas de selección de modelos han sido dadas por Berger y Pericchi (1997), Moreno (1997), Moreno, Bertolino y Racugno (1998). En este capítulo se construyen y se calculan los correspondientes factores de Bayes del modelo M_2 frente a M_1. Esta aproximación tiene interesantes propiedades: (i) es un procedimiento bayesiano completamente "automático" para comparación de modelos, (ii) estos factores de Bayes dependen de la muestra a través de estadísticos suficientes, (iii) no dependen de muestras de entrenamiento específicas de tal manera que la estabilidad no representa un problema, (iv) la igualdad $B_{21}(z) = 1/B_{12}(z)$ es válida y consecuentemente se satisface la igualdad $P(M_2|z) = 1 - P(M_1|z)$.

5.1. El método intrínseco

La idea es dividir la muestra aleatoria x en dos partes, $x = (x(l), x(n-l))$. La muestra de entrenamiento $x(l) = (x_1, x_2, ..., x_l)$ será usada para convertir la distribución a priori impropia a una distribución a priori propia

$$\pi_i(\theta_i|x(l)) = \frac{f_1(x(l)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(l))},$$

donde $m_i^N(x(l)) = \int f_i(x(l)|\theta_i)\pi_i^N(\theta_i)d\theta_i$, $i = 1, 2$. El factor de Bayes es entonces calculado usando los datos $x(n-l)$ y $\pi_i(\theta_i|x(l))$ como la distribución a priori

propia.

Resultando el factor de Bayes parcial

$$B_{21}(x(n-l)|x(l)) = B_{21}^N(x) \cdot B_{12}^N(x(l)),$$

donde

$$B_{12}^N(x(l)) = \frac{m_1^N(x(l))}{m_2^N(x(l))}.$$

Note que $B_{21}(x(n-l)|x(l))$ esta bien definido sólo si $x(l)$ es tal que $0 < m_i^N(x(l)) < \infty$, $i = 1, 2$. Si no hay una submuestra de $x(l)$ para la cual se cumple la desigualdad anterior, $x(l)$ es llamada una muestra de entrenamiento minimal.

5.2. El factor de Bayes intrínseco

El factor de Bayes parcial depende de la muestra de entrenamiento. Para evitar la dificultad de elegir $x(l)$, Berger y Pericchi (1996) propusieron usar una muestra de entrenamiento minimal para calcular $B_{21}(x(n-l)|x(l))$. Entonces, un promedio sobre todas las posibles muestras de entrenamiento minimal es considerada. De esto resulta el factor de Bayes intrínseco aritmético (FBIA) de M_2 frente a M_1 :

$$B_{21}^{IA}(x) = B_{21}^N(x) \cdot \frac{1}{L} \sum_{l=1}^{L} B_{12}^N x(l) \tag{5.6}$$

donde L es el número de muestras de entrenamiento minimal $x(l)$ contenida en x.

5.3. Distribuciones a priori intrínsecas

Sea $M_1 : \{f_1(x|\theta_1), \pi_1(\theta_1)\}$ y $M_2 : \left\{f_2(z|\theta_2), \pi_2^N(\theta_2)\right\}$ dos modelos generales donde

(i) $f_1(x|\theta_1)$ es anidado en $f_2(x|\theta_2)$,

(ii) $\pi_1(\theta_1)$ es una a priori propia,

(iii) $\pi_2^N(\theta_2)$ es una a priori impropia.

Asúmase que para alguna muestra de tamaño n la función de verosimilitud $f_2(x_1, x_2, ..., x_n|\theta_2)$ es una función integrable con respecto a $\pi_2^N(\theta_2)$; un argumento asintótico muestra que la metodología intrínseca genera una única densidad de probabilidad $\pi_2^I(\theta_2)$, llamada distribución a priori intrínseca (Berger y Pericchi 1996). Esta distribución a priori esta dada por

$$\pi_2^I(\theta_2) = \pi_2^N(\theta_2) E^{M_2}_{x(l)|\theta_2} B_{12}^N(x(l)), \tag{5.7}$$

donde la esperanza es tomada con respecto a la densidad de la muestra de entrenamiento minimal $x(l)$ bajo el modelo M_2.

El modelo muestreado en (5.1) esta anidado en (5.2), y para la distribución a priori dada en (5.4), la muestra de entrenamiento minimal es un vector aleatorio $z(l) = (x_{ij}, i = 1, 2, ..., k, j = 1, 2)$, donde x_{i1}, x_{i2} son independientes y distribuidas bajo M_1 como $\mathcal{N}(x|\mu, \tau_i^2)$ y bajo M_2 como $\mathcal{N}(x|\mu_i, \sigma_i^2)$.

Teorema 5.1 *Las distribuciones a priori intrínsecas para comparar los modelos M_1 frente a M_2, son $\left\{\pi_1^N(\theta_1), \pi_2^I(\theta_2)\right\}$, donde $\pi_2^I(\theta_2) = \int \pi_2^I(\theta_2|\theta_1)\pi_1^N(\theta_1)d\theta_1$ y*

$$\pi_2^I(\theta_2|\theta_1) = \prod_{i=1}^{k} \mathcal{N}(\mu_i|\mu, \frac{\tau_i^2 + \sigma_i^2}{2}) HC^+(\sigma_i|0, \tau_i), \tag{5.8}$$

$HC^+(\sigma_i|0, \tau_i)$ denota la media densidad de Cauchy.

Demostración *Consideremos los modelos*

$$M_1 : f_1(z|\theta_1) = \prod_{i=1}^{k} \mathcal{N}(x_i|\mu, \tau_i^2), \quad \pi_1(\theta_1) = \delta(\mu, \tau_1, ..., \tau_k),$$

y

$$M_2 : f_2(z|\theta_2) = \prod_{i=1}^{k} \mathcal{N}(x_i|\mu_i, \sigma_i^2), \quad \pi_2^N(\theta_2) = \frac{c_2}{\prod_{i=1}^{k} \sigma_i},$$

donde $\delta(\mu, \tau_1, ..., \tau_k)$ es la delta de Dirac en el punto $\theta_1 = (\mu, \tau_1, ..., \tau_k)$ y c_2 es una constante positiva arbitraria. Luego tenemos

$$B_{12}^N(z(l)) = \frac{\prod_{i=1}^{k} \prod_{j=1}^{2} \mathcal{N}(x_{ij}|\mu, \tau_i^2)}{m_2^N(z(l))}$$

donde

$$m_2^N(z(l)) = \frac{c_2}{2^k} \frac{1}{\prod_{i=1}^{k} |x_{i1} - x_{i2}|}.$$

Mostraremos el resultado anterior para (x_1, x_2) y lo generalizaremos para (x_{i1}, x_{i2}).

Sean X_1 y X_2 observaciones independientes de una densidad de localización-escala $\sigma^{-1} g((x_i - \mu)/\sigma)$ y $\pi^N(\mu, \sigma) = 1/\sigma$, entonces para $x_1 \neq x_2$,

$$m(x_1, x_2) = \frac{1}{2\,|x_1 - x_2|}.$$

Sin perdida de generalidad consideremos que $x_2 > x_1$ *y se hacen los cambios de variables* $(\mu, \sigma) \rightarrow (v, w) \equiv ((x1 - \mu)/\sigma, (x2 - \mu)/\sigma)$, *luego despejando se tiene*

$$\sigma = \frac{x_1 - x_2}{v - w} \qquad y \qquad \mu = \bar{x} - \frac{(x1 - x2)}{2}\frac{(v + w)}{(v - w)},$$

y el jacobiano de la transformación es de la forma

$$\begin{aligned} J &= \left| \begin{array}{cc} \frac{\partial \mu}{\partial v} & \frac{\partial \mu}{\partial w} \\ \frac{\partial \sigma}{\partial v} & \frac{\partial \sigma}{\partial w} \end{array} \right| = \left| \begin{array}{cc} \frac{(x_1 - x_2)w}{(v-w)^2} & \frac{-(x_1 - x_2)v}{(v-w)^2} \\ \frac{-(x_1 - x_2)}{(v-w)^2} & \frac{x_1 - x_2}{(v-w)^2} \end{array} \right| = -\frac{(x_1 - x_2)^2}{(v - w)^3} \\ &= \frac{-\sigma^2}{v - w} = \frac{-\sigma^2}{\frac{x1 - x2}{\sigma}} = \frac{-\sigma^3}{x1 - x2} = \frac{\sigma^3}{|x1 - x2|}, \end{aligned}$$

donde $v \in (-\infty, \infty)$ *y* $w \in (v, \infty)$.*Considerando el jacobiano y las nuevas variables se tiene que*

$$\begin{aligned} m(x_1, x_2) &= \frac{1}{|x_1 - x_2|} \int_{-\infty}^{\infty} \int_{v}^{\infty} g(v) g(w) dw dv \\ &= \frac{1}{|x_1 - x_2|} \cdot P(V < W), \end{aligned}$$

donde V *y* W *son variables aleatorias independientes con densidad* $g(\cdot)$. *Puesto que* $P(V < W) = P(W < V) = 1/2$, *se concluye que*

$$m(x_1, x_2) = \frac{1}{2\,|x_1 - x_2|},$$

y por tanto

$$m_2^N(z(l)) = \frac{c_2}{2^k} \frac{1}{\prod_{i=1}^{k} |x_{i1} - x_{i2}|}.$$

Entonces se tiene que la distribución a priori intrínseca condicional de θ_2 *viene*

dada por la siguiente ecuación

$$\begin{aligned}
\pi_2^I(\theta_2|\theta_1) &= \pi_2^N(\theta_2)E^{M_2}_{z(l)|\theta_2}B_{12}^N(z(l)) \\
&= \frac{c_2}{\prod_{j=1}^k \sigma_j} \times \prod_{i=1}^k \int\int \left(\frac{2^k\,|x_{i1}-x_{i2}|}{c_2}\right)\mathcal{N}(x_{i1}|\mu,\tau_i^2)\mathcal{N}(x_{i2}|\mu,\tau_i^2) \\
&\quad \times\mathcal{N}(x_{i1}|\mu_i,\sigma_i^2)\mathcal{N}(x_{i2}|\mu_i,\sigma_i^2)dx_{i1}dx_{i2} \\
&= \frac{1}{2^k\pi^{2k}\prod_{j=1}^k \sigma_j^3\tau_j^2} \times \prod_{i=1}^k \int\int (|x_{i1}-x_{i2}| \\
&\quad \times \exp\left\{\frac{-\left[(x_{i1}-\mu)^2+(x_{i2}-\mu)^2\right]}{2\tau_i^2}\right\} \\
&\quad \times \exp\left\{\frac{-\left[(x_{i1}-\mu_i)^2+(x_{i2}-\mu_i)^2\right]}{2\sigma_i^2}\right\})dx_{i1}dx_{i2} \\
&= \frac{1}{2^k\pi^{2k}\prod_{j=1}^k \sigma_j^3\tau_j^2} \times \prod_{i=1}^k \int\int (|x_{i1}-x_{i2}| \\
&\quad \times \exp\left\{\frac{-(x_{i1}^2+x_{i2}^2)}{2}\left(\tau_i^{-2}+\sigma_i^{-2}\right)+\frac{(x_{i1}+x_{i2})\mu}{\tau_i^2}\right. \\
&\quad \left.+\frac{(x_{i1}+x_{i2})\mu_i}{\sigma_i^2}-\frac{\mu^2}{\tau_i^2}-\frac{\mu_i^2}{\sigma_i^2}\right\}dx_{i1}dx_{i2},
\end{aligned}$$

haciendo $d_i^2 = \left(\frac{x_{i1}^2-x_{i2}^2}{2}\right)^2$ *y* $m_i = \frac{x_{i1}+x_{i2}}{2}$, *obtenemos*

$$\begin{aligned}
&= \frac{1}{2^k\pi^{2k}\prod_{j=1}^k \sigma_j^3\tau_j^2} \times \prod_{i=1}^k \int\int (|x_{i1}-x_{i2}| \\
&\quad \exp\left\{-d_i^2\left(\tau_i^{-2}+\sigma_i^{-2}\right)-\frac{(m_i-\mu)^2}{\tau_i^2}-\frac{(m_i-\mu_i)^2}{\sigma_i^2}\right\})dx_{i1}dx_{i2}. \qquad (5.9)
\end{aligned}$$

Considerando las nuevas variables (u_i, v_i), $i = 1, 2, ..., k$ *definidas como*

$$u_i = x_{i1} - x_{i2}, \quad v_i = x_{i1} + x_{i2}$$

y despejando, se tiene que $x_{i1} = \frac{u_i+v_i}{2}$ *y* $x_{i2} = \frac{v_i-u_i}{2}$ *para* $i = 1, 2, ..., k$. *Además el jacobiano de la transformación es*

$$J = \begin{vmatrix} \frac{\partial x_{i1}}{\partial u_i} & \frac{\partial x_{i1}}{\partial v_i} \\ \frac{\partial x_{i2}}{\partial u_i} & \frac{\partial x_{i2}}{\partial v_i} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{1}{2}$$

así

$$\begin{aligned}\pi_2^I(\theta_2|\theta_1) &= \frac{1}{2^k\pi^{2k}\prod_{j=1}^k\sigma_j^3\tau_j^2}\times\prod_{i=1}^k\int\int(\frac{|u_i|}{2}\\ &\quad \exp\left\{-\frac{u_i^2}{4}\left(\frac{\tau_i^2+\sigma_i^2}{\tau_i^2\sigma_i^2}\right)-\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}-\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right\})du_idv_i\\ &= \frac{1}{2^{2k}\pi^{2k}\prod_{j=1}^k\sigma_j^3\tau_j^2}\times\prod_{i=1}^k\int\left[\exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\}dv_i\right.\\ &\quad \left.\int|u_i|\exp\left\{-\frac{(\tau_i^2+\sigma_i^2)\,u_i^2}{4\tau_i^2\sigma_i^2}\right\}du_i\right],\end{aligned}$$

donde

$$\begin{aligned}&\int|u_i|\exp\left\{-\frac{(\tau_i^2+\sigma_i^2)\,u_i^2}{4\tau_i^2\sigma_i^2}\right\}du_i\\ &= 2\int_0^\infty u_i\exp\left\{-\frac{(\tau_i^2+\sigma_i^2)\,u_i^2}{4\tau_i^2\sigma_i^2}\right\}du_i\\ &= 2\left[\frac{1}{-2\left(\frac{\tau_i^2+\sigma_i^2}{4\tau_i^2\sigma_i^2}\right)}\right]\cdot\exp\left\{-\frac{(\tau_i^2+\sigma_i^2)\,u_i^2}{4\tau_i^2\sigma_i^2}\right\}_{u_i=0}^{u_i=\infty}\\ &= \frac{4\tau_i^2\sigma_i^2}{\tau_i^2+\sigma_i^2}.\end{aligned}\tag{5.10}$$

Sustituyendo el resultado (5.10) tenemos

$$\begin{aligned}\pi_2^I(\theta_2|\theta_1) &= \frac{1}{2^{2k}\pi^{2k}\prod_{j=1}^k\sigma_j^3\tau_j^2}\times\prod_{i=1}^k\frac{4\tau_i^2\sigma_i^2}{\tau_i^2+\sigma_i^2}\\ &\quad\times\prod_{i=1}^k\int\exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\}dv_i\\ &= \frac{1}{\pi^{2k}\prod_{j=1}^k\sigma_j}\times\prod_{i=1}^k\frac{1}{\tau_i^2+\sigma_i^2}\times\\ &\quad\prod_{i=1}^k\int\exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\}dv_i.\end{aligned}\tag{5.11}$$

Puesto que $A(z-a)^2+B(z-b)^2=(A+B)(z-c)^2+\frac{AB}{A+B}(a-b)^2$, *donde* $c=\frac{1}{A+B}(Aa+Bb)$, *haciendo* $z=\frac{v_i}{2}$, $a=\mu$, $b=\mu_i$, $A=\tau_i^{-2}$ *y* $B=\sigma_i^{-2}$ *se tiene*

que

$$\begin{aligned}
&\int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i \\
= \;& \exp\left\{-\frac{1}{2}\left[\frac{2\tau_i^{-2}\sigma_i^{-2}}{\tau_i^{-2}+\sigma_i^{-2}}\left(\mu_i-\mu\right)^2\right]\right\} \\
& \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{\sqrt{2}\frac{v_i}{2}-\sqrt{2}c}{\left(\tau_i^{-2}+\sigma_i^{-2}\right)^{-1/2}}\right)^2\right\} dv_i \\
= \;& \exp\left\{-\frac{1}{2}\left[\frac{2\tau_i^{-2}\sigma_i^{-2}}{\tau_i^{-2}+\sigma_i^{-2}}\left(\mu_i-\mu\right)^2\right]\right\} \\
& \times \int_{-\infty}^{\infty} \exp\left\{-\frac{\left(\tau_i^{-2}+\sigma_i^{-2}\right)}{2}\left(\frac{v_i}{\sqrt{2}}-\sqrt{2}c\right)^2\right\} dv_i,
\end{aligned}$$

considerando el cambio de variable $t=\frac{v_i}{\sqrt{2}}$ *tal que* $dt=\frac{1}{\sqrt{2}}dv_i$,

$$\begin{aligned}
&\int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i \\
= \;& \exp\left\{-\frac{1}{2}\left[\frac{2\tau_i^{-2}\sigma_i^{-2}}{\tau_i^{-2}+\sigma_i^{-2}}\left(\mu_i-\mu\right)^2\right]\right\} \\
& \times\sqrt{2}\int_{-\infty}^{\infty} \exp\left\{-\frac{\left(\tau_i^{-2}+\sigma_i^{-2}\right)}{2}\left(t-\sqrt{2}c\right)^2\right\} dt \\
= \;& \exp\left\{-\frac{1}{2}\left[\frac{2\tau_i^{-2}\sigma_i^{-2}}{\tau_i^{-2}+\sigma_i^{-2}}\left(\mu_i-\mu\right)^2\right]\right\} \\
& \times\left[\sqrt{2}\sqrt{2\pi}(\tau_i^{-2}+\sigma_i^{-2})^{-1/2}\right] \\
= \;& 2\sqrt{\pi}\left(\frac{\tau_i^2+\sigma_i^2}{\tau_i^2\sigma_i^2}\right)^{-1/2}\times\exp\left\{-\frac{1}{2}\left[\frac{2\left(\frac{1}{\tau_i^2\sigma_i^2}\right)}{\frac{\tau_i^2+\sigma_i^2}{\tau_i^2\sigma_i^2}}\right]\left(\mu_i-\mu\right)^2\right\} \\
= \;& 2\sqrt{\pi}\frac{1}{\left(\frac{\tau_i^2+\sigma_i^2}{\tau_i^2\sigma_i^2}\right)^{1/2}}\times\exp\left\{-\frac{1}{2}\left[\frac{2}{\tau_i^2+\sigma_i^2}\right]\left(\mu_i-\mu\right)^2\right\} \\
= \;& \frac{2\sqrt{\pi}}{\sqrt{2}}\left(\frac{2\tau_i^2\sigma_i^2}{\tau_i^2+\sigma_i^2}\right)^{1/2}\times\exp\left\{-\frac{1}{2}\left[\frac{2}{\tau_i^2+\sigma_i^2}\right]\left(\mu_i-\mu\right)^2\right\} \\
= \;& \frac{\sqrt{2\pi}\tau_i\sigma_i}{\left(\frac{\tau_i^2+\sigma_i^2}{2}\right)^{1/2}}\times\exp\left\{-\frac{1}{2}\left[\frac{2}{\tau_i^2+\sigma_i^2}\right]\left(\mu_i-\mu\right)^2\right\}
\end{aligned}$$

$$= \frac{2\pi\tau_i\sigma_i}{\sqrt{2\pi}\left(\frac{\tau_i^2+\sigma_i^2}{2}\right)^{1/2}} \times \exp\left\{-\frac{1}{2}\left[\frac{2}{\tau_i^2+\sigma_i^2}\right](\mu_i-\mu)^2\right\}$$

$$= 2\pi\tau_i\sigma_i \cdot \mathcal{N}(\mu i|\mu, \frac{\tau_i^2+\sigma_i^2}{2}).$$

Sustituyendo este último resultado en (5.11), se tiene que

$$\begin{aligned}\pi_2^I(\theta_2|\theta_1) &= \frac{1}{\pi^{2k}\prod_{j=1}^k \sigma_j} \times \prod_{i=1}^k \frac{1}{\tau_i^2+\sigma_i^2} \times \prod_{i=1}^k 2\pi\tau_i\sigma_i \cdot \mathcal{N}(\mu i|\mu, \frac{\tau_i^2+\sigma_i^2}{2}) \\ &= \prod_{i=1}^k \frac{2}{\pi}\frac{\tau_i}{\tau_i^2+\sigma_i^2} \cdot \mathcal{N}(\mu i|\mu, \frac{\tau_i^2+\sigma_i^2}{2}) \\ &= \prod_{i=1}^k HC^+(\sigma_i|0,\tau_i) \cdot \mathcal{N}(\mu i|\mu, \frac{\tau_i^2+\sigma_i^2}{2}).\blacksquare\end{aligned}$$

Este teorema indica que, en la distribución a priori intrínseca $\pi_2^I(\theta_2|\theta_1)$, las $\mu_i^{'}s$ son condicionalmente independientes y normalmente distribuidas y $\sigma_i|\tau_i$ son independientes de $\sigma_j|\tau_j$, $i \neq j$, y se distribuye como una media de Cauchy.

5.4. Análisis de varianza bajo heterocedasticidad

Dada la muestra $z = (x_{11}, ..., x_{1n_1}, ..., x_{k_1}, ..., x_{kn_k})$, el factor Bayes para las distribuciones a priori $\{\pi_1^N(\theta_1), \pi_2^I(\theta_2)\}$ dadas en el teorema anterior, es

$$B_{21}^I(z) = \frac{\int \prod_{i=1}^k \left\{\prod_{j=1}^{n_i} \mathcal{N}(x_{ij}|\mu_i, \sigma_i^2)\right\} \pi_2^I(\mu_i, \sigma_i) d\mu_i d\sigma_i}{\int \left\{\prod_{i=1}^k \{\prod_{j=1}^{n_i} \mathcal{N}(xij|\mu, \tau_i^2)\}\tau_i^{-1} d\tau_i\right\} d\mu}. \tag{5.12}$$

Trabajando con el numerador tenemos

$$\begin{aligned}&\int\int \prod_{i=1}^k \left\{\prod_{j=1}^{n_i} \mathcal{N}(x_{ij}|\mu_i, \sigma_i^2)\right\} \pi_2^I(\mu_i, \sigma_i) d\mu_i d\sigma_i \\ = &\int\int \left[\prod_{i=1}^k \left\{\prod_{j=1}^{n_i} \mathcal{N}(x_{ij}|\mu_i, \sigma_i^2)\right\}\right. \\ &\left.\times \left(\int\int \prod_{i=1}^k HC^+(\sigma_i|0,\tau_i) \times \mathcal{N}(\mu i|\mu, \frac{\tau_i^2+\sigma_i^2}{2})\tau_i^{-1} d\mu d\tau_i\right)\right] d\mu_i d\sigma_i\end{aligned}$$

$$= \int\int \left[\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\int\int \left\{ \prod_{i=1}^{k} (2\pi)^{-1/2} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{-1/2} \right.$$
$$\left. \left. \times \exp\left[-\frac{1}{2} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{-1} (\mu_i - \mu)^2 \right] \frac{2\tau_i}{\pi(\tau_i^2 + \sigma_i^2)} \times \tau_i^{-1} \right\} d\mu d\tau_i \right] d\mu_i d\sigma_i$$

tomando en cuenta (B.2) y (B.3) se cumple

$$\int \exp\left[-\frac{1}{2} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{-1} (\mu_i - \mu)^2 \right] d\mu = \sqrt{2\pi} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{1/2},$$

y

$$\int \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] d\mu_i = \frac{(2\pi)^{1/2} \sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{2\sigma_i^2} \right],$$

luego

$$= \int\int \left[\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\int \left\{ \prod_{i=1}^{k} (2\pi)^{-1/2} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{-1/2} \right.$$
$$\left. \left. \sqrt{2\pi} \left(\frac{\tau_i^2 + \sigma_i^2}{2} \right)^{1/2} \times \frac{2\tau_i}{\pi(\tau_i^2 + \sigma_i^2)} \times \tau_i^{-1} \right\} d\tau_i \right] d\mu_i d\sigma_i$$
$$= \int\int \left[\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\left. \int \left\{ \prod_{i=1}^{k} \frac{2}{\pi(\tau_i^2 + \sigma_i^2)} \right\} d\tau_i \right] d\mu_i d\sigma_i$$

$$
\begin{aligned}
&= \int \left[\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \frac{(2\pi)^{1/2}\sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2} \right] \right\} \right. \\
&\quad \left. \int \left\{ \prod_{i=1}^{k} \frac{2}{\pi(\tau_i^2+\sigma_i^2)} \right\} d\tau_i \right] d\sigma_i \\
&= \frac{2^k (2\pi)^{k/2} \left(\prod_{i=1}^{k} n_i \right)^{-1/2}}{(2\pi)^{n/2}\pi^k} \int \left[\prod_{i=1}^{k} \left\{ \sigma_i^{-n_i+1} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2} \right] \right\} \right. \\
&\quad \left. \int \left\{ \prod_{i=1}^{k} \frac{1}{(\tau_i^2+\sigma_i^2)} \right\} d\tau_i \right] d\sigma_i \\
&= 2^{\frac{3k-n}{2}} \pi^{\frac{-(n+k)}{2}} \left(\prod_{i=1}^{k} n_i \right)^{-1/2} \int \left[\prod_{i=1}^{k} \left\{ \sigma_i^{-n_i+1} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2} \right] \right\} \right. \\
&\quad \left. \int \left\{ \prod_{i=1}^{k} \frac{1}{(\tau_i^2+\sigma_i^2)} \right\} d\tau_i \right] d\sigma_i \\
&= 2^{\frac{3k-n}{2}} \pi^{\frac{-(n+k)}{2}} \left(\prod_{i=1}^{k} n_i \right)^{-1/2} \\
&\quad \times \int_0^\infty \left\{ \prod_{i=i}^{k} \int_0^\infty \frac{1}{(\tau_i^2+\sigma_i^2)\sigma_i^{n_i-1}} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2} \right] d\sigma_i \right\} d\tau_i \\
&= 2^{\frac{3k-n}{2}} \pi^{\frac{-(n+k)}{2}} \left(\prod_{i=1}^{k} n_i \right)^{-1/2} \times I_{2.} \qquad (5.13)
\end{aligned}
$$

Donde $I_2 = \int_0^\infty \left\{ \prod_{i=i}^{k} \int_0^\infty \frac{1}{(\tau_i^2+\sigma_i^2)\sigma_i^{n_i-1}} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2} \right] \right\} d\sigma_i d\tau_i$.

Ahora trabajando con el denominador, se tiene que

$$
\begin{aligned}
&\int \int \prod_{i=1}^{k} \{ \prod_{j=1}^{n_i} \mathcal{N}(xij|\mu, \tau_i^2) \} \tau_i^{-1} d\tau_i d\mu \\
&= \int \int \left(\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \tau_i^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu)^2}{2\tau_i^2} \right] \tau_i^{-1} \right\} \right) d\tau_i d\mu
\end{aligned}
$$

$$= (2\pi)^{-n/2} \int\int \left(\prod_{i=1}^{k} \left\{ \tau_i^{-n_i-1} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu)^2}{2} \tau_i^{-2} \right] \right\} \right) d\tau_i d\mu,$$

pero

$$\begin{aligned}
& -\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu)^2}{2} \\
= & -\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\mu)^2}{2} \\
= & -\frac{\sum_{j=1}^{n_i}[(x_{ij}-\bar{x}_i)^2+2(x_{ij}-\bar{x}_i)(\bar{x}_i+\mu)+(\bar{x}_i-\mu)^2]}{2} \\
= & -\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}2(x_{ij}-\bar{x}_i)(\bar{x}_i+\mu)+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2} \\
= & -\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2}.
\end{aligned}$$

Así

$$\begin{aligned}
& (2\pi)^{-n/2} \int\int \left(\prod_{i=1}^{k} \left\{ \tau_i^{-n_i-1} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu)^2}{2} \tau_i^{-2} \right] \right\} \right) d\tau_i d\mu \\
= & (2\pi)^{-n/2} \int\int \left(\prod_{i=1}^{k} \left\{ \tau_i^{-(n_i+1)} \right. \right. \\
& \left. \left. \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2} \tau_i^{-2} \right] \right\} \right) d\tau_i d\mu,
\end{aligned}$$

tomando $p = n_i$ y $a = \frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2}$ en (B.1), tenemos que

$$
\begin{aligned}
& (2\pi)^{-n/2}\int\int\left(\prod_{i=1}^{k}\left\{\tau_i^{-(n_i+1)}\right.\right. \\
& \left.\left.\times\exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2}\tau_i^{-2}\right]\right\}\right)d\tau_i d\mu \\
= \;& (2\pi)^{-n/2}\int\prod_{i=1}^{k}\left\{\frac{1}{2}\left[\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_i}(\bar{x}_i-\mu)^2}{2}\right]^{-n_i/2}\Gamma\left(\frac{n_i}{2}\right)\right\}d\mu \\
= \;& (2\pi)^{-n/2}\int\prod_{i=1}^{k}\left\{\frac{1}{2}\left[\frac{s_i^2+n_i(\bar{x}_i-\mu)^2}{2}\right]^{-n_i/2}\Gamma\left(\frac{n_i}{2}\right)\right\}d\mu \\
= \;& (2\pi)^{-n/2}2^{n/2}2^{-k}\int\prod_{i=1}^{k}\left\{\left[s_i^2+n_i(\bar{x}_i-\mu)^2\right]^{-n_i/2}\Gamma\left(\frac{n_i}{2}\right)\right\}d\mu \\
= \;& \pi^{-n/2}2^{-k}\prod_{i=1}^{k}\left\{\Gamma\left(\frac{n_i}{2}\right)\right\} \\
& \times\int\prod_{i=1}^{k}\left\{\left[s_i^2+n_i(\bar{x}_i-\mu)^2\right]^{-n_i/2}\right\}d\mu \\
= \;& \pi^{-n/2}2^{-k}\prod_{i=1}^{k}\left\{\Gamma\left(\frac{n_i}{2}\right)\right\}\times I_3. \qquad (5.14)
\end{aligned}
$$

donde $I_3 = \int\prod_{i=1}^{k}\left\{[s_i^2+n_i(\bar{x}_i-\mu)^2]^{-n_i/2}\right\}d\mu$.

Efectuando el cociente entre (5.13) y (5.14), se tiene

$$
\begin{aligned}
B_{21}^{I}(z) &= \frac{\int\prod_{i=1}^{k}\left\{\prod_{j=1}^{n_i}\mathcal{N}(x_{ij}|\mu_i,\sigma_i^2)\right\}\pi_2^I(\mu_i,\sigma_i)d\mu_i d\sigma_i}{\int\prod_{i=1}^{k}\{\prod_{j=1}^{n_i}\mathcal{N}(xij|\mu,\tau_i^2)\}\tau_i^{-1}d\tau_i d\mu} \\
&= \frac{2^{\frac{3k-n}{2}}\pi^{\frac{-(n+k)}{2}}\left(\prod_{i=1}^{k}n_i\right)^{-1/2}\times I_2}{\pi^{-n/2}2^{-k}\prod_{i=1}^{k}\left\{\Gamma\left(\frac{n_i}{2}\right)\right\}\times I_3}.
\end{aligned}
$$

Así

$$B_{21}^{I}(z) = \frac{I_2}{2^{-(\frac{5k-n}{2})}\pi^{k/2}\left(\prod_{i=1}^{k} n_i\right)^{1/2}\prod_{i=1}^{k}\left\{\Gamma\left(\frac{n_i}{2}\right)\right\} \times I_3}.$$

5.5. El factor de Bayes para contraste bajo homocedasticidad

En la sección anterior el factor de Bayes es derivado bajo la condición de heterocedasticidad. Cuando se asume la condición de homocedasticidad los modelos a comparar son

$$f_1(z|\theta_1) = \mathcal{N}(x_1|\eta_1, \tau^2)...\mathcal{N}(x_k|\eta_k, \tau^2) \text{ y } f_2(z|\theta_2) = \mathcal{N}(x_1|\mu_1, \sigma^2)...\mathcal{N}(x_k|\mu_k, \sigma^2).$$

El contraste de homocedasticidad frente al de heterocedasticidad puede hacerse en un contexto frecuentista usando la aproximación de Bartlett. Desde el punto de vista bayesiano los modelos anidados a ser comparados son:

$$M_1 : f_1(z|\theta_1) = \prod_{i=1}^{k}\mathcal{N}(x_i|\eta_{i,}\tau^2), \quad \pi_1^N(\theta_1) = \frac{c_1}{\tau} \tag{5.15}$$

y

$$M_2 : f_2(z|\theta_2) = \prod_{i=1}^{k}\mathcal{N}(x_i|\mu_{i,}\sigma_i^2), \quad \pi_2^N(\theta_2) = \frac{c_2}{\prod_{i=1}^{k}\sigma_i} \tag{5.16}$$

donde $\theta_1 = (\eta_1, ..., \eta_k, \tau)$, y $\theta_2 = (\mu_1, ..., \mu_k, \sigma_1, ...\sigma_k)$ y c_1, c_2 son constantes positivas arbitrarias. Se supone que los parámetros de localización y escala en cada densidad son a priori independientes.

Teorema 5.2 *Las distribuciones a priori intrínseca para comparar el modelo (5.15) frente al (5.16) son* $\{\pi_1^N(\theta_1), \pi_2^{HI}(\theta_2)\}$, *donde*

$$\pi_2^{HI}(\theta_2) = \int \pi_2^{HI}(\theta_2|\theta_1)\pi_1^N(\theta_1)d\theta_1$$

y

$$\pi_2^{HI}(\theta_2|\theta_1) = \prod_{i=1}^{k}\mathcal{N}(\mu_i|\eta_i, \frac{\tau^2+\sigma_i^2}{2})HC^+(\sigma_i|0,\tau).$$

Demostración *Consideremos los modelos*

$$M_1 : f_1(z|\theta_1) = \prod_{i=1}^{k} \mathcal{N}(x_i|\eta_{i,}\tau^2), \quad \pi_1(\theta_1) = \delta(\eta_1, ..., \eta_{k,}\tau)$$

y

$$M_2 : f_2(z|\theta_2) = \prod_{i=1}^{k} \mathcal{N}(x_i|\mu_{i,}\sigma_i^2), \quad \pi_2^N(\theta_2) = \frac{c_2}{\prod_{i=1}^{k}\sigma_i},$$

donde $\delta(\mu, \tau_1, ..., \tau_k)$ *es la delta de Dirac en el punto* $\theta_1 = (\mu, \tau_1, ..., \tau_k)$ *y* c_2 *es una constante positiva arbitraria. Luego tenemos que*

$$B_{12}^N(z(l)) = \frac{\prod_{i=1}^{k}\prod_{j=1}^{2}\mathcal{N}(x_{ij}|\eta_i, \tau^2)}{m_2^N(z(l))},$$

donde

$$m_2^N(z(l)) = \frac{c_2}{2^k}\frac{1}{\prod\limits_{i=1}^{k}|x_{i1} - x_{i2}|}.$$

La distribución a priori intrínseca condicional de θ_2 *es*

$$\begin{aligned}
\pi_2^I(\theta_2|\theta_1) &= \pi_2^N(\theta_2)E_{z(l)|\theta_2}^{M_2}B_{12}^N(z(l)) \\
&= \frac{c_2}{\prod_{j=1}^{k}\sigma_j} \times \prod_{i=1}^{k}\int\int\left(\frac{2^k\,|x_{i1} - x_{i2}|}{c_2}\right)\mathcal{N}(x_{i1}|\eta_i, \tau^2)\mathcal{N}(x_{i2}|\eta_i, \tau^2) \\
&\quad \times\mathcal{N}(x_{i1}|\mu_i, \sigma_i^2)\mathcal{N}(x_{i2}|\mu_i, \sigma_i^2)dx_{i1}dx_{i2} \\
&= \frac{1}{2^k\pi^{2k}\prod_{j=1}^{k}\sigma_j^3\tau^2} \times \prod_{i=1}^{k}\int\int(|x_{i1} - x_{i2}| \\
&\quad \times \exp\left\{\frac{-\left[(x_{i1} - \eta_i)^2 + (x_{i2} - \eta_i)^2\right]}{2\tau^2}\right\} \\
&\quad \times \exp\left\{\frac{-\left[(x_{i1} - \mu_i)^2 + (x_{i2} - \mu_i)^2\right]}{2\sigma_i^2}\right\})dx_{i1}dx_{i2} \\
&= \frac{1}{2^k\pi^{2k}\tau^{2k}\prod_{j=1}^{k}\sigma_j^3} \times \prod_{i=1}^{k}\int\int(|x_{i1} - x_{i2}| \times \exp\left\{-\frac{(x_{i1}^2 + x_{i2}^2}{2}\right. \\
&\quad \left.\times\left(\tau^{-2} + \sigma_i^{-2}\right) - \frac{(x_{i1} + x_{i2})\eta_i}{\tau^2} - \frac{(x_{i1} + x_{i2})\mu_i}{\sigma_i^2} + \frac{\eta_i^2}{\tau^2} + \frac{\mu_i^2}{\sigma_i^2}\right\})dx_{i1}dx_{i2}
\end{aligned}$$

$$= \frac{1}{2^k\pi^{2k}\tau^{2k}\prod_{j=1}^{k}\sigma_j^3}\times\prod_{i=1}^{k}\int\int(|x_{i1}-x_{i2}| \exp\left\{-d_i^2\left(\tau^{-2}+\sigma_i^{-2}\right)-\frac{(m_i-\eta_i)^2}{\tau^2}-\frac{(m_i-\mu_i)^2}{\sigma_i^2}\right\})dx_{i1}dx_{i2}, \quad (5.17)$$

donde $d_i^2=\frac{(x_{i1}-x_{i2})^2}{4}$ *y* $m_i=\frac{x_{i1}+x_{i2}}{2}$, *respectivamente. Considerando las nuevas variables* (u_i,v_i), $i=1,2,...,k$ *definidas como*

$$u_i=x_{i1}-x_{i2}, \quad v_i=x_{i1}+x_{i2},$$

despejando se tiene que $x_{i1}=\frac{u_i+v_i}{2}$, $x_{i2}=\frac{v_i-u_i}{2}$, $i=1,2,...,k$, *y el jacobiano de la transformación es:*

$$J=\begin{bmatrix}\frac{\partial x_{i1}}{\partial u_i} & \frac{\partial x_{i1}}{\partial v_i}\\ \frac{\partial x_{i2}}{\partial u_i} & \frac{\partial x_{i2}}{\partial v_i}\end{bmatrix}=\begin{bmatrix}1/2 & 1/2\\ -1/2 & 1/2\end{bmatrix}=1/2.$$

Sustituyendo las nuevas variables en (5.17) tenemos que

$$\pi_2^I(\theta_2|\theta_1) = \frac{1}{2^k\pi^{2k}\tau^{2k}\prod_{j=1}^{k}\sigma_j^3}\times\prod_{i=1}^{k}\int\int(\frac{|u_i|}{2} \exp\left\{-\frac{u_i^2}{4}\left(\frac{\tau^2+\sigma_i^2}{\tau^2\sigma_i^2}\right)-\frac{(\frac{v_i}{2}-\eta_i)^2}{\tau^2}-\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right\})du_idv_i$$

$$= \frac{1}{2^{2k}\pi^{2k}\tau^{2k}\prod_{j=1}^{k}\sigma_j^3}\times\prod_{i=1}^{k}[\int\exp\left\{-\left[\frac{(\frac{v_i}{2}-\eta_i)^2}{\tau^2}+\frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\}dv_i \int|u_i|\exp\left\{-\frac{(\tau^2+\sigma_i^2)\,u_i^2}{4\tau^2\sigma_i^2}\right\}du_i],$$

donde,

$$\int|u_i|\exp\left\{-\frac{(\tau^2+\sigma_i^2)\,u_i^2}{4\tau^2\sigma_i^2}\right\}du_i = 2\int_0^\infty u_i\exp\left\{-\frac{(\tau^2+\sigma_i^2)\,u_i^2}{4\tau^2\sigma_i^2}\right\}du_i$$

$$= 2\left[\frac{1}{-2\left(\frac{\tau^2+\sigma_i^2}{4\tau^2\sigma_i^2}\right)}\right] \times\exp\left\{-\frac{(\tau^2+\sigma_i^2)\,u_i^2}{4\tau^2\sigma_i^2}\right\}_{u_i=0}^{u_i=\infty}$$

$$= \frac{4\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}.$$

Sustituyendo este último resultado tenemos

$$\begin{aligned}\pi_2^I(\theta_2|\theta_1) &= \frac{1}{2^{2k}\pi^{2k}\tau^{2k}\prod_{j=1}^k \sigma_j^3} \times \prod_{i=1}^k \frac{4\tau^2\sigma_i^2}{\tau^2+\sigma_i^2} \\ &\quad \times \prod_{i=1}^k \int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\eta_i)^2}{\tau^2} + \frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i \\ &= \frac{1}{\pi^{2k}\tau^{2k}\prod_{j=1}^k \sigma_j^3} \times \prod_{i=1}^k \frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2} \\ &\quad \times \prod_{i=1}^k \int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\eta_i)^2}{\tau^2} + \frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i. \qquad (5.18)\end{aligned}$$

Puesto que $A(z-a)^2 + B(z-b)^2 = (A+B)(z-c)^2 + \frac{AB}{A+B}(a-b)^2$, *donde* $c = \frac{1}{A+B}(Aa+Bb)$, *haciendo* $z = \frac{v_i}{2}$, $a = \eta_i$, $b = \mu_i$, $A = \tau^{-2}$ *y* $B = \sigma_i^{-2}$ *se tiene que*

$$\begin{aligned}&\int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\eta_i)^2}{\tau^2} + \frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i \\ =\ & \exp\left\{-\frac{1}{2}\left[\frac{2\tau^{-2}\sigma_i^{-2}}{\tau^{-2}+\sigma_i^{-2}}(\mu_i-\eta_i)^2\right]\right\} \\ &\times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{\sqrt{2}\frac{v_i}{2} - \sqrt{2}c}{\left(\tau^{-2}+\sigma_i^{-2}\right)^{-1/2}}\right)^2\right\} dv_i \\ =\ & \exp\left\{-\frac{1}{2}\left[\frac{2\tau^{-2}\sigma_i^{-2}}{\tau^{-2}+\sigma_i^{-2}}(\mu_i-\eta_i)^2\right]\right\} \\ &\times \int_{-\infty}^{\infty} \exp\left\{-\frac{\left(\tau^{-2}+\sigma_i^{-2}\right)}{2}\left(\frac{v_i}{\sqrt{2}} - \sqrt{2}c\right)^2\right\} dv_i,\end{aligned}$$

considerando el cambio de variable $t = \frac{v_i}{\sqrt{2}}$ *tal que* $dt = \frac{1}{\sqrt{2}} dv_i$,

$$\begin{aligned}&\int \exp\left\{-\left[\frac{(\frac{v_i}{2}-\mu)^2}{\tau_i^2} + \frac{(\frac{v_i}{2}-\mu_i)^2}{\sigma_i^2}\right]\right\} dv_i \\ =\ & \exp\left\{-\frac{1}{2}\left[\frac{2\tau^{-2}\sigma_i^{-2}}{\tau^{-2}+\sigma_i^{-2}}(\mu_i-\eta_i)^2\right]\right\} \\ &\times\sqrt{2}\int_{-\infty}^{\infty} \exp\left\{-\frac{\left(\tau^{-2}+\sigma_i^{-2}\right)}{2}\left(t - \sqrt{2}c\right)^2\right\} dt\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{-\frac{1}{2}\left[\frac{2\tau^{-2}\sigma_i^{-2}}{\tau^{-2}+\sigma_i^{-2}}\left(\mu_i-\eta_i\right)^2\right]\right\} \\
&\quad \times\left[\sqrt{2}\sqrt{2\pi}(\tau^{-2}+\sigma_i^{-2})^{-1/2}\right] \\
&= \frac{2\pi\tau\sigma_i}{\sqrt{2\pi}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{1/2}} \times \exp\left\{-\frac{1}{2}\left[\frac{2}{\tau^2+\sigma_i^2}\right]\left(\mu_i-\eta_i\right)^2\right\} \\
&= 2\pi\tau\sigma_i \cdot \mathcal{N}(\mu_i|\eta_{i,}\frac{\sigma_i^2+\tau^2}{2}),
\end{aligned}
$$

y por lo tanto

$$
\begin{aligned}
\pi_2^I(\theta_2|\theta_1) &= \frac{\tau^{2k}}{\pi^{2k}\tau^{2k}\prod_{j=1}^k\sigma_j^3} \times \prod_{i=1}^k \frac{\sigma_i^2}{\tau^2+\sigma_i^2} \times \prod_{i=1}^k 2\pi\tau\sigma_i \cdot \mathcal{N}(\mu i|\mu, \frac{\tau^2+\sigma_i^2}{2}) \\
&= \prod_{i=1}^k \frac{2}{\pi}\frac{\tau}{\tau^2+\sigma_i^2} \cdot \mathcal{N}(\mu i|\eta_i, \frac{\tau^2+\sigma_i^2}{2}) \\
&= \prod_{i=1}^k HC^+(\sigma_i|0,\tau) \cdot \mathcal{N}(\mu i|\eta_i, \frac{\tau^2+\sigma_i^2}{2}).\blacksquare
\end{aligned}
$$

Para una muestra z, el factor Bayes con las distribuciones a priori intrínsecas $\{\pi_1^N(\theta_1), \pi_2^{HI}(\theta_2)\}$ es de la forma

$$
B_{21}^{HI}(z) = \frac{1}{2^{\frac{n-3k}{2}-1}\pi^k\Gamma(\frac{n-k}{2})}S^{n-k}I_1, \tag{5.19}
$$

donde $S^2 = \sum_{i=1}^k s_i^2$, y $\quad I_1 = \int_0^\infty \tau^{k-1}\left\{\prod_{i=1}^k \int_0^\infty \frac{\exp\{-s_i^2/2\sigma_i^2\}}{(\sigma_i^2+\tau^2)\sigma_i^{n_i-1}}d\sigma_i\right\}d\tau.$

Por definición se tiene lo siguiente

$$
B_{21}^{HI}(z) = \frac{\int\left[\prod_{i=1}^k\left\{\prod_{j=1}^{n_i}\mathcal{N}(x_{ij}|\mu_i,\sigma_i^2)\right\}\left(\int\tau^{-1}\prod_{i=1}^k\mathcal{N}(\mu_i|\eta_i,\frac{\tau^2+\sigma_i^2}{2})HC^+(\sigma_i|0,\tau)d\eta_i\right)\right]d\mu_i d\sigma_i}{\int\prod_{i=1}^k\prod_{j=1}^{n_i}\mathcal{N}(xij|\eta_i,\tau^2)\tau^{-1}d\eta_i d\tau}.
$$

Trabajando con el numerador se tiene

$$\begin{aligned}
&\int\int\left[\prod_{i=1}^{k}\left\{\prod_{j=1}^{n_i}\mathcal{N}(x_{ij}|\mu_i,\sigma_i^2)\right\}\right.\\
&\left.\times\left(\int\int\tau^{-1}\prod_{i=1}^{k}\mathcal{N}(\mu_i|\eta_i,\frac{\tau^2+\sigma_i^2}{2})HC^{+}(\sigma_i|0,\tau)d\eta_i d\tau\right)\right]d\mu_i d\sigma_i\\
=&\int\int\left(\prod_{i=1}^{k}\left\{(2\pi)^{-n_i/2}\sigma_i^{-n_i}\times\exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{2\sigma_i^2}\right]\right\}\right.\\
&\times\int\int\left[\tau^{-1}\prod_{i=1}^{k}\left\{(2\pi)^{-1/2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1/2}\right.\right.\\
&\left.\left.\left.\times\exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}(\mu_i-\eta_i)^2\right]\frac{2\tau}{\pi(\tau^2+\sigma_i^2)}\right\}\right]d\eta_i d\tau\right)d\mu_i d\sigma_i.
\end{aligned}$$

Tomando en cuenta (B.2) y (B.3) se tiene que

$$\int\exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{2\sigma_i^2}\right]d\mu_i=\frac{(2\pi)^{1/2}\sigma_i}{(n_i)^{1/2}}\times\exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right]$$

y

$$\int\exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}(\mu_i-\eta_i)^2\right]d\eta_i=\sqrt{2\pi}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{1/2},$$

así

$$\begin{aligned}
&\int\int\left(\prod_{i=1}^{k}\left\{(2\pi)^{-n_i/2}\sigma_i^{-n_i}\times\exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{2\sigma_i^2}\right]\right\}\right.\\
&\times\int\int\left[\tau^{-1}\prod_{i=1}^{k}\left\{(2\pi)^{-1/2}\times\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1/2}\right.\right.\\
&\left.\left.\left.\times\exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}(\mu_i-\eta_i)^2\right]\times\frac{2\tau}{\pi(\tau^2+\sigma_i^2)}\right\}\right]d\eta_i d\tau\right)d\mu_i d\sigma_i
\end{aligned}$$

$$= \int\int \left(\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\left. \int \left[\tau^{-1} \prod_{i=1}^{k} \left\{ (2\pi)^{-1/2} \times \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{-1/2} \times \sqrt{2\pi} \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{1/2} \right. \right. \right.$$
$$\left. \left. \left. \times \frac{2\tau}{\pi(\tau^2 + \sigma_i^2)} \right\} \right] d\tau \right) d\mu_i d\sigma_i$$

$$= \int \left(\prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \sigma_i^{-n_i} \times \frac{(2\pi)^{1/2}\sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\left. \int \left[\tau^{-1} \prod_{i=1}^{k} \left\{ (2\pi)^{-1/2} \times \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{-1/2} \times \sqrt{2\pi} \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{1/2} \right. \right. \right.$$
$$\left. \left. \left. \times \frac{2\tau}{\pi(\tau^2 + \sigma_i^2)} \right\} \right] d\tau \right) d\sigma_i$$

$$= \frac{(2\pi)^{-n/2}(2\pi)^{\frac{k}{2}} 2^k}{\pi^k} \int \left(\prod_{i=1}^{k} \left\{ \sigma_i^{-n_i} \times \frac{\sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\left. \int \left[\tau^{-1} \prod_{i=1}^{k} \left\{ \frac{\tau}{(\tau^2 + \sigma_i^2)} \right\} \right] d\tau \right) d\sigma_i$$

$$= 2^{\frac{3k-n}{2}} \pi^{-\left(\frac{n+k}{2}\right)} \int \left(\prod_{i=1}^{k} \left\{ \sigma_i^{-n_i} \times \frac{\sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{2\sigma_i^2} \right] \right\} \right.$$
$$\left. \int \left[\tau^{-1} \prod_{i=1}^{k} \left\{ \frac{\tau}{(\tau^2 + \sigma_i^2)} \right\} \right] d\tau \right) d\sigma_i.$$

Haciendo $s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

$$= 2^{\frac{3k-n}{2}} \pi^{-\left(\frac{n+k}{2}\right)} \left(\prod_{i=1}^{k} n_i \right)^{-1/2} \int_0^\infty \tau^{k-1} \left\{ \prod_{i=1}^{k} \int \frac{1}{\sigma_i^{n_i - 1}(\tau^2 + \sigma_i^2)} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2} \right] d\sigma_i \right\} d\tau.$$

Denotando $I_1 = \int_0^\infty \tau^{k-1} \left\{ \prod_{i=1}^{k} \int_0^\infty \frac{1}{(\tau^2+\sigma_i^2)\sigma_i^{n_i-1}} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2}\right] d\sigma_i \right\} d\tau$ tenemos

$$\begin{aligned} & 2^{\frac{3k-n}{2}} \pi^{-\frac{(n+k)}{2}} (\prod_{i=1}^{k} n_i)^{-1/2} \\ & \times \int \tau^{k-1} \left\{ \prod_{i=1}^{k} \int_0^\infty \frac{1}{(\tau^2+\sigma_i^2)\sigma_i^{n_i-1}} \times \exp\left[-\frac{s_i^2}{2\sigma_i^2}\right] d\sigma_i \right\} d\tau \\ = \; & 2^{\frac{3k-n}{2}} \pi^{-\frac{(n+k)}{2}} (\prod_{i=1}^{k} n_i)^{-1/2} \times I_1. \end{aligned} \tag{5.20}$$

Ahora trabajando con el denominador y tomando en cuenta (B.2)

$$\int \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\eta_i)^2}{2\tau^2}\right] d\eta_i = \frac{(2\pi)^{1/2}\tau}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\tau^2}\right]$$

se tiene que

$$\begin{aligned} & \int\int \prod_{i=1}^{k} \left\{ \prod_{j=1}^{n_i} \mathcal{N}(x_{ij}|\eta_i,\tau^2) \right\} \tau^{-1} d\eta_i d\tau \\ = \; & \int\int \tau^{-1} \prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \tau^{-n_i} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\eta_i)^2}{2\tau^2}\right] \right\} d\eta_i d\tau \\ = \; & \int \tau^{-1} \prod_{i=1}^{k} \left\{ (2\pi)^{-n_i/2} \tau^{-n_i} \frac{(2\pi)^{1/2}\tau}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\tau^2}\right] \right\} d\tau \\ = \; & \int \tau^{-(n-k+1)} (2\pi)^{\frac{k-n}{2}} (\prod_{i=1}^{k} n_i)^{-1/2} \times \exp\left[-\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\tau^2}\right] d\tau. \end{aligned}$$

Tomando $p = n - k$ y $a = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2}$ en (B.1)

$$\int \tau^{-(n-k+1)}(2\pi)^{\frac{k-n}{2}}(\prod_{i=1}^{k} n_i)^{-1/2} \times \exp\left[-\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\tau^2}\right] d\tau$$

$$= \frac{(2\pi)^{\frac{k-n}{2}}}{2}(\prod_{i=1}^{k} n_i)^{-1/2}\Gamma(\frac{n-k}{2})\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2}\right)^{-\frac{(n-k)}{2}}$$

$$= \Gamma(\frac{n-k}{2})\pi^{\frac{k-n}{2}}2^{-1}S^{-(n-k)}(\prod_{i=1}^{k} n_i)^{-1/2}, \tag{5.21}$$

donde $S^2 = \sum_{i=1}^{k} s_i^2$.

Efectuando el cociente de las ecuaciones (5.20) y (5.21), tenemos que

$$B_{21}^{HI}(z) = \frac{2^{\frac{3k-n}{2}}\pi^{-\frac{(n+k)}{2}}(\prod_{i=1}^{k} n_i)^{-1/2} \times I_1}{\Gamma(\frac{n-k}{2})\pi^{\frac{k-n}{2}}2^{-1}S^{-(n-k)}(\prod_{i=1}^{k} n_i)^{-1/2}} = \frac{1}{2^{\frac{n-3k}{2}-1}\pi^k\Gamma(\frac{n-k}{2})}S^{n-k}I_1. \tag{5.22}$$

A continuación veamos un ejemplo de los resultados obtenidos.

Ejemplo 5.1 *(Contraste de homocedasticidad frente a heterocedasticidad) Como ilustración vamos a considerar un ejemplo, que aparece en Welch(1951). Supóngase que los datos proceden de tres poblaciones normales, y son obtenidos los siguientes resultados:* $(n_1, n_2, n_3) = (20, 10, 10)$, $(\bar{x}1, \bar{x}2, \bar{x}3) = (27.8, 24.1, 22.2)$, $(s_1^2, s_2^2, s_3^2) = (1141.9, 56.7, 138.6)$.

Se contrastará si todas las varianzas son todas iguales. Se calcularán los factores de Bayes con las distribuciones a priori intrínsecas B_{21}^{HI}. *Además, con el objeto de ilustrar mejor su comportamiento, se calculan el valor-p y los factores de Bayes para diferentes valores de* s_1^2 *y los restantes valores muestrales fijos. Estos valores se presentan en la segunda y tercera columna. Para aquellos que prefieren la interpretación de la probabilidad a posteriori frente a los factores de Bayes, se han calculado las probabilidades a posteriori de* M_1 *correspondientes a los valores de los factores de Bayes y la distribución a priori* $P(M_1) = P(M_2) = 1/2$.

Tabla 5.1.
Factores de Bayes intrínsecos

s_1^2	$valor - p$	B_{21}^{HI}	$P^{HI}(M_1\|x)$
1141.9	0.001	19.20	0.05
600	0.04	0.65	0.61
300	0.31	0.08	0.93

Para valores de s_1^2 lejanos a los de s_2^2 y s_3^2 el modelo bajo heterocedasticidad se encuentra claramente favorecido como indica la primera fila; en la columna del factor de Bayes intrínseco podemos observar quel número es mucho mayor que 1, por lo cual es favorecido el modelo 2 (modelo bajo heterocedasticidad); y en la columna de la probabilidad a posteriori podemos observar que el modelo 1 (modelo bajo homocedasticidad) es muy poco favorecido. Cuando s_1^2 se aproxima a los valores de s_2^2 y s_3^2 las probabilidades a posteriori aumentan favoreciendo el modelo bajo homocedasticidad. Mientras que los factores de Bayes intrínsecos disminuyen, favoreciendo también el modelo bajo homocedasticidad.

Capítulo 6

Conclusiones

En la mayoría de los problemas de decisión, aparece de forma natural la incertidumbre. Lo que observamos, y podemos medir, es solo una posibilidad entre muchas —como en el lanzamiento de una moneda— y, de algún modo, necesitamos de una escala que nos represente la verosimilitud de lo realmente observado en relación con el resto. En otras ocasiones, ni siquiera existe la posibilidad de observar la realización de un suceso. Así, si nos preguntamos sobre la posibilidad de que el hombre ponga el pie en el planeta Marte antes del año 2020, éste es un suceso incierto de tipo no repetitivo, pero al que podemos asignar una mayor o menor verosimilitud, de modo puramente personal o subjetivo y que dependerá, en gran medida, de lo informado que esté uno en el tema. Un experto en el área daría una respuesta muy distinta que la que nos ofrecería una persona ajena al tema.

Lo más interesante de este tipo de resultados es que cada individuo, dependiendo de la información que tenga sobre los sucesos inciertos, cuantifica su incertidumbre con una medida de probabilidad personal o subjetiva.

La conclusión que se sigue es que todas las probabilidades son siempre condicionadas a un cierto estado de información, no hay, por consiguiente, probabilidades absolutas.

El teorema de Bayes es válido en todas las aplicaciones de la teoría de la probabilidad. Sin embargo, hay una controversia sobre el tipo de probabilidades que emplea. En esencia, los seguidores de la estadística tradicional sólo admiten probabilidades basadas en experimentos repetibles y que tengan una confirmación empírica mientras que los llamados estadísticos bayesianos permiten probabilidades subjetivas. El teorema de Bayes puede servir entonces para indicar cómo debemos modificar nuestras probabilidades subjetivas cuando recibimos información adicional de un experimento. La estadística bayesiana está demostrando su utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori y

permite revisar esas estimaciones en función de la evidencia.

Ahora bien nuestro estudio gira alrededor de los factores de Bayes los cuales utilizan normalmente una distribución a priori impropia que depende de una constante arbitraria cuya eliminación puede ser costosa en los problemas de contraste de hipótesis y selección de modelos.

De todas las metodologías previamente ensayadas la desarrollada por Berger y Pericchi (1996) con la introducción de factores de Bayes intrínsecos y distribuciones a priori intrínsecas ha tenido un desarrollo extenso en multitud de problemas.

Lo que podemos concluir después de haber hecho un estudio minucioso a los factores de Bayes intrínsecos es lo siguiente:

Se puede apreciar que los factores de Bayes intrínsecos son completamente automáticos, en el sentido que son basados solo en los datos y en distribuciones a priori no informativas estándares. Sin embargo, existe el problema de la elección óptima de la muestra de entrenamiento minimal.

Otra ventaja es que corresponden a factores de Bayes actuales para distribuciones a priori intrínsecas razonables; además que los factores de Bayes pueden ser aplicados para comparación de modelos y predicción.

Los factores de Bayes pueden ser aplicados en situaciones en las cuales los métodos asintóticos bayesianos no son aplicables. También pueden usarse para contraste de hipótesis bayesianos estándar.

Una de las desventajas de los factores de Bayes intrínsecos es que computacionalmente son muy costosos, ya que su cálculo puede ser bastante intensivo.

Otra desventaja es que no son invariantes para transformaciones multivariadas de los datos.Los factores de Bayes intrínsecos pueden ser inestables para muestras de tamaño pequeño.

Ahora con respecto al quinto capítulo, podemos decir que tanto el contraste de hipótesis bajo homocedasticidad como el contraste de hipótesis bajo heterocedasticidad tienen una fundamentación bayesiana, pues fueron resueltos estrictamente bajo los principios de cálculo de probabilidades y teoría de decisión.

La obtención de los factores de Bayes intrínsecos para el contraste de hipótesis bajo homocedaticidad así como bajo heterocedasticidad es muy laborioso, pero este procedimiento puede ser omitido utilizando las ecuaciones (5.19) y (5.22), las cuales fueron demostradas minuciosamente.

Las ecuaciones antes mencionadas así como la utilización de algún paquete computacional, reduce ampliamente el trabajo para la selección de modelos. Pero desgraciadamente para algunas distribuciones a priori, los factores de Bayes intrínsecos son difíciles de calcular, ya que surgen integrales que no pueden ser resueltas analíticamente, para resolver éste problema existen métodos númericos.

Así la conclusión general a que se llega es que bajo el enfoque bayesiano los

problemas de selección de modelos y contraste de hipótesis se resuelven a tráves de los factores de Bayes, cuando la distribución a priori es propia. En el caso en que al definir los factores de Bayes, la distribución apriori sea una distribución a priori impropia, se utilizarán los "factores de Bayes intrínsecos" (FBI), los cuales fueron definidos por Berger y Pericchi (1996).

Al final del capítulo 5 se lleva a cabo una selección de modelos, contrastando homocedasticidad frente a heterocedasticidad. Aunque este ejemplo es muy breve ilustra claramente como utilizar los factores de bayes intrínsecos.

Apéndice A

Notación

μ	Media poblacional
σ^2	Varianza poblacional
Θ	Espacio muestral
$\exp(\theta)$	Exponencial de θ
$\pi(\theta)$	Distribución a priori del parámetro θ
$f(x\|\theta)$	Función de densidad de los datos
$I(\theta)$	Matriz de información esperada de Fisher
$\pi(\theta\|x)$	Distribución a posteriori del parámetro θ
$h(x,\theta)$	Distribución conjunta de x y θ
$\mathcal{N}(\theta,\sigma^2)$	Distribución normal con media θ y varianza σ^2
$\Gamma(\alpha)$	La función gamma
$\pi_i^N(\theta)$	Distribución a priori no informativa de θ
$\pi_i^I(\theta_i)$	Distribución a priori intrínseca de θ_i
$\wp(\theta)$	Distribución Poisson
F	Estadistico de prueba
$F_{m,n}$	Distribución F con m y n grados de libertad
$x(l)$	Muestra de entrenamiento minimal
$det\ \hat{I}$	Determinante de la matriz observada de Fisher
$t(\alpha,\mu,\sigma^2)$	Distribución t con α grados de libertad, parámetro de localización μ y parámetro de escala σ^2
$\mathcal{C}(\mu,\sigma)$	Distribución de Cauchy con $-\infty < \mu < \infty$ y $\beta > 0$
$Ga(\alpha,\beta)$	Distribución gamma con parámetros $\alpha > 0$, $\beta > 0$

$\mathcal{N}_p(\theta, \sum)$	Normal p-variada, con media μ y matriz de covarianza $\sum$
$i.i.d$	Independiente e idénticamente distribuida
$m(x)$	Densidad marginal de X
I_p	La matriz idéntidad de $(p \times p)$.
$l_x(\theta)$	La función de verosimilitud
$\bar{x}$	La media de la muestra; esto es $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$
$E_{\hat{\theta}_2}^{M_2}$	Esperanza a $\hat{\theta}_2$ bajo el modelo M_2
H_0	Hipótesis nula
H_1	Hipótesis alternativa

ABREVIACIONES

Fte. de var.	Fuente de variación
Trat.	Tratamiento
G.l	Grados de libertad
S.C	Suma de cuadrados
C.M	Cuadrado medio
CME	Cuadrado medio esperado
Bloq. Tot.	Bloque total
Med. del bloque	Media del bloque
Tot. del trat.	Total del tratamiento
Med. del trat.	Media del tratamiento

Apéndice B

Integrales

Para $a > 0$, $p > 0$

$$\int_0^\infty x^{-(p+1)} e^{-ax^{-2}} dx = \frac{1}{2} a^{-p/2} \Gamma(p/2) \tag{B.1}$$

$$\int \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2}{2\sigma_i^2} \right] d\mu_i = \frac{(2\pi)^{1/2} \sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{2\sigma_i^2} \right] \tag{B.2}$$

$$\int \exp\left[-\frac{1}{2} \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{-1} (\mu_i - \eta_i)^2 \right] d\eta_i = \sqrt{2\pi} \left(\frac{\tau^2 + \sigma_i^2}{2} \right)^{1/2}. \tag{B.3}$$

Para $a > 0$

$$\int_0^\infty e^{-ax^2} dx = \frac{1}{2} \sqrt{\frac{\pi}{a}} \tag{B.4}$$

(B.2) es obtenida de la siguiente forma

$$\int \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{2\sigma_i^2}\right] d\mu_i$$

$$= \int \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\mu_i)^2}{2\sigma_i^2}\right] d\mu_i$$

$$= \int \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+2\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\mu_i)+\sum_{j=1}^{n_i}(\bar{x}_i-\mu_i)^2}{2\sigma_i^2}\right] d\mu_i$$

$$= \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right] \int \exp\left[-\frac{\sum_{j=1}^{n_i}(\bar{x}_i-\mu_i)^2+2\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\mu_i)}{2\sigma_i^2}\right] d\mu_i$$

$$= \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right] \int \exp\left[-\frac{\sum_{j=1}^{n_i}(\bar{x}_i-\mu_i)^2}{2\sigma_i^2}\right] d\mu_i$$

$$= \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right] \int \exp\left[-\frac{n_i(\bar{x}_i-\mu_i)^2}{2\sigma_i^2}\right] d\mu_i$$

tomando $t = \mu_i - x_i$ y $dt = d\mu_i$

$$= \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right] \int \exp\left[-\frac{n_i t^2}{2\sigma_i^2}\right] dt = \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right] \cdot \sqrt{\frac{\pi}{\frac{n_i}{2\sigma_i^2}}}$$

$$= \frac{(2\pi)^{1/2}\sigma_i}{(n_i)^{1/2}} \times \exp\left[-\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{2\sigma_i^2}\right].$$

(B.3) es encontrada de la siguiente manera

$$\int \exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}(\mu_i-\eta_i)^2\right]d\eta_i$$

tomando $t=(\eta_i-\mu_i)$, $dt=d\eta_i$ y usando (B.4)

$$\begin{aligned}
&\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}(\mu_i-\eta_i)^2\right]d\eta_i \\
= \;& 2\int_0^{\infty}\exp\left[-\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}t^2\right]dt \\
= \;& 2\left[\frac{1}{2}\sqrt{\frac{\pi}{\frac{1}{2}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{-1}}}\right]=\sqrt{\pi(\tau^2+\sigma_i^2)} \\
= \;& \sqrt{\pi(\tau^2+\sigma_i^2)}=\sqrt{2\pi(\frac{\tau^2+\sigma_i^2}{2})} \\
= \;& \sqrt{2\pi}\left(\frac{\tau^2+\sigma_i^2}{2}\right)^{1/2}.
\end{aligned}$$

Bibliografía

[1] Alamilla López Norma Edith (2004). *Selección bayesiana de modelos: Una aplicación a los modelos de regresión lineal heterocedásticos usando factores de Bayes intrínsecos.* Tesis de maestría. IMASS-UNAM.

[2] *An Intrinsic Limiting Procedure for Model Selection and Hypothesis Testing.* Journal of the American Statistical Association, 93, 1451-1460.

[3] Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Segunda edición, Springer-Verlag.

[4] Berger, James O. (1996). *The Intrinsic Bayes Factor for model selection and Prediction.* Journal of the American Statistical Asociation, 91, 109-122.

[5] Berger, J. O. and Pericchi, L.R. (1998). *On Cristicism and Comparison of default Bayes Factors for Model Selection and Hypothesis Testing. Proceedings of the Workshop on Model Selection.* Ed. W. Racugno, Pitágora Ed., Bologna, 1-50.

[6] Bernardo, José M. y Smith Adrian F. M. (1994), *Bayesian Theory.* John Wiley & Sons Ltd.

[7] Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, M. A: Addison Wesley.

[8] Box, G. E. P., and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* Wiley Classics Library Edition.

[9] Casella, George y Berger, Roger I. (2002). *Statistical Inference.* Duxbury/Tomson Learning.

[10] Moreno, E. and Giron, J. (1999). *Model selection with vague priori information.* Technical Report. University of Granada.

[11] Moreno, E., Bertolino F. and Racugno W. *Bayesian Model selection Approach to Analysis of variance under heterocedasticy.* The Statiscian Journal of the Royal Stadistic Society - Series D. Vol 49, issue 4. December 2000.

[12] Morris, H. Degroot. *Probabilidad y Estadística.* Segunda edición. Addisson Wesley.

[13] Susan J. and Arnold J. *Probabilidad y estadística con aplicaciones para ingeniería y ciencias computacionales.*Cuarta edición. McGrawHill.

[14] Zacarías Santiago Adriana (2006). *Métodos de integración y simulación Monte Carlo en la teoría bayesiana.* Tesis de licenciatura. Universidad Tecnológica de la Mixteca.