

63385

# **METODO BAYESIANO: MODELO LINEAL BAYESIANO**

Especialidad en Estadística  
Orientación: Diseño Experimental



Autora: Silvia Cabrera  
Director: Mg. Héctor Agnelli  
Año: 2006

Universidad Nacional de Río Cuarto  
Facultad de Ciencias Exactas, Físico-Químicas y Naturales  
Departamento de Matemática

CABRERA, S.  
Método Bayesiano: Mo



2006

63385

# MÉTODO BAYESIANO: MODELO LINEAL BAYESIANO

Especialidad en Estadística  
Orientación: Diseño Experimental

Autora: **Silvia Cabrera**

Director: **Mg. Héctor Agnelli**

Universidad Nacional de Río Cuarto  
Facultad de Ciencias Exactas, Físico-Químicas y Naturales  
Departamento de Matemática

Año: 2006

65385

<b>MFN:</b>
<b>Clasif:</b> 5.470

## Agradecimientos:

Finalizado este trabajo, deseo manifestar mi más sincero agradecimiento a todas aquellas personas que, de una u otra forma, han colaborado en su elaboración.

Así, y en primer lugar, quiero expresar mi gratitud a mi Director el **Mg. Héctor Agnelli**, la confianza y ánimo que siempre ha depositado en mí, al valorar este trabajo, en el que el rasgo más significativo se centra en las orientaciones docentes que se presentan para desarrollar un contenido educativo y pedagógico.

Del mismo modo, deseo agradecer a la **Facultad de Ciencias Económicas** de la Universidad de Río Cuarto, en la que a través de sus docentes, y mi colaboración en la misma durante estos años como profesora, me han permitido recopilar materiales que he podido coordinar y depurar para utilizarlos en la elaboración del presente trabajo.

Un lugar destacado en este apartado de agradecimientos queda para mi **familia**, en donde mi madre, mi marido y mi hijo Pablo, así como mis hermanas, mis cuñados y mis sobrinos, han sido la luz que me ha permitido seguir en aquellos momentos más grises.

Por último, no puedo olvidar a mi compañera de estudio, de labor docente y amiga **Nancy Scattolini**, puesto que siempre me ha transmitido ilusión, energía y ganas de aprender, junto con la creencia de que es posible mejorar.

A todos, simplemente gracias por haber contribuido a concluir esta etapa.

*Silvia Cabrera*





## Rev. THOMAS BAYES

Thomas nació en Londres, Inglaterra en el año 1702, aunque el acta de su nacimiento definitivamente no existe y murió el 17 de Abril de 1761 en Tunbridge Wells, Kent, Inglaterra, en su tumba se halla la siguiente inscripción:

*“ Rev. Thomas Bayes. Hijo de Joshua y Ann Bayes. 7 de Abril de 1761.  
En reconocimiento del importante trabajo sobre probabilidad  
realizado por Thomas Bayes. La cripta fue restaurada en 1969 con  
contribuciones recibidas de estadísticos de todo el mundo.”*

El padre de Thomas Bayes fue uno de los seis primeros ministros disidentes de la Iglesia Anglicana que fueron ordenados en Inglaterra. Thomas fue educado en forma privada, esto tal vez fue necesario por tratarse del hijo de un ministro disidente. Nada se conoce de quienes fueron sus maestros, aunque *Banard* en uno de los puntos de su libro, deja entrever la intrigante posibilidad de que uno de ellos haya sido *de Moivre*, quien en ese tiempo impartía enseñanza privada en Londres.

Thomas Bayes fue ordenado ministro disidente, al igual que su padre y fue uno de sus primeros asistentes en Holborn. Al final de 1720 es nombrado ministro de la capilla presbiteriana en Tunbridge Wells, a 35 millas al sudeste de Londres. Aparentemente Bayes procuraba retirarse como ministro en 1749, hecho que se produjo recién en 1752, aunque siguió viviendo en Tunbridge Wells.

Thomas, teólogo y matemático part-time, fue quien primero usó la probabilidad inductivamente y quién estableció una matemática básica para inferencia probabilística, esto es, un medio para calcular la probabilidad de un evento que ocurrirá en ensayos futuros, dado que el evento ha ocurrido en ensayos anteriores.

Bayes pone en consideración su teoría de probabilidad con un artículo titulado “ *Essay towards solving a problem in the doctrine of chances*” que fue publicado post-mortem en 1764 en el *Philosophical Transactions of the Royal Society of London*, el artículo fue enviado a la Royal Society por Richard Price, un amigo de Bayes, quién escribió:

*“Le envío un ensayo que he encontrado entre los papeles de nuestro amigo difunto Sr. Bayes, y qué, en mi opinión, tiene gran mérito.... En una introducción que él tiene escrito a este Ensayo, dice, que su plan al principio era, averiguar un método por el cual se podía juzgar acerca de la probabilidad que un evento tiene que pasar, en circunstancias dadas, suponiendo que no sabemos nada e ir involucrándolo bajo las mismas circunstancias, pasado un cierto número de tiempos, y faltando otro número de tiempos.”*

El artículo, en cuestión, sienta la técnica básica de la estadística, ahora llamada *estimación Bayesiana*. Como dato curioso, *Laplace* aceptó las conclusiones de Bayes recién en 1781 y sólo dos trabajos fueron publicados mientras él vivió:

- “*Divine Providence and Government Is the Happiness of His Creatures*” en 1731
- “*An Introduction to the Doctrine of Fluxions, and a Defence of the Analyst*” en 1736.

Este último es un contraataque a las declaraciones hechas por Bishop *Berkeley* sobre las fundamentaciones lógicas del cálculo de *Newton*. Bayes le escribe a *Berkeley*:

*“...representadas las disputas y controversias entre matemáticos como desacreditar la evidencia de sus métodos: y... sus representadas Lógicas y Metafísicas como apropiado para abrir sus ojos, y los saca desde sus dificultades. ... Si las disputas de los profesores de cualquier ciencia desacreditan la propia ciencia, Lógicas y Metafísicas son muchas más desacreditadas que la Matemática, ¿por qué, por consiguiente, si yo soy medio ciego, debo tomar yo para mi guía, uno que no puede ver en absoluto?”*

Thomas, fue electo miembro de la Royal Society en 1742, a pesar que en ese momento no tenía trabajos publicados en matemática, de hecho, ninguno se publicó en su vida bajo su propio nombre, el artículo sobre diferenciales, referido anteriormente, fue publicado anónimamente. Otra publicación matemática fue sobre series asintóticas que apareció después de su muerte.

---

# Índice

---

Prefacio	1
<b>Capítulo I: Conceptos Generales</b>	<b>1</b>
Introducción	1
Notación general: Parámetros, datos y predicciones.	3
<b>1.1 Probabilidad</b>	<b>3</b>
1.1.1 Probabilidad Subjetiva	3
<b>1.2 Teorema de Bayes</b>	<b>4</b>
1.2.1 Generalización del Teorema de Bayes	5
<b>1.3 Proceso Bayesiano</b>	<b>7</b>
1.3.1 Teorema de Bayes y la función de Verosimilitud	9
1.3.2 Principio de Verosimilitud	9
1.3.3 La Verosimilitud estandarizada	10
1.3.4 Naturaleza secuencial del Teorema de Bayes	11
<b>1.4 Distribución a priori informativa.</b>	<b>13</b>
1.4.1 Prioris Conjugadas	16
<b>1.5 Distribuciones a priori no informativas</b>	<b>18</b>
1.5.1 Distribuciones a priori propias e impropias	18
1.5.2 Principio de invarianza de Jeffreys	21
1.5.3 Dificultades con distribuciones no informativas	23
<b>1.6 Predicción</b>	<b>24</b>
1.6.1 Odds y Razón de Verosimilitud	25
<b>1.7 Diferencias entre la Inferencia Bayesiana y la Clásica</b>	<b>26</b>
<b>1.8 CONCLUSIONES</b>	<b>27</b>
<b>Capítulo II : Inferencia Bayesiana en la toma de decisiones</b>	<b>29</b>
Introducción	29
<b>2.1 Teoría de la decisión.</b>	<b>30</b>
<b>2.2 Estimación puntual</b>	<b>32</b>
2.2.1 Función de pérdida cuadrática	32

2.2.2 Función de pérdida valor absoluto	33
2.2.3 Función de pérdida lineal	33
2.2.4 Función de pérdida 0 - 1	33
<b>2.3 Estimación por intervalos: Intervalos de Credibilidad</b>	<b>34</b>
2.3.1 Regiones de Máxima Densidad (HPD)	34
<b>2.4 Test de Hipótesis</b>	<b>36</b>
<b>2.5 CONCLUSIONES</b>	<b>40</b>
<b>Capítulo III: El modelo lineal general y la inferencia Bayesiana</b>	<b>42</b>
Introducción	42
<b>3.1 El Modelo Lineal General</b>	<b>43</b>
3.1.1 Distribución a priori no informativa	46
3.1.2 Distribución a posteriori con información a priori conjugada	48
<b>3.2 La distribución predictiva a posteriori para nuevos datos</b>	<b>51</b>
<b>3.3 Variancias Heterogéneas y Correlacionadas</b>	<b>52</b>
<b>3.4 Inferencia Bayesiana en el Modelo Lineal</b>	<b>54</b>
3.4.1 Estimación Puntual	54
3.4.2 Regiones HPD	55
3.4.3 Test de Hipótesis	55
<b>3.5 CONCLUSIONES</b>	<b>59</b>
<b>Capítulo IV: Aplicación del método Bayesiano a un problema de Decisión de Inversión</b>	<b>61</b>
Introducción	61
<b>4.1 Decisión de inversión</b>	<b>63</b>
4.1.1 Planteo del problema	64
4.1.2 <i>Seleccionar un criterio de decisión: Ganancia Esperada</i>	65
4.1.3 <i>Prueba de Confiabilidad</i>	70
<b>4.2 Función de utilidad</b>	<b>73</b>
<b>4.3 Modelo de Regresión para la función de utilidad</b>	<b>77</b>
4.3.1 Construcción de una distribución a priori	78
4.3.2 Cálculo de la distribución a posteriori	79
<b>4.4 Inferencia a posteriori</b>	<b>80</b>
4.4.1 Análisis de variancia	80
4.4.2 Estimación puntual	81
4.4.3 Estimación por intervalos	82
4.4.4 Test de hipótesis para los coeficientes de regresión	84
<b>4.5 CONCLUSIONES</b>	<b>86</b>
<b>Síntesis y Reflexión Final</b>	<b>87</b>
<b>Bibliografía Consultada</b>	<b>89</b>

---

## Prefacio

---

El objeto de estudio de este trabajo está constituido, fundamentalmente, por el conjunto de los conceptos básicos que configuran la *teoría Bayesiana*, las condiciones que caracterizan dichos conceptos tanto en el desarrollo y formulación del modelo lineal general *Bayesiano*, como en la teoría de toma de decisiones.

En el Capítulo I exponemos algunos aspectos generales importantes del *Método Bayesiano*, incluyendo el papel de la inferencia *Bayesiana* dentro de la estadística, la opción de distribuciones *a priori* informativas y, en particular, de distribuciones *a priori* no informativas. Presentamos las inferencias *a posteriori*, la *predicción* desde éste punto de vista y por último, remarcamos las diferencias que existen con la estadística Clásica. En el Capítulo II nos introducimos a la *teoría de toma de decisiones* desde el punto de vista de *Bayes* aplicado a estudios cargados de una onda incertidumbre, y que por ello el proceso *Bayesiano* se torna tan relevante, además mostramos las estimaciones puntuales, por intervalos y test de hipótesis teniendo en cuenta la estimación *Bayesiana* aplicada a la teoría de toma de *decisiones*. Estos capítulos nos permiten construir el Capítulo III, en el cual analizamos al modelo lineal Normal clásico justificado desde el punto de vista *Bayesiano*, damos algunas interpretaciones de este importante tema, adicionamos información *a priori* para el modelo Lineal Normal, describiendo las distribuciones *a priori* y *a posteriori* de los parámetros involucrados en él y aplicamos



las estimaciones *Bayesianas* a los coeficientes de regresión. Finalmente, analizamos, en el Capítulo IV un problema simulado sobre *decisión de inversión*, en el cual utilizamos los conceptos de la metodología *Bayesiana*, desde el planteo mismo del problema hasta el modelado matemático de la función de utilidad implicada en dicho problema.

El objetivo de este trabajo es explorar el uso y relevancia del *Teorema de Bayes* aplicado a los problemas que residen en la investigación científica, en la que deben hacerse inferencias acerca de valores del parámetro sobre los que poco se conoce *a priori*.

Mi inicial formación como profesora en Matemática y posteriores estudios de Estadística y Docencia Universitaria, y fundamentalmente como colaboradora en la Facultad de Ciencias Económicas de ésta universidad como profesora e investigadora, pueden explicar el interés que me acerca y me ha animado a este estudio, con un razonable temor por la envergadura de los temas a analizar, reflexionar y afrontar las dificultades que un trabajo de este tipo conlleva. La aproximación a los contenidos de este trabajo tuvo que ver con la percepción de que conceptos, como probabilidad frecuentista e intervalos de confianza chocan con los conceptos “intuitivos” y esto en ocasiones deriva al rechazo progresivo de utilizar métodos estadísticos en problemas de gestión y administración.

La idea con que se elaboró este trabajo, parte de considerar que la estadística construida sobre la base de una perspectiva *Bayesiana* es una aproximación tangible del uso del método científico. Este punto de vista -el análisis *Bayesiano*- tiene varias ventajas, por ejemplo, permite incorporar evidencia desde experiencias previas y conclusiones globales extraídas de experimentos previos. Otra es la subjetividad, aunque esta es la objeción que muchos estadísticos realizan al método *Bayesiano*, desde una óptica científica, esta subjetividad es la norma en ciencia.

La estadística no es meramente un conjunto de métodos para analizar datos, también es una vía o camino para integrar esos datos dentro del proceso científico. La experimentación y la cuantificación de la incertidumbre son fundamentales en ciencias. Los científicos aprenden usando el método científico, el que queda descripto con los siguientes pasos:

**Método científico.**

- 1- Preguntarse cuál es la cuestión o planteo del problema.
- 2- Reunir y evaluar la información relevante.
- 3- Basado en la información actual, realizar una investigación o un experimento en dirección a la cuestión planteada en el paso 1. Considerar los costos y los beneficios disponibles para el experimento, incluyendo el valor de alguna información relevante de contenido.
- 4- Llevar a cabo la investigación o experimento.
- 5- Usar la evidencia obtenida en el paso 4, para actualizar la información previa disponible; sacar conclusiones.
- 6- Repetir los pasos 3 a 5 las veces que sea necesario.

Creemos que investigar un problema siguiendo los pasos anteriores, proporciona evidencia suficiente del valor que tiene el método *Bayesiano*.

La idea básica del método *Bayesiano* es la siguiente:

*Esquema Bayesiano*



donde  $\theta$ , es el parámetro desconocido que se quiere estudiar, es decir que es el objeto de estudio de cualquier inferencia estadística.

Además, la incertidumbre es un aspecto básico de nuestro mundo. Hay muchos eventos en nuestras vidas en los que los resultados son inciertos. Por ejemplo, no estamos seguros sobre si mañana viviremos diez años más o si tendremos una enfermedad “rara”. Las *probabilidades* proporcionan los medios para medir la incertidumbre. Una *probabilidad* de un evento particular es un número entre 0 y 1, eso mide la opinión subjetiva del sujeto sobre la verosimilitud del evento. Las *probabilidades*, aquí, son condicionales en el sentido que un evento es dependiente en nuestro actual estado de conocimiento. Cuando obtenemos nueva información sobre el evento, nuestras *probabilidades* cambian. Luego, la teoría *Bayesiana*, provee un mecanismo para cambiar esas *probabilidades* cuando obtenemos nuevos datos.

En el proceso de elaboración de este trabajo han ido apareciendo, como es lógico, nuevos temas vinculados al argumento central de nuestro trabajo. Estos temas, entre los

cuales se podría citar una discusión profunda sobre el carácter metodológico de la teoría de *toma de decisiones*, o bien, un análisis exhaustivo de las aplicaciones de la metodología *Bayesiana* en la *teoría de la decisión* ( no sólo del modelo de regresión), han recibido un tratamiento incompleto, a veces marginal. Ello ha sido así, tanto por la necesidad de centrarnos en el objeto de nuestro estudio, ya de por sí suficientemente complejo, como por nuestras propias limitaciones con relación a los nuevos temas. No obstante se ha procurado que el texto final presente la mayor coherencia que nos ha sido posible y asuma, decorosamente, sus carencias.

Con respecto a algunos aspectos formales, habría que señalar la inclusión en el texto de preguntas que motivan a los contenidos subsiguientes; conclusiones y resúmenes de cada capítulo a modo de cierres parciales como lo hacemos en nuestras prácticas docentes, de forma tal de permitir la construcción paulatina pero significativa de los temas tratados.

Con relación a las notas de pie de página creemos que la exposición será más fluida, en la medida que su contenido pueda ser incluido en el texto. Por último, debemos señalar que la bibliografía final recoge, por un lado, los libros o artículos de los autores mencionados o citados a lo largo del texto, y por otro comentarios sobre los mismos que motiven a indagar exhaustivamente acerca de ésta importante teoría.

En definitiva, este trabajo pretende ser una modesta contribución al análisis de la utilización de la metodología *Bayesiana* en la construcción del *modelo lineal general* y en las *estimaciones* de sus *coeficientes*.



---

# Capítulo I

## Conceptos generales

---

### Introducción

---

Hacia mediados del siglo XVIII, precisamente en 1764, fue publicado el *Teorema de Bayes*, así llamado en conmemoración al ministro disidente de la Iglesia Anglicana que lo desarrolló en respuesta a los postulados de la inferencia gaussiana.

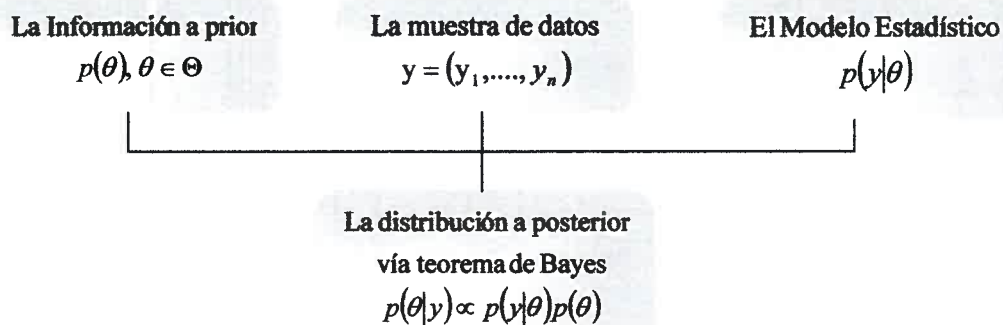
El estudio clásico de las distribuciones de probabilidad o estadística gaussiana, supone funciones de densidad simétricas y bien definidas, así como la ausencia de cualquier tipo de conocimiento previo por parte del investigador. Según Box y Tiao (1992), *Bayes*, en la justificación de su teoría argumentaba que los datos no necesariamente provenían de tales funciones de densidad, sino, todo lo contrario, que probablemente eran generados por leyes probabilísticas sujetas a formas asimétricas y sesgadas. Afirmando que, en tanto que el investigador conociera estas características, el procedimiento correcto de inferencia estadística debería incorporar esta información y, de esa forma, contar con un marco probabilístico más apropiado para la inferencia estadística.

Pero debió pasar más de dos siglos desde la aparición del *Teorema de Bayes*, para verlo resurgir y esto se correspondió a la mayor disponibilidad de recursos informáticos.

Es así como en el siglo pasado, Box y Tiao (1992) publican un texto sobre esta teoría: *Bayesian Inference in Statistical Analysis*. Desde entonces es posible observar cada vez más artículos sobre el *método Bayesiano* en publicaciones Estadísticas, Matemáticas y de Ciencias Aplicadas.

El objetivo de este capítulo es presentar el marco teórico de los conceptos básicos insertos en la *teoría Bayesiana*, como andamiaje a la *inferencia Bayesiana*. Para ello desarrollaremos el *Teorema de Bayes* desde su versión más elemental; considerándolo como una simple consecuencia de la definición de probabilidad condicional de un evento dado otro hasta su versión en términos de la función de densidad (marginales, condicionales y conjuntas).

La estructura de este capítulo es resumida en el siguiente esquema:



Asimismo, presentaremos las diferencias entre la metodología Clásica y Bayesiana.

Las discusiones en este capítulo están basadas sobre comentarios de estudiosos de la temática acerca de la interpretación de las componentes involucradas en el proceso *Bayesiano*, a saber, *a priori* informativas y no informativas y la *predicción*.

Por ello, en la construcción del *método Bayesiano*, se hace imprescindible comenzar con la interpretación del *concepto de probabilidad*, ya que la misma marca las diferencias entre *inferencia Bayesiana* y *Gausiana*.

Sin embargo, antes de comenzar con la conceptualización de probabilidad en términos *Bayesiano*, consideramos, para estructurar convenientemente éste trabajo, importante presentar las notaciones simbólicas que utilizaremos a lo largo del mismo.

*Notación general: Parámetros, datos y predicciones.*

Con  $\theta$  denotamos al vector de cantidades no observables o parámetros de la población de interés. Con  $y$  denotamos a los datos observados, la muestra o evidencia muestral, por último con  $\tilde{y}$  a una futura observación del proceso. En general estos símbolos representan cantidades multivariadas.

Preparando el terreno para ulteriores generalizaciones usamos la palabra *función de densidad* o simplemente *densidad* en un sentido amplio, que coincide con el de función de probabilidad en el caso discreto, y usamos los términos *distribuciones* y *densidades* indistintamente.

## 1.1 Probabilidad

Intuitivamente al término *probabilidad* se lo utiliza para describir numéricamente a la *incertidumbre* asociada con nuestras afirmaciones. Así es, como asiduamente escuchamos “La probabilidad de que ascienda a jefe es 1 de 10” o “Hay un 70% de probabilidad que Bush gane las próximas elecciones de USA”. Estas son aseveraciones probabilísticas, que de ninguna manera describen una gran repetición de esas situaciones.

Las probabilidades, aquí, se interpretan como la convicción personal con respecto del problema, antes de que se encuentre disponible la evidencia muestral.

Esta interpretación *subjetiva* de *probabilidad* es el sustento natural del *método Bayesiano*.

### 1.1.1 Probabilidad Subjetiva

La probabilidad subjetiva se refiere a los juicios que tiene una persona acerca de hechos o proposiciones inciertas.

Así, si  $A$  es un evento,  $P(A)$  mide el grado de creencia razonable de que  $A$  ocurrirá. Si  $P(A)=1$ , la persona está segura de que  $A$  ocurre, y si  $P(A)=0$ ; ella está segura de que  $A$  no ocurre. Cuando  $P(A)$  crece de 0 a 1, ésta indica el aumento de creencia de que el evento  $A$  ocurra. Así  $P(A)$  describe la *incerteza* que la persona tiene al respecto del



evento  $A$  según la información o experiencia que el tiene, digamos  $H$  (historia), luego  $P(A)$  es, siendo rigurosos,  $P(A|H)$ .

Luego podemos decir que, una *probabilidad subjetiva* es una medida del grado de *incerteza* personal que un individuo tiene con respecto a cierta proposición, basado en la información disponible. Como tal información varía de individuo a individuo, sus grados de creencia serán inicialmente diferentes. En este sentido es que Bruno de Finetti (Iglesias Zuazola, 1995) dice: “*Probabilidades no existen*”. La única restricción es que estas probabilidades no sean inconsistentes o sea satisfagan los *axiomas de probabilidad*.

Ahora bien, el paso siguiente es ¿cómo actualizamos esta información al disponer de nueva información?, la respuesta nos la da el *Teorema de Bayes*.

## 1.2 Teorema de Bayes

El *Teorema de Bayes* está basado en la definición de probabilidad condicional.

Es decir dados  $B$  y  $A$  dos sucesos (con  $P(A) \neq 0$  y  $P(B) \neq 0$ ), entonces se tiene que:

$$P(B|A) = \frac{P(AB)}{P(A)} \quad \text{y} \quad P(A|B) = \frac{P(AB)}{P(B)} \Rightarrow P(B|A) \cdot P(A) = P(A \cdot B) = P(A|B) \cdot P(B)$$

$$\Rightarrow P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} = \frac{P(A|B)}{P(A)} \cdot P(B) \quad (1.2.1)$$

que es la forma mas simple y conocida del *Teorema de Bayes*.

### Ahora bien, ¿Cómo interpretamos esta expresión?

La expresión (1.2.1) puede ser interpretada de la siguiente forma: interesa inferir con respecto a la ocurrencia del evento  $B$  a partir de una probabilidad inicial o probabilidad *a priori*  $P(B)$  para la ocurrencia de este evento y la observación de un evento  $A$ , relacionada de alguna manera con  $B$ . La descripción de cuan probable es  $B$  una vez que  $A$  ocurrió es representada por  $P(B|A)$ , y esta expresión se la denomina probabilidad *a posteriori* de  $B$ . El *Teorema de Bayes* entrega una forma coherente para actualizar la

probabilidad *a priori*,  $P(B)$ . Tal actualización consiste en multiplicar esta última probabilidad por  $\frac{P(A|B)}{P(A)}$ .

En otras palabras, esta fórmula nos permite actualizar y revisar la información inicial cuando disponemos de nueva información.

### 1.2.1 Generalización del Teorema de Bayes

Sean  $B_1, B_2, \dots, B_k$  una colección de eventos mutuamente excluyentes y exhaustivo<sup>1</sup> entonces :

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{\sum_i P(B_i) \cdot P(A|B_i)} \quad (1.2.2)$$

La colección de  $B_i$ 's es interpretado como un conjunto de hipótesis de las cuales una y sólo una es verdadera, es decir que  $B_i$ 's son hipótesis acerca del modelo generador de los datos y  $A$  denota el conjunto de datos observados, es decir que  $A$  es la evidencia experimental. Luego, la ocurrencia de  $A$  transforma la probabilidad *a priori*  $P(B_i)$  en probabilidades a posteriori  $P(B_i|A)$ , ya que se determinan una vez obtenida la evidencia experimental, luego  $P(B_i|A)$  reflejan el grado de *creencia corregido* con respecto a las hipótesis  $B_1, B_2, \dots, B_k$  después de obtener los datos experimentales.

Las probabilidades  $P(A|B_i)$  son conocidas con el nombre de *verosimilitud*, es decir  $P(A|B_i)$  es la verosimilitud dada a  $B_i$  por A o la verosimilitud de  $B_i$  dado A.

$P(B_i)$  es en realidad  $P(B_i|H)$ , donde  $H$  (historia) representa la información que se tiene para la descripción inicial de la *incerteza* con respecto al conjunto de hipótesis.

Por último el denominador de la expresión (1.2.2) es un promedio ponderado de las probabilidades  $P(A|B_i)$ , con  $i = 1, \dots, k$ .

<sup>1</sup> Mutuamente Excluyente: Si  $B_i \cap B_j = \emptyset$  siempre que  $i \neq j$

Exhaustivo:  $\cup B_i = \Omega, \forall i$

**Ilustremos estas actualizaciones, con un ejemplo que nos brinda Mendenhall, (1990):**

Una compañía estudia la comercialización de un nuevo producto. El presidente de la compañía desea que el producto sea superior al de su más cercano competidor. Con base en una evaluación preliminar que realizó personal experimentado en comercialización, se decide asignar una posibilidad del 50% de que el producto sea superior al ofrecido por el competidor, 30% de que tenga la misma calidad y un 20% de que sea inferior.

Un estudio de mercado sobre el producto concluye que éste es superior al del competidor. Sobre la base de la experiencia en resultados de las encuestas, se determina que si el producto realmente es superior, la probabilidad de que la encuesta alcance la misma conclusión es 0.7. Si el producto tiene la misma calidad que el del competidor, la probabilidad de que la encuesta dé como resultado un producto superior es 0.4. Si el producto es inferior, la probabilidad de que la encuesta indique un producto superior es de 0.2. Dado el resultado de la encuesta, ¿cuál es la probabilidad, corregida, de obtener un producto superior?.

**Resolución:**

Sean  $B_1, B_2$  y  $B_3$  los eventos de que el producto sea *superior*, tiene la *misma calidad* y es *inferior* al del competidor, respectivamente. Las probabilidades *a priori* correspondientes son 0.5, 0.3 y 0.2. Sea  $A$  el evento “*la encuesta revelará un producto superior*”. Las probabilidades condicionales que involucran una *evidencia experimental* son  $P(A|B_1)=0.7$   $P(A|B_2)=0.4$  y  $P(A|B_3)=0.2$  (Verosimilitudes).

Luego, la probabilidad *a posteriori*  $P(B_1|A)$  deseada es:

$$P(B_1|A) = \frac{P(B_1) \cdot P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)} = 0.6863$$

En otras palabras, *la posibilidad corregida de obtener un producto superior es del 69%*.

### 1.3 Proceso Bayesiano

El proceso *Bayesiano* esta compuesto por:

La información *a priori*, la muestra de datos, el cálculo de la distribución *a posteriori* y en algunos casos el cálculo de la distribución de futuras observaciones, esto es, la *predicción*.

Re - escribimos el *Teorema de Bayes* en términos de las funciones de *densidad*, esta forma diferente del *Teorema de Bayes*, resultará especialmente útil en el Capítulo II, cuando ilustremos el uso el mismo en el contexto de la teoría de decisión.

Asiduamente, a un problema lo ponemos en términos de  $k$  cantidades desconocidas, (donde  $k$  puede ser 1 o mayor que 1):

$$\theta = (\theta_1, \theta_2, \dots, \theta_k)$$

con valores en un conjunto  $\Theta$ , esto es,  $\Theta$  una *variable aleatoria* definida de tal manera que sus valores representan las posibles opciones en que puede ocurrir un fenómeno aleatorio antes de llevar a cabo un experimento.

El grado de creencia del investigador con respecto a esas posibilidades se encuentra expresado en términos de una función de *densidad*,  $p(\theta|H)$ ,  $\theta \in \Theta$ , que recibe el nombre de: función de *densidad* de probabilidad *a priori* de  $\theta$ , recordemos que esta función de densidad describe la *incerteza* personal con respecto a  $\theta$ , esta incerteza personal se puede basar en cualquier tipo de información que se encuentre disponible, incluyendo el *juicio subjetivo*.

Con el objetivo de *actualizar* nuestra convicción o *incerteza personal* observamos  $n$  valores de un vector aleatorio  $y = (y_1, y_2, \dots, y_n)$  relacionados con  $\theta$  a través de una familia de modelos probabilísticos,  $p(y|\theta, H)$ , donde la dependencia de  $\theta$  es fundamental.

La función,  $p(y|\theta, H)$ , recibe el nombre de función de *verosimilitud* debido a que representa el grado de concordancia del resultado muestral  $y$ , dado el valor  $\theta$  de  $\Theta$ .

**Nota:** Algunas veces, escribimos  $l(\theta|y) = p(y|\theta)$

Cuando combinamos la distribución *a priori* con la función de *verosimilitud*, ello da lugar a una distribución condicional para  $\theta$  dada la evidencia muestral, que lleva por nombre distribución *a posteriori* de  $\theta$ .

Esta combinación se hace de acuerdo con el *Teorema de Bayes*, que lo replanteamos de la siguiente manera:

Sea  $p(\theta)$  la función de densidad de probabilidades *a priori* de  $\theta$ , que es el conocimiento sobre  $\theta$  sin el conocimiento de la muestra y sea  $p(y|\theta)$  la función de *verosimilitud*. Entonces la función de densidad de probabilidad *a posteriori* de  $\theta$  dada la evidencia muestral “y” es:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} ; \theta \in \Theta \quad (1.3.1),$$

donde  $p(\theta, y)$  es la distribución conjunta de  $\theta$  e  $y$ .

$$y \quad p(y) = E(p(y|\theta)) = c^{-1} = \begin{cases} \int p(y|\theta) \cdot p(\theta) \cdot d(\theta) & \text{si } \theta \text{ es continuo} \\ \sum_{\theta} p(y|\theta) \cdot p(\theta) & \text{si } \theta \text{ es discreto} \end{cases}$$

(la suma y la integral es realizada sobre el rango de posibles valores de  $\theta$ , y  $E(p(y|\theta))$  es la esperanza matemática)

La cantidad “c” es simplemente una constante normalizante necesaria para garantizar que la integral o la suma de la distribución *a posteriori*,  $p(\theta|y)$ , sea 1.

Así, la expresión (1.3.1) puede ser expresada alternativamente como:

$$p(\theta|y) = cp(y|\theta)p(\theta) \quad (1.3.2)$$

Las expresiones (1.3.1) o (1.3.2) se refieren usualmente al *Teorema de Bayes*.

**Aclaración:** En éste trabajo, a veces, para referirnos a la distribución *a priori* y a la distribución *a posteriori*, simplemente diremos *a priori* y *a posteriori*, respectivamente.

### 1.3.1 Teorema de Bayes y la función de Verosimilitud

La función de *Verosimilitud*,  $p(y|\theta)$ , juega un papel muy importante en el proceso *Bayesiano*. Ésta es, la función a través de la cual, la muestra “y” modifica el conocimiento *a priori* de  $\theta$ ; por consiguiente se la considera como la representación de la información de  $\theta$  que viene desde la muestra.

La función de *Verosimilitud* definida en (1.3.2) esta multiplicada por una constante,  $c$ , que deja a esa *verosimilitud invariante*. Luego, en la fórmula de *Bayes*, esa constante “c” no afecta a la distribución *a posteriori* de  $\theta$ . Entonces, se la puede cancelar, ya que es sólo el valor relativo de la *verosimilitud* el que es de importancia. Este hecho produce la densidad *a posteriori* no normalizada, la cual tiene por expresión:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \quad (1.3.3)$$

donde  $\propto$  denota *proporcional a*

El símbolo ( $\propto$ ) indica que la relación expresada en (1.3.3) es una proporción y no una igualdad. Es decir, la distribución de probabilidad de  $\theta$  posterior a la información de “y” es proporcional al producto de la distribución *a priori* de  $\theta$  y la *verosimilitud* de  $\theta$  dado y.

Luego el proceso de actualización de la información lo resumimos como:

$$\text{Posteriori} \propto \text{Priori} \times \text{Verosimilitud}$$

Esta relación resume la manera en la que debemos modificar nuestras creencias para tener en cuenta los datos que tenemos disponible

Esta relación es la interpretación del siguiente principio:

### 1.3.2 Principio de Verosimilitud

Sean  $X$  e  $Y$  cantidades aleatorias con función de densidad  $p_1(\cdot|\theta)$  y  $p_2(\cdot|\theta)$ ,  $\theta \in \Theta$ , es decir estas cantidades están relacionadas con el objeto de la inferencia,  $\theta$ .

Si  $x$  e  $y$  son realizaciones de  $X$  e  $Y$  respectivamente, tal que  $p(x|\theta) = c p(y|\theta)$  para alguna constante  $c$  y para cada  $\theta \in \Theta$ , entonces los dos resultados experimentales



poseen la misma información sobre  $\theta$ , consecuentemente deben conducir a la misma inferencia.

El *método Bayesiano* de inferencia no viola el *principio de verosimilitud* para una *priori* dada  $p$  sobre  $\theta$ .

En efecto:

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto p(y|\theta)p(\theta) \\ &\propto p(\theta|y) \end{aligned}$$

$$\text{Luego, } p(\theta|x) \propto p(\theta|y), \forall \theta \in \Theta$$

El *principio de verosimilitud* nos dice que la/s conclusiones que obtenemos a partir del experimento sólo dependen de la observación que *ocurrió* y no de las que podrían *haber ocurrido*.

### 1.3.3 La Verosimilitud estandarizada

Notemos que la forma en que escribimos anteriormente el *Teorema de Bayes* con un signo de proporcionalidad, es debido a que el resultado no se altera si multiplicamos a  $p(y|\theta)$  por una constante o más generalmente por algo que es exclusivamente una función de  $y$ . Acordando que, podemos considerar la definición de la verosimilitud como cualquier múltiplo de  $p(y|\theta)$  en lugar de necesariamente la misma  $p(y|\theta)$ .

Cuando la integral  $\int p(y|\theta) \cdot p(\theta) \cdot d(\theta)$ , (**interpretada** como una integral múltiple,  $\iint \dots \int \dots d\theta_1 d\theta_2 \dots d\theta_k$  si  $k > 1$ , e **interpretada** como una sumatoria o sumatoria múltiple en el caso discreto), realizada sobre los valores posibles de  $\theta$ , es finita, entonces a veces es conveniente referirse a la cantidad:

$$\frac{p(y|\theta)}{\int p(y|\theta)p(\theta)d\theta} \quad (1.3.4)$$

Llamamos a esta cantidad la *verosimilitud estandarizada*, esto es, la verosimilitud estandarizada no depende de la escala en que sea medida.

### 1.3.4 Naturaleza secuencial del Teorema de Bayes

La distribución *a posteriori* como ya hemos mencionado en oportunidades previas, refleja el grado de creencia modificado del investigador con respecto a la variable aleatoria  $\theta$  después de obtener información muestral. Es decir, que el *Teorema de Bayes* nos permite poner al día información continuamente sobre un conjunto de parámetros  $\theta$  cuando más observaciones se toman. Así, dentro de este contexto decimos que el *Teorema de Bayes* es un proceso secuencial, iterativo e interactivo.

*Interactivo*, porque existe un diálogo real entre la fórmula de *Bayes* y la información inicial del investigador, éste es partícipe necesario en el proceso de actualización de la información mediante el bagaje intelectual acerca del problema en cuestión.

*Iterativo*, porque una vez obtenida la distribución *a posteriori*, ésta puede convertirse, en un futuro, en una distribución *a priori* cuando lleve a cabo otra revisión con respecto a la variable aleatoria. Y esta revisión de las probabilidades se hace posible mediante el empleo sucesivo del *Teorema de Bayes*.

*Secuencial*, porque la actualización de estas probabilidades se realizan paso a paso.

Veamos:

Sea  $y_1$  una muestra inicial de observaciones, entonces por *Bayes* tenemos:

$$p(\theta|y_1) \propto p(\theta) \cdot p(y_1|\theta) \quad (1.3.5)$$

y sea  $y_2$  un segundo conjunto de observaciones tal que  $y_1$  e  $y_2$  tienen densidad conjunta,  $p(y_1, y_2|\theta)$ ,  $\theta \in \Theta$ , entonces,

$$p(\theta|y_1, y_2) = \frac{p(y_1, y_2, \theta)}{p(y_1, y_2)} \propto p(y_2|y_1, \theta)p(y_1|\theta)p(\theta),$$

$$\text{luego } p(\theta|y_1, y_2) \propto p(y_2|y_1, \theta)p(\theta|y_1)$$

**Observamos** que el segundo factor de esta última expresión es la nueva distribución *a priori* cuando se incorpora el bagaje inicial a la información contenida en  $y_1$ . El primer factor es la densidad atribuida a  $y_2$  cuando  $y_1$  fue observado y  $\theta$  es el valor del parámetro.

Ahora si  $y_2$  es independiente de  $y_1$  dado  $\theta$ , entonces

$$p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \quad (1.3.6)$$

Esto es, que bajo la hipótesis de independencia condicional la distribución *a posteriori* para  $\theta$  dado  $y_1$  e  $y_2$  puede ser encontrada tratando la distribución *a posteriori* de  $\theta$  dado  $y_1$  como la distribución *a priori* antes de observar  $y_2$ .

Es decir, la expresión (1.3.6) es precisamente de la misma forma que la expresión (1.3.5), excepto que  $p(\theta|y_1)$ , que es la distribución *a posteriori* de  $\theta$  dado  $y_1$ , juega el rol de la distribución *a priori* para el segundo conjunto de observaciones.

Obviamente este proceso puede ser repetido un número finito de veces.

En particular, si se tiene  $n$  observaciones independientes, la distribución *a posteriori* puede ser recalculada después de cada nueva observación, hasta el  $m$  estado de la *verosimilitud* asociado con la observación  $m$ -ésima, es combinada con la distribución *a posteriori* de  $\theta$  después de  $m-1$  observaciones dada la nueva distribución *a posteriori*, es decir:

$$p(\theta|y_1, \dots, y_n) \propto p(\theta|y_1, \dots, y_{n-1})p(y_n|\theta) \quad \text{con } m = 2, \dots, n \text{ donde } p(\theta|y_1) \propto p(\theta)p(y_1|\theta)$$

Así, el Teorema de *Bayes* describe, el proceso científico desde la experiencia, y muestra cuanto conocimiento sobre el estado natural representado por  $\theta$  es continuamente modificado por nuevos datos disponibles.

Luego de esta discusión, resumimos a los componentes del *proceso Bayesiano* en:

- 1- Suponer que el parámetro  $\theta$  es *aleatorio*.
- 2- Describir el comportamiento *aleatorio* de  $\theta$  mediante su distribución de densidad  $p(\cdot)$ , distribución *a priori*.
- 3- Muestra de datos.
- 4- Actualización del comportamiento de  $\theta$ , distribución *a posteriori*.

## 1.4 Distribución a priori Informativa

En la exposición anterior, se observa que uno de los componentes principales del *método Bayesiano* es la distribución *a priori*, que se le atribuye a los parámetros del modelo estadístico. Quizás, la cuestión más importante concerniente a la distribución *a priori* es: ¿Qué representa la probabilidad *a priori*?, y ¿Cuánto puede ser derivado de ella?. La forma de cómo encontrar una distribución *a priori* usando la información inicial entorno del problema es uno de los puntos que genera más polémica en el uso del *método Bayesiano*. Una de las causas puede ser que hemos desarrollado nuestro pensamiento en términos de variables observables, dejando de lado el significado real de hechos no tangibles.

Aquí, consideramos dos interpretaciones básicas que pueden ser dadas para distribuciones *a priori*.

Dentro de la interpretación de *población*, la distribución *a priori* representa una población de posibles valores del parámetro, desde el cual  $\theta$ , de interés actual, puede ser muestreado. Y dentro de la interpretación del estado de conocimiento más *subjetivo*, la guía principal es que debemos expresar nuestro conocimiento e incertidumbre sobre  $\theta$  como si su valor podría ser estimado como una realización aleatoria desde la distribución *a priori*. (Box y Tiao, 1992)

Según Gelman *et al* (1998), para algunos problemas, tal como estimar la probabilidad de fallos en un nuevo proceso industrial, no hay perfecta adecuación de poblaciones de  $\theta$ 's desde las cuales puede ser muestreado, excepto en una contemplación hipotética. Típicamente, la distribución *a priori* debería incluir todos los posibles valores de  $\theta$ , pero la distribución necesaria no es realísticamente agrupada entorno del verdadero valor, porque a menudo la información sobre  $\theta$  contenida en la muestra puede ser altamente superada por alguna especificación de probabilidad *a priori* razonable.

Para abordar el problema de la construcción de distribuciones *a priori* y la justificación de tal especificación, tomaremos el simple modelo binomial<sup>2</sup>.

---

<sup>2</sup> Modelo binomial:  $p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$ ,  $p(\theta|y) \propto \theta^y (1-\theta)^{n-y}$ , donde  $\theta$  representa la proporción de sucesos en la población o equivalentemente la probabilidad del suceso en cada tirada

Este modelo tiene como objetivo, estimar una proporción de la población desconocida a partir de los resultados de una secuencia de ensayos de Bernoulli; esto es, cada uno de los datos,  $y_1, \dots, y_n$ , puede ser 0 o 1.

Según Gelman *et al* (1998), las primeras contribuciones de importancia, sobre la *a priori* razonable para el modelo binomial se dan al final del siglo XVII y principios del XVIII., cuando los estudiosos de aquel tiempo se concentraron en la cuestión de los predatos. Es decir: Dado  $\theta$ , se preguntaron ¿Cuáles son las probabilidades de varios posibles resultados de la variables aleatoria  $y$ ? En un intento por contestarla daremos un resumen que nos aporta éste mismo autor:

En su famoso escrito de 1764, Bayes buscó, la probabilidad,  $\Pr(\theta \in (\theta_1, \theta_2) | y)$ , su solución fue basada en una analogía física de un espacio de probabilidad como una mesa rectangular (tal como una mesa de billar):

- 1- Distribución *a priori*: Una bola  $W$  es tirada aleatoriamente (según una distribución uniforme sobre la mesa). La posición horizontal de la bola en la mesa es  $\theta$ .
- 2- *Verosimilitud*: Una bola  $O$  es tirada aleatoriamente  $n$  veces. El valor de  $y$  es el número de veces que  $O$  arribó al lado derecho de  $W$ .

Es decir,  $\theta$  es asumida como que tiene una distribución *a priori* uniforme sobre  $[0,1]$ .

Usando reglas elementales de la teoría de probabilidad, Bayes, entonces obtiene:

$$\begin{aligned} \Pr(\theta \in (\theta_1, \theta_2) | y) &= \frac{\Pr(\theta \in (\theta_1, \theta_2), y)}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} p(y|\theta)p(\theta)d\theta}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta}{p(y)} \end{aligned} \quad (1.4.1)$$



Bayes subsiguientemente en la evaluación del denominador, muestra que:

$$\begin{aligned}
 p(y) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta & (1.4.2) \\
 &= \frac{1}{n+1} \quad \text{para } y = 0, \dots, n
 \end{aligned}$$

Consecuentemente, todos los valores posibles de  $y$  son igualmente probables *a priori*.

El numerador de (1.4.1) es una integral Beta incompleta sin una expresión de forma cerrada para valores grandes de  $(n-y)$  e  $y$ , un factor que aparentemente presentaba algunas dificultades para Bayes. Laplace, sin embargo, independientemente del Teorema descubierto por Bayes, y desarrollando analíticamente nuevas herramientas para calcular integrales; extendió la función  $\theta^y (1-\theta)^{n-y}$  alrededor de su máximo como  $\theta = \frac{y}{n}$  y evaluó la integral Beta incompleta usando lo que ahora conocemos como la aproximación Normal.

Laplace, al analizar el modelo binomial usó también la distribución *a priori uniforme*. Su primera aplicación fue para estimar la proporción de nacimientos femeninos en una población: Un total de 241945 mujeres y 251527 hombres nacidos en París entre 1745 a 1770. Siendo  $\theta$  la probabilidad de cualquier nacimiento femenino, Laplace mostró que:

$$\Pr(\theta \geq 0.5 | y = 241925, n = 251527 + 241945) \approx 1.15 \times 10^{-42},$$

y así demostró lo que era “moralmente cierto” (Gelman *et al* 1998), que  $\theta < 0.5$ .

La justificación de Bayes de estas distribuciones *a priori uniformes*, están basadas sobre observaciones. Gelman *et al* (1998), argumenta que tales justificaciones son apelables, dado que las mismas están expresadas en términos de cantidades observables “ $y$ ” y “ $n$ ”. Y considera que, la racionalidad de Laplace para densidades *a priori uniformes*, ha sido menos clara. Pero es a Laplace a quien se le atribuye el llamado “principio de razón insuficiente”, el cual dice que si algo no es conocido sobre  $\theta$ , entonces, la especificación *uniforme* es apropiada.

Luego, en la construcción de distribuciones *a priori* se debe tener en cuenta no sólo el significado de la misma, es decir asignar distribuciones *a priori* que reflejen



información substantiva sobre  $\theta$ , sino también que ésta permita obtener expresiones *a posteriori* analíticamente admisibles, es decir que sea función de densidad.

#### 1.4.1 Prioris Conjugadas

La definición formal es la siguiente: Si  $F$  es una clase de distribuciones muestrales (verosimilitud),  $p(y|\theta)$ , y  $P$  es una clase de distribuciones *a priori* de  $\theta$ , entonces la clase  $P$  es *conjugada* con respecto a  $F$  si

$$p(\theta|y) \in P \text{ para todo } p(\cdot|\theta) \in F \text{ y } p(\cdot) \in P.$$

En otras palabras si las distribuciones *a priori* y *a posteriori* pertenecen a la misma familia de distribuciones; la distribución *a priori* se llama *conjugada a priori*.

Esta definición puede resultar imprecisa desde que si  $P$  es escogida como la clase de todas las distribuciones, entonces  $P$  será siempre *conjugada* no importando que clase de distribuciones de muestreo sea usada. El interés fundamental aquí, sin embargo, estará en las familias de distribuciones *a priori conjugadas naturales*, las cuales se definen al tomar a  $P$  como el conjunto de todas las *densidades* que tienen la misma forma funcional que la *verosimilitud*.

Una densidad *a priori conjugada natural* es altamente conveniente ya que su aplicación resulta en una densidad *a posteriori* de la misma forma. Es decir, es una información muy flexible.

La ventaja de la *conjugada natural a priori* es que siempre que observemos una nueva muestra, la revisión de informaciones sobre  $\theta$  puede llevarse a cabo mediante el mismo procedimiento analítico. Esto es, si la verosimilitud,  $p(\cdot|y)$  es de la misma forma que  $p(\cdot)$  (distribución *a priori*), se la puede usar como la función *a priori* para analizar la nueva evidencia muestral, y las fórmulas obtenidas anteriormente se pueden usar para obtener la nueva *a posteriori*.

**Para afianzar ideas, consideremos el ejemplo de la distribución binomial (Box y Tiao (1992):**

La distribución binomial provee un modelo de información proveniente de una secuencia de  $n$  ensayos o extracciones intercambiables obtenidas desde una gran

población donde cada ensayo puede tomar uno de dos posibles resultados, convencionalmente llamados “éxito” y “fracaso”. Debido a la intercambiabilidad, los datos pueden ser resumidos por el número total de éxitos en los  $n$  ensayos, al cual se denotará por  $x$ . Se puede pasar de una formulación basada en términos de ensayos intercambiables a una basada en variables aleatorias independientes e idénticamente distribuidas definiendo al parámetro  $\theta$  como la proporción de éxitos en la población o, equivalentemente, como la probabilidad de éxito en cada ensayo.

El modelo muestral binomial queda definido por:

$$p(x, n|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \left( \frac{n!}{(n-x)!x!} \right) \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

$$\propto \theta^x (1-\theta)^{n-x}$$

Si la distribución *a priori* para  $\theta$  es de la forma:

$$p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (0 \leq \theta \leq 1)$$

Esto es, una distribución *beta*,  $\theta \sim \text{Beta}(\alpha, \beta)$  con una función de densidad de probabilidad:

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}; \quad \text{con } \theta \in [0, 1]$$

Entonces, la distribución *a posteriori*:

$$p(\theta|x) = \frac{p(x, n|\theta)p(\theta)}{\int_0^1 p(x, n|\theta)p(\theta)d\theta}$$

$$= \frac{\frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\frac{n!}{(n-x)!x!} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1} d\theta}$$

$$= \frac{\theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1}}{\int_0^1 \theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1} d\theta}$$

$$\text{como } \int \theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1} d\theta = \text{Beta}(x+\alpha, n+\beta-x)$$

Luego:

$$p(\theta|x) = \frac{\theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1}}{\text{Beta}(x+\alpha, n+\beta-x)} \propto \theta^{x+\alpha-1} (1-\theta)^{n+\beta-x-1}, \text{ con } 0 \leq \theta \leq 1$$

esto es,

$$\theta|x \sim \text{Beta}(x+\alpha, n+\beta-x)$$

Que es una densidad beta con parámetros  $(x+\alpha)$  y  $(n+\beta-x)$ . Luego, la familia *conjugada* para la distribución binomial es la familia de distribuciones *beta*.

Si bien es deseable tratar con familias conjugadas, no siempre es posible aproximar la información *a priori* por un miembro de estas familias.

## 1.5 Distribuciones a priori no informativas

Cuando carecemos de información *a priori* sobre  $\theta$ , la construcción de distribuciones *a priori* resulta dificultosa, estas distribuciones deben ser tales que garanticen un rol o efecto mínimo en la distribución *a posteriori*. Tales distribuciones a menudo son llamadas “*distribuciones a priori referenciales*” y la densidad *a priori* es descripta como vaga, difusa o *no informativa*.

La razón para utilizar distribuciones *a priori no informativas* es frecuentemente explicada como la intención de “*dejar que los datos hablen por sí mismos*”, de modo que las inferencias no estén afectadas por información externa a los datos. (Gelman *et al* 1998)

### 1.5.1 Distribuciones a priori propias e impropias

Una propiedad básica de una función de densidad de probabilidad,  $p(\theta)$ , es que la integral o suma sobre el rango de los parámetros sea 1. Es decir:

$$\left. \begin{array}{l} \int p(\theta) d\theta \\ \sum p(\theta) \end{array} \right\} = 1 \left\{ \begin{array}{l} \theta \text{ es continua} \\ \theta \text{ es discreta} \end{array} \right.$$

Si por ejemplo, tomamos a  $p(\theta) = k$ ,  $-\infty < \theta < \infty$  (1.5.1), para cualquier  $k$  posiblemente no represente una densidad de probabilidad, por ejemplo si tomamos  $k = 0$ .

Sin embargo, en algunas ocasiones es útil extender el concepto de densidad de probabilidad, para los casos en donde  $\int_{-\infty}^{\infty} p(\theta) d\theta = \infty$ , las cuales son llamadas *distribuciones impropias*.

La densidad  $p(\theta) = k$  es considerada como la representación de una densidad normal de variancia infinita. Otro ejemplo de distribución impropia es:

$$p(\theta) = k/\theta, \quad 0 < \theta < \infty \quad (1.5.2).$$

A veces resulta que, cuando tenemos una densidad *a priori impropia*, ésta puede combinarse con una verosimilitud común y obtener como resultado una densidad *a posteriori propia*. Así, si usamos la distribución uniforme sobre toda la recta real  $p(\theta) = k$  para  $k \neq 0$ , y la combinamos con la verosimilitud normal obtenemos la verosimilitud estandarizada o normalizada como densidad *a posteriori*; esto dice que el rasgo dominante de la posteriori es la verosimilitud. En síntesis, debemos pensar en una densidad impropia como una aproximación que es válida para algún rango grande de valores pero no será considerada como verdaderamente válida a lo largo de todo su rango.

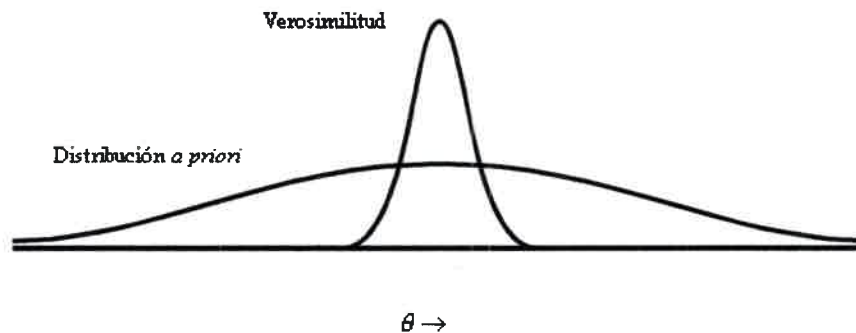
Las funciones de densidad del tipo (1.5.1) y (1.5.2) son frecuentemente empleadas para representar el comportamiento local de la distribución *a priori* en la región donde la *verosimilitud* es estimable pero no sobre todo el rango admisible.

Pero suponiendo que esas formas son una aproximación suficiente de la densidad *a priori* deducida desde las formas (1.5.1) y (1.5.2) solo sobre el rango en donde la *verosimilitud* es estimable y ser cero en el exterior de ese rango, esto asegura que la *a priori* actualmente usada es *propia*.

En general, llamamos densidad *a priori*  $p(\theta)$  *propia* si esta no depende de los datos y la integral es 1. (Si la integral de  $p(\theta)$  es un valor finito positivo, esta es llamada densidad *no normalizada* y puede normalizarse, esto es multiplicando por una constante obtenemos que la integral sea 1).

La distribución *a posteriori* obtenida desde una distribución *impropia a priori* debe ser interpretada con mucho cuidado, uno debe siempre chequear esta distribución *a posteriori* ver que tenga integral finita y que su forma sea sensible.

Su interpretación más razonable es una aproximación a situaciones donde las *verosimilitudes* dominan a la densidad *a priori*:



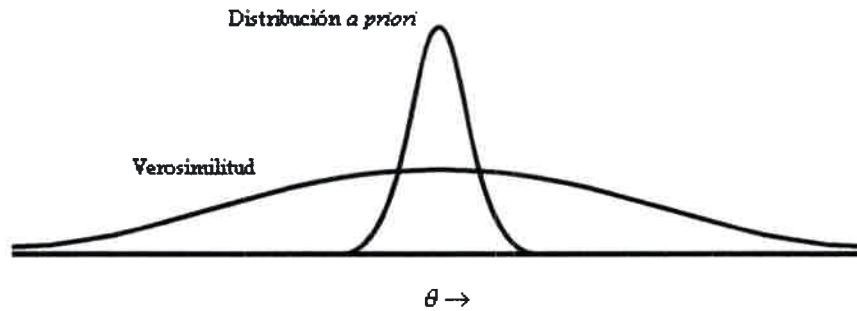
**Verosimilitud dominante** (a menudo apropiado para el análisis de datos científicos)

Cuando hay más observaciones que parámetros, la distribución *a priori no informativa* es satisfactoria, ésta tiene resultados aceptables y toma menos esfuerzo que especificar una conocida *a priori* desde la forma probabilística. (Gelman *et al* 1998)

Es decir que los resultados asintóticos nos permiten formalizar que, la noción de la importancia que tiene la distribución *a priori* disminuye cuando el tamaño de la muestra crece. Una consecuencia de este resultado es que en situaciones de análisis, con tamaño de muestra grande, no necesitamos investigar o trabajar duramente para formular una distribución *a priori* que refleje correctamente toda la información disponible. En otras palabras la *verosimilitud domina a la distribución a priori*.

En cambio cuando el tamaño de la muestra es pequeña, la distribución *a priori* es la parte crítica de la especificación del modelo.

Esto es, porque cuando el tamaño de la muestra es pequeño o el número de parámetros es grande, la verosimilitud es más suave:



*A priori dominante* (raramente apropiado para el análisis de datos científicos)

En estas situaciones Gelman *et al* (1998) aconseja trabajar con *modelos jerárquicos*<sup>3</sup>.

### 1.5.2 Principio de invarianza de Jeffreys

Una aproximación que es algunas veces usada para definir *distribuciones a priori no informativas* fue desarrollada por Jeffreys, quien se basó en considerar transformaciones uno a uno del parámetro:  $\phi = h(\theta)$ . Por transformación de variables, la densidad *a priori*  $p(\theta)$  es equivalente, en términos de expresar la misma información, a la siguiente densidad *a priori* para  $\phi$ :

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta) |h'(\theta)|^{-1} \quad (1.5.3)$$

El principio general de Jeffreys es que cualquier regla para determinar la densidad *a priori*,  $p(\theta)$ , debe conducir a un resultado equivalente si se aplica al parámetro transformado; esto es,  $p(\phi)$  calculado a partir de  $p(\theta)$  y aplicando la transformación

<sup>3</sup> Muchas aplicaciones estadísticas involucran parámetros múltiples que pueden considerarse como relacionados o conectados de alguna manera por la estructura del problema, implicando que el modelo de probabilidad conjunta para estos parámetros debe reflejar la dependencia entre ellos. Este tipo de modelos surgen frecuentemente, en aplicaciones como: (i) En el estudio de datos longitudinales, incluyendo el análisis de medidas repetidas y de curvas de crecimiento, en los que cada individuo es observado en diferentes puntos en el tiempo. (ii) En áreas como las ciencias de la salud y de la educación, donde es común que los datos se recaben por estratos. Por ejemplo, pueden observarse pacientes dentro de hospitales (o estudiantes dentro de escuelas), los cuales a su vez se toman como representativos de ciertas zonas de la ciudad o de la región. En estos casos no es raro que además se cuente con información adicional en forma de covariables tanto a nivel de pacientes como de hospitales. Ver Gelman *et al* 1998, Capítulo 5 pág 119.



anterior (1.5.3) debe coincidir con la distribución que sería obtenida determinando  $p(\phi)$  directamente usando el modelo transformado:  $p(y, \phi) = p(\theta)p(y|\theta)$ .

La selección de Jeffreys para una densidad *a priori no informativa* es:

$$p(\theta) \propto [J(\theta)]^{1/2}$$

donde  $J(\theta)$  es la *información de Fisher* para  $\theta$ :

$$J(\theta) = E \left[ \left( \frac{d \log p(y|\theta)}{d\theta} \right)^2 \middle| \theta \right] = -E \left[ \frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta \right] \quad (1.5.4)$$

Para verificar que el modelo a priori de Jeffreys es invariante a la reparametrización, evaluamos  $J(\phi)$  como  $\theta = h^{-1}(\phi)$ :

$$\begin{aligned} J(\phi) &= -E \left[ \frac{d^2 \log p(y|\phi)}{d\phi^2} \right] \\ &= -E \left[ \frac{d^2 \log p(y|\theta = h^{-1}(\phi))}{d\theta^2} \left| \frac{d\theta}{d\phi} \right|^2 \right] \\ &= J(\theta) \left| \frac{d\theta}{d\phi} \right|^2 \end{aligned}$$

Así,  $J(\phi)^{1/2} = J(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|$ , tal como es requerido.

El principio de Jeffreys puede ser extendido a modelos multiparamétricos, pero los resultados son más controvertidos. Aproximaciones simples basadas al asumir *distribuciones a priori no informativas* para los componentes de un vector de parámetros  $\theta$  pueden dar diferentes resultados que los obtenidos con el principio de Jeffreys.

### 1.5.3 Dificultades con distribuciones no informativas

La búsqueda de distribuciones *a priori no informativa* tiene serios problemas, algunos de ellos son, según Box y Tiao (1992):

- 1- Buscar siempre tales distribuciones puede conducir a conclusiones erróneas: Si la *verosimilitud* es verdaderamente dominante dentro del problema, entonces la elección entre un rango de densidades *a priori* relativamente llanas o difusas no puede tener importancia. Esto parece animar al establecimiento de tales distribuciones *a priori* pero, se debe tener cuidado ya que su uso puede ser inapropiado.
- 2- Para muchos problemas, esta elección no es clara ya que buscar una distribución *a priori impropia* para una densidad que sea difusa o uniforme en una parametrización, puede no ser en otra. Esto es la dificultad esencial con el principio de *Laplace* de razón insuficiente: ¿sobre que escala debería aplicarse el principio?. Por ejemplo, la razonable densidad *a priori* en la normal con media  $\theta$ , es uniforme, mientras que para  $\sigma^2$ , la densidad  $p(\sigma^2) \propto \frac{1}{\sigma^2}$  parece razonable.

Sin embargo, si nosotros definimos  $\phi = \log \sigma^2$ , entonces la densidad *a priori* sobre  $\phi$  es  $p(\phi) = p(\sigma^2) \left| \frac{d\sigma^2}{d\phi} \right| \propto \frac{1}{\sigma^2} \sigma^2 = 1$ , esto es, uniforme sobre  $\phi = \log \sigma^2$ .

Con distribuciones discretas se tiene la misma dificultad de decisión, como la subdivisión de las respuestas en “átomos” con igual probabilidad.

En casi todos los problemas reales, el análisis de los datos tienen más información que podría ser convenientemente incluida en el modelo estadístico. Esto es un tema para la *verosimilitud* así como lo mismo que la distribución *a priori*.

En la práctica, esto es siempre zanjado por numerosas razones:

- 1- Para describir el modelo más conveniente, por que esto puede deberse a la dificultad de expresar el conocimiento correctamente en forma probabilística.

- 2- Para simplificar cálculos.
- 3- o quizás para evitar usar una posible fuente de información no confiable.

Excepto por esta última razón, todos estos argumentos son convenientes y son mejores justificados por el reclamo, "la respuesta no tendría muchos cambios si nosotros hubiéramos sido más exactos".

## 1.6 Predicción.

Para hacer inferencias sobre una observación desconocida, a menudo llamada *inferencia predictiva*, seguimos los siguientes pasos.

Antes que los datos  $y$  sean observados, la distribución de los "y" desconocidos pero observables es:

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta) p(y|\theta) d\theta.$$

Esta distribución marginal de  $y$ , es llamada *distribución predictiva a priori*:

*a priori* debido a que no está condicionada por ninguna observación previa, y *predictiva* porque es la distribución de una cantidad que es observable.

Después que los datos  $y$  han sido observados, se pueden predecir nuevos valores,  $\tilde{y}$ , desde el mismo proceso.

La distribución de  $\tilde{y}$  es llamada la *distribución predictiva a posteriori*, debido a que es condicionada por los datos ya observados  $y$ :

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) d\theta \end{aligned}$$

La segunda y tercera línea muestra la distribución predictiva *a posteriori* como un promedio de la predicción condicional sobre la distribución a posteriori de  $\theta$ .

La última ecuación se deduce porque  $y$  e  $\tilde{y}$  son condicionalmente independientes, dado  $\theta$  en este modelo.

Esta distribución se usa a menudo en métodos empíricos *Bayesianos* los cuales se pueden ver en sección 7.8 de Peter M. Lee (1997).

El nombre de *distribución predictiva* se debe a que ésta representa las actuales predicciones del valor de  $y$  conteniendo ambas incertidumbres, a saber, la del valor de  $\theta$  y la incertidumbre residual sobre  $y$  cuando  $\theta$  es conocida..

El valor de usar la *distribución predictiva* reside en que estamos verificando los supuestos subyacentes. Si, por ejemplo, el cambio de  $p(y)$  fuera pequeño (en algún sentido) para el valor observado de  $y$ , esto puede sugerir que la forma de la verosimilitud es sospechosa. Algunos estudiosos sugieren la revisión en la distribución *a priori*, diciendo que esta es la sospechosa, aunque las dificultades sobre  $p(\theta)$  son lógicas, ya que simplemente representan nuestras creencias *a priori*.

### 1.6.1 Odds y Razón de Verosimilitud

La razón de la densidad de  $\theta$  evaluada en dos puntos  $\theta_1$  y  $\theta_2$  es llamada *odds*:

$$\text{Odds a priori} = \frac{p(\theta_1)}{p(\theta_2)}$$

$$\text{Odds a posteriori} = \frac{p(\theta_1|y)}{p(\theta_2|y)}$$

Los *odds* son de gran aplicación en el caso discreto en donde  $\theta_1$  y  $\theta_2$  son dos eventos complementarios (por ejemplo los eventos sano y enfermo, con tratamiento y sin tratamiento, vivo y muerto, etc.), ya que permiten evaluar cuan más probable es un evento que el otro.

Note que el *odds a posteriori* puede escribirse como:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)}$$

por lo que el *odds a posteriori* es igual al *odds a priori* multiplicado por la *razón de verosimilitud*. A esta *razón de verosimilitud* se le conoce como el *Factor de Bayes*. La importancia del *Factor de Bayes* es que es una medida de la información contenida exclusivamente en los datos.

Los *odds* y el *Factor de Bayes* son utilizados en *inferencia Bayesiana* para contrastar dos modelos de probabilidad (pruebas de hipótesis sobre los parámetros).

## 1.7 Diferencias entre la Inferencia Bayesiana y la Clásica

De acuerdo con O'Hagan (1994), se pueden identificar cuatro aspectos fundamentales que diferencian la inferencia *Bayesiana* a la inferencia Clásica:

- **Información *a priori*.** Todos los problemas son únicos y tienen su propio contexto. De tal contexto se deriva información *a priori*, y es la formulación y uso de esta información *a priori* la que diferencia la inferencia *Bayesiana* de la estadística Clásica.
- **Probabilidad Subjetiva.** La estadística *Bayesiana* formaliza la noción de que todas las probabilidades son *subjetivas*, dependiendo de las creencias individuales y la información disponible. Así, el análisis *Bayesiano* resulta personal, único de acuerdo con las creencias individuales de cada uno.
- **Auto consistente.** Al tratar al parámetro  $\theta$  como aleatorio, la inferencia *Bayesiana* se basa completamente en la teoría de Probabilidad. Esto tiene muchas ventajas y significa que toda inferencia puede ser tratada en términos de declaraciones probabilísticas para  $\theta$ .
- **No “adhockery”.** Debido a que la inferencia clásica no puede hacer declaraciones probabilísticas acerca de  $\theta$ , varios criterios son desarrollados para juzgar si un estimador particular es en algún sentido “bueno”. Estos criterios utilizados por la estadística Clásica son llamados “adhockery” por De Finetti y han conducido a una proliferación de procedimientos, frecuentemente en conflicto unos con otros. La inferencia *Bayesiana* deja de lado esta tendencia a inventar criterios ad hoc para juzgar y comparar estimadores al basarse exclusivamente en la distribución *a posteriori* para expresar en términos exclusivamente probabilísticos toda inferencia referente al parámetro.

## 1.8 CONCLUSIONES

La distinción fundamental entre la inferencia clásica o Gausiana y la *Bayesiana* es que en ésta, los parámetros de interés son *variables aleatorias*, luego, tiene tanto *a priori* como *a posteriori distribuciones*.

En la inferencia clásica los parámetros tienen valores únicos, aunque estos valores son desconocidos, y esto no permite tratarlos como aleatorios o como probabilidades asociadas.

Además, el método *Bayesiano* necesariamente adopta una interpretación de probabilidad menos restrictiva y más intuitiva, que en la estadística frecuencial, como se ve en la sección 1.1

El método *Bayesiano*, brevemente comprende los siguientes pasos principales:

- 1- **Verosimilitud**: Obtener la función de verosimilitud,  $p(y|\theta)$ . Este paso describe simplemente la importancia que el proceso da a la muestra,  $y$ , en términos del parámetro desconocido  $\theta$ .
- 2- **A priori**: Obtener la densidad a priori,  $p(\theta)$ . La distribución *a priori* expresa el conocimiento sobre  $\theta$  antes de haber observado la muestra,  $y$ .
- 3- **A posteriori**: Aplicando el *Teorema de Bayes* para derivar la densidad a posteriori,  $p(\theta|y)$ . Ésta, ahora expresa cuál es el conocimiento sobre  $\theta$  después de haber observado la muestra,  $y$ .
- 4- **Inferencia**: Las inferencias apropiadas se derivan desde la distribución *a posteriori*. Generalmente toda la información está expresada a través de la distribución *a posteriori*, y puede incluir inferencias específicas tales como estimadores puntuales, estimación por intervalos de probabilidad o hipótesis de probabilidad.

La principal objeción a la inferencia *Bayesiana*, es que las conclusiones dependen de la selección específica de la distribución *a priori*. Aunque para otros esto es lo



interesante de la inferencia *Bayesiana*, este es un debate aún no cerrado. Sin embargo, antes de dejar esta característica, se debe señalar que inclusive en inferencia clásica, y además en investigación científica en general, estos conocimientos *a priori* son utilizados implícitamente. Así por ejemplo, el conocimiento *a priori* es utilizado para formular un modelo de verosimilitud apropiado. En pruebas de hipótesis, las creencias *a priori* acerca de la plausibilidad de una hipótesis son frecuentemente utilizadas para ajustar el nivel de significancia de la prueba. Así, si se cree que los datos pueden conducir al rechazo de la hipótesis, esto se puede ajustar escogiendo un nivel de significancia bastante alto. En este sentido entonces, la inferencia *Bayesiana* formaliza la incorporación de la información *a priori*, la cual es incorporada implícitamente en el análisis clásico.

---

## Capítulo II

# Inferencia Bayesiana en la toma de decisiones

---

### Introducción

---

En este capítulo trataremos la inferencia *Bayesiana* a través de temas tales como teoría de la *Decisión*, estimación puntual, por intervalos y pruebas de Hipótesis.

Las discusiones, aquí presentadas, girarán entorno a los conexiones entre los métodos *Bayesianos* y no *Bayesianos*. Sin embargo, se debe tener en cuenta que estos temas, salvo teoría de la *Decisión*, no tienen la trascendencia que sí tienen en la estadística Clásica o Gausiana. Para algunos autores, temas como la estimación por intervalos y las pruebas o Test de Hipótesis carecen de importancia como temas de inferencia ya que en estadística *Bayesiana* la *distribución a posteriori* “*es*” la inferencia.

El desarrollo de los temas mencionados anteriormente dentro de la estadística *Bayesiana*, en realidad obedece más que todo a la necesidad de presentar metodologías y resultados análogos a los de la estadística Clásica. Ya que, desde el punto de vista *Bayesiano*, todas estas estimaciones, son realmente resúmenes de información acerca del parámetro  $\theta$ .

El objetivo de éste Capítulo es exponer el marco teórico de la estimación puntual, por intervalos y test de hipótesis de los parámetros involucrados en la teoría de la *Decisión*.



## 2.1 Teoría de la decisión.

El origen de la teoría estadística de las *decisiones* es relativamente reciente, pues se remonta a los primeros años de la década de 1950. Desde entonces esta rama de la estadística ha ganado rápidamente popularidad, lo que ha determinado un desarrollo cada vez mayor de la correspondiente teoría y sus aplicaciones.

La teoría de la *decisión* es un área de suma importancia en estadística ya que muchos problemas del mundo real pueden tomar la forma de un problema de decisión, por ejemplo:

- ¿A quién votar en las próximas elecciones?
- ¿Debo aceptar este trabajo o esperar por si aparece una oferta mejor?,
- Un gerente de una empresa debe decidir si agrega una o dos nuevas líneas de producción.

Son preocupaciones reales que se transforman en problemas que estudia la teoría de las *decisiones*.

Por otro lado, toda la inferencia estadística puede ser entendida como un *proceso de decisión*. Por ejemplo, luego de observar un conjunto de datos, la pregunta que usualmente se formula es:

- ¿por qué valor debo decidir para estimar el parámetro?.

Existen muchos enfoques en teoría de la *decisión*, pero tal vez, el más coherente sea el enfoque *Bayesiano*. Éste es el único que considera toda la información disponible para la toma de *decisiones* (Box & Tiao, 1992) y este argumento es frecuentemente usado para privilegiar el uso de los métodos *Bayesianos* sobre los clásicos.

Los elementos necesarios para plantear *un problema de decisión* son los siguientes:

1. Un espacio paramétrico  $\Theta$  que describa todos los posibles *estados de la naturaleza* (los ambientes o escenarios económicos mutuamente

excluyentes y colectivamente exhaustivo<sup>1</sup> son los llamados *estados de la naturaleza*)

2. Un conjunto de acciones,  $A$ , sobre las cuales tomar la decisión.
3. Una *función de pérdida*  $L$ , donde  $L(\theta, a)$  es la pérdida en la que se incurre al tomar la decisión  $a \in A$ , cuando el verdadero estado de la naturaleza es  $\theta$ . Es decir, debe especificarse una función de pérdida  $L$  que a cada par  $(\theta, a)$  asocie la pérdida  $L(\theta, a)$ .
4. Y una *distribución a posteriori*  $p(\theta|y)$  que contenga toda la información disponible sobre el parámetro.

Dada la distribución *a posteriori*, la idea es optar por la acción que minimice la *pérdida esperada a posteriori* (o maximice la ganancia), esto es, que minimice la siguiente expresión:

$$\rho(a, y) = \int L(\theta, a) p(\theta|y) d\theta$$

En esta expresión,  $\rho(a, y)$  es la pérdida en la que se espera incurrir al tomar la acción  $a \in A$ , habiendo observado el valor  $y$ .

Distintos valores de “ $y$ ” determinarán distintos valores de “ $a$ ” que minimicen  $\rho(a, y)$ . El conjunto de estos valores, denotados por  $d(y)$ , constituyen la *regla de decisión de Bayes*.

Por último, dada una regla de Bayes  $d(y)$ , se puede calcular el riesgo asociado a este conjunto de acciones promediando las pérdidas esperadas posteriores correspondientes, ponderadas por las probabilidades de “ $y$ ”, y así obtener el *riesgo de Bayes*:

$$RB(d) = \int \rho(d(y), y) p(y) dy$$

<sup>1</sup> **Mutuamente excluyentes** significa que la naturaleza sólo puede estar en un estado a la vez en cualquier momento.

**Colectivamente exhaustivo** quiere decir que la lista de estados está completa. Siempre se encontrará en uno y solamente en un estado.

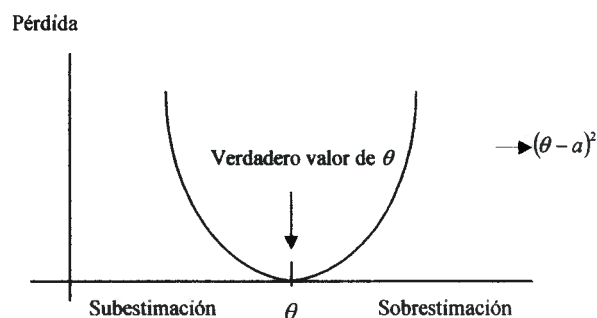
## 2.2 Estimación Puntual.

La *distribución a posteriori* ofrece un resumen completo de la inferencia sobre el parámetro  $\theta$ . En cierto modo, la *distribución a posteriori* es la inferencia. Sin embargo, en algunos casos es deseable resumir esta información de alguna manera. En particular, uno podría querer dar un simple valor estimado del parámetro desconocido (en vez de una función de densidad). La pregunta que surge inmediatamente es, **¿qué valor elegir como estimador de  $\theta$ ?** Desde este punto de vista, el problema de estimación se convierte en un problema de *decisión*, y para resolverlo se especificará una función  $L(\theta, a)$  que mida la pérdida percibida en estimar  $\theta$  por  $a$ . Así, para determinar un estimador *Bayesiano*, debe especificarse una función de *pérdida*, la cual no es una tarea fácil, ya que las consecuencias económicas no son fácilmente medibles. La selección de ésta dependerá del contexto.

**Ilustremos estas ideas, considerando distintas funciones de pérdida:**

### 2.2.1 Función de pérdida cuadrática

Sea  $\theta$  escalar y  $a$  la decisión. La función de pérdida cuadrática tiene como expresión algebraica a:  $L(\theta, a) = (\theta - a)^2$  y su representación gráfica es:



Bajo esta función de pérdida, la *pérdida esperada a posteriori* será:

$$\begin{aligned}
 \rho(a, y) &= \int L(\theta, a) p(\theta|y) d\theta \\
 &= \int (\theta - a)^2 p(\theta|y) d\theta \\
 &= \int [\{\theta - E(\theta|y)\} + \{E(\theta|y) - a\}]^2 p(\theta|y) d\theta \\
 &= \int \{\theta - E(\theta|y)\}^2 p(\theta|y) d\theta + 2 \int \{\theta - E(\theta|y)\} \{E(\theta|y) - a\} p(\theta|y) d\theta \\
 &\quad + \int \{E(\theta|y) - a\}^2 p(\theta|y) d\theta \\
 &= \text{Var}(\theta|y) + \{E(\theta|y) - a\}^2
 \end{aligned}$$

la cual es minimizada cuando  $a = E(\theta|y)$ . Por lo tanto, el estimador *Bayes* de  $\theta$  bajo una función de pérdida cuadrática es la media *a posteriori*, y la pérdida de este estimador es la variancia *a posteriori*

### 2.2.2 Función de pérdida valor absoluto

Esta función de pérdida es definida por:  $L(\theta, a) = |\theta - a|$ , en tal caso:

$$\rho(a, y) = \int |\theta - a| p(\theta|y) d\theta$$

Y el estimador de *Bayes* de  $\theta$  es la mediana *a posteriori*. Berger (1988) presenta la demostración de este resultado.

### 2.2.3 Función de pérdida lineal

Esta función de pérdida es definida por:

$$L(\theta, a) = \begin{cases} k_0(\theta - a) & \text{si } \theta \geq a \\ k_1(a - \theta) & \text{si } \theta < a \end{cases}$$

para  $k_0$  y  $k_1 > 0$ . El estimador de *Bayes* de  $\theta$  es el cuantil  $\frac{k_0}{k_0 + k_1}$  de la distribución *a posteriori*. La demostración de este resultado es similar a la de la función de pérdida valor absoluto. De hecho notar que, la función de *pérdida valor absoluto* es el caso particular de la función de *pérdida lineal* cuando  $k_0 = k_1 = 1$

### 2.2.4 Función de pérdida 0 - 1

Esta función de pérdida es definida por:

$$L(\theta, a) = \begin{cases} 0 & \text{si } |\theta - a| \leq \varepsilon \\ 1 & \text{si } |\theta - a| > \varepsilon \end{cases}$$

luego resulta que:

$$\rho(a, y) = P(|\theta - a| > \varepsilon | y) = 1 - P(|\theta - a| \leq \varepsilon | y)$$

Claramente el estimador de *Bayes* de  $\theta$  es la *moda a posteriori*



## 2.3 Estimación por intervalos: Intervalos de Credibilidad

El objetivo de los intervalos de *credibilidad* es dar un resultado análogo al de los intervalos de confianza clásicos y su justificación radica en que los estimadores puntuales no dan una medida de la exactitud de la estimación.

Los intervalos de *credibilidad* pueden ser interpretados como los intervalos dentro de los cuales se encuentra el parámetro con cierta probabilidad. Esta interpretación, en los intervalos clásicos no es posible y representa un problema en dichos intervalos, en los que el parámetro es considerado un valor constante (en este caso, la interpretación es frecuencial).

La definición rigurosa de estos intervalos es la siguiente:

Sea  $p(\theta|y)$  la función *a posteriori* de  $\theta$  condicionada sobre el resultado muestral  $y = (y_1, y_2, \dots, y_n)$ , sean  $a$  y  $b$  límites tales que:

$$\text{Credibilidad de } [a, b] = P(a < \theta < b | y) = \int_a^b p(\theta|y) d\theta = F(b|y) - F(a|y) = 1 - \alpha$$

En donde  $a$  y  $b$  son funciones del resultado muestral  $y$ .

Entonces el intervalo  $(a, b)$  es un *intervalo Bayesiano* tal que la probabilidad de que  $\theta$  se encuentre contenido en  $(a, b)$  es  $1 - \alpha$ .

Una dificultad con los intervalos de *credibilidad Bayesianos* es que estos pueden ser definidos de varias maneras. Entre las alternativas existentes, se encuentran las *regiones de máxima densidad a posteriori* (HPD).

### 2.3.1 Regiones de Máxima Densidad (HPD)

Sea  $p(\theta|y)$  la función de densidad *a posteriori*. Una región  $R$  en el espacio paramétrico de  $\theta$  es llamada una región de magnitud  $(1 - \alpha)$  HPD si:

a)  $\Pr\{\theta \in R | y\} = 1 - \alpha$

b) para todo  $\theta_i \in R$  y  $\theta_j \notin R$ , entonces  $p(\theta_i | y) \geq p(\theta_j | y)$

Estas regiones tienen las siguientes propiedades que se deducen obviamente de su definición:

- 1- La región debe ser tal que la densidad de probabilidad de cada punto interior a esa región es mayor que cualquier punto fuera de la región.
- 2- La región debe ser tal que para una dada probabilidad, su volumen debe ser lo más pequeño en el espacio paramétrico. Es decir, si suponemos que  $p(\theta|y)$  es no uniforme sobre cada región en el espacio de  $\theta$ , entonces la región HPD de magnitud  $(1-\alpha)$  es única. Considerando que, si  $\theta_1$  y  $\theta_2$  son dos puntos tales que  $p(\theta_1|y) = p(\theta_2|y)$ , entonces estos dos puntos son incluidos simultáneamente en  $(1-\alpha)$  regiones HPD. La recíproca es siempre verdadera. Esto es, si  $p(\theta_1|y) \neq p(\theta_2|y)$ , entonces existe  $(1-\alpha)$  regiones HPD las cuales incluye un punto pero no el otro.

#### Observación:

A diferencia de los intervalos o regiones de confianza de la estadística clásica, un intervalo de *credibilidad Bayesiano* o región HPD es, en efecto, un intervalo de probabilidad. Es decir, que la probabilidad de que  $\theta$  se encuentre contenido en la región es  $1-\alpha$ , mientras que con las regiones de confianza sólo puede decirse que una cantidad de  $100(1-\alpha)\%$  de estos intervalos o regiones contendrá el valor real de  $\theta$ . En síntesis, decimos que  $1-\alpha$  es el nivel de *credibilidad* y  $\alpha$  es el nivel de *incredibilidad*.

Aunque, existe un problema con las regiones HPD, el cual es, que no son invariantes a las transformaciones 1 a 1, a no ser que la transformación sea lineal. Gelman *et al* (1998) propone como solución, intervalos centrales de probabilidad *a posteriori*, cuyos límites corresponden en el caso de un intervalo de probabilidad  $(1-\alpha)$  a los cuantiles  $\frac{\alpha}{2}$  y  $1-\frac{\alpha}{2}$  de la distribución *a posteriori*.

Por último, aún cuando lleguemos a resultados idénticos de aquellos que se derivan de la teoría Clásica, se debe considerar que, estos difieren considerablemente de los intervalos de probabilidad *Bayesiana*.

## 2.4 Test de Hipótesis

En pruebas de hipótesis clásicas, dada una hipótesis planteada  $H_0 : \theta \in \Theta_0$  y una alternativa  $H_1 : \theta \in \Theta_1$ , un procedimiento de prueba o test es evaluado en términos de las probabilidades de cometer un error de tipo I y de tipo II. En el método *Bayesiano*, será necesario calcular las probabilidades *a posteriori*  $\alpha_0 = P\{\theta \in \Theta_0 | y\}$  y  $\alpha_1 = P\{\theta \in \Theta_1 | y\}$  correspondientes a  $H_0$  y  $H_1$ , y tomar una decisión de acuerdo a estos valores.

Es decir, la inferencia *Bayesiana* básica sobre test de hipótesis reside en su probabilidad *a posteriori* en contraste con la inferencia clásica donde las hipótesis son rechazadas o no rechazadas.

Luego todo lo que necesitamos es calcular las probabilidades *a posteriori*:

$$\alpha_0 = P(\theta \in \Theta_0 | y) \quad \alpha_1 = P(\theta \in \Theta_1 | y)$$

donde  $\theta$  es el parámetro desconocido el cual es una realización de un conjunto  $\Theta$  y,  $y = (y_1, y_2, \dots, y_n)$  un conjunto de observaciones.

Y además necesitamos decidir entre  $H_0$  y  $H_1$ . (se debe notar que  $\alpha_0 + \alpha_1 = 1$  de la misma manera que en la teoría clásica  $\Theta = \Theta_0 \cup \Theta_1$  y  $\emptyset = \Theta_0 \cap \Theta_1$ ).

Aunque las probabilidades *a posteriori* de la hipótesis son nuestros últimos objetivos antes necesitamos encontrar las probabilidades *a priori*:

$$\pi_0 = P(\theta \in \Theta_0), \quad \pi_1 = P(\theta \in \Theta_1)$$

que denotan las probabilidades *a priori* para  $H_0$  y  $H_1$ , respectivamente

**Notar que,**  $\pi_0 + \pi_1 = 1$  simplemente como  $\alpha_0 + \alpha_1 = 1$ .

Es útil también, considerar los *odds*<sup>2</sup> *a priori* sobre  $H_0$  en contra de  $H_1$ , es decir

$$\frac{\pi_0}{\pi_1}$$

<sup>2</sup> Ver Capítulo I, sección 1.6.1

Y los *odds a posteriori* sobre  $H_0$  en contra de  $H_1$ , o sea

$$\frac{\alpha_0}{\alpha_1}$$

Obsérvese que si los *odds a priori* están cerca de 1 entonces consideramos a  $H_0$  como más o menos probable que  $H_1$  *a priori*, mientras que si la razón (ratio) es grande consideramos  $H_0$  como relativamente más probable y cuando esa razón es pequeña consideramos que es relativamente menos probable. Similares comentarios podríamos aplicar para la interpretación de los *odds a posteriori*.

También es útil definir el *Factor de Bayes*<sup>3</sup>, B, a favor de  $H_0$  en contra de  $H_1$  cómo:

$$B = \frac{\frac{\alpha_0}{\alpha_1}}{\frac{\pi_0}{\pi_1}} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0}$$

El interés en el *Factor de Bayes* es que este puede algunas veces ser interpretado como “*los odds*” a favor de  $H_0$  en contra de  $H_1$  que son dados por los datos. Esto es de valor, porque  $\frac{\alpha_0}{\alpha_1} = B \frac{\pi_0}{\pi_1}$  y  $\alpha_1 = 1 - \alpha_0$ , y podemos encontrar la probabilidad *a posteriori*  $\alpha_0$  de  $H_0$  desde su probabilidad *a priori* y el *Factor de Bayes*, en éste caso, es:

$$\alpha_0 = \frac{1}{[1 + (\pi_1 / \pi_0) B^{-1}]} = \frac{1}{[1 + \{(1 - \pi_0) / \pi_0\} B^{-1}]}$$

La interpretación anterior es claramente válida cuando las hipótesis son simples, esto es:

$$\Theta_0 = \{\theta_0\} \quad \text{y} \quad \Theta_1 = \{\theta_1\}$$

<sup>3</sup> Ver Capítulo I, sección 1.6.1

Para algún  $\theta_0$  y  $\theta_1$ . En ese caso, entonces  $\alpha_0 \propto \pi_0 p(x|\theta_0)$  y  $\alpha_1 \propto \pi_1 p(x|\theta_1)$  y entonces:

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 p(y|\theta_0)}{\pi_1 p(y|\theta_1)}$$

Y así el *Factor de Bayes* es

$$B = \frac{p(y|\theta_0)}{p(y|\theta_1)}$$

Se sigue que  $B$  es la razón de verosimilitud de  $H_0$  a  $H_1$ , para la mayoría de los estadísticos (sean *Bayesianos* o no) estos son los *odds* a favor de  $H_0$  en contra de  $H_1$  que son dados por los datos.

Aunque, la interpretación no es realmente simple cuando  $H_0$  y  $H_1$  son compuestas, es decir, cuando ellas contienen más de un elemento. En tal caso es conveniente escribir

$$\rho_0(\theta) = \frac{p(\theta)}{\pi_0} \quad \text{para} \quad \theta \in \Theta_0$$

y

$$\rho_1(\theta) = \frac{p(\theta)}{\pi_1} \quad \text{para} \quad \theta \in \Theta_1$$

Donde  $p(\theta)$  es la densidad *a priori* de  $\theta$ , tal que  $\rho_0(\theta)$  es la restricción de  $p(\theta)$  para  $\Theta_0$  renormalizada dada la densidad de probabilidad sobre  $\Theta_0$ , y similarmente por  $\rho_1(\theta)$ .

Entonces tenemos:

$$\begin{aligned} \alpha_0 &= P(\theta \in \Theta_0 | y) \\ &= \int_{\theta \in \Theta_0} p(\theta | y) d\theta \\ &\propto \int_{\theta \in \Theta_0} p(\theta) p(y | \theta) d\theta \\ &= \pi_0 \int_{\theta \in \Theta_0} p(y | \theta) \rho_0(\theta) d\theta \end{aligned}$$

La constante de proporcionalidad depende solamente de  $y$ . Similarmente

$$\alpha_1 \propto \pi_1 \int_{\theta \in \Theta_1} p(y|\theta) \rho_1(\theta) d\theta$$

Y así el *Factor de Bayes* es:

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{\int_{\theta \in \Theta_0} p(y|\theta) \rho_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(y|\theta) \rho_1(\theta) d\theta}$$

que es la razón de verosimilitud ponderada ( por  $\rho_0$  y  $\rho_1$  ) de  $\Theta_0$  y  $\Theta_1$ .

Como esta expresión del *Factor de Bayes* envuelve tanto a  $\rho_0$  y a  $\rho_1$  como a la función de verosimilitud  $p(y|\theta)$ , el *Factor de Bayes* no puede ser considerado como una medida de apoyo relativo para las hipótesis proporcionada solamente por los datos. Algunas veces, sin embargo, **B** es relativamente poco afectado dentro de límites razonables para las opciones de  $\rho_0$  y  $\rho_1$ , y entonces podemos considerar a **B** como una medida de apoyo relativo para las hipótesis proporcionadas por los datos. Cuando esto es así, el *Factor de Bayes* es razonablemente objetivo, por ejemplo, en un informe científico podría haber sido incluido para que usuarios diferentes de los datos pudieran determinar sus *odds a posteriori* personales, multiplicando sus *odds a priori* personales por el factor.

Muchos autores usan este *Factor de Bayes* simplemente como un factor. Notablemente, Peirce (1878) e independientemente Good (1950, 1983) en Box & Tiao (1992) se refieren al logaritmo del *Factor de Bayes* como el *peso de la evidencia*. El hecho de tomar logaritmos es, que si tenemos varios experimentos sobre dos simples hipótesis, entonces multiplican los *factores de Bayes*, y así agregan peso a la evidencia.



## 2.5 CONCLUSIONES

---

Desde el punto de vista *Bayesiano* existen tres ingredientes en la obtención de la decisión óptima:

- a) La distribución de *probabilidad a priori*, representa un estado de incertidumbre antes de haber visto cualquiera evidencia muestral.
- b) Los *costos* (o pérdidas), la función de pérdida, en este contexto, la definimos como una pérdida de oportunidad. Así, debemos especificar para cada estado posible de la naturaleza una función que describa la pérdida que resulta de no tomar la decisión óptima.
- c) La razón de *verosimilitud*. Una verosimilitud indica lo posible que es cada resultado muestral, dado los diversos estados de la naturaleza.

En el análisis de decisiones *Bayesianas*, se supone que una opción tiene que ser hecha de un conjunto de acciones posibles  $(a_1, a_2, \dots, a_r)$ , donde las ganancias o utilidad de una acción dada dependen de un estado de la naturaleza, digamos  $\theta$ , el cual es desconocido.

El conocimiento del decisor sobre  $\theta$  es representado por una distribución *a posteriori* que combina el conocimiento *a priori* de  $\theta$  con la información proporcionada por un experimento o *evidencia muestral*, y se supone entonces que el decisor escoge aquella acción que *maximiza* la esperanza de las ganancias sobre la distribución *a posteriori*.

Una aplicación importante de tal análisis, es a los problemas de decisiones económicos, como a sí también para introducir un nuevo producto industrial.

El hecho, que en tales situaciones las decisiones diferentes pueden ser el resultado de las opciones diferentes de distribuciones *a priori* ha preocupado a algunos estadísticos. Sin embargo, Box & Tiao (1992) argumentan a favor de la inferencia *Bayesiana* en *teoría de la decisión*, diciendo: "...para los problemas de esta índole, cualquier procedimiento que no tome en cuenta tales opiniones diferentes *a prioris* parecería necesariamente mal concebido".

Del mismo modo, la mayoría de las dificultades para interpretar los test de hipótesis están dadas por la dicotomía artificial,  $\theta = \theta_0$  y  $\theta \neq \theta_0$ . En problemas que involucran un parámetro  $\theta$  continuo, por ejemplo, la diferencia entre dos medias, la hipótesis que  $\theta$  es exactamente cero es raramente razonable, y es de más interés estimar su distribución *a posteriori* o una estimación del intervalo *a posteriori* correspondiente a  $\theta$ . Para un parámetro  $\theta$  continuo, la pregunta: ¿Se hace  $\theta$  igual a 0?, generalmente puede ser más útil referirse como: después de haber observado la muestra,  $y$ , ¿Qué conocimiento tenemos ahora acerca del parámetro  $\theta$ ?. La respuesta a ésta, esta dada por la distribución *a posteriori* de  $\theta$ .

Gelman *et al* (1998), cuestionan el análisis clásico de test de hipótesis, aún cuando, los p-valores pueden ser correspondidos con probabilidades *a posteriori* bajo distribuciones *a priori no informativa*. Exponiendo el siguiente ejemplo: Supongamos que observamos  $y = 1$  con  $y \sim N(\theta, 1)$ , con una densidad *a priori uniforme* sobre  $\theta$ . Siendo el p-valor unilateral de 0.16 y el p-valor bilateral de 0.32, ambos mayores que el límite máximo aceptado convencionalmente de 0.05 para estadísticas significativas, no se puede rechazar la hipótesis que  $\theta = 0$ . Pero sí consideramos el análisis *Bayesiano*, obtenemos que la probabilidad *a posterior* de que  $\theta > 0$  es de 0.84 u 84%, siendo ésta última, una conclusión más informativa y satisfactoria que el veredicto dicotómico de rechazar o no a la hipótesis nula.

---

## Capítulo III

# El modelo lineal general y la inferencia Bayesiana

---

### Introducción

---

La regresión lineal es ciertamente una de las herramientas estadísticas más usada. Las aplicaciones de regresión lineal se dan en casi todas las áreas científicas, incluyendo genética, física, economía, ciencias médicas, investigación de mercados, etc.

En este capítulo presentaremos la metodología *Bayesiana* para el tratamiento de la inferencia en el modelo básico de regresión lineal múltiple. Analizaremos el modelo a partir de las distribuciones *a priori* no informativas y conjugadas haciendo énfasis en la naturaleza secuencial del teorema de *Bayes*. Además, abordaremos brevemente, el caso en que se tenga variancias heterogéneas y correlacionadas.

El objetivo de éste Capítulo es exponer el marco teórico de la estimación puntual, por intervalos y test de hipótesis de los parámetros involucrados en el modelo de regresión lineal múltiple y de la predicción de nuevas observaciones.

### 3.1 El Modelo Lineal General

En muchos estudios científicos se investiga acerca de relaciones entre dos cantidades observables, que pueden ser representadas por una recta. Tales relaciones son llamadas *lineales* y sus modelos son llamados *modelos de regresión*.

Así, la cuestión que surge naturalmente es:

¿Cómo una cantidad digamos,  $y$ , varía en función de otra cantidad o vector de cantidades,  $X$ ? Esta pregunta nos lleva a reflexionar a que en realidad lo que nos interesa es la distribución condicional de  $y$ , dado  $X$ , cuya parametrización es:  $p(y|\theta, X)$ , donde  $\theta$  es el parámetro de tal distribución, bajo un modelo en el cual  $n$  observaciones  $(X, y)_i$  son intercambiables<sup>1</sup>.

Esta cantidad,  $y$ , es nuestro interés primario, y es llamada variable *respuesta* que la suponemos continua.

El vector o matriz  $X$  esta formado por variables llamadas *explicativas* y pueden ser discretas o continuas.

En algunos casos elegimos una sola variable *explicativa* de interés que la llamamos variable *tratamiento*<sup>2</sup>, y al resto de las integrantes de  $X$  las llamamos variables de *control*.

En los problemas de regresión hay dos situaciones ligeramente diferentes. En la primera los experimentadores son libres de poner los valores de la variable explicativa (por ejemplo en un experimento diseñado), y en la segunda ambos valores son aleatorios, aun cuando consideremos en hallar una relación causal o explicativa con el otro. El análisis, sin embargo, resulta ser el mismo en ambos casos.

Con el propósito de facilitar la aplicación de la inferencia *Bayesiana* consideraremos, aquí, el *modelo lineal general* de la forma usual:

<sup>1</sup> El usual punto de partida para el análisis estadístico es asumir que los  $n$  valores de  $y_i$  pueden ser considerados como intercambiables, significando que la densidad de probabilidad conjunta  $p(y_1, \dots, y_n)$  debería ser invariante por permutaciones de los índices.

<sup>2</sup> Las *variables tratamiento* son aquellas que describen atributos que se manipulan o por lo menos potencialmente manipulables por el investigador (como las dosis de drogas aplicadas a un paciente en un tratamiento médico experimental), considerando que *las variables de control* miden otras características de la unidad experimental o ambiente experimental, como el peso del paciente, medidas antes del tratamiento.

$$y = X\beta + \varepsilon$$

donde:

- $y$  es un vector  $n \times 1$  de observaciones, es decir un vector de observaciones  $(y_1, \dots, y_n)$ ;
- $X$  es una matriz de orden  $n \times k$  de constantes conocidas; es decir  $n$  observaciones y  $k$  variables, con rango  $k$
- $\beta$  es un vector  $k \times 1$  de parámetros, es decir de coeficientes de regresión desconocidos  $(\beta_1, \dots, \beta_k)$ ,
- $y$  y  $\varepsilon$  un vector de  $n \times 1$  de errores aleatorios, además se asume que los elementos de  $\varepsilon$  tienen media cero, no correlacionados y tienen una matriz de variancias y covariancias equivalente a  $\sigma^2 I_n$ , donde  $I$  es la matriz identidad de orden  $n \times n$ , esto es:  $\varepsilon \sim N(0, \sigma^2 I)$  (distribución Normal multivariante)

Con la definición de este *modelo lineal* el vector de parámetros  $\theta$  usual, se convierte en  $\theta = (\beta, \sigma^2)$ . La desviación estándar es un escalar.

Tal modelo expresa simplemente que la distribución condicional de  $y$  dado el parámetro  $\theta = (\beta_1, \dots, \beta_k, \sigma^2)$  es la distribución normal multivariada  $N(X\beta, \sigma^2 I)$ , así  $Ey = X\beta$ , luego en notación matricial el modelo es:

$$y | \beta, \sigma^2, X \sim N(X\beta; \sigma^2 I)$$

Explícitamente, el *modelo* sería :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$



en particular, para la  $i$ -ésima observación:

$$y_i = \beta_1 x_{j1} + \dots + \beta_k x_{jk} + \varepsilon_j$$

La relación entre las variables *explicativas* y la de *respuesta*, considerando que  $x$  es un vector, puede ser escrita de la forma:  $y = x_1 \beta_1 + \dots + x_k \beta_k + \varepsilon$ , donde  $x_j = (x_{1j}, \dots, x_{nj})^T$ ,  $\forall j = 1, \dots, k$ , son las columnas de la matriz  $X$  la cual, es llamada *matriz derivativa* o de *diseño*.

Con estas definiciones, el postulado matemático del *Teorema de Bayes* al *modelo lineal general* toma la siguiente forma:

$$p(\beta, \sigma^2 | y) \propto p(\beta, \sigma^2) p(y | \beta, \sigma^2) \quad (3.1.1)$$

Como el teorema se aplica sobre los parámetros que serán objetos de inferencia, la ecuación (3.1.1) postula que:

la función de probabilidad conjunta de los parámetros  $\beta$  y  $\sigma^2$ , condicionada (*a posteriori*) por la muestra "y", es proporcional a la *verosimilitud* de haber obtenido los valores muestrales con los parámetros  $\beta$  y  $\sigma^2$ , ponderada por la distribución de probabilidad *a priori*.

Bajo el supuesto de que los errores tienen una distribución normal, la función de verosimilitud se define:

$$p(y | \beta, \sigma^2) \propto \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\right] \quad (3.1.2)$$

La forma cuadrática,  $(y - X\beta)^T (y - X\beta)$ , la podemos escribir de la siguiente manera:

$$(y - X\beta)^T (y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta$$

Si la matriz  $n \times n$ ,  $X^T X$  es no singular, completamos cuadrados y obtenemos:

$$(y - X\beta)^T (y - X\beta) = (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) + s$$

donde:

➤  $\hat{\beta} = (X^T X)^{-1} X^T y$  (3.1.3), que es el clásico **Estimador de Mínimos Cuadrados Ordinarios** de  $\beta$ ,

➤ y  $s = (y - X\hat{\beta})^T (y - X\hat{\beta})$  es la suma de cuadrados de los residuos y  $s^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{v}$ ; con  $v = n - k$  grados de libertad (3.1.4)

Como  $s$  es una función que no depende de  $\beta$ , entonces, la verosimilitud puede ser expresada como:

$$p(y|\beta, \sigma^2) \propto \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left[-\frac{1}{2\sigma^2} \left( vs^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right)\right] \quad (3.1.5)$$

Hasta este punto se han planteado los elementos básicos del proceso *Bayesiano* para el Modelo Lineal General Normal, quedando por discutir el tipo de información *a priori* que posee el investigador.

Como ya hemos visto en el Capítulo I, la teoría *Bayesiana* propone dos posibilidades, a saber:

- el análisis con información *a priori* no informativa
- y con información conocida o *a priori* informativa.

A continuación discutiremos estas posibilidades para el Modelo lineal Normal analizado precedentemente.

### 3.1.1 Distribución *a priori* no informativa

Recordemos que cuando no se tiene información previa sobre los parámetros, la selección *a priori* adecuada adquiere una connotación especial porque es necesario elegir tal distribución de manera que no influya sobre ninguno de los posibles valores de los parámetros. Algunos criterios importantes para su selección lo constituyen el principio de la razón insuficiente de Laplace<sup>3</sup> y el método de Jeffreys<sup>4</sup>. Estos criterios

<sup>3</sup> Ver Capítulo I Sección 1.4 y 1.5.3

<sup>4</sup> Ver Capítulo I Sección 1.5.2



conducen a considerar *a priori* uniformes en el caso que el parámetro sea de posición y uniformes sobre el logaritmo del parámetro en caso que sea un escalar.

En el modelo de regresión normal; dado que  $\beta$  es un parámetro de posición y  $\sigma$  es un escalar o parámetro de escala, las distribuciones *a priori no informativa* serán:

$$p(\beta) \propto 1 \quad \text{y} \quad p(\sigma^2) \propto \frac{1}{\sigma^2}$$

y se asume independencia *a priori* entre  $\beta$  y  $\sigma$ , es decir que para los parámetros  $(\beta, \sigma^2)$  en el Modelo lineal Normal, se supone que:

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2) \propto \frac{1}{\sigma^2}$$

Ahora, por (3.1.1) y (3.1.5) la distribución *a posteriori*  $p(\beta, \sigma^2|y)$  es:

$$p(\beta, \sigma^2|y) \propto \frac{1}{(2\pi)^{n/2} \sigma^{n+2}} \exp\left[-\frac{1}{2\sigma^2} \left( v s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right)\right] \quad (3.1.6)$$

Esta expresión puede presentarse como el siguiente producto:

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y) p(\sigma^2|y)$$

con:

$$p(\beta|\sigma^2, y) \propto \frac{\exp\left\{-\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})\right\}}{\sigma^k} \quad \text{y} \quad p(\sigma^2|y) \propto \frac{\exp\left\{-\frac{v s^2}{2\sigma^2}\right\}}{(\sigma^2)^{v/2+1}}$$

donde  $p(\beta|\sigma^2, y)$  es la distribución *a posteriori* condicional de  $\beta$  dado  $\sigma^2$  y  $p(\sigma^2|y)$  es la distribución *a posteriori* marginal de  $\sigma^2$ .

Se puede notar que:

$$p(\sigma^2|y) = \text{Gamma} - \text{Inv}\left(\frac{v}{2}, \frac{v s^2}{2}\right) \quad \text{y} \quad p(\beta|\sigma^2, y) = N\left(\hat{\beta}, (X^T X)^{-1} \sigma^2\right) \quad (3.1.7)$$

Ahora, integrando sobre  $\sigma$  la distribución *a posteriori* conjunta dada en (3.1.6), se obtiene la distribución *a posteriori* marginal de  $\beta$  que es:

$$p(\beta|y) = t_k(v, \hat{\beta}, s^2(X^T X)^{-1})$$

Para derivar la distribución *a posteriori* marginal de un conjunto de  $k_1$  elementos de  $\beta$ , se integrará  $p(\beta|y)$  con respecto a los  $k_2$  elementos restantes ( $k_1 + k_2 = k$ ).

Sean las matrices  $H = \frac{X^T X}{s^2} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$  y  $H^{-1} = \begin{bmatrix} H_{11}^* & H_{12}^* \\ H_{21}^* & H_{22}^* \end{bmatrix} = (X^T X)^{-1} s^2$ ,

- $\beta_1$  el vector que contiene a los  $k_1$  parámetros de interés,
- $\beta_2$  el vector que contiene a los  $k_2$  parámetros restantes
- y  $H_{11}^{-1}$  la matriz que contiene las variancias y covariancias de los  $k_1$  parámetros  $\beta_1$  dados los  $k_2$  parámetros de  $\beta_2$ .

Tras integrar  $p(\beta|y)$  con respecto a  $\beta_2$  se obtiene que  $p(\beta_1|y) = t_{k_1}(v, \hat{\beta}_1, H_{11}^*)$ .

Por último, para obtener la distribución *a posteriori* de un elemento de  $\beta$ , digamos  $\beta_i$ , bastará con hacer  $k_1 = 1$  y entonces  $p(\beta_i|y) = t(v, \hat{\beta}_i, h_{ii}^*)$  que es una distribución *t*-Student univariada con  $v$  grados de libertad, parámetro de posición  $\hat{\beta}$  y parámetro de escala o escalar  $h_{ii}^*$ , donde  $h_{ii}^*$  es el elemento  $(i, i)$  de  $s^2(X^T X)^{-1}$

### 3.1.2 Distribución *a posteriori* con información *a priori* conjugada

Una densidad *a priori* con información conjugada es altamente conveniente ya que su aplicación resulta en una densidad *a posteriori* de la misma forma. Es decir, es una información muy flexible.

En esta situación  $\beta$  y  $\sigma^2$  tienen cierta distribución *a priori* conocida. Pues, la distribución Normal-Gamma Inversa es conjugada con la verosimilitud Normal correspondiente al vector de datos “ $y$ ”, la metodología para hallarla parte de los resultados obtenidos anteriormente. Así, las distribuciones  $p(\sigma^2|y)$  y  $p(\beta|\sigma^2, y)$

indicadas en (3.1.7), calculadas a partir de una muestra inicial  $(y_1, X_1)$ , serán las distribuciones *a priori* en esta sección a las que se denotará como:

$$p(\sigma^2) = \text{Gamma-Inv}\left(\frac{\nu_0}{2}, \frac{\nu_0 s_0^2}{2}\right) \quad \text{y} \quad p(\beta|\sigma^2) = N_k(\hat{\beta}_0, \sigma^2 M_0^{-1})$$

con los parámetros definidos según lo expuesto en (3.1.4).

Ahora, para una segunda muestra  $(y_2, X_2)$  de  $n_2$  observaciones independientes de las  $n_1$  iniciales, la función de verosimilitud de los datos es:

$$p(y_2|\beta, \sigma^2) = \frac{1}{(2\pi)^{n_2/2} \sigma^{n_2}} \exp\left\{\frac{-1}{2\sigma^2}(y_2 - X_2\beta)^T(y_2 - X_2\beta)\right\}$$

y por el *Teorema de Bayes* la distribución *a posteriori* conjunta de  $\beta$  y  $\sigma^2$  es:

$$p(\beta, \sigma^2|y_2) \propto p(y_2|\beta, \sigma^2)p(\beta, \sigma^2)$$

$$p(\beta, \sigma^2|y_2) \propto \frac{\exp\left\{\frac{-\nu s^2}{2\sigma^2}\right\}}{(\sigma^2)^{\nu/2-1}} \times \frac{\exp\left\{\frac{-1}{2\sigma^2}(\beta - \hat{\beta})^T M(\beta - \hat{\beta})\right\}}{\sigma^k}$$

donde:

- $\nu = n_1 + n_2 - k$ ;
- $M = X_1^T X_1 + X_2^T X_2 = M_0 + X_2^T X_2$ ;
- $\hat{\beta} = (X_1^T X_1 + X_2^T X_2)^{-1}(X_1^T y_1 + X_2^T y_2) = M^{-1}(X_1^T y_1 + X_2^T y_2)$  y
- $s^2 = \frac{(y_1 - X_1 \hat{\beta})^T (y_1 - X_1 \hat{\beta}) + (y_2 - X_2 \hat{\beta})^T (y_2 - X_2 \hat{\beta})}{\nu}$

con  $s^2$  y  $\hat{\beta}$  estadísticas conjuntamente suficientes para  $\sigma^2$  y  $\beta$ . En esta expresión el primer término corresponde a la distribución *a posteriori* marginal  $p(\sigma^2|y_2)$  y el segundo término corresponde a la distribución *a posteriori* condicional  $p(\beta|\sigma^2, y_2)$ , las cuales quedan definidas como:

$$p(\sigma^2|y_2) = \text{Gamma-Inv}\left(\frac{v}{2}, \frac{v s^2}{2}\right) \quad \text{y} \quad p(\beta|\sigma^2, y_2) = N_k(\hat{\beta}, \sigma^2 M^{-1})$$

Análogamente al caso anterior, se puede concluir que:

- La distribución a posteriori marginal de  $\beta$ ,  $p(\beta|y_2)$  es una  $t_{k_1}(v, \hat{\beta}, s^2 M^{-1})$ .
- La distribución a posteriori marginal para un conjunto  $\beta_1$  de  $k_1$  elementos de  $\beta$ ,  $p(\beta_1|y_2)$  es una  $t_{k_1}(v, \hat{\beta}_1, H_{11}^{-1})$  donde  $H_{11}^{-1}$  contiene los elementos de  $s^2 M^{-1}$  correspondientes a las variancias y covariancias de  $\beta_1$ .
- La distribución a posteriori marginal para un elemento  $\beta_i$  de  $\beta$ ,  $p(\beta_i|y_2)$  es una  $t(v, \hat{\beta}_i, h_{11}^{-1})$  donde  $h_{11}^{-1}$  es el elemento  $(i, i)$  de  $s^2 M^{-1}$ .

Debido a la naturaleza secuencial del *Teorema de Bayes*, como se mostró en el Capítulo I sección 1.3.4, la distribución *a posteriori* puede ser corregida indefinidamente conforme nueva información muestral sea disponible. Según lo expuesto en esta sección, esta corrección se realiza sólo mediante una actualización de los parámetros de las distribuciones Gamma Inversa y Normal iniciales.

Sean las siguientes distribuciones para  $\beta$  y  $\sigma^2$ :

$$p(\sigma^2|y_0) = \text{Gamma-Inv}\left(\frac{v_0}{2}, \frac{v_0 s_0^2}{2}\right) \quad \text{y} \quad p(\beta|\sigma^2, y_0) = N_k(\hat{\beta}_0, \sigma^2 M_0^{-1})$$

Las expresiones a continuación presentan la forma en la que los parámetros  $v_0$ ,  $M_0$ ,  $\hat{\beta}_0$  y  $s_0^2$  son actualizados por la nueva información muestral  $(X, y, n)$  para obtener los valores a posteriori de  $v$ ,  $M$ ,  $\hat{\beta}$  y  $s^2$ .

- $v = v_0 + n$
- $M = M_0 + X^T X$
- $\hat{\beta} = (M_0 + X^T X)^{-1} (M_0 \hat{\beta}_0 + X^T y)$
- $s^2 = \frac{v_0 s_0^2 + y^T y + \hat{\beta}_0^T M_0 \hat{\beta}_0 - \hat{\beta}^T (M_0 \hat{\beta}_0 + X^T y)}{v}$

### 3.2 La distribución predictiva a posteriori para nuevos datos

Supongamos que aplicamos el modelo de regresión a un conjunto nuevo de datos, para los cuales tenemos observada la matriz  $\tilde{X}$  de variables explicativas, y queremos predecir las respuestas,  $\tilde{y}$ .

Es decir queremos calcular la función de densidad de probabilidad *predictiva* para un vector de  $q$  observaciones futuras  $\tilde{y} = (\tilde{y}_{n+1}, \tilde{y}_{n+1}, \dots, \tilde{y}_{n+q})^T$  las cuales pueden ser presentadas por el siguiente modelo:

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon} \quad (3.2.1)$$

Asumiendo que  $\tilde{\varepsilon} \sim N_q[0, \sigma^2 I]$  y por (3.2.1), se tiene que:

$$p(\tilde{y}|\beta, \sigma^2) = \frac{1}{(2\pi)^{q/2} \sigma^q} \exp\left\{-\frac{1}{2\sigma^2} (\tilde{y} - \tilde{X}\beta)^T (\tilde{y} - \tilde{X}\beta)\right\}$$

y a partir de este resultado y de la distribución *a posteriori* conjunta de  $\beta$  y  $\sigma^2$ , se puede calcular la distribución *a posteriori* conjunta para  $\tilde{y}$ ,  $\beta$  y  $\sigma^2$ :

$$p(\tilde{y}, \beta, \sigma^2|y) = p(\tilde{y}|\beta, \sigma^2) p(\beta, \sigma^2|y)$$

Ahora, para obtener la distribución *a posteriori* de  $\tilde{y}$  será necesario integrar  $p(\beta, \sigma^2|y)$  con respecto a  $\beta$  y  $\sigma^2$ , tras lo cual se obtiene:

$$p(\tilde{y}|y) \propto \left[1 + \frac{1}{v} (\tilde{y} - \tilde{X}\hat{\beta})^T H (\tilde{y} - \tilde{X}\hat{\beta})\right]^{-(v+q)/2}$$

que tiene la forma de una  $t_q(v, \tilde{X}\hat{\beta}, H^{-1})$ ,

$$\text{donde } H = \frac{[I - \tilde{X}M^{-1}\tilde{X}^T]}{s^2} = \frac{[I - \tilde{X}M^{-1}\tilde{X}^T]^{-1}}{s^2},$$

$$\text{con } \tilde{M} = M + \tilde{X}^T \tilde{X} M = X^T X.$$

### 3.3 Variaciones Heterogéneas y Correlacionadas

El modelo tratado hasta aquí parte del supuesto de que los errores son independientes y normalmente distribuidos con media cero y variancia común  $\sigma^2$ .

Pero cuando las variancias son heterogéneas y los errores correlacionados, la verosimilitud de los datos es:

$$y \sim N_n[X\beta, \Sigma_y] \quad (3.3.1)$$

donde  $\Sigma_y$  es una matriz  $n \times n$  simétrica definida positiva.

Cuando la matriz de variancias  $\Sigma_y$  es conocida y partiendo de una distribución *a priori* uniforme no informativa para  $\beta$ , la distribución *a posteriori* podrá ser obtenida multiplicando a ambos lados de (3.3.1) por el factor de Cholesky<sup>5</sup> (raíz cuadrada de la matriz)  $\Sigma_y^{-1/2}$ , con lo cual se obtiene:

$$\Sigma_y^{-1/2} y \sim N_n[\Sigma_y^{-1/2} X\beta, I]$$

y la distribución *a posteriori*  $p(\beta | \Sigma_y, y)$  será una  $N_k[\hat{\beta}, (X^T \Sigma_y^{-1} X)^{-1}]$  con  $\hat{\beta} = (X^T \Sigma_y^{-1} X)^{-1} X^T \Sigma_y^{-1} y$ .

Si la matriz de covariancias  $\Sigma_y$  es desconocida, el análisis consistirá en dividir el problema de inferencia *a posteriori* en dos partes, a saber: La inferencia *a posteriori* para  $\beta$  dada  $\Sigma_y$  que ya fue tratada precedentemente y la inferencia *a posteriori* para  $\Sigma_y$ . Sin embargo, el tratamiento de este problema será complicado dado que tanto  $\hat{\beta}$  como  $(X^T \Sigma_y^{-1} X)^{-1}$  dependen de  $\Sigma_y$  y a la vez determinan una distribución *a priori* para  $\Sigma_y$ .

El problema de **heteroscedasticidad** es un caso particular en el que la matriz de covariancias  $\Sigma_y = \sigma^2 Q_y$ , con la matriz  $Q_y$  conocida y el parámetro  $\sigma^2$  desconocido. A partir de la distribución *a priori* no informativa  $p(\beta, \sigma^2) \propto \sigma^{-2}$ , será necesario

<sup>5</sup> Ver Gelman *et al* 1998 pág 238 y 478





calcular  $Q_y^{-1/2}$  y repetir el procedimiento planteado cuando  $\sum_y$  es conocida, esto es, multiplicar a ambos lados de (3.3.1) por  $Q_y^{-1/2}$ . En este caso, las estadísticas conjuntamente suficientes para  $\beta$  y  $\sigma^2$  son:

$$\hat{\beta} = (X^T Q_y^{-1} X)^{-1} X^T Q_y^{-1} y \quad \text{Estimador de Mínimos Cuadrados generalizados}$$

y

$$s^2 = \frac{(y - X\hat{\beta})^T Q_y^{-1} (y - X\hat{\beta})}{v} \quad \text{con } v = n - k \quad (3.3.2)$$

y un análisis similar al expuesto para hallar la distribución *a priori* no informativa resultará en las siguientes distribuciones *a posteriori*:

$$p(\sigma^2 | y) = \text{Gamma - Inv} \left( \frac{v}{2}, \frac{vs^2}{2} \right)$$

y

$$p(\beta | \sigma^2, y) = N_k \left( \beta | \hat{\beta}, (X^T Q_y^{-1} X)^{-1} \sigma^2 \right) \quad (3.3.3)$$

Por último, el modelo de regresión lineal **ponderada** es un caso particular del modelo de regresión lineal general tratado en esta sección en el que  $\sum_y$  es una matriz diagonal con sus elementos dados por:

$$\sum_{ii} = \frac{\sigma^2}{w_i}$$

$w_1, w_2, \dots, w_n$  son conocidos como pesos o ponderaciones. Luego, los resultados dados en (3.3.2) y (3.3.3) pueden ser adecuados a esta situación haciendo:

$$Q_y = \text{diag} \left( \frac{1}{w_1}, \dots, \frac{1}{w_n} \right) \quad \text{y} \quad Q_y^{-1} = \text{diag}(w_1 \cdots w_n)$$



### 3.4 Inferencia Bayesiana en el Modelo Lineal

#### 3.4.1 Estimación Puntual

Desde la perspectiva *Bayesiana*, el estimador de máxima verosimilitud de un parámetro  $\theta$ , como se discutió en el Capítulo II, puede ser la moda más grande de la distribución *a posteriori*  $p(\theta|y)$  y también la media y la mediana de  $p(\theta|y)$  son buenos estimadores de  $\theta$  y su cálculo a partir de una forma funcional conocida para  $p(\theta|y)$  no reviste complejidad.

En este caso, como la distribución *a posteriori* de  $\beta$  dado  $\sigma^2$ ,  $p(\beta|\sigma^2, y)$ , es una  $N_k(\hat{\beta}, \sigma^2 M^{-1})$  y la distribución *a posteriori* de  $\beta$ ,  $p(\beta|y)$ , es una  $t_k(v, \hat{\beta}, s^2 M^{-1})$ , la media de  $\beta$  es  $\hat{\beta}$ , valor que coincide con la moda y la mediana debido a la simetría de las distribuciones Normal y *t*-student.

Para el caso de  $\sigma^2$ , la moda de la distribución Gamma-Inv  $\left(\frac{v}{2}, \frac{vs^2}{2}\right)$  es:

$$\sigma_{\text{mod}}^2 = \frac{vs^2}{(v+2)}$$

Dado que  $p(\sigma^2|y)$  no es simétrica, la media *a posteriori*  $E^{p(\sigma^2|y)}[\sigma^2]$  difiere del valor de la moda. Si bien este estimador no es el valor más probable de  $\sigma^2$ , sí es el de menor variancia *a posteriori* por lo que sería conveniente observar ambos estimadores junto a sus respectivas variancias *a posteriori*.

Teniendo en cuenta las propiedades de la distribución Gamma Inversa, se pueden calcular la media y la variancia *a posteriori* de  $\sigma^2$ :

$$\mu^p = E^{p(\sigma^2|y)}(\sigma^2) = \frac{vs^2}{v-2} \quad \text{para } \frac{v}{2} > 1$$

$$V^p = E^{p(\sigma^2|y)}[(\sigma^2 - \mu^p)^2] = \frac{2v^2s^4}{(v-2)^2(v-4)} \quad \text{para } \frac{v}{2} > 2$$

donde  $V^P = V_{\mu^P}^P$  es la variancia *a posteriori* del estimador  $\mu_p$ . La variancia *a posteriori* para la moda es:

$$V_{\sigma_{\text{mod}}^2}^P = V^P + (\mu^P - \sigma_{\text{mod}}^2)^2$$

### 3.4.2 Regiones HPD

Como se discutió en el Capítulo II sección 2.3.1, la característica principal de los intervalos HPD es que cualquier punto contenido en la región tiene una densidad de probabilidad mayor o igual a la de cualquier punto fuera del intervalo.

En este caso, teniendo en cuenta la definición dada en la sección 2.3.1 del Capítulo II y dado que la distribución *a posteriori* para un  $\beta_i$  con  $\sigma^2$  desconocido es una  $t(v, \hat{\beta}_i, h_{ii})$ , los límites inferior y superior para una región HPD con probabilidad  $(1-\alpha)$  para  $\beta_i$  están dados por:

$$\beta_{il} = t\left(\frac{\alpha}{2}\right)\sqrt{h_{ii}} + \hat{\beta}_i \dots\dots\dots(3.4.1)$$

$$\beta_{is} = t\left(1-\frac{\alpha}{2}\right)\sqrt{h_{ii}} + \hat{\beta}_i \dots\dots\dots(3.4.2)$$

donde  $t\left(\frac{\alpha}{2}\right)$  y  $t\left(1-\frac{\alpha}{2}\right)$  son los cuantiles  $\left(\frac{\alpha}{2}\right)$  y  $\left(1-\frac{\alpha}{2}\right)$  de la distribución  $t$  de

Student con  $v$  grados de libertad

Para el cálculo de una región HPD para  $\sigma^2$ , dado que la distribución *a posteriori* Gamma Inversa no es simétrica, Gelman (1998) propone que en estos casos, se debe recurrir a métodos numéricos.

### 3.4.3 Test de Hipótesis

Recordando brevemente, lo visto en el Capítulo II sección 2.4, se sabe que en un test de hipótesis clásico, dada una hipótesis planteada  $H_0: \theta \in \Theta_0$  y una alternativa  $H_1: \theta \in \Theta_1$ , un procedimiento de test es evaluado en términos de las probabilidades de cometer un error de tipo I y de tipo II.

En el análisis *Bayesiano*, es necesario calcular las probabilidades *a posteriori*  $\alpha_0 = P\{\theta \in \Theta_0 | y\}$  y  $\alpha_1 = P\{\theta \in \Theta_1 | y\}$  correspondientes a  $H_0$  y  $H_1$ , y tomar una decisión de acuerdo a estos valores. Así, encontrábamos el *factor de Bayes*, cuyo interés reside en que representa la razón de verosimilitud de  $H_0$  a  $H_1$  dada por los datos muestrales.

Como siempre  $\pi_0$  y  $\pi_1$  denotan las probabilidades *a priori*  $H_0$  y  $H_1$ , luego el *factor de Bayes* a favor de  $H_0$  es:

$$\mathbf{B} = \frac{(\alpha_0 / \alpha_1)}{(\pi_0 / \pi_1)} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{p(y|\theta_0)}{p(y|\theta_1)}$$

De esta manera, si  $\mathbf{B}$  es mayor que 1, entonces la información muestral habrá favorecido a la hipótesis planteada.

Ahora bien, en nuestro caso, un test de hipótesis unilateral para un  $\beta_i$  puede ser evaluado fácilmente en términos de estas probabilidades.

Sean las hipótesis  $H_0 : \beta_i \in \Theta_0$  y  $H_1 : \beta_i \in \Theta_1$ , donde  $\Theta_0$  está enteramente a un lado de  $\Theta_1$ . Se sabe que la distribución *a posteriori* marginal de  $\beta_i$ ,  $p(\beta_i | y)$ , es una  $t(v, \hat{\beta}_i, h_{ii})$ , por lo que estandarizando, las probabilidades *a posteriori* pueden ser calculadas de la siguiente manera:

$$\alpha_0 = \int_{\Theta_0^*} t(v, \hat{\beta}_i, h_{ii}) \delta \beta_i = \int_{\Theta_0^*} t(v) \delta t \quad \text{y} \quad \alpha_1 = 1 - \alpha_0$$

donde  $\Theta_0^*$  es el espacio paramétrico transformado tras la estandarización.

Si la distribución *a priori*  $p(\beta_i)$  es conocida,  $\pi_0$  y  $\pi_1$  pueden calcularse de manera similar.

En el caso de una hipótesis planteada de la forma  $H_0 : \beta = \beta_0$  es más complicado, pues resulta poco razonable desde un punto de vista *Bayesiano* que  $\beta$ , que es una variable aleatoria continua, sea igual exactamente a  $\beta_0$ . Así, las probabilidades *a posteriori*  $\alpha_0$  y  $\alpha_1$  serían 0 y 1 respectivamente sin importar los valores de las hipótesis a probar y la distribución *a posteriori*.

Berger (1988), expone que más razonablemente sería plantear  $H_0 : \beta \in \Theta_0$ , con  $\Theta_0 = (\beta - b, \beta + b)$ , donde  $b$  es un vector de constantes escogido tal que, si  $\beta$  está en

$\Theta_0$  pueda ser considerado indistinguible de  $\beta_0$ . Luego, el problema consiste en determinar que tan grande debe ser  $b$ .

Por su lado Box & Tiao (1992), proponen una solución interesante al problema planteado. El mismo consiste en calcular una región HPD de contenido  $(1-\alpha)$  para  $\beta$  y verificar si  $\beta_0^* = (\beta_{10}, \beta_{20}, \dots, \beta_{k0})$  está dentro o fuera de la región y dada.

Si aplicamos lo visto en el Capítulo 2 sección 2.3.1 para una dada región HPD,  $R$ , en este caso tenemos que, para todo  $\beta_i \in R$  y  $\beta_j \notin R$ , se cumple que:

$$p(\beta_i | y) \geq p(\beta_j | y)$$

Entonces, dado  $\beta_0^* = (\beta_{10}, \beta_{20}, \dots, \beta_{k0})$  y dada  $R$ , una región HPD de contenido  $(1-\alpha)$  para  $\beta$  se tiene que:

$$\text{si } \Pr\{p(\beta | y) > p(\beta_0 | y)\} < 1 - \alpha \quad \text{entonces } \beta_0 \in R$$

y

$$\text{si } \Pr\{p(\beta | y) > p(\beta_0 | y)\} \geq 1 - \alpha \quad \text{entonces } \beta_0 \notin R$$

Dado que  $p(\beta | y) = N_k(\hat{\beta}, \sigma^2 M^{-1})$  es una función monótona decreciente de la cantidad:

$$\frac{(\beta - \hat{\beta})^T M(\beta - \hat{\beta})}{ks^2}$$

$p(\beta | y)$  será mayor que  $p(\beta_0 | y)$  si y sólo si el valor  $F(k, v)$  es menor que  $\frac{(\beta_0 - \hat{\beta})^T M(\beta_0 - \hat{\beta})}{ks^2}$ .

Por lo tanto,  $\beta_0$  estará dentro de la región HPD si:

$$\Pr\left\{F(k, v) < \frac{(\beta_0 - \hat{\beta})^T M(\beta_0 - \hat{\beta})}{ks^2}\right\} < 1 - \alpha$$



$$\frac{(\beta_0 - \hat{\beta})^T M (\beta_0 - \hat{\beta})}{ks^2} < F(k, v, 1 - \alpha)$$

Supongamos, ahora, que el interés está en un conjunto de  $k_1$  elementos de  $\beta$  agrupados en el vector  $\beta_1$ , cuya distribución *a posteriori* marginal es una  $t_k(v, \hat{\beta}_1, H_{11}^{-1})$ .

Entonces, siguiendo el mismo desarrollo anterior,  $\beta_{10}^* = (\beta_{10}, \beta_{20}, \dots, \beta_{k_1,0})$  estará dentro de la región HPD de contenido  $(1 - \alpha)$  si:

$$(\beta_{10} - \hat{\beta}_1)^T H_{11}^{-1} (\beta_{10} - \hat{\beta}_1) < k_1 F(k_1, v, 1 - \alpha)$$

Este procedimiento constituye una justificación *Bayesiana* para el análisis de variancia clásico. A partir de estos resultados Box & Tiao (1992) presentan la descomposición de la variabilidad para una hipótesis planteada de la forma  $\beta = \beta_0$ .

A continuación presentamos el cuadro de ANOVA para la hipótesis usual de regresión,  $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ , sin considerar el intercepto  $\beta_1$ .

### Análisis de variancia para evaluar la significancia del modelo de regresión

Fuentes de Variabilidad	Grados de Libertad	Sumas de Cuadrados	Cuadrados Medios	Razón de Cuadrados Medios
Parámetros	$k - 1$	$\hat{\beta}_1^T s^2 H_{11}^{-1} \hat{\beta}_1$	$\frac{\hat{\beta}_1^T s^2 H_{11}^{-1} \hat{\beta}_1}{k - 1}$	$\frac{\hat{\beta}_1^T s^2 H_{11}^{-1} \hat{\beta}_1}{k - 1} \sim F(k - 1, v)$
Residual	$v$	$vs^2$	$s^2$	

### 3.5 CONCLUSIONES

Los objetivos característicos o comunes que tienen los modelos de regresión, resumidamente son:

- 1- Entender el comportamiento de  $y$ , dado  $x$ ;
- 2- Prediciendo  $y$ , dado  $x$ , para las observaciones futuras;
- 3- La inferencia causal, o prediciendo cómo “ $y$ ” cambiaría si se cambiara  $x$  de una manera especificada.

Todos estos objetivos, deberían ser llevados a cabo a través de modelos de *probabilidad* cuando nos encontramos en una situación de incertidumbre, es decir, deberíamos determinar la inferencia *a posteriori*, y *predicción*.

Puesto que la reacción natural de cualquiera que se encuentre en una situación de incertidumbre será eliminar cuanta incertidumbre le sea posible obteniendo más información, el enfoque *Bayesiano* se constituye en un método más natural e “intuitivo” para el análisis estadístico.

La metodología *Bayesiana* hace uso tanto de la información de que dispone inicialmente el investigador como de la información resultante de una investigación o experimentación para obtener, mediante el *Teorema de Bayes*, la información final sobre la cual se tomará la decisión. Esta es la mayor diferencia entre los métodos *Bayesianos* y los clásicos.

Además, el enfoque *Bayesiano* justifica el uso del conocimiento subjetivo del investigador, en regresión o en cualquier otra herramienta estadística.

Ahora, cuando no se cuenta con información *a priori*, observamos que existe una fuerte similitud entre los **resultados Bayesianos** y los **resultados** clásicos o gaussianos. Sin embargo, existe una diferencia sustancial entre ambos métodos, y ésta se encuentra en el enfoque y análisis del problema.

Asimismo, la contrastación de la información *a priori* con la información muestral mediante la comparación de los *odds*, resulta no sólo fácil en cuanto a los cálculos sino también a la interpretación de los resultados.

Por último, el peso de la información *a priori* y muestral en la distribución *a posteriori* es directamente proporcional a la cantidad de información con que se cuente en cada caso. Así, si se cuenta con información muestral significativa, la función de verosimilitud dominará a la distribución *a priori*, lo que a menudo resulta apropiado, como vimos en el capítulo I, para el análisis de datos científicos.



---

## Capítulo IV

# Aplicación del método Bayesiano a un problema de Decisión de Inversión

---

### Introducción

---

Según Hossein Arsham<sup>1</sup>, el análisis de decisión proporciona un soporte cuantitativo a los tomadores de decisiones en todas las áreas tales como, analistas en las oficinas de planificación, ingenieros, analistas financieros y de economía, consultores en proyectos de gerencia, planificadores de procesos de producción, política monetaria, etc. Y tales decisiones son generalmente del tipo económica, aunque no es excluyente. Koontz *et al* (1990) define a la toma de decisiones “*como el acto de seleccionar un curso de acción entre distintas alternativas y para ser eficaz, el mismo debe ser racional*”.

Por otra parte, Levin (1998) reconoce que en este proceso de toma de decisión, existen elementos que definen el ambiente, y ellos son:

1. Un objetivo que se busca alcanzar.

---

<sup>1</sup> El Dr. Hossein Arsham, es Profesor en Ciencia de la decisión, Simulación y Estadística en la University of Baltimore USA

2. Varios cursos de acción.
3. Un cálculo del beneficio o valor de las diversas alternativas.
4. Incertidumbre sobre el resultado o estado de naturaleza que habrá de ocurrir, debido a eventos que escapan al control.

El análisis de decisiones, entonces, es una metodología que trata de diagnosticar la estructura de un problema de decisión y, por lo tanto, de llegar a una mejor decisión.

Aquí, consideramos a la toma de decisiones *bajo riesgo*, pues se admite que no se conocen anticipadamente con certeza las consecuencias o resultados de las decisiones.

En la toma de *decisiones bajo riesgo*, la estrategia *Bayesiana* permite incorporar a la información anterior que se posee, otra más actualizada; por ejemplo, proveniente de una investigación; y así se beneficia la decisión, dado que no desecha el conocimiento previo sino por el contrario lo combina con el nuevo.

En este Capítulo ilustraremos la teoría de la decisión utilizando la metodología *Bayesiana*, con la simulación de un problema de decisión de Inversión. Analizaremos un modelo de regresión cuadrático con el enfoque *Bayesiano*, para una función de *utilidad* simulada, a partir de distribuciones *a priori* conjugadas. Además, contrastaremos las distribuciones *a priori* y *a posteriori* para evaluar el efecto de información adicional. Por último realizaremos, una comparación entre la metodología Clásica y *Bayesiana*.

El objetivo de este Capítulo es ilustrar con datos simulados, la aplicación de la metodología *Bayesiana* expuesta en los Capítulos anteriores y observar similitudes y diferencias entre los resultados obtenidos con la metodología *Bayesiana* y aquellos que se derivan de la estadística clásica.



## 4.1 Decisión de inversión

El análisis de decisiones es la disciplina que consiste en evaluar alternativas complejas en términos de valores y de incertidumbre. Éste análisis brinda información sobre las diferencias entre las alternativas definidas, y genera sugerencias de nuevas y mejores alternativas. Para comprender la situación de decisión, dice Hossein Arsham<sup>2</sup>, se debe usar números para cuantificar valores e incertidumbres subjetivas, lo cual nos permite comprender la situación de decisión. Agregando, que los resultados numéricos deben reconvertirse para generar información cualitativa.

Toda decisión necesita de un decisor responsable. Éste tiene varias alternativas, y debe elegir una. El objetivo, entonces del decisor es elegir la mejor alternativa.

Para ello, el decisor debe realizar el siguiente proceso:

1. Definir el problema; considerar todas las alternativas confiables y evaluar los resultados posibles para cada alternativa.
2. Discutir los resultados de acuerdo a su reembolso monetario o de acuerdo a la ganancia neta en activos o con respecto al tiempo.
3. Cuantificar los valores inciertos en términos de probabilidad.

El análisis de decisiones es un proceso que le permite al decisor seleccionar una decisión (sólo una) entre un conjunto de alternativas posibles de decisión, cuando existe incertidumbre con respecto al futuro, con el objetivo de optimizar el retorno resultante, en términos de algún tipo de criterio de decisión numérico.

**Ilustremos estas ideas con un problema simulado de inversión.**

---

<sup>2</sup> Ob. Cit

#### 4.1.1 Planteo del problema

Supongamos que el Gerente de una empresa debe rendir cuenta ante los accionistas de sus decisiones de inversión, para ello cuenta con los siguientes elementos:

Un **número finito de alternativas** posibles de decisión, por ejemplo debe escoger entre las tres acciones o alternativas de inversión siguientes:

$a_1$ : Bonos.

$a_2$ : Acciones.

$a_3$ : Depósitos a Plazo Fijo.

La pregunta, ahora, que surge con naturalidad es:

**¿Qué acción lleva a la mejor decisión?**

Pero decidir la acción depende también de cuál de los **ambientes o escenarios económicos**, mutuamente excluyentes y colectivamente exhaustivo, llamados *estados de la naturaleza*, se enfrentará el próximo año.

Estos *estados de la naturaleza* son los estados de la economía durante un año, en este caso el decide que ellos son:

$\theta_1$ : Un estado de crecimiento económico alto.

$\theta_2$ : Un estado de crecimiento económico moderado.

$\theta_3$ : Un estado de crecimiento económico igual .

$\theta_4$ : Un estado de crecimiento económico bajo.

La mejor decisión (acción) para un cierto estado de la naturaleza no es la mejor para otro. Si el crecimiento económico para el próximo año es de moderado a alto, puede ser beneficioso invertir en acciones. Pero si el estado de crecimiento económico es bajo o se encuentra en recesión, invertir en acciones puede ser una mala decisión.

Obviamente estamos ante un problema de *incertidumbre bajo riesgo*. Es conveniente, por tanto, disponer de un procedimiento formal para analizar cada una de las acciones posibles en una situación dada y seleccionar la mejor.

Para ello, formulamos el problema de decisión de inversión construyendo una matriz de beneficios monetarios,  $X(a, \theta)$ . Es decir, cada curso de acción o alternativa de inversión tiene asociado un determinado beneficio o ganancia con cada uno de los estados de la naturaleza. Y este beneficio es cuantificado en términos a su reembolso monetario o de acuerdo a la ganancia.

A continuación se ilustra la matriz de *Beneficio monetario*:

		Estados de la Naturaleza			
		Crecimiento alto $\theta_1$	Crecimiento moderado $\theta_2$	Sin cambio $\theta_3$	Bajo $\theta_4$
Alternativas de Inversión	Bonos $a_1$	12%	8%	6%	3%
	Acciones $a_2$	15%	7%	3%	-2%
	Depósitos Plazo Fijo $a_3$	7%	7%	7%	7%

Donde cada  $x_{ij} \in X$ , es la ganancia. Es decir, por ejemplo si estamos en un escenario económico de alza o crecimiento alto e invertimos en Bonos por cada \$100 invertido nuestro beneficio es de \$12.

El problema entonces es decidir que acción tomar entre las tres alternativas posibles, con las ganancias dadas tal y como son mostradas en la tabla.

#### 4.1.2 Seleccionar un criterio de decisión: Ganancia Esperada

Los criterios de decisión se clasifican, teniendo en cuenta los grados de información que se tiene sobre los *estados de la naturaleza*, así:

- **Decisión tomada bajo certeza:** Los estados de la naturaleza que ocurrirán se asumen conocidos.

- **Decisión tomada bajo incertidumbre:** La probabilidad de que ocurra un estado de la naturaleza es absolutamente desconocida.
- **Decisión tomada bajo riesgo:** Existe conocimiento de la probabilidad que un estado de la naturaleza ocurra.

En el caso de **Decisión tomada bajo incertidumbre:** El criterio de decisión, se toma basándose en la experiencia de quién toma la decisión. Éste incluye un punto de vista optimista o pesimista, agresivo o conservador. Y los criterios son Maximin<sup>3</sup>, Mínimax<sup>4</sup>, Máximax<sup>5</sup> y el Principio de Razonamiento Insuficiente<sup>6</sup>.

Ahora bien, en general en el análisis de decisión se asume que el tomador de decisiones se enfrenta a un problema de riesgo. El riesgo implica cierto grado de incertidumbre y la habilidad para controlar plenamente los resultados o consecuencias de dichas acciones. Según Hossein Arsham (1996), “...el manejo efectivo del riesgo requiere la evaluación y el análisis del impacto subsiguiente del proceso de decisión. Este proceso permite al tomador de decisiones evaluar las estrategias alternativas antes de tomar cualquier decisión”.

Siempre que un decisor tiene cierto conocimiento sobre los estados de la naturaleza puede asignar una probabilidad subjetiva a la ocurrencia de cada estado. Y cuando lo hace, el problema se clasifica como *toma de decisiones bajo riesgo*.

En muchos casos, el decisor puede necesitar la opinión de un especialista para limitar sus incertidumbres con respecto a la probabilidad de cada estado de la naturaleza. En tal caso, el decisor puede comprar información relevante a especialistas, para poder tomar una mejor decisión.

<sup>3</sup> Criterio Maximin: Para encontrar una decisión óptima: Marcar la mínima ganancia a través de todos los estados de la naturaleza posibles. Identificar la decisión que tiene máximo de las mínimas ganancias.

<sup>4</sup> Criterio Mínimax: Para encontrar la decisión óptima se debe, determinar la mejor ganancias de todas las decisiones para cada estado de la naturaleza. Calcular el costo de oportunidad para cada alternativa de decisión como la diferencia entre su ganancia y la mejor ganancia calculada. -Para cada decisión encontrar el máximo costo de oportunidad para todos los estados de la naturaleza y seleccionar la alternativa de decisión que tiene el mínimo costo de oportunidad.

<sup>5</sup> Criterio Máximax: Para encontrar la decisión óptima: Encontrar la máxima ganancia para cada alternativa de decisión. Seleccionar la decisión que tiene la máxima de las máximas ganancias.

<sup>6</sup> El Principio de Razonamiento Insuficiente o Criterio de Laplace consiste en asumir que todos los estados de la naturaleza son equiprobables. El procedimiento para encontrar una decisión óptima, consiste en calcular la ganancia esperada para cada decisión y seleccionar la decisión con la mayor ganancia esperada.



Es decir, tendremos que asignar probabilidades,  $P(\theta_1), P(\theta_2), \dots, P(\theta_4)$ , a los estados de la naturaleza y habrá que decidir qué acción es mejor, frente a la duda del *estado de la naturaleza* que se presentará.

Luego, para abordar el problema el decisor se debe responder a la siguiente pregunta:

**¿En qué estado estará la economía el año próximo?**

Para ello el Gerente de la empresa decide comprar información relevante a especialistas, y así limitar sus incertidumbres con respecto a la probabilidad de cada estado de la naturaleza.

La compra de dicha información la realiza a una consultora, digamos  $C_p$ , obteniendo así, las siguientes probabilidades *a priori* de los *estados de la naturaleza* para el próximo año:

$$P(\theta_1) = 0,4; P(\theta_2) = 0,2; P(\theta_3) = 0,3; P(\theta_4) = 0,1$$

Así la **matriz de Beneficio monetario** queda definida de la siguiente manera:

		Estados de la Naturaleza			
		Crecimiento alto $\theta_1$ $P(\theta_1) = 0,4$	Crecimiento moderado $\theta_2$ $P(\theta_2) = 0,2$	Sin cambio $\theta_3$ $P(\theta_3) = 0,3$	Bajo $\theta_4$ $P(\theta_4) = 0,1$
Alternativas de acción	Bonos $a_1$	12%	8%	6%	3%
	Acciones $a_2$	15%	7%	3%	-2%
	Depósitos Plazo Fijo $a_3$	7%	7%	7%	7%

La función de *Beneficio*,  $X(a, \theta)$  representa cuán buena podría ser la decisión  $a$  si el valor del parámetro fuera  $\theta$ . La decisión *óptima*,  $a^*$ , es el máximo de la esperanza de  $X(a, \theta)$  sobre todos los posibles valores de  $\theta$ , es decir el **máximo del Beneficio esperado**.

Luego calculamos *el Beneficio Esperado*, esto es:



$$BE(a) = E(X(a, \theta)) = \sum X(a, \theta)p(\theta)$$

En estas condiciones el *Beneficio esperado* de cada curso de acción o alternativa es:

		Beneficio esperado
Alternativas de acción	Bonos $a_1$	$0,4 \cdot 12 + 0,2 \cdot 8 + 0,3 \cdot 6 + 0,1 \cdot 3 = 8,5^*$
	Acciones $a_2$	$0,4 \cdot 15 + 0,2 \cdot 7 + 0,3 \cdot 3 + 0,1 \cdot (-2) = 8,1$
	Depósitos Plazo Fijo $a_3$	$0,4 \cdot 7 + 0,2 \cdot 7 + 0,3 \cdot 7 + 0,1 \cdot 7 = 7$

Luego la decisión óptima es la de invertir en Bonos (8,5), lo que significa que en un gran número de inversiones similares el **mejor valor esperado** proviene de la alternativa de invertir en Bonos, esto es 8,5% de ganancia ( $VE_{MAX} = 8,5$ )

Algunas observaciones que se realiza al **criterio de la ganancia esperada**, es que:

- es factible de usar en situaciones donde es posible hacer una planificación apropiada, y las situaciones de decisión son repetitivas.
- un problema de este criterio es que no considera las situaciones ante posibles pérdidas.

En la teoría estadística de la decisión, es usual formular ese *beneficio* en términos de una función de *pérdida*  $L(\theta, a)$  (como se vio en el Capítulo II). En otras palabras, especificar, para cada estado de la naturaleza  $\theta$ , la pérdida que ocasiona cada decisión  $a \in A$  ( $A$  es el conjunto de acciones). Es decir, debe especificarse una función de pérdida  $L$  que a cada par  $(\theta, a)$  asocie la pérdida  $L(\theta, a)$ .

Formalmente, la *pérdida*  $L(\theta, a) = -X(a, \theta)$ ; o lo que es lo mismo definir la *pérdida relativa* de la decisión óptima  $a^*$ , que ésta dada por  $L(\theta, a) = X(a^*, \theta) - X(a, \theta)$ . (Box & Tiao, 1992)

La pérdida aquí es entendida como la *pérdida de oportunidad*, es decir, la *pérdida* que representa la consecuencia económica resultante de la decisión  $a$  cuando  $\theta$  es el



estado real de la naturaleza. Generalmente, en la teoría de la decisión el principal componente es la función de *pérdida*.

En nuestro caso teniendo en cuenta que  $L(\theta, a) = X(a^*, \theta) - X(a, \theta)$ , la matriz de *pérdida* queda definida así:

		Estados de la Naturaleza			
		Crecimiento alto $\theta_1$	Crecimiento moderado $\theta_2$	Sin cambio $\theta_3$	Bajo $\theta_4$
Alternativas de acción	Bonos $a_1$	15-12=3	8-8=0	7-6=1	7-3=4
	Acciones $a_2$	15-15=0	8-7=1	7-3=4	7-(-2)=10
	Depósitos Plazo Fijo $a_3$	15-7=8	8-7=1	7-7=0	7-7=0

Ahora en lugar de *maximizar* el *Beneficio esperado*, *minimizamos* la *pérdida esperada*.

La *pérdida esperada* de cada curso o alternativa de acción es:

		Pérdida esperada
Alternativas de acción	Bonos $a_1$	$0,4 \cdot 3 + 0,2 \cdot 0 + 0,3 \cdot 1 + 0,1 \cdot 4 = 1,9^*$
	Acciones $a_2$	$0,4 \cdot 0 + 0,2 \cdot 1 + 0,3 \cdot 4 + 0,1 \cdot 10 = 2,3$
	Depósitos Plazo Fijo $a_3$	$0,4 \cdot 8 + 0,2 \cdot 1 + 0,3 \cdot 0 + 0,1 \cdot 0 = 3,4$

Nuevamente la decisión óptima es la de invertir en Bonos, pero ¿que significa 1,9%?

Veamos, ubicados en la posición de contar con la información perfecta respecto de los escenarios futuros, se determina un nuevo valor, que es denominado *Beneficio Esperado con Información Perfecta* (BEIP), es la suma del resultado económico más alto, para cada escenario, por cada probabilidad de que ocurra ese estado de la naturaleza, en nuestro caso:

Estados de la naturaleza	
Crecimiento alto: $\theta_1$	$15 \cdot 0,4 = 6,0$
Crecimiento moderado: $\theta_2$	$8 \cdot 0,2 = 1,6$
Sin cambio: $\theta_3$	$7 \cdot 0,3 = 2,1$
Bajo: $\theta_4$	$7 \cdot 0,1 = 0,7$

A la suma de esos productos se la denomina *Beneficio Esperado con Información Perfecta*:

$$BEIP = 6,0 + 1,6 + 2,1 + 0,7 = 10,4$$

Y el *Costo Esperado de Información Perfecta (CEIP)*, es el máximo costo que se está dispuesto a incurrir para obtener la información que nos permita un conocimiento exacto de lo que acontecerá, el cual se define como:

$$CEIP = BEIP - VE_{MAX} = 10,4 - 8,5 = 1,9$$

1,9% es el máximo costo que se está dispuesto a incurrir para obtener la información que nos permita un conocimiento exacto de lo que acontecerá, es decir, *CEIP* corresponde al costo de oportunidad de la decisión seleccionada usando el criterio de la *ganancia esperada*. Esta decisión es la que genera una menor pérdida para el tomador de decisiones.

**Notar que *CEIP* es igual al *mínimo de la pérdida esperada***

Luego, si la información cuesta más de 1,9% de la inversión, no la deberíamos comprar. Por ejemplo, si se va a invertir \$1.000.000, el máximo que se debería pagar por la información es de \$19.000.

#### 4.1.3 Prueba de Confiabilidad

Sin embargo, el Gerente se siente algo reacio a tomar esta decisión y por ello solicita los registros de la consultora  $C_p$ , es decir, los datos históricos, del desempeño alcanzado en relación con las predicciones anteriores que hubieren formulado.

		Lo que resultó en el pasado			
		Crecimiento alto $\theta_1$	Crecimiento moderado $\theta_2$	Sin cambio $\theta_3$	Bajo $\theta_4$
Lo que predijo $C_P$	Resultado Positivo: $R_E$	80%	70%	50%	60%
	Resultado Negativo: $R_F$	20%	30%	50%	40%

Observar que, la tabla nos indica las siguientes probabilidades:

$$P(R_E|\theta_1)=0,80, P(R_E|\theta_2)=0,70, P(R_E|\theta_3)=0,50, P(R_E|\theta_4)=0,60$$

$$P(R_F|\theta_1)=0,20, P(R_F|\theta_2)=0,30, P(R_F|\theta_3)=0,50, P(R_F|\theta_4)=0,40$$

Para poder decidir si la consultora  $C_P$  es confiable debemos calcular las siguientes probabilidades:

$P(\theta_i|R_E)$  con  $i=1;\dots;4$  y  $P(\theta_i|R_F)$  con  $i=1;\dots;4$ , para ello utilizamos el teorema de Bayes visto en el Capítulo I:

$$P(\theta_i|R_E)=\frac{P(\theta_i)\cdot P(R_E|\theta_i)}{\sum_i P(\theta_i)\cdot P(R_E|\theta_i)} \quad \text{y} \quad P(\theta_i|R_F)=\frac{P(\theta_i)\cdot P(R_F|\theta_i)}{\sum_i P(\theta_i)\cdot P(R_F|\theta_i)}$$

Recordar que  $\theta_1, \theta_2, \theta_3$  y  $\theta_4$  es una colección de eventos mutuamente excluyentes y exhaustivo. La ocurrencia de  $R_E$  y  $R_F$  transforma la probabilidad *a priori*  $P(\theta_i)$  en probabilidades a posteriori  $P(\theta_i|R_E)$  y  $P(\theta_i|R_F)$  respectivamente, ya que se determinan una vez obtenida la evidencia experimental, luego  $P(\theta_i|R_E)$  y  $P(\theta_i|R_F)$  reflejan el grado de *creencia corregido* con respecto a las hipótesis  $\theta_1, \theta_2, \theta_3$  y  $\theta_4$  después de obtener los registros de desempeño de la consultora  $C_P$ .

Con fines ilustrativos mostraremos el proceso de cálculo de  $P(\theta_i|R_E)$ :

$$P(\theta_1|R_E) = \frac{P(\theta_1) \cdot P(R_E|\theta_1)}{\sum_{i=1}^4 P(\theta_i) \cdot P(R_E|\theta_i)} = \frac{0,4 \cdot 0,8}{0,4 \cdot 0,8 + 0,2 \cdot 0,7 + 0,3 \cdot 0,5 + 0,1 \cdot 0,6} = \frac{0,32}{0,67} = 0,47761194$$

Y así sucesivamente se obtiene las siguientes *probabilidades corregidas* de los estados de la naturaleza dado que  $C_p$  predijo positivamente:

Estados de la naturaleza	Probabilidad a priori	Verosimilitud	Probabilidad conjunta	Probabilidad a posteriori
Crecimiento alto $\theta_1$	0,4	0,8	0,32	0,47761194
Crecimiento moderado $\theta_2$	0,2	0,7	0,14	0,20895522
Sin cambio $\theta_3$	0,3	0,5	0,15	0,2238806
Bajo $\theta_4$	0,1	0,6	0,06	0,08955224
			Sum = 0,67	

Y las siguientes son las *probabilidades corregidas* de los estados de la naturaleza dado que  $C_p$  predijo negativamente:

Estados de la naturaleza	Probabilidad a priori	Verosimilitud	Probabilidad conjunta	Probabilidad a posteriori
Crecimiento alto $\theta_1$	0,4	0,2	0,08	0,24242424
Crecimiento moderado $\theta_2$	0,2	0,3	0,06	0,18181818
Sin cambio $\theta_3$	0,3	0,5	0,15	0,45454545
Bajo $\theta_4$	0,1	0,4	0,04	0,12121212
			Sum = 0,33	

Ahora calculamos el *Beneficio esperado* de cada alternativa con información adicional, es decir calculamos el **Beneficio esperado dado que el análisis resulto positivo** y el **Beneficio esperado dado que el análisis resulto negativo**:



		Beneficio esperado $BE(a_i)$	Beneficio esperado/res.Positivo $BE(a_i R_E)$	Beneficio esperado/res.Negativo $BE(a_i R_F)$
Alternativas de acción	Bonos $a_1$	8,5	9,01492537	7,45454545
	Acciones $a_2$	8,1	9,11940299	6,03030303
	Depósitos Plazo Fijo $a_3$	7	7	7

Así, de la lectura de la tabla resulta que; cuando el análisis es **positivo** invertimos en **Acciones** y cuando el análisis es **negativo** invertimos en **Bonos**.

Ahora, calculamos el *Beneficio Esperado con información adicional*, es decir:

$$BEIA = 0,67 \cdot 9,11940299 + 0,33 \cdot 7,45454545 = 8,57$$

La diferencia entre el *Beneficio esperado con información adicional* (8,57) y el *Beneficio esperado* (8,5) nos da el **Costo esperado** de la compra de la información a la consultora  $C_p$ , con información adicional:

$$\text{Y así, } CEIA = 8,57 - 8,5 = 0,07$$

Que nuevamente expresa el costo que el Gerente debe estar dispuesto a pagar por obtener información adicional, y que es sensiblemente menor que el costo por información perfecta que alcanzaba a 1,9%

Luego, parece ser confiable la consultora  $C_p$

## 4.2 Función de utilidad

Sin embargo hasta aquí hemos trabajado con la *matriz de Beneficio monetario*, es decir con una tabla de redistribución expresadas en términos del valor monetario esperado. Pero el valor del dinero varía de situación a situación y de una decisión a otra. Generalmente, el valor del dinero no es una función lineal de la cantidad de dinero, por

ejemplo, hay personas motivadas a comprar un boleto de lotería por \$10 sabiendo que el mismo tiene una ganancia esperada negativa.

Luego, el *criterio de la ganancia esperada* para la toma de decisiones puede no ser apropiado cuando se tenga una única oportunidad para tomar la decisión y ésta tiene riesgos considerables. En tal caso, determinar la **utilidad** monetaria puede ser un criterio más razonable. Ya que la **utilidad** transforma el uso de un resultado en un valor numérico que mide la valoración personal del resultado

El proceso de toma de decisiones involucra factores psicológicos y económicos, entre otros y el concepto de **utilidad** es un intento de medir el provecho que tiene el dinero para el decisor en lo individual.

Por lo tanto, para tomar una decisión acertada considerando la actitud que tiene el decisor con respecto al riesgo, debemos convertir la *matriz de beneficios monetarios* en una *matriz de utilidad*.

Ahora la pregunta sería:

#### ¿Cómo se mide la utilidad con cada decisor?

La utilidad de un resultado puede estar dentro de una escala comprendida entre 0 y 1. El menor resultado obtenido tiene un valor de utilidad de 0 y el mayor resultado obtenido tiene un valor de utilidad de 1.

Para determinar la **utilidad**, se procede de la siguiente manera (Hossein Arsham<sup>7</sup>):

Debemos transformar la matriz de *Beneficio monetario* en una *matriz de Utilidad*, para ello, asignamos una utilidad 0 al valor más bajo y un valor 1 al más alto.

Para todas las otras posibles ganancias deberíamos pedirle al decisor que elija entre las siguientes hipótesis:

“suponga que Ud. podría recibir esa ganancia en forma segura o recibiría, ya sea la mayor ganancia con probabilidad  $p$  y la menor ganancia con probabilidad  $(1 - p)$ .”

Cambiando el valor de  $p$ , y repitiendo un proceso similar, existe un valor de  $p$  con el que el decisor es indiferente ante las dos hipótesis.

---

<sup>7</sup> Ob. Cit



Supongamos que el valor  $p$  de indiferencia es  $p = 0,48$  para una ganancia de \$12 por cada \$100

La respuesta a esta elección son las probabilidades de “indiferencia” con respecto a la ganancia y se usan como valores para la **utilidad**.

Supongamos que nuestra matriz de *Beneficios monetarios* se transforma en la siguiente *Matriz de utilidad*:

	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$
Bonos $a_1$	12	8	6	3
Acciones $a_2$	15	7	3	-2
Depósitos $a_3$	7	7	7	7

	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$
Bonos $a_1$	0,48	0,34	0,28	0,13
Acciones $a_2$	1	0,19	0,13	0
Depósitos $a_3$	0,19	0,19	0,19	0,19

El valor de la utilidad refleja la perspectiva del tomador de decisiones.

Pero en la teoría de decisiones no sólo, interesa determinar la utilidad óptima sino **predecir** la utilidad del tomador de decisiones para un valor monetario dado. Y para ello en primer lugar, debemos determinar la *función de utilidad*.

Obviamente que la predicción y precisión aumentan desde la **representación** de una función a través de **tablas** hasta la construcción de un **modelo matemático** que la represente.

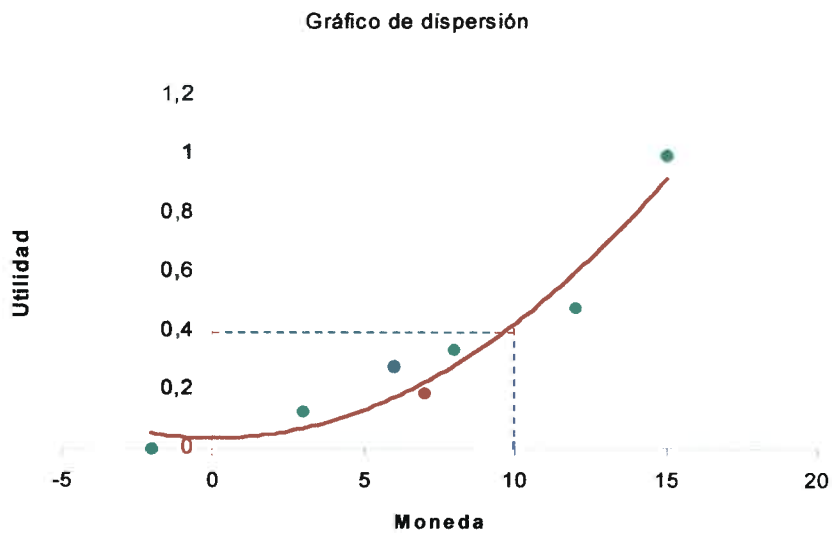
Y en ese intento por construir el **modelo**, presentamos en primer lugar la representación en tabla de la función de utilidad.

Con  $U$  representamos la utilidad y con  $M$  los valores monetarios, así entonces quedan representados los pares ordenados  $(M,U)$ , en la siguiente tabla:

M	U
12	0,48
8	0,34
6	0,28
3	0,13
15	1
7	0,19
3	0,13
-2	0
7	0,19
7	0,19
7	0,19
7	0,19

Como podemos observar, la representación en tablas esta limitada a los valores numéricos dentro de la misma. Supongamos que deseamos obtener la utilidad de \$10. Para ello podríamos interpolar pero como la **función de utilidad** generalmente no es lineal, el resultado no representaría adecuadamente la utilidad del decisor.

Para poder resolver esta dificultad podemos graficar la función de utilidad a través de un **gráfico de dispersión**, para ello decidimos la forma aproximada que debería tener la **función de utilidad**. En general, una función cuadrática ajusta relativamente bien a la función de utilidad cuando la variable explicativa es la monetaria (Hossein Arsham, 1996)<sup>8</sup>.



<sup>8</sup> Ob. Cit

Ahora sí, al tener el gráfico podemos leer la **utilidad** de \$10 directamente del gráfico, como se muestra en la figura anterior, el resultado es aproximadamente  $U = 0,40$ .

Pero como observamos éste es aproximado luego, para los procesos de **predicciones**, un **modelo matemático** es el que mejor funciona.

### 4.3 Modelo de Regresión para la función de utilidad

Para construir un modelo matemático para la función de utilidad, utilizamos su representación gráfica anterior y como mencionamos anteriormente una función cuadrática parece ser un buen ajuste. Para ello, utilizamos el análisis de regresión para estimar los coeficientes de la **función utilidad** que mejor se ajusta a los datos. Recordando que el modelo de regresión es de la forma  $y = X\beta + \varepsilon$  (Capítulo III).

Así nuestro modelo es el siguiente:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon, \text{ con } \varepsilon \sim N(0, \sigma^2 I)$$

siendo,

$y =$  la utilidad de cada ganancia

$x =$  el Beneficio Monetario medido en porcentaje

y recordando que nuestros datos son:

<b>x</b>	-2	3	3	6	7	7	7	7	7	8	12	15
<b>y</b>	0	0,13	0,13	0,28	0,19	0,19	0,19	0,19	0,19	0,34	0,48	1

En notación matricial el modelo es:

$$y = \begin{bmatrix} 0 \\ 0,13 \\ 0,13 \\ 0,28 \\ 0,19 \\ 0,19 \\ 0,19 \\ 0,19 \\ 0,19 \\ 0,34 \\ 0,48 \\ 1 \end{bmatrix} \quad X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ 1 & 12 & 144 \\ 1 & 15 & 225 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

El clásico **Estimador de Mínimos Cuadrados Ordinarios** de  $\beta$  es:

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 12 & 80 & 736 \\ 80 & 736 & 7592 \\ 736 & 7592 & 88936 \end{bmatrix}^{-1} \begin{bmatrix} 3,31 \\ 32,59 \\ 374,85 \end{bmatrix} = \begin{bmatrix} 0,03618 \\ -0,000344 \\ 0,003945 \end{bmatrix}$$

Por lo tanto el modelo de regresión es:

$$\hat{y} = 0,03618 - 0,000344x + 0,003945x^2$$

Obviamente que este modelo proporciona información más útil que el gráfico y la tabla vista anteriormente. Por ejemplo, colocando  $x = 10$ , se obtiene el valor predicho de utilidad  $y = 0,42724$ .

#### 4.3.1 Construcción de una distribución *a priori*

La construcción de una distribución *a priori* para los parámetros del modelo la realizamos a partir de la construcción de la **función de utilidad** anterior y los conceptos vistos en el Capítulo III, así las distribuciones *a priori* obtenidas son las siguientes:

➤ **Distribución condicional**  $p(\beta | \sigma^2, y) = N_k(\hat{\beta}, \sigma^2 M^{-1})$



➤ **Distribución marginal**  $p(\beta|y) = t_k(v, \beta, s^2 M^{-1})$

➤ **Distribución marginal**  $p(\sigma^2|y) = \text{Gamma-Inv}\left(\frac{v}{2}, \frac{vs^2}{2}\right)$

Con los siguientes valores para sus parámetros:

$$n = 12 \qquad k = 3 \qquad v = n - k = 12 - 3 = 9$$

$$M = X^T X = \begin{bmatrix} 12 & 80 & 736 \\ 80 & 736 & 7592 \\ 736 & 7592 & 88936 \end{bmatrix}$$

$$\hat{\beta} = M^{-1} X^T y = \begin{bmatrix} 0,03618 \\ -0,000344 \\ 0,003945 \end{bmatrix}$$

$$s^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{v} = \frac{0,051454}{9} = 0,005717$$

Además,

$$s^2 M^{-1} = \begin{bmatrix} 0,0022375 & -0,00043706 & 0,000018792 \\ -0,00043706 & 0,0001504 & -0,0000092221 \\ 0,000018792 & -0,0000092221 & 0,00000069601 \end{bmatrix} \text{ que es la matriz de}$$

variancias y covariancias de los  $\beta_i$

#### 4.3.2 Cálculo de la distribución a posteriori

Dadas las distribuciones *a priori* calculadas en la sección anterior, se calcularán ahora las distribuciones *a posteriori*. El procedimiento consistirá en actualizar los parámetros de las distribuciones *a priori* con nueva información según los resultados expuestos en el Capítulo III sección 3.1.2.

Para ello el Gerente con información brindada por sus asesores financieros obtiene la siguiente **función de utilidad**:

**Cuadro 1:**

<b>x</b>	-2	3	3	6	7	7	7	7	7	8	12	15
<b>y</b>	0	0,15	0,15	0,16	0,21	0,21	0,21	0,21	0,21	0,31	0,45	1

Los parámetros actualizados por la nueva información son:

$$v = n_1 + n_2 - k = v_0 + n = 9 + 12 = 21$$

$$M = X_1^T X_1 + X_2^T X_2 = M_0 + X_2^T X_2 = \begin{bmatrix} 24 & 160 & 1472 \\ 160 & 1472 & 15184 \\ 1472 & 15184 & 177872 \end{bmatrix}$$

$$\hat{\beta} = M^{-1}(X_1^T y_1 + X_2^T y_2) = \begin{bmatrix} 0,038754 \\ -0,0011174 \\ 0,0039597 \end{bmatrix}$$

$$s^2 = \frac{(y_1 - X_1 \hat{\beta}_1)^T (y_1 - X_1 \hat{\beta}_1) + (y_2 - X_2 \hat{\beta})^T (y_2 - X_2 \hat{\beta})}{v} = 0,0046649$$

## 4.4 Inferencia a posteriori

### 4.4.1 Análisis de variancia

El análisis de variancia clásico para el modelo de regresión, a partir de la nueva información para la **utilidad** dada en el **Cuadro 1** y el análisis de variancia *Bayesiano* a partir de la distribución *a posteriori* se presentan a continuación:

**Cuadro 2:** Análisis de variancia **Clásico** para el modelo de regresión:

Fuente de Variación	DF	SS	MS	F	P
Regresión	2	0,65221573	0,32610787	63,24105	0,0000
Residual	9	0,04640927	0,00515659		
Total	11				

Para la construcción del siguiente cuadro ver Capítulo III pág. 58

**Cuadro 3:** Análisis de variancia **Bayesiano** para el modelo de regresión:

Fuentes de Variabilidad	DF	SS	MS	F	P
Regresión	2	1,3262	0,6631	142,14666	0,0000
Residual	21	0,097963	0,0046649		
Total	23				

Observando ambos cuadrados, concluimos que en ambos casos se acepta la significancia del modelo con un nivel de significación del 5%. Sin embargo en el análisis *Bayesiano* se cuenta con información adicional lo que se traduce en más grados de libertad para el error y un valor F mayor.

#### 4.4.2 Estimación puntual

Realizamos la estimación puntual Clásica para el modelo de regresión, a partir de la nueva información para la **utilidad** dada en el **Cuadro 1** y la estimación puntual *Bayesiana* a partir de los resultados mostrados en el Capítulo III sección 3.4.1:

**Cuadro 4:** Estimación **Clásica** para los parámetros del modelo de regresión

Parámetro	Valor estimado	Desv. Estándar
$\beta_1$	0,041328	0,044924
$\beta_2$	-0,001891	0,011647
$\beta_3$	0,003975	0,000792
$\sigma^2$	0,00515659	0,0718094



**Cuadro 5:** Estimación *Bayesiana* para los parámetros del modelo de regresión

Parámetro	Valor estimado		Desv. Estándar	
$\beta_1$	0,038754		0,030214	
$\beta_2$	-0,0011174		0,0078334	
$\beta_3$	0,0039597		0,00053288	
$\sigma^2$	Media $\frac{vs^2}{v-2}$	0,0051559	$\sqrt{\frac{2v^2s^4}{(v-2)^2(v-4)}}$	0,0017685
	Moda $\frac{vs^2}{(v+2)}$	0,0042593	$\frac{2v^2s^4}{(v-2)^2(v-4)} + \left( \frac{vs^2}{(v-2)} - \frac{vs^2}{(v+2)} \right)^2$	0,000003931

Observamos que en la estimación *Bayesiana* la moda de  $\sigma^2$  tiene el menor valor y con desviación estándar también menor. Es conveniente, entonces, reportar el valor de la moda como estimador de  $\sigma^2$ .

Se puede apreciar, además, que tanto en las estimaciones de los  $\beta_i$  como para  $\sigma^2$ , las desviaciones estándar de los estimadores desde la distribución *a posteriori* son considerablemente menores a las obtenidas con la estimación Clásica. Es evidente que el uso de información *a priori* adecuada trae consigo una ganancia en precisión.

#### 4.4.3 Estimación por intervalos

Calcularemos las regiones HPD con contenido de probabilidad de 0,95 o nivel de *credibilidad* 0,95, a partir de los resultados de la sección 3.5 del Capítulo III, teniendo en cuenta los resultados (3.4.1) y (3.4.2) para los  $\beta_i$ , así se obtiene:

- $\hat{\beta}_1 \pm t_{(0,975,21)}\sqrt{h_{11}}$

$$\left[ \hat{\beta}_1 + t_{(0,975,21)}\sqrt{h_{11}}; \hat{\beta}_1 - t_{(0,975,21)}\sqrt{h_{11}} \right]$$

$$= [0,038754 - (2,080)(0,030214); 0,038754 + (2,080)(0,030214)]$$

$$= [-0,024091; 0,101599]$$

$$\bullet \hat{\beta}_2 \pm t_{(0,025;21)} \sqrt{h_{22}^*}$$

$$\left[ \hat{\beta}_2 + t_{(0,975;21)} \sqrt{h_{22}^*}; \hat{\beta}_2 - t_{(0,975;21)} \sqrt{h_{22}^*} \right]$$

$$= [-0,0011174 - (2,080)(0,0078334); -0,0011174 + (2,080)(0,0078334)]$$

$$= [-0,017411; 0,015176]$$

$$\bullet \hat{\beta}_3 \pm t_{(0,975;21)} \sqrt{h_{33}^*}$$

$$\left[ \hat{\beta}_3 + t_{(0,975;21)} \sqrt{h_{33}^*}; \hat{\beta}_3 - t_{(0,975;21)} \sqrt{h_{33}^*} \right]$$

$$= [0,0039597 - (2,080)(0,00053288); 0,0039597 + (2,080)(0,00053288)]$$

$$= [0,0028513; 0,0050681]$$

Para todos los casos anteriores, existe una **probabilidad** del 95% de que el parámetro se encuentre en el intervalo calculado

#### **Cálculo de un intervalo de credibilidad para $\sigma^2$ :**

A los fines de ilustrar los contenidos del Capítulo III, aquí se calculará un intervalo de *credibilidad* para  $\sigma^2$  propuesto por Box & Tiao (1992):

Siendo  $\sigma^2|y \sim \text{Gamma-Inv} \left( \frac{v}{2}, \frac{vs^2}{2} \right)$ , un intervalo de *credibilidad* de 95% para  $\sigma^2$

es:

$$\left[ \frac{vs^2}{\chi_v^2(0,975)}; \frac{vs^2}{\chi_v^2(0,025)} \right] = \left[ \frac{21 \cdot 0,0046649}{35,4789}; \frac{21 \cdot 0,0046649}{10,2829} \right]$$

$$= [0,0027612 ; 0,0095268]$$

#### 4.4.4 Test de hipótesis para los coeficientes de regresión

Se contrastarán las hipótesis  $H_0: \beta_3 = 0$  vs.  $H_1: \beta_3 \neq 0$  comparando sus probabilidades *a priori*  $\pi_0$  y  $\pi_1$ , a posteriori  $\alpha_0$  y  $\alpha_1$  y el factor de Bayes, teniendo en cuenta los resultados obtenidos en la sección 3.4.3 del Capítulo III

**Cálculo de  $\pi_0$  y  $\pi_1$ :**

Se sabe que los  $\beta_i \sim t(v; \hat{\beta}_i; h_{ii})$ , en éste caso  $\beta_3 \sim t(9; 0,003945; 0,00000069601)$  entonces estandarizando las variables obtenemos que el estadístico:

$$t = \frac{\hat{\beta}_3}{\text{desv.est de } \hat{\beta}_3} = \frac{0,003945}{\sqrt{0,00000069601}} = 4,7287$$

Luego:

$$\pi_0 = P(\beta_3 = 0) = 0,0008$$

$$\pi_1 = P(\beta_3 \neq 0) = 1 - \pi_0 = 0,9992$$

**Cálculo de  $\alpha_0$  y  $\alpha_1$ :**

La distribución *a posteriori* marginal de  $\beta_i$ ,  $p(\beta_i|y)$ , es una  $t(v, \hat{\beta}_i, h_{ii})$ , por lo que estandarizando, las probabilidades *a posteriori* pueden ser calculadas como las anteriores.

Así, dado que  $\beta_3 \sim t(21; 0,0039597; 0,00000028396)$

$$\text{El estadístico } t = \frac{\hat{\beta}_3}{\text{desv.est de } \hat{\beta}_3} = 7,4308$$

Luego:

$$\alpha_0 = p(\beta_3 = 0|y) = 0$$

y

$$\alpha_1 = 1 - \alpha_0 = 1$$

**Cálculo del factor de Bayes:**

$$B = \frac{(\alpha_0 / \alpha_1)}{(\pi_0 / \pi_1)} = \frac{0}{0,00080064} = 0$$

**Resumiendo:**

	<i>A priori</i>	<i>A posteriori</i>
A favor de $H_0$	$\pi_0 = 0,0008$	$\alpha_0 = 0$
A favor de $H_1$	$\pi_1 = 0,9992$	$\alpha_1 = 1$
Razón Odds	0,00080064	0
<b>Factor de Bayes a favor de <math>H_0</math>: <math>B = 0</math></b>		

Observamos que, tanto *a priori* como *a posteriori* las probabilidades de que el parámetro  $\beta_3 = 0$  son menores que las probabilidades de que  $\beta_3 \neq 0$ , es decir, es más probable que los datos estén ajustados por un polinomio de grado 2 que de un polinomio de grado 1.

*A priori* la probabilidad de que  $\beta_3 = 0$  es 0,00080064 veces la probabilidad de que  $\beta_3 \neq 0$  y *a posteriori* 0. El **factor de Bayes**,  $B = 0$  implica, por ser menor que 1, que la información ha favorecido a la hipótesis alterna.

## 4.5 CONCLUSIONES

---

La principal ventaja del método *Bayesiano* es que permite utilizar toda la información disponible, aprovechando información *a priori* que se tenga, investigaciones anteriores, conocimiento subjetivo, información *a priori* dogmática y muestral; proporcionando de esta manera una muy buena herramienta para la toma de decisión en situación de riesgo.

Al contar con más información *a priori*, los estimadores obtenidos con la metodología *Bayesiana* serán más precisos (menor variancia).

Por otro lado, la metodología *Bayesiana* permite contrastar la información *a priori* y *a posteriori* y el *factor de Bayes*, todos ellos en términos de *probabilidades*. De ahí la facilidad de su interpretación.

La implementación del análisis *Bayesiano*, nos permite formalizar en un modelo estadístico la información no muestral sobre los parámetros de un modelo. La información no muestral puede corresponder a juicios de valor, argumentos teóricos, etc, no representados en los datos disponibles. Por esta razón, esta forma de pensar puede enriquecer considerablemente el universo metodológico y pragmático del análisis de toma de decisiones en problemas de inversión, o en cualquier otro problema de incertidumbre bajo riesgo.

Además, al tratar parámetros y otras fuentes de incertidumbre como variables aleatorias, el análisis *Bayesiano* permite aprehender de una manera más precisa los riesgos (*pérdida esperada*) de cada acción de inversión. En este sentido, la interpretación *Bayesiana* del concepto de *probabilidad* permite responder a preguntas tales como, cuál es la *probabilidad* de que una u otra sea la consecuencia de una decisión de inversión.

---

## Síntesis y Reflexión Final

---

La controversia entre la estadística *Clásica* y *Bayesiana*, reside en la interpretación del concepto de *probabilidad*, lo que ha generado un debate continuo desde la publicación del *Teorema de Bayes* en 1764, hasta nuestros días. La distinción fundamental entre una inferencia y la otra es que, en la *Bayesiana*, los parámetros de interés son *variables aleatorias*, luego, tienen tanto *a priori* como *a posteriori distribuciones*. Es decir, en la inferencia clásica los parámetros tienen valores únicos, aunque estos valores son desconocidos, y esto no permite tratarlos como aleatorios, por lo tanto no se le puede asignar una distribución de *probabilidades*.

Del estudio efectuado en el ámbito de la teoría *Bayesiana*, consideramos que se pueden sintetizar las siguientes conclusiones de interés. Estando dichas conclusiones muy interrelacionadas las dividiremos, para una mayor claridad expositiva, del siguiente modo:

- a) El enfoque *Bayesiano* justifica el uso del conocimiento subjetivo del investigador, en cualquier herramienta estadística. En consecuencia, esta metodología aprovecha todas las fuentes de información; a saber: Información *a priori* proveniente de investigaciones anteriores, información dogmática y muestral.
- b) Cuando la información *a priori* es no informativa, la metodología *Bayesiana* y *Clásica* proponen resultados similares cuantitativos, mas no cualitativos.

- c) Dado que con el uso de distribuciones *a priori conjugadas*, los resultados *Bayesianos* difieren de los obtenidos con la metodología Clásica, los estudiosos del tema sugieren que el investigador debe ser muy cuidadoso en su elección.
- d) La metodología *Bayesiana* permite contrastar la información *a priori* con la información muestral mediante la comparación de los *Odds a priori* y *a posteriori* y el *factor de Bayes*, todos en términos de *probabilidad*.
- e) Al adicionar información sobre los parámetros de interés, los estimadores obtenidos con metodología *Bayesiana* tendrán menor variancia, lo que representa un estimación más precisa.
- f) El enfoque *Bayesiano* en regresión es aplicable en cualquier caso en el que la metodología Clásica sea aplicable.
- g) La inferencia *Bayesiana* se basa exclusivamente en términos de *probabilidades a posteriori*. De ahí la facilidad de su interpretación.

Por último, la aplicación de la metodología *Bayesiana* a la teoría de la *Decisión*, refuerza el valor de dicha teoría. En problemas de investigación científica, la situación común es que, la distribución *a priori* es dominada por la *verosimilitud*. Sin embargo, en la generalidad de los problemas de *Decisión*, la distribución *a priori* puede ser influyente en la determinación de la distribución *a posteriori*. El hecho que en tales situaciones las decisiones diferentes puede ser el resultado de las opciones diferentes de distribuciones *a priori*, puede ser una ventaja para el análisis *Bayesiano*, ya que éste combina moderadamente la experiencia con la intuición para llegar a una decisión llamada “óptima”.

El método *Bayesiano*, permite formalizar en un modelo estadístico la información no muestral sobre los parámetros de un modelo. La información no muestral puede corresponder a juicios de valor, argumentos teóricos, etc, no representados en los datos disponibles. Por esta razón, dice Hossein Arsham, esta forma de pensar puede enriquecer considerablemente el universo metodológico y pragmático del análisis de *toma de decisiones* en problemas de incertidumbre bajo riesgo.





---

## Bibliografía consultada

---

La literatura sobre estadística *Bayesiana*, en los últimos años, es vasta. Aquí, listamos y comentamos los libros que utilizamos y consultamos en la elaboración de este trabajo, incluyendo referencias de libros *no Bayesianos* que permitieron realizar una comparación con la estadística *Bayesiana*.

En primer lugar el libro de *Gelman* (6), el cual consideramos que es una excelente obra sobre teoría *Bayesiana*, nos proporcionó la notación utilizada aquí, es importante destacar el lenguaje accesible que utiliza, en cuanto a las interpretaciones en lenguaje coloquial que realiza de la simbología estadística, además de realizar permanentemente recomendaciones sobre las herramientas que se deben utilizar según el tipo de problema que en la práctica pueden surgir.

Asimismo, la obra de *Box* (4), trata el tema de la inferencia concerniente a la media y la dispersión de una distribución Normal  $N(\theta, \sigma)$  de una forma exhaustiva y define los intervalos HPD estandarizados para  $\sigma$ .

En el libro de *O'Hagan* (17), destacamos el análisis comparativo que realiza entre la inferencia *Bayesiana* y la Clásica; como así también la introducción que realiza a la teoría de la *Decisión*.

La inferencia *Bayesiana* para los parámetros del Modelo Normal, es ampliamente tratado por *Berger* (2). Como así también el problema de la selección de una

distribución *a priori* para el modelo de regresión lineal múltiple. Este mismo autor presenta de un modo ameno la teoría de la *Decisión* con metodología *Bayesiana*.

El libro de *Lee* (13), presenta un lenguaje sencillo para explicar la *distribución predictiva*.

Para el tema del modelo de regresión lineal simple con errores normales e independientemente distribuidos con media 0 y variancia  $\sigma^2$  destacamos el libro de *Press* (18).

*Levin* (14) y *Koontz* (12), nos permitieron acercarnos a los elementos que configuran el acto de decidir en la ciencia de la economía y administración. Y *Mendenhall* (16), nos proporcionó ejemplos simples y didácticos de las mismas desde una visión *Bayesiana*.

Fundamentalmente, a través de los escritos de *Hossein Arsham* (9), sobre teoría de la *Decisión*, pudimos construir la aplicación de la metodología *Bayesiana* al problema de inversión expuesto en el Capítulo IV.

Así también, los textos de *Guido del Pino* (7) y de *Iglesias Zuazola* (10) resultan ser una importante alternativa de consulta para la interpretación subjetiva de *probabilidad*.

Por último, destacamos la obra de *Berry* (3), porque consideramos que es un tratado didáctico acabado para enseñar la estadística *Bayesiana* en el ámbito Universitario, en carreras de grado.

- 
- 1- Aaker, D; Day, G. 1995. *Investigación de Mercados*. México. McGraw Hill.
  - 2- Berger, J. O. 1988. *Statistical Decision Theory and Bayesian Analysis*. Segunda edición. Springer-Verlag, New York.
  - 3- Berry, Donald A. 1996. *Statistics A Bayesian Perspective*. U.S.A: Duxbury Press.
  - 4- Box, G; Tiao, G.C. 1992. *Bayesian inference in Statistical Analysis*. New York. Wiley
  - 5- Canavos, G. 1984. *Probabilidad y Estadística. Aplicaciones y métodos*. MacGraw-Hill.

- 6- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 1998. *Bayesian Data Analysis*, Segunda edición. Chapman & Hall, Londres
- 7- Guido del Pino, M. 1995. *Estadística teoría y métodos*. Chile: Editorial Universitaria S.A.
- 8- Harnett, D; Murphy, J. L. 1987. *Introducción al Análisis Estadístico*. U.S.A.: Addison - Wesley Iberoamericana, S.A.
- 9- Hossein Arsham. 1994. *Herramientas para el Análisis de Decisión: Análisis de Decisiones Riesgosas*. Papers publicado en version electrónica:  
URL:<http://www.mirror-service.org/sites/home.ubalt.edu/ntsbarsh/Business-stat>
- 10- Iglesias Zuazola, P. 1995. *Fundamentos de Inferencia Bayesiana: Teorema del tipo de Finetti*. Universidad Católica de Chile.
- 11- Iversen, Gudmund R. 1984. *Bayesian Statistical Inference*. U.S.A: Sage Publications Ltd.
- 12- Koontz-Wehrich. 1990. *Administración*. México- McGraw Hill-
- 13- Lee Peter M. 1997. *Bayesian Statistics: An Introduction*. Editorial Arnold Gran Bretaña Segunda edición.
- 14- Levin, Richard. 1998. *Estadística para Administradores*. México .Prentice Hall-
- 15- Maddala, G.S. 1985. *Econometría*. MacGraw-Hill.
- 16- Mendenhall, William.1990.*Estadística para Administradores*. Grupo editorial Iberoamérica.
- 17- O'Hagan, A. 1994. *Kendall's Advanced Theory of Statistics (volumen 2B). Bayesian Inference*. New York: Wiley.
- 18- Press, S.J. 1989. *Bayesian Statistics: Principles, Models and Applications*. John Wiley & Sons. New York.

## Referencia sobre la vida y trabajos de Bayes:

---

Finalmente, creemos que es importante conocer las referencias históricas acerca de la vida de *Bayes*. La historia en sentido amplio, nos permite comprender la realidad que nos rodea, y en sentido estricto, la historia de vida de los grandes autores, como sin lugar a dudas es *Bayes*, nos proporciona una idea del impacto cultural que su teoría representó y representa en la estadística.

La biografía de *Bayes* fue traducida del artículo publicado en versión electrónica. URL. <http://www.history.mes.st-andrews.ac.uk/history/mathematicians/bayes.html>.

Para profundizar la historia del *Teorema de Bayes* y discusiones sobre su impacto en la estadística listamos algunas obras de autores reconocidos y citados por *Gelman* (6) como referencias:

- Barnard. 1958. *Thomas Bayes. A biographical note*, *Biometrika* 45,293-295.
- Pearson. 1978. *The history of statistics in the 17<sup>th</sup> and 18<sup>th</sup> centuries*. London: Macmillan.
- Dale. 1991. *History of inverse probability: from Thomas Bayes to Karl Pearson*. Berlin: Springer.

U.N.R.C.  
Biblioteca Central



63385

63385



U.N.R.C.  
Biblioteca Central



63395