Estadística Bayesiana

Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

February 13, 2022

Índice

1 Introducción al enfoque Bayesiano **1**

1.1 Introducción 1

1.2 Limitaciones de la estadística frecuentista 2

1.3 Enfoque de la Probabilidad 3

1.4 Teorema de Bayes 4

1.5 Teoría de decisión 6

1.6 Paradigma Bayesiano 6

2 Paradigma Bayesiano **8**

2.1 Distribución a priori 8

2.2 Análisis Conjugado 10

2.3 Predicción 15

2.3.1 Actualización secuencial de la información 23

3 Teoría de decisiones **24**

3.1 Fundamentos 24

3.2 Cómo debe tomarse una decisión trascendente? 25

3.3 Axiomas de coherencia 27

3.4 Utilidad esperada 29

3.4.1 Criterio General de Decisión 29
3.5 Otros criterios de decisión 38
3.5.1 Criterio Pesimista: 39
3.5.2 Criterio Optimista: 39
3.5.3 Criterio de Laplace: 40

4 De la información a priori a la distribución *a priori* 41

4.1 Introducción 41
4.2 Información histórica 41
4.3 Distribuciones poco informativas y muy informativas 42
4.3.1 Familia conjugada 43
4.4 Distribuciones No informativas 45
4.4.1 Prior Laplace 46
4.4.2 Familia Conjugada No Informativa o Poco Informativa 47
4.4.3 Criterio de Jeffreys 48
4.5 Análisis de Sensibilidad - Robustez 53

Capítulo 1

Introducción al enfoque Bayesiano

1.1 Introducción

En los últimos años ha habido un mayor interés en el desarrollo de métodos de inferencia Bayesiana. La razón principal es que la metodología Bayesiana proporciona un paradigma completo para la inferencia estadística bajo incertidumbre, que permite combinar información derivada de observaciones con información obtenida de expertos (Berger, 1985; Bernardo and Smith, 1994; Robert, 1994). Una gran ventaja del enfoque Bayesiano es que se puede usar información inicial, la cual se incorpora al modelo a través de distribuciones de probabilidad (O'Hagan et al., 2006). Esto puede ser muy útil en situaciones específicas donde se puede obtener conocimiento de expertos o información histórica.

El paradigma Bayesiano ha sido reconocido desde hace mucho tiempo como conceptualmente atractivo, sin embargo, en la práctica su implementación puede ser compleja debido a las dificultades de cálculo. Esencialmente se requiere integración en altas dimensiones para calcular distribuciones de probabilidad. Recientemente, los métodos Bayesianos se han vuelto populares con la aparición de nuevos algoritmos computacionales que abordan esta integración de manera directa. El desarrollo de métodos como los de Monte Carlo vía cadenas de Markov (MCMC) han permitido dar soluciones numéricas para problemas basados en modelos verdaderamente complejos (Gilks et al., 1996; Chen et al., 2000; Gamerman and Lopes, 2006). Algunas veces se incluyen en el modelo variables aleatorias auxiliares o variables latentes para facilitar el proceso de generación, a pesar del hecho de que la dimensionalidad aumenta (Tanner and Wong, 1987; Tan et al., 2010). Los métodos MCMC han sido decisivos en el crecimiento de la literatura Bayesiana en general (Chen et al., 2000; Gamerman and Lopes, 2006) y, en particular, para el desarrollo de nuevos métodos para el análisis de datos categóricos (Johnson and Albert, 1999; Congdon, 2005) y modelos lineales generalizados (Dey et al., 2000). Recientemente ha habido un desarrollo en la estimación de modelos Bayesiados usando el método INLA (*Integrated Nested Laplace Approximations*)

(Wang et al., 2018; Gomez-Rubio, 2020), el cual es una alternativa para los métodos MCMC, resultando ser más rápidos.

La implementación de los algoritmos juega un papel fundamental para la metodología Bayesiana. Para implementar los algoritmos, el software debe ser lo suficientemente flexible y potente para realizar los cálculos necesarios. El software `R` es uno de los lenguajes de programación más utilizados para el cálculo y gráficas en estadística (R Development Core Team, 2008; Albert, 2009). Otro proyecto interesante es BUGS (*Bayesian inference Using Gibbs Sampling*), el cual sirve para análisis Bayesiano de modelos complejos utilizando métodos MCMC. WinBUGS y su versión de código abierto OpenBUGS son parte de este proyecto y son ampliamente utilizados por la comunidad Bayesiana (Lunn et al., 2000; Ntzoufras, 2009; Lunn et al., 2012). Y recientemente desarrollado los paquetes JAGS (Plummer, 2003) y STAN (Stan Development Team, 2015).

1.2 Limitaciones de la estadística frecuentista

Algunas limitaciones de la estadística frecuentista, y que le dan una posible ventaja a la estadística Bayesiana, son:

- No permite incorporar en el análisis estadístico información extra muestral disponible.
- Muchos de los métodos se apoyan en resultados asintóticos, la ley de los grandes números, el teorema central del límite, etc.
- Se necesitan muestras grandes para que los resultados sean confiables.
- Si no hay datos, la estadística frecuentista no es posible.
- Si hay pocos datos, la estadística frecuentista presenta problemas.
- La estadística Bayesiana usa información disponible a partir de los datos de muestras, y a partir de la información extra muestral.
- La estadística frecuentista solo usa dos hipótesis, mientras que en la Bayesiana es posible considerar k hipótesis a la vez.

1.3 Enfoque de la Probabilidad

Estadística es un conjunto de técnicas para describir un fenómeno, a partir de un conjunto de datos que presentan variabilidad. Es un conjunto de métodos para alcanzar conclusiones acerca de una o varias características de interés de una población a partir de información parcial provista por una muestra de dicha población.

La estadística se ocupa de los métodos científicos para recolectar, organizar, resumir, presentar y analizar datos usando modelos, así como de obtener conclusiones válidas y tomar desiciones con base en ese análisis. La estadística es la rama de la matemática que utiliza conjuntos de datos para obtener inferencias basadas en el cálculo de probabilidades.

La probabilidad tiene diferentes enfoques.

- *Enfoque clásico.* Sea $(\Omega, \mathcal{F}, \mathbb{P})$ espacio de probabilidad, la probabilidad de que ocurra el evento A se define como:
$$\mathbb{P}(A) = \frac{|A|}{n},$$
donde $|A|$ denota la cardinalidad de A. Si un experimento o un fenómeno puede ocurrir de n maneras diferentes, mutuamente excluyentes e igualmente probables.

  Ej: lanzar un dado. Sea A el evento de obtener un 3, donde $n = 6$, entonces $\mathbb{P}(A) = \frac{1}{6}$.

- *Enfoque frecuentista.* La probabilidad de que ocurra el evento A se define como:
$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{f_A(n)}{n},$$
donde $f_A(n)$ es el número de veces que ocurre el evento A en n repeticiones idénticas e independientes. Se basa en datos observados. El límite $\lim_{n\to\infty} \frac{f_A(n)}{n}$ equivale a decir que $\forall \epsilon > 0$ existe k tal que si $n > k$ entonces $\left| \frac{f_A(n)}{n} - \mathbb{P}(A) \right| < \epsilon$.

  Ej: lanzar un dado. Sea A el evento de obtener un 3. Para calcular $\mathbb{P}(A)$ se lanzan $n = 600$ veces el dado, y se observa que A ocurre $f_A(600) = 106$ veces, entonces $\mathbb{P}(A) = \frac{106}{600}$.

- *Enfoque subjetivo.* La probabilidad de que ocurra el evento A es una medida del grado de creencia que tiene un individuo de la creencia de A, con base en la información K que dicho individuo posee, es decir,
$$\mathbb{P}(A) = \mathbb{P}(A|K).$$
  Ej: sea A el evento de que hoy llueva en la ciudad de México. Considerando que tenemos dos posibles escenarios de información: K_1 es un individuo que vive en Wuhan (China) y No tiene acceso a internet (No tiene información); K_2 es un individuo que vive en la Ciudad de México (tiene información, puede dar un vistazo).

1.4 Teorema de Bayes

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ espacio de probabilidad. Sean $A, B \in \mathcal{F}$, tal que $\mathbb{P}(B) > 0$. La probabilidad condicional del evento A dado el evento B se define como

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ espacio de probabilidad. Sean $A, B \in \mathcal{F}$, tal que $\mathbb{P}(A) > 0$. El teorema de Bayes toma la forma:

$$\begin{aligned} \mathbb{P}(B|A) &= \frac{\mathbb{P}(B)\mathbb{P}(A|B)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} \end{aligned}$$

De manera general, sea $\{B_1, \ldots, B_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos contenidos en $\mathcal{F}$, tales que $\mathbb{P}(B_j) > 0 \ \forall j = 1, \ldots, n.$. Sea $A \in \mathcal{F}$ un evento tal que $\mathbb{P}(A) > 0$. La Regla de Bayes se define como:

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

Ejemplo 1.1 (Teorema de Bayes: COVID-19)**.** Suponga que un grupo de ingenieros biomédicos han diseñado una prueba para el diagnóstico del COVID-19. Suponga que se elige a un individuo para la prueba, y consideremos los eventos de interés:

A	el evento de que la prueba diagnostique presencia de COVID-19
B	el evento de tener efectivamente COVID-19

Los científicos presumen de que la prueba es muy buena. Probaron con un grupo de 100 portadores de COVID-19 y el 99% dio positivo: $\mathbb{P}(A|B) = \frac{99}{100}$

Probaron con un grupo de 100 sanos y el 99% dio negativo: $\mathbb{P}(A^c|B^c) = \frac{99}{100}$

Sea p la proporción de mexicanos con COVID-19: $p = \mathbb{P}(B)$. Utilizando la regla de Bayes, buscamos $\mathbb{P}(B|A) = \mathbb{P}($ *tener COVID-19* $|$ *diagnóstico positivo* $)$.

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{\frac{99}{100}p}{\frac{99}{100}p + \frac{1}{100}(1-p)}$$

Tabulando, para distintos valores de p:

p	$\mathbb{P}(B\|A)$
0.0001	0.0098
0.001	0.0902
0.01	0.5000
0.1	0.9167
0.5	0.9900

¿Qué sucedió? La prueba fue probada con personas que conocíamos su estado de salud. En la práctica esto NO sucede.

> *Existen 2 tipos de pruebas de utilidad diagnóstica: las basadas en la detección del virus (RNA o antígeno viral) y las basadas en la detección de anticuerpos (IgM o IgG) frente al virus.* [1]
>
> *La tasa de detección positiva aumenta significativamente (98.6%) cuando se combina la IgM con PCR para cada paciente en comparación con una sola prueba (Guo L, 2020).*

Repetir la prueba de manera independiente:

$$\begin{aligned}
\mathbb{P}(B|AA) &= \frac{\mathbb{P}(AA|B)\mathbb{P}(B)}{\mathbb{P}(AA|B)\mathbb{P}(B)+\mathbb{P}(AA|B^c)\mathbb{P}(B^c)} \\
&= \frac{\mathbb{P}(A|B)\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(A|B)\mathbb{P}(B)+\mathbb{P}(A|B^c)\mathbb{P}(A|B^c)\mathbb{P}(B^c)} \\
&= \frac{\frac{99}{100}\frac{99}{100}p}{\frac{99}{100}\frac{99}{100}p+\frac{1}{100}\frac{1}{100}(1-p)}
\end{aligned}$$

`Código R:` Bayes1_1ReglaBayes.R

p	$\mathbb{P}(B\|AA)$
0.0001	0.495
0.001	0.907
0.01	0.990
0.1	0.999
0.5	1.000

[1] `https://www.fisterra.com/guias-clinicas/covid-19/`

1.5 Teoría de decisión

El *enfoque Bayesiano* se basa en diseñar una teoría estadística basada en una pequeña serie de principios básicos (axiomas) que permiten estructurar la solución a cualquier problema de inferencia usando teoría de decisión.

En el contexto de Estadística, los elementos de un problema de decisión en ambiente de incertidumbre son:

- Espacio de acciones potenciales disponibles $\mathcal{A}$.
- Espacio parametral Θ, que contiene los posibles estados de la naturaleza.
- Espacio de consecuencias $\mathcal{C} = \mathcal{A} \times \Theta$.

Para resolver un problema de decisión es necesario cuantificar la incertidumbre sobre Θ como las consecuencias en $\mathcal{C}$.

Los axiomas implican que la única forma racional de cuantificar incertidumbre es a través de una medida de probabilidad $f(\theta)$ y que las consecuencias deben cuantificarse por medio de una función de pérdida $\mathcal{L}(a, \theta)$.

La mejor acción será la que minimice la pérdida esperada final:

$$\mathcal{L}_x^*(a) = \int_\Theta \mathcal{L}(a, \theta) f(\theta|x) \mathrm{d}\theta.$$

1.6 Paradigma Bayesiano

Sea $X_1, \ldots, X_n$ una muestra de variables aleatorias (v.a.) de una función de densidad de probabilidad $P(X|\boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, $\Theta \subset \mathbb{R}^d$, donde $\boldsymbol{\theta}$ es el vector de parámetros y Θ es el espacio paramétrico, cuya función de densidad conjunta o verosimilitud es $P(\boldsymbol{X}|\boldsymbol{\theta})$. El problema en estadística se reduce a hacer inferencias sobre el supuesto valor de $\boldsymbol{\theta}$.

En estadística Bayesiana (Robert, 1994; Berger, 1985; Bernardo and Smith, 1994) se modela la incertidumbre de $\boldsymbol{\theta}$ usando métodos probabilísticos, y se considera $\boldsymbol{\theta}$ como aleatoria, en específico, se usa el Teorema de Bayes:

$$P(\boldsymbol{\theta}|\boldsymbol{X}) \;=\; \frac{P(\boldsymbol{\theta})P(\boldsymbol{X}|\boldsymbol{\theta})}{P(\boldsymbol{X})},$$

donde $P(\boldsymbol{X}) = \int_\Theta P(\boldsymbol{\theta})P(\boldsymbol{X}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ no depende de $\boldsymbol{\theta}$, por lo que es común escribir

$$P(\boldsymbol{\theta}|\boldsymbol{X}) \;\propto\; P(\boldsymbol{\theta})P(\boldsymbol{X}|\boldsymbol{\theta}).$$

La distribución de probabilidad inicial (*a priori*) $P(\boldsymbol{\theta})$ describe la información inicial que se tiene de $\boldsymbol{\theta}$. Esta distribución se basa en experiencia previa (información histórica, experiencia de expertos en los datos u otra información adicional).

Entonces, la distribución de probabilidad final (*a posteriori*) $P(\boldsymbol{\theta}|\boldsymbol{X})$ permite incorporar información contenida en la muestra a través de $P(\boldsymbol{X}|\boldsymbol{\theta})$ e información inicial a través de $P(\boldsymbol{\theta})$. Por tanto, en estadística Bayesiana las inferencias sobre el parámetro $\boldsymbol{\theta}$ se basan en calcular la distribución final $P(\boldsymbol{\theta}|\boldsymbol{X})$.

En ocasiones el propósito de un análisis estadístico es hacer predicciones de una observación futura X^*. La distribución de probabilidad que describe el comportamiento de una observación futura X^* es $P(X^*|\boldsymbol{\theta})$, sin embargo $\boldsymbol{\theta}$ es desconocido, así que es necesario estimarlo.

Desde la perspectiva Bayesiana, la distribución marginal de X^*

$$P(X^*) \;=\; \int_{\Theta} P(X^*|\boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

conocida como distribución predictiva inicial (*predictiva a priori*), describe la información acerca de X^* dada la información inicial disponible por $P(\boldsymbol{\theta})$.

Una vez obtenida la muestra $\boldsymbol{X}$, la distribución final del parámetro $P(\boldsymbol{\theta}|\boldsymbol{X})$ junto con el modelo $P(X^*|\boldsymbol{\theta})$, inducen un distribución conjunta de $(X^*, \boldsymbol{\theta})$ condicional en los valores observados

$$P(X^*, \boldsymbol{\theta}|\boldsymbol{X}) \;=\; P(X^*|\boldsymbol{\theta}, \boldsymbol{X})P(\boldsymbol{\theta}|\boldsymbol{X}) \;=\; P(X^*|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{X}),$$

donde la igualdad $P(X^*|\boldsymbol{\theta}, \boldsymbol{X}) = P(X^*|\boldsymbol{\theta})$ se cumple siempre que haya independencia condicional entre la observación futura X^* y la muestra observada $\boldsymbol{X}$. Por tanto, la distribución de probabilidad

$$P(X^*|\boldsymbol{X}) \;=\; \int_{\Theta} P(X^*|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{X})\mathrm{d}\boldsymbol{\theta},$$

conocida como distribución predictiva final (*predictiva a posteriori*), describe el comportamiento de X^* dada toda la información disponible de $\boldsymbol{X}$ y de $\boldsymbol{\theta}$.

Capítulo 2

Paradigma Bayesiano

Estudiamos conceptos básicos del paradigma Bayesiano, análisis conjugado, y predicción.

Para más detalles se puede revisar Berger (1985), Bernardo and Smith (1994), Box and Tiao (1973), Hoff (2009), Gelman et al. (2004), Robert (1994), Robert (2007).

2.1 Distribución a priori

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución desconocida F. En Estadística es común considerar una familia paramétrica de densidades,

$$\mathcal{P} = \{p(x|\theta) : \theta \in \Theta\},$$

y proceder como si la distribución F correspondiera a alguno de los modelos en $\mathcal{P}$. De esta manera, el problema se reduce a hacer inferencias sobre el supuesto valor del parámetro θ que corresponde al "modelo verdadero". Desde el punto de vista Bayesiano, la información previa sobre el valor desconocido de θ se describe a través de una *distribución inicial* (*prior distribution*) $p(\theta)$. El teorema de Bayes,

$$p(\theta|x_1, \ldots, x_n) = \frac{p(\theta)p(x_1, \ldots, x_n|\theta)}{\int p(\theta)p(x_1, \ldots, x_n|\theta)d\theta},$$

permite entonces incorporar la información contenida en la muestra, produciendo una descripción de la incertidumbre sobre el valor del parámetro a través de la *distribución final* (*posterior distribution*) $p(\theta|x_1, \ldots, x_n)$.

Es común escribir el teorema de Bayes como

$$p(\theta|x_1, \ldots, x_n) \propto p(\theta)\mathcal{L}(\theta|x_1, \ldots, x_n),$$

donde

$$\mathcal{L}(\theta|x_1,\ldots,x_n) \propto p(x_1,\ldots,x_n|\theta),$$

es la *función de verosimilitud* (*likelihood*).

Ejemplo 2.1 (Proporción de enfermos)**.** Se examina una población, y θ es la proporción de enfermos. Un día, se examinan 10 sujetos, denotados por $X_1,\ldots,X_{10}$, donde $X_i = 1$ indica que el sujeto i está enfermo y $X_i = 0$ está sano. Esto puede verse como una muestra aleatoria con distribución Bernoulli de parámetro θ, con fdp $f_X(x;\theta) = \theta^x(1-\theta)^{1-x}I_{\{0,1\}}(x)$ para $0 \leq \theta \leq 1$.
¿Cuál es el valor de θ?

Personas enfermas: $X_i \sim \text{Bernoulli}(\theta)$, con $\theta \in [0,1]$.

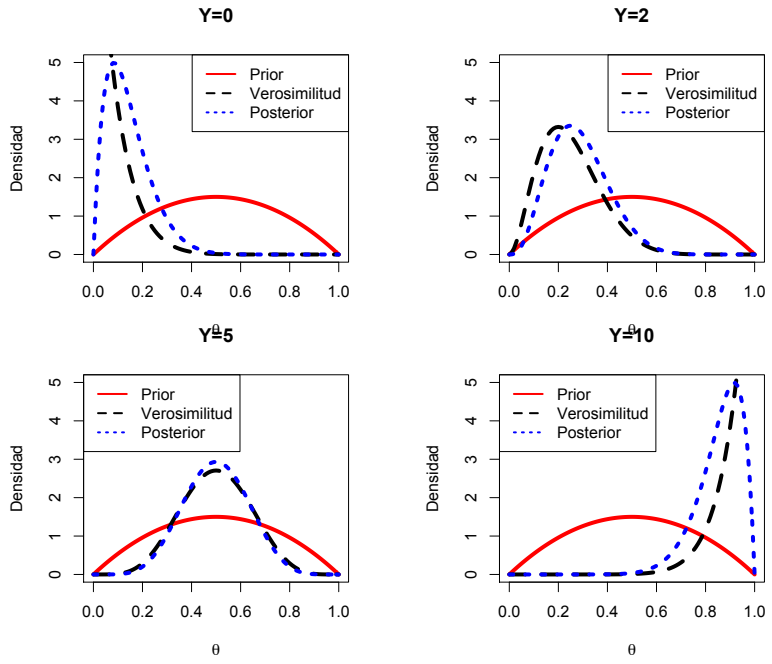
Suponga que el experto tiene información inicial acerca de θ, y que ha observado que a lo largo de los días la proporción de enfermos cambia como una v.a. con fdp $\pi(\theta) = 6\theta(1-\theta)I_{[0,1]}(\theta)$, es decir, $\theta \sim Beta(2,2)$ (distribución a priori).

Entonces la distribución a posteriori de θ es:

$$\begin{aligned} P(\theta|x_1,\ldots,x_n) &= \frac{P(\boldsymbol{\theta})P(x_1,\ldots,x_n|\boldsymbol{\theta})}{P(x_1,\ldots,x_n)} \\ &= \frac{\prod_{i=1}^{10}\theta^{x_i}(1-\theta)^{1-x_i} \times 6\theta(1-\theta)}{\int_0^1 \prod_{i=1}^{10}\theta^{x_i}(1-\theta)^{1-x_i} \times 6\theta(1-\theta)d\theta} \\ &\propto \theta^{\sum_{i=1}^{10}x_i+1}(1-\theta)^{10-\sum_{i=1}^{10}x_i+1} \end{aligned}$$

que tiene una densidad $Beta\left(\sum_{i=1}^{10}x_i + 2,\ 12 - \sum_{i=1}^{10}x_i\right)$.

`Código R:` Bayes2_1DistribucionBernoulli.R



2.2 Análisis Conjugado

Tanto $p(\theta)$ como $p(\theta|x_1, \ldots, x_n)$ son distribuciones de probabilidad sobre el parámetro θ. La primera distribución sólo describe la información inicial y la segunda distribución actualiza dicha información usando también la información muestral que se pueda obtener. Resulta conveniente tanto para el análisis como desde el punto de vista computacional, que $p(\theta)$ y $p(\theta|x_1, \ldots, x_n)$ pertenezcan a la misma familia paramétrica.

Definición 2.1. Sea $\mathcal{P} = \{p(x_1, \ldots, x_n|\theta) : \theta \in \Theta\}$ una familia paramétrica. Una clase (o colección) de distribuciones de probabilidad $\mathcal{F}$ es una familia conjugada para $\mathcal{P}$ si para todo $p(x_1, \ldots, x_n|\theta) \in \mathcal{P}$ y $p(\theta) \in \mathcal{F}$ se cumple que $p(\theta|x_1, \ldots, x_n) \in \mathcal{F}$.

Para garantizar que $p(\theta)$ y $p(\theta|x_1, \ldots, x_n)$ pertenezcan a la misma familia general de funciones de distribución, se elige a $p(\theta)$ de tal manera que tenga la misma "estructura"de $p(x_1, \ldots, x_n|\theta)$ vista como una función de θ.

A continuación se estudiarán casos particulares de familias conjugadas de las distribuciones más usadas en el análisis de datos categóricos y tablas de contingencia.

Ejemplo 2.2 (Familia conjugada: Bernoulli)**.** Muchas aplicaciones hacen referencia a un número fijo n de observaciones binarias. Sean $y_1, \ldots, y_n$ los resultados de n ensayos independientes e idénticos tal que $p(Y_i = 1) = \pi$ y $p(Y_i = 0) = 1 - \pi$. Generalmente se etiquetan como "éxitos" y "fracasos" los resultados 1 y 0 respectivamente. Las variables aleatorias $\{Y_i\}$

independientes e idénticamente distribuidas se conocen como variables aleatorias Bernoulli. El número total de éxitos, $Y = \sum_{i=1}^{n} Y_i$, tiene una distribución binomial con índice n y parámetro π, y se denota por $Bin(y|n, \pi)$. La función de masa de probabilidad de Y es

$$p(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \qquad y = 0, 1, \ldots, n,\ 0 < \pi < 1,$$

donde el coeficiente binomial es

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}.$$

La verosimilitud es

$$\mathcal{L}(\pi|y) \propto \pi^y(1-\pi)^{n-y}.$$

Una distribución inicial conjugada para una distribución binomial es la distribución beta con parámetros α y β (ambos positivos), $\pi \sim Beta(\pi|\alpha, \beta)$, cuya función de densidad de probabilidad de π está dada por

$$p(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad 0 < \pi < 1,\ \alpha > 0,\ \beta > 0,$$

donde $\Gamma(\cdot)$ es la función gamma dada por

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

Note que una distribución inicial simétrica sobre π se obtiene haciendo $\alpha = \beta$ y cuando $\alpha = \beta = 1$ se reduce a una distribución inicial uniforme. La distribución final de π también es una distribución beta, con parámetros $\alpha + y$ y $\beta + n - y$, cuya función de densidad de probabilidad es

$$p(\pi|y, n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}\pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1}, \qquad 0 < \pi < 1.$$

■

Ejemplo 2.3 (Familia conjugada: Binomial)**.** Suponga que $Y_1, \ldots, Y_G$ son variables aleatorias independientes con distribución $Bin(y_i|n_i, \pi)$ para $i = 1, \ldots, G$. Entonces, si la distribución inicial para π es $Beta(\pi|\alpha, \beta)$, con α y β conocidas, la distribución final de π es

$$\pi \sim Beta\left(\pi|\alpha + \sum_{i=1}^{G} y_i, \beta + \sum_{i=1}^{G} n_i - \sum_{i=1}^{G} y_i\right).$$

■

Ejemplo 2.4 (Familia conjugada: Poisson)**.** Algunas veces los datos de conteo no resultan de un número fijo de ensayos y su rango son los números enteros no negativos. En estos casos, y bajo ciertas condiciones, un modelo que puede utilizarse es la distribución Poisson.

Sea Y una variable aleatoria Poisson con parámetro $\mu > 0$; entonces su función de masa de probabilidad es

$$p(Y = y) = e^{-\mu}\frac{\mu^y}{y!}, \qquad y = 0, 1, \ldots, \ \mu > 0.$$

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ son variables aleatorias con distribución Poisson, con media y varianza común μ. La distribución inicial de μ se puede tomar como una distribución gamma con parámetros α y β (ambos positivos), con media α/β, y cuya función de densidad de probabilidad está dada por

$$p(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}, \qquad \mu > 0, \ \alpha > 0, \ \beta > 0.$$

La verosimilitud de la distribución Poisson es

$$\mathcal{L}(\mu|y_1, \ldots, y_n) = \mathcal{L}(\mu|\mathbf{y}) \propto \prod_{i=1}^{n} e^{-\mu}\mu^{y_i}.$$

La distribución final de μ es

$$\begin{aligned} p(\mu|\mathbf{y}) &\propto \left[\prod_{i=1}^{n} \exp(-\mu)\mu^{y_i}\right] \mu^{\alpha-1}\exp(-\beta\mu) \\ &= \mu^{\alpha+\sum_{i=1}^n y_i - 1}\exp[-\mu(\beta + n)], \end{aligned}$$

de tal manera que la distribución final para μ es también una distribución gamma, específicamente,

$$p(\mu|y) = Gamma\left(\mu|\alpha + \sum_{i=1}^{n} y_i, \beta + n\right).$$

Debido a que la distribución inicial y la final pertenecen a la misma familia de distribuciones, entonces la familia de distribuciones gamma es conjugada para la familia de distribuciones Poisson. ■

Ejemplo 2.5 (Familia conjugada: Multinomial)**.** Algunos ensayos tienen más de dos posibles categorías. Suponga que cada uno de N ensayos idénticos e independientes pueden clasificarse en cualquiera de k categorías; en el caso binomial $k = 2$. Sea $y_{ij} = 1$ si el ensayo i se clasifica en la categoría j y sea $y_{ij} = 0$ en otro caso. Entonces $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ representa un ensayo multinomial con $\sum_j y_{ij} = 1$; por ejemplo, $(0, 0, 1, 0)$ denota una clasificación en la categoría 3 de cuatro posibles categorías. Note que y_{ik} es redundante, debido a que es linealmente dependiente de y_{ij} con $j = 1, \ldots, k-1$.

Sea $n_j = \sum_i y_{ij}$ el número de ensayos que se clasifican en la categoría j. Se dice que los conteos $(n_1, n_2, \ldots, n_k)$ tienen una distribución multinomial. Sea $\pi_j = p(Y_{ij} = 1)$ la probabilidad de que se clasifique en la categoría j, con $j = 1, \ldots, k$, para cada ensayo; necesariamente $0 < \pi_j < 1 \ \forall j$, y $\sum_{j=1}^k \pi_j = 1$.

La función de probabilidad de la distribución multinomial está dada por

$$p(n_1,\ldots,n_k|\pi_1,\ldots,\pi_k,N) = N!\prod_{j=1}^{k}\frac{\pi_j^{n_j}}{n_j!}, \qquad n_j = 0,1,\ldots,\ 0<\pi_j<1,$$

con $N=\sum_{j=1}^{k} n_j$ y $\sum_{j=1}^{k}\pi_j = 1$.

Note que si $n_j,\ j=1,\ldots,k$, fueran variables independientes Poisson con medias $\mu_j,\ j=1,\ldots,k$, entonces su distribución condicional, dada $N=\sum_{j=1}^{k} n_j$, sería multinomial con parámetro $\pi_j = \mu_j/\sum_{j=1}^{k}\mu_j$. La demostración es inmediata, dado que N tiene una distribución Poisson con media $\sum_{j=1}^{k}\mu_j$ y de aquí

$$\begin{aligned} p(n_1,\ldots,n_k|N) &= p(n_1,\ldots,n_k)/p(N) \\ &= \frac{e^{-\sum\mu_i}\prod_{j=1}^{k}(\mu^{n_j}/n_j!)}{e^{-\sum\mu_j}(\sum\mu_j)^N/N!} \\ &= N!\prod_{j=1}^{k}\frac{\pi_j^{n_j}}{n_j!}. \end{aligned} \tag{2.1}$$

Una distribución inicial conjugada para la distribución multinomial con vector de parámetros $(\pi_1,\ldots,\pi_k)$ puede ser la distribución Dirichlet, con función de probabilidad

$$p(\pi_1,\ldots,\pi_k|\alpha) = \frac{\Gamma(\alpha_*)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}\prod_{l=1}^{k}\pi_l^{\alpha_l-1}, \qquad 0<\pi_l<1,\ \alpha_j>0,$$

con $\sum_{j=1}^{k}\pi_j = 1$, $\alpha=(\alpha_1,\ldots,\alpha_k)$ y $\alpha_* = \sum_{i=1}^{k}\alpha_i$.

La distribución está parametrizada por un vector $\alpha=(\alpha_1,\ldots,\alpha_k)$ tal que $E(\pi_i)=\alpha_i/\alpha_*$, $Var(\pi_i) = E(\pi_i)(1-E(\pi_i))/(1+\alpha_*)$ y $Cov(\pi_i,\pi_j) = -E(\pi_i)E(\pi_j)/(1+\alpha_*)$. El valor de α_* se interpreta como el "tamaño de muestra inicial hipotético", y determina la cantidad de información contenida en la distribución inicial: una α_* pequeña implica información vaga mientras que una α_* grande indica una distribución inicial robusta para $(\pi_1,\ldots,\pi_k)$.

La distribución final de $(\pi_1,\ldots,\pi_k)$ es

$$\begin{aligned} p(\pi_1,\ldots,\pi_k|n_1,\ldots,n_k,\alpha,N) &\propto p(n_1,\ldots,n_k|\pi_1,\ldots,\pi_k,N)p(\pi_1,\ldots,\pi_k|\alpha) \\ &= N!\prod_{i=1}^{k}\frac{\pi_i^{n_i}}{n_i!}\frac{\Gamma(\alpha_*)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}\prod_{l=1}^{k}\pi_l^{\alpha_l-1} \\ &\propto \prod_{i=1}^{k}\pi_i^{n_i+\alpha_i-1}. \end{aligned}$$

Este último es el kernel de una distribución Dirichlet con vector de parámetros $(n_1 + \alpha_1, \ldots, n_k + \alpha_k)$, por lo tanto esa es la distribución final de $(\pi_1, \ldots, \pi_k)$. Esta distribución contiene toda la información disponible sobre las probabilidades $(\pi_1, \ldots, \pi_k)$ de las celdas, condicional a las observaciones $(n_1, \ldots, n_k)$.

A falta de información inicial se usa una distribución inicial no informativa. Una de las distribuciones iniciales más usadas para parámetros multinomiales es precisamente la distribución de Dirichlet con vector de parámetros $\alpha = (1/2, \ldots, 1/2)$.

Teniendo en cuenta que α_* se interpretó como el tamaño de muestra inicial, la cantidad $I = \alpha_*/(N + \alpha_*)$ puede considerarse como la proporción de la información total que contribuye a la distribución inicial. De esta manera, un valor de α_* que permita obtener $I = 0.01$ produciría alrededor del 1% de la información total, mientras que $I \approx 1$ implicaría que los datos están completamente dominados por la distribución inicial. ■

2.3 Predicción

En muchas ocasiones el propósito de un análisis estadístico es *predecir* el valor de una observación futura X con base en la información disponible. El problema de inferencia sobre θ puede considerarse como un paso intermedio en la solución al problema de predicción, aunque en ciertas situaciones puede ser de interés en sí mismo. Por otro lado, debido a resultados de consistencia, un parámetro puede verse como el límite de una sucesión de *estadísticas* (*i.e.* funciones de las observaciones) cuando el tamaño de la muestra tiende a infinito (ver teorema **??**). De esta manera, hacer inferencias acerca del valor del parámetro θ puede considerarse como una forma límite de hacer inferencias predictivas acerca de las observaciones.

Dado el valor del parámetro θ, la distribución que describe el comportamiento de la observación futura x es $p(x|\theta)$; sin embargo, el valor de θ generalmente es desconocido. Algunos métodos estadísticos tradicionales atacan este problema *estimando* a θ con base en la muestra observada, y en muchos casos simplemente sustituyen el valor de θ con la estimación resultante.

Desde la perspectiva Bayesiana, el modelo paramétrico $p(x|\theta)$, junto con la distribución inicial $p(\theta)$, inducen una distribución conjunta para (X, θ), dada por

$$p(x, \theta) = p(x|\theta)p(\theta).$$

La distribución marginal

$$p(x) = \int p(x|\theta)p(\theta)d\theta,$$

describe nuestro conocimiento acerca de X dada la información inicial disponible. Dicha distribución se conoce comúnmente como la *distribución predictiva (inicial)* (*prior predictive distribution*).

De manera similar, una vez obtenida la muestra, el modelo $p(x|\theta)$ y la distribución final inducen una distribución conjunta para (X, θ) *condicional en los valores observados* $x_1, \ldots, x_n$; *i.e.*

$$p(x, \theta|x_1, \ldots, x_n) = p(x|\theta, x_1, \ldots, x_n)p(\theta|x_1, \ldots, x_n) = p(x|\theta)p(\theta|x_1, \ldots, x_n),$$

donde la última igualdad se da siempre y cuando haya independencia condicional de X y $(X_1, \ldots, X_n)$ dado θ. Así, la distribución

$$p(x|x_1, \ldots, x_n) = \int p(x|\theta)p(\theta|x_1, \ldots, x_n)d\theta,$$

describe el comportamiento de X dada toda la información disponible y se conoce como la *distribución predictiva (final)* (*posterior predictive distribution*).

Ejemplo 2.6 (Predicción: Bernoulli)**.** Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución Bernoulli con función de masa de probabilidad

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}, \qquad x \in \{0,1\},\ 0 < \theta < 1,$$

y supongamos que θ tiene una distribución inicial Beta(α_0, β_0).

Como ya se mencionó, la distribución beta es conjugada para el modelo Bernoulli, por lo que la distribución final de θ también es beta. De hecho,

$$p(\theta|x_1, \ldots, x_n) = Beta(\theta|\alpha_1, \beta_1),$$

donde $\alpha_1 = \alpha_0 + \sum_{i=1}^n x_i$ y $\beta_1 = \beta_0 + n - \sum_{i=1}^n x_i$. La distribución predictiva final está dada por

$$p(x|x_1, \ldots, x_n) = \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(\alpha_1+x)\Gamma(\beta_1+1-x)}{\Gamma(\alpha_1+\beta_1+1)}, \qquad x = 0, 1.$$

Esta distribución se conoce como *Beta-Binomial*. ■

Ejemplo 2.7 (Predicción: Normal)**.** Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución normal con función de densidad

$$p(x|\theta) = N(x|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}, \qquad x \in \Re,\ \theta \in \Re,$$

con $\sigma > 0$ conocido. Supongamos que θ tiene una distribución inicial conjugada,

$$p(\theta) = N(\theta|\mu_0, \tau_0^2).$$

Entonces la distribución final de θ está dada por

$$p(\theta|x_1, \ldots, x_n) = N(\theta|\mu_1, \tau_1^2),$$

donde

$$\mu_1 = (1/\tau_0^2 + n/\sigma^2)^{-1}(\mu_0/\tau_0^2 + n\bar{x}/\sigma^2)$$

y

$$\tau_1^2 = (1/\tau_0^2 + n/\sigma^2)^{-1}.$$

La distribución predictiva final es entonces

$$p(x|x_1, \ldots, x_n) = N(x|\mu_1, \tau_*^2),$$

con

$$\tau_*^2 = \sigma^2\tau_1^2(1/\tau_1^2 + 1/\sigma^2).$$

`Código R:` Bayes2_2DistribucionNormal.R

Los cálculos se presentan a continuación.

La distribución final de θ está dada por:

$$
\begin{aligned}
& p(\theta|x_1,\ldots,x_n) \\
&= \frac{p(\theta)p(x_1,\ldots,x_n|\theta)}{\int p(\theta)p(x_1,\ldots,x_n|\theta)d\theta} \\
&= \frac{(2\pi\tau_0^2)^{-1/2}\exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\prod_{i=1}^n(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\}}{\int (2\pi\tau_0^2)^{-1/2}\exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\prod_{i=1}^n(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\}d\theta} \\
&= \frac{\exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right\}}{\int \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right\}d\theta} \quad \text{simplificar lo que No depende de } \theta \\
&= \frac{\exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right)\right\}d\theta}
\end{aligned}
$$

Simplificar la potencia de la exponencial:

$$
\begin{aligned}
& \frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2 \\
&= \frac{1}{\tau_0^2}(\theta^2-2\theta\mu_0+\mu_0^2)+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i^2-2x_i\theta+\theta^2) \\
&= \frac{1}{\tau_0^2}\theta^2-2\frac{1}{\tau_0^2}\theta\mu_0+\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2-2\frac{1}{\sigma^2}\sum_{i=1}^n x_i\theta+\frac{1}{\sigma^2}n\theta^2 \quad \text{asociar con } \theta \\
&= \theta^2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)-2\theta\left(\frac{\mu_0}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^n x_i\right)+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right) \\
&= \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta^2-2\theta\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{1}{\sigma^2}n\bar{x}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right) \quad \text{completar binomio para } \theta \\
&= \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta^2-2\theta\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}+\left[\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right]+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right) \\
&\quad -\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2 \\
&= \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right)-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)^2}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}
\end{aligned}
$$

Sustituir en la distribución final de θ:

$$
\begin{aligned}
& p(\theta|x_1,\ldots,x_n) \\
= \; & \frac{p(\theta)p(x_1,\ldots,x_n|\theta)}{\int p(\theta)p(x_1,\ldots,x_n|\theta)d\theta} \\
= \; & \frac{\exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right)\right\}d\theta} \\
= \; & \frac{\exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right)-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)^2}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2\right)-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)^2}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right)\right\}d\theta} \\
= \; & \frac{\exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right)\right\}}{\int \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right)\right\}d\theta} \quad \text{simplificar lo que No depende de } \theta \\
= \; & \frac{\exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\}}{\int \exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\}d\theta} \quad \text{Kernel de v.a. Normal} \\
= \; & \frac{\frac{1}{\sqrt{2\pi\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}}\exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\}}{\frac{1}{\sqrt{2\pi\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}}\int \exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\}d\theta} \quad \text{completar integral} \\
= \; & \frac{1}{\sqrt{2\pi\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}}\exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\} \\
= \; & \text{Normal}\left(\mu_1=\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)},\ \tau_1^2=\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)
\end{aligned}
$$

La distribución final de θ también podría calcularse omitiendo el denominador y usando la proporcionalidad con respecto a θ, de la siguiente forma:

$$
\begin{aligned}
& p(\theta|x_1,\ldots,x_n) \\
& \quad \propto \quad p(\theta)p(x_1,\ldots,x_n|\theta) \\
& \quad \propto \quad (2\pi\tau_0^2)^{-1/2}\exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\} \\
& \quad \propto \quad \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} \qquad \text{simplificar lo que No depende de } \theta \\
& \quad \propto \quad \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\right\} \\
& \quad \propto \quad \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2+\left(\frac{1}{\tau_0^2}\mu_0^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2\right)-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)^2}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right)\right\} \\
& \quad \propto \quad \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right)\right\} \qquad \text{simplificar lo que No depende de } \theta \\
& \quad \propto \quad \exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\} \qquad \text{Kernel de v.a. Normal} \\
& \quad \propto \quad \frac{1}{\sqrt{2\pi\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}}\exp\left\{-\frac{1}{2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\right]^2\right\} \qquad \text{completar distribución Normal} \\
& \quad = \quad \text{Normal}\left(\mu_1=\frac{\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)}\ ,\ \tau_1^2=\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)
\end{aligned}
$$

La distribución predictiva final de x es:

$$\begin{aligned}
p(x|x_1,\ldots,x_n) &= \int p(x|\theta)p(\theta|x_1,\ldots,x_n)d\theta \\
&= \int (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}(2\pi\tau_1^2)^{-1/2}\exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\}d\theta \\
&= (2\pi\sigma^2)^{-1/2}(2\pi\tau_1^2)^{-1/2}\int \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(x-\theta)^2+\frac{1}{\tau_1^2}(\theta-\mu_1)^2\right)\right\}d\theta
\end{aligned}$$

Simplificar la potencia de la exponencial:

$$\begin{aligned}
&\left(\frac{1}{\sigma^2}(x-\theta)^2+\frac{1}{\tau_1^2}(\theta-\mu_1)^2\right) \\
&= \frac{x^2}{\sigma^2}-2\frac{x}{\sigma^2}\theta+\frac{1}{\sigma^2}\theta^2+\frac{1}{\tau_1^2}\theta^2-2\frac{\mu_1}{\tau_1^2}\theta+\frac{\mu_1^2}{\tau_1^2} \quad \text{asociar con } \theta \\
&= \theta^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)-2\theta\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)+\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right) \quad \text{completar binomio para } \theta \\
&= \left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)\left[\theta-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]^2+\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}
\end{aligned}$$

Sustituir en la distribución predictiva final de x:

$$\begin{aligned}
p(x|x_1,\ldots,x_n) &= \int p(x)p(\theta|x_1,\ldots,x_n)d\theta \\
&= (2\pi\sigma^2)^{-1/2}(2\pi\tau_1^2)^{-1/2}\int \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(x-\theta)^2+\frac{1}{\tau_1^2}(\theta-\mu_1)^2\right)\right\}d\theta \\
&= \frac{\displaystyle\int \exp\left\{-\frac{1}{2}\left\{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)\left[\theta-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]^2+\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\}d\theta}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\tau_1^2}} \\
&= \frac{\exp\left\{-\frac{1}{2}\left\{\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\}}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\tau_1^2}} \\
&\times \int \exp\left\{-\frac{1}{2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}}\left[\theta-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]^2\right\}d\theta
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp\left\{-\frac{1}{2}\left\{\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\}}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\tau_1^2}} \\
&\times \sqrt{2\pi\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}}\int \frac{1}{\sqrt{2\pi\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}}}\exp\left\{-\frac{1}{2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}}\left[\theta^2-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]^2\right\}d\theta \\
&= \frac{\exp\left\{-\frac{1}{2}\left\{\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\}}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\tau_1^2}} \times \sqrt{2\pi\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}} \quad \text{simplificar} \\
&= \frac{\sqrt{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)^{-1}}}{\sqrt{2\pi\sigma^2\tau_1^2}}\exp\left\{-\frac{1}{2}\left\{\left(\frac{x^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)-\frac{\left(\frac{x^2}{\sigma^4}+2\frac{x}{\sigma^2}\frac{\mu_1}{\tau_1^2}+\frac{\mu_1^2}{\tau_1^4}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\} \quad \text{asociar con } x \\
&= \frac{\exp\left\{-\frac{1}{2}\left\{x^2\left[\frac{1}{\sigma^2}-\frac{1}{\sigma^4\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]-2x\left[\frac{\mu_1}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]+\left[\frac{\mu_1^2}{\tau_1^2}-\frac{\frac{\mu_1^2}{\tau_1^4}}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]\right\}\right\}}{\sqrt{2\pi\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}}
\end{aligned}
$$

Simplificando algunos términos:

$$
\begin{aligned}
\left(\frac{1}{\sigma^2}-\frac{1}{\sigma^4\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right) &= \frac{1}{\sigma^2}-\frac{1}{\sigma^4\left(\frac{\sigma^2+\tau_1^2}{\sigma^2\tau_1^2}\right)} \\
&= \frac{1}{\sigma^2}-\frac{1}{\sigma^2\left(\frac{\sigma^2+\tau_1^2}{\tau_1^2}\right)} = \frac{1}{\sigma^2}-\frac{\tau_1^2}{\sigma^2\left(\sigma^2+\tau_1^2\right)} \\
&= \frac{\left(\sigma^2+\tau_1^2\right)-\tau_1^2}{\sigma^2\left(\sigma^2+\tau_1^2\right)} = \frac{\sigma^2}{\sigma^2\left(\sigma^2+\tau_1^2\right)} \\
&= \frac{1}{\left(\sigma^2+\tau_1^2\right)} \\
&= \frac{1}{\sigma^2\tau_1^2\frac{\left(\sigma^2+\tau_1^2\right)}{\sigma^2\tau_1^2}} = \frac{1}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}
\end{aligned}
$$

$$\begin{aligned}
\left(\frac{\mu_1^2}{\tau_1^2} - \frac{\frac{\mu_1^2}{\tau_1^4}}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right) &= \frac{\mu_1^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\left(\frac{1}{\tau_1^2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right) - \frac{1}{\tau_1^4}\right) \\
&= \frac{\mu_1^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\left(\frac{1}{\sigma^2\tau_1^2} + \frac{1}{\tau_1^4} - \frac{1}{\tau_1^4}\right) \\
&= \frac{\mu_1^2}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\left(\frac{1}{\sigma^2\tau_1^2}\right)
\end{aligned}$$

Sustituir en la distribución predictiva final de x:

$$\begin{aligned}
p(x|x_1,\ldots,x_n) &= \int p(x)p(\theta|x_1,\ldots,x_n)d\theta \\
&= \frac{\exp\left\{-\frac{1}{2}\left\{x^2\left[\frac{1}{\sigma^2} - \frac{1}{\sigma^4\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right] - 2x\left[\frac{\mu_1}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right] + \left[\frac{\mu_1^2}{\tau_1^2} - \frac{\frac{\mu_1^2}{\tau_1^4}}{\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right]\right\}\right\}}{\sqrt{2\pi\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}} \\
&= \frac{\exp\left\{-\frac{1}{2}\left\{x^2\frac{1}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)} - 2x\left[\frac{\mu_1}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right] + \frac{\mu_1^2}{\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\right\}\right\}}{\sqrt{2\pi\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\left[x^2 - 2x\mu_1 + \mu_1^2\right]\right\}}{\sqrt{2\pi\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}\left(x-\mu_1\right)^2\right\}}{\sqrt{2\pi\sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)}} \\
&= \text{Normal}\left(\mu_1\ ,\ \tau_*^2 = \sigma^2\tau_1^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)\right)
\end{aligned}$$

■

Otros ejemplos se muestran en los apéndices **??** y **??**.

2.3.1 Actualización secuencial de la información

El Teorema de Bayes proporcional el mecanismo para actualizar nuestro estado de la información, llevando de la distribución inicial a la final. La distribución final se convierte entonces en la 'nueva' distribución inicial antes de observar nuevos datos.

Dado $f(\theta)$ la distribución a priori (inicial). Suponga que se observa $X_1 = x_1$ de la densidad $f(x|\theta)$, por el teorema de Bayes:

$$f(\theta|x_1) \quad \propto \quad f(x_1|\theta)f(\theta)$$

Ésta es la 'nueva' distribución inicial antes de observar $X_2 = x_2$ de la densidad $f(x|\theta)$, independiente de X_1. Aplicando el teorema de Bayes:

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

esto se debe a que

$$\begin{aligned} f(\theta|x_1, x_2) &= \frac{f(x_1, x_2, \theta)}{f(x_1, x_2)} \\ &= \frac{f(x_2|\theta, x_1)f(\theta|x_1)f(x_1)}{f(x_1, x_2)} \end{aligned}$$

y además, como x_1 y x_2 son independientes,

$$\begin{aligned} f(x_2|\theta, x_1) &= \frac{f(x_1, x_2, \theta)}{f(x_1, \theta)} \\ &= \frac{f(x_1, x_2, \theta)/f(\theta)}{f(x_1, \theta)/f(\theta)} \\ &= \frac{f(x_1, x_2|\theta)}{f(x_1|\theta)} \\ &= \frac{f(x_1|\theta)f(x_2|\theta)}{f(x_1|\theta)} \\ &= f(x_2|\theta). \end{aligned}$$

Este es el resultado obtenido de haber actualizado de 'un solo golpe' la distribución inicial $f(\theta)$ con base en la muestra $\{x_1, x_2\}$.

Capítulo 3

Teoría de decisiones

Para más detalles se puede revisar Robert (2007) y Bernardo (1981).

3.1 Fundamentos

Una parte importante dentro del estudio de la estadística bayesiana se centra en la teoría de decisión, para ello es necesario entender los siguientes puntos

- El objetivo de la estadística, y en particular de la estadística Bayesiana, es proporcionar una metodología para analizar adecuadamente la información con la que se cuenta (análisis de datos) y decidir de manera razonable sobre la mejor forma de actuar (teoría de decisión).

- La metodología bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes.

- La teoría de decisión propone un método de tomar decisiones basado en unos principios básicos sobre la elección coherente entre opciones alternativas.

Nos hallamos frente a un problema de decisión cuando debemos elegir entre dos o más formas de actuar. En general hay dos tipos de decisiones:

- **Triviales:**
    1. Peliculas

2. Conectarme a clase
3. Comida

- **Trascendentales**

1. Carrera profesional
2. Casarse
3. Tener hijos

Es evidente que tomar una decisión trivial no traera mayores consecuencias, sin embargo, en las trascedentales habrá consecuencias con las que hay que vivir, entonces la pregunta sería...

3.2 Cómo debe tomarse una decisión trascendente?

Lo mejor sería seguir los siguientes pasos:

Paso 1: Elaborar un conjunto de posibles alternativas al que llamaremos, **espacio de decisiones o acciones** y denotaremos por D.

Para ello hay que tomar en cuenta lo siguiente:

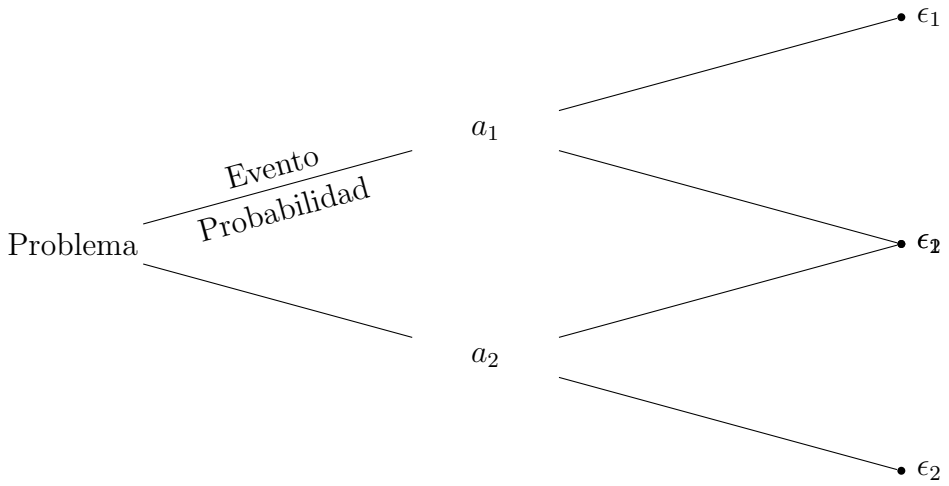
- La elección de algún elemento debe excluir la elección de cualquier otra.
- D puede ser infinito, pero en la practica suele ser finito.
- El problema de decisión se planetea en un **ambiente de incertidumbre.**

Una gran dificultad para la toma de decisiones consiste en la falta de información sobre lo que sucederá según se actúe de una forma o de otra.

Por ello uno de los principales objetivos es el desarrollo de procesos lógicos para la toma de decisiones bajo condiciones de incertidumbre.

Paso 2: Considerar para cada posible decisión al conjunto de sucesos inciertos que determinan sus eventuales consecuencias.

Esquematicamente, la situación, en el caso de un número finito de alternativas y un número finito de sucesos inciertos puede represntarse mediante un árbol de decisión de la forma:



Definición: Un **problema de decisión** esta definido conjuntamente por los elementos $(\epsilon, A, \zeta, \preceq)$, donde:

- ϵ es un álgebra de eventos relevantes (ϵ_j).
- A es un conjunto de opciones o acciones potenciales cuyos elementos denotaremos por a_i.
- ζ es el conjunto de las consecuencias de haber elegido la acción a_i bajo la ocurrencia del evento ϵ_j.
- $\preceq$ es una relación (binaria) de preferencia para algunos de los elementos de A.

Ejemplo: Decidir salir a la calle con o sin paraguas.

Entonces $A := \{a_1, a_2\}$, donde:

$$\begin{aligned} a_1 &= \textit{Llevar paraguas} \\ a_2 &= \textit{No llevar paraguas} \end{aligned}$$

Ahora $\epsilon := \{\epsilon_1, \epsilon_2\}$, donde:

$$\begin{aligned} \epsilon_1 &= Llueve \\ \epsilon_2 &= No\ llueve \end{aligned}$$

Por lo que: $\zeta = \{C_{ij} = (a_i, \epsilon_j) : a_i \in A, \epsilon_j \in \epsilon\}$, con lo que tendriamos:

$$\begin{aligned} C_{11} &: Llevar\ paraguas\ y\ llueve. \\ C_{12} &: Llevar\ paraguas\ y\ no\ llueve. \\ C_{21} &: No\ llevar\ paraguas\ y\ llueve. \\ C_{22} &: No\ llevar\ paraguas\ y\ no\ llueve. \end{aligned}$$

Nota: A lo controlamos mientras que ϵ es incierto.

3.3 Axiomas de coherencia

Son una serie de principios que establecen las condiciones para tomar decisiones coherentemente y para aclarar las posibles ambigedades en el proceso de toma de decisión.

Denotareamos por $\ell = \{c_1|\epsilon_1, c_2|\epsilon_2, \ldots, c_k|\epsilon_k\}$ a una situación en la que se obtiene la consecuencia c_i si sucede ϵ_i, para $i = 1, \ldots, k$ con la condición de que los sucesos sean exahustivos y mutuamente excluyentes. ℓ_i es una forma de actuar.

Primer axioma, comparabilidad: Para todo par de opciones ℓ_1 y ℓ_2 es cierto una de las tres relaciones:

$$\begin{aligned} \ell_1 &< \ell_2 \\ \ell_2 &< \ell_1 \\ \ell_1 &\sim \ell_2 \end{aligned}$$

Además, es posible encontrar dos consecuencias c^* y c_* tales que $c^* > c_*$ y que para toda consecuencia c:

$$c_* \leq c \leq c^*$$

Este axioma refleja el hecho observable de que en la practica se comparan opciones de muy distinto tipo, forzados por la necesidad de actuar.

Segundo axioma, transitividad: Si $\ell_1 > \ell_2$ y $\ell_2 > \ell_3$, entonces:

$$\ell_1 > \ell_3$$

Análogamente si $\ell_1 \sim \ell_2$ y $\ell_2 \sim \ell_3$, entonces:

$$\ell_1 \sim \ell_3$$

Tercer axioma, sustitución y dominación: Si $\ell_1 > \ell_2$ cuando sucede A y $\ell_1 > \ell_2$ cuando sucede A^c entonces:

$$\ell_1 > \ell_2$$

De igual forma cuando: $\ell_1 \sim \ell_2$ cuando sucede A y $\ell_1 \sim \ell_2$ cuando sucede A^c entonces:

$$\ell_1 \sim \ell_2$$

En virtud del principio de sustitución, en una opción puede reemplazarse una consecuencia por otra opción equivalente a ella.

Cuarto axioma, sucesos de referencia: La persona que decide puede concebir un procedimiento de generar un punto aleatorio κ en el cuadrado unitario, esto es, un número $\kappa = (x, y)$, $0 \leq x \leq 1$, $0 \leq y \leq 1$ tal que para cualquier par de regiones R_1, R_2 del cuadrado unitario el suceso $\kappa \in R_1$ le resulta menos verosimil que el suceso $\kappa \in R_2$ sii el area de R_1 es menor que la de R_2.

Este axioma nos permite construir una gama de opciones con la que medir, por comparación, la deseabilidad de las demas. Especificamente, consideremos opciones de la forma:

$$\{c^*|R, c_*|R^c\}, \qquad R \subset (0,1)^2$$

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

De la misma manera, las preferencias del decisor entre las distintas consecuencias debe cuantificarse a través de una **función de utilidad.**

3.4 Utilidad esperada

Los postulados de coherencia también permiten definir formalmente una medida de las preferencias del decisor entre las consecuencias.

En efecto, definiremos la utilidad de una consecuencia c como un número $u(c)$ de tal forma que $u(c) \in [0, 1]$ que mide la deseabilidad relativa de la consecuencia c.

Evidentemente, la utilidad de la consecuencia menos preferida c_* será $u(c_*) = 0$ y de la más preferida $u(c^*) = 1$.

Definición: Entenderemos por **espacio de resultados** y lo denotaremos por Θ, a una partición de Ω de eventos relevantes.

Definición: En un problema de decisión $(\epsilon, \ell, A, \prec)$ con espacio de estados relevantes $\Theta = \{\epsilon_1, \ldots, \epsilon_n\} \subset \epsilon$, función de probabilidad P sobre Θ y función de utilidad u sobre C, la **utilidad esperada** de la acción $a_i \in A = \{a_1 \ldots, a_k\}$ se denota por $\bar{u}(a_i)$ y se define como:

$$\bar{u}(a_i) := \sum_{j=1}^{m} u(a_i, \epsilon_j) P(\epsilon_j) \quad i = 1, \ldots, k$$

3.4.1 Criterio General de Decisión

En un problema de decisión, la relación de preferencia $\prec$ sobre A queda definida por:

$$a_1 \prec a_2 \leftrightarrow \bar{u}(a_1) \leq \bar{u}(a_2)$$

Esto implica que:

$$\begin{aligned} a_1 \sim a_2 &\leftrightarrow \bar{u}(a_1) = \bar{u}(a_2) \\ a_1 < a_2 &\leftrightarrow \bar{u}(a_1) < \bar{u}(a_2) \end{aligned}$$

Finalmente, la acción óptima de A, misma que denotaremos por a_* será aquella que satisfaga:

$$\bar{u}(a_*) = \max_i \bar{u}(a_i)$$

Puede ocurrir que a_* no sea única. Entonces hablaremos del Conjunto de acciones óptimas:

$$A^* \subset A$$

Ejemplo maximizar la utilidad esperada:

Un empresario adquiere pescado fresco en el mercado central para su posterior venta. Cada caja de pescado la identifica como excelente o no excelente en función del porcentaje de pescado que se considere de calidad excelente. Una caja de pescado excelente contiene un 90% de pescado de alta calidad, mientras que una caja de pescado no excelente contiene solo un 20% de pescado de alta calidad. Una caja de pescado excelente genera un beneficio de 100, mientras que una caja de pescado no excelente causa unas pérdidas de 100 por la mala imagen de la empresa que se llevan los clientes. Antes de comprar una caja el empresario puede comprobar la calidad de la misma extrayendo un ejemplar de pescado con el objetivo de verificar si se trata o no de pescado de alta calidad. Establezca la estrategia que debe seguir el empresario, así como el coste de la información.

Vamos a resolver el problema enumerando todos los pasos:

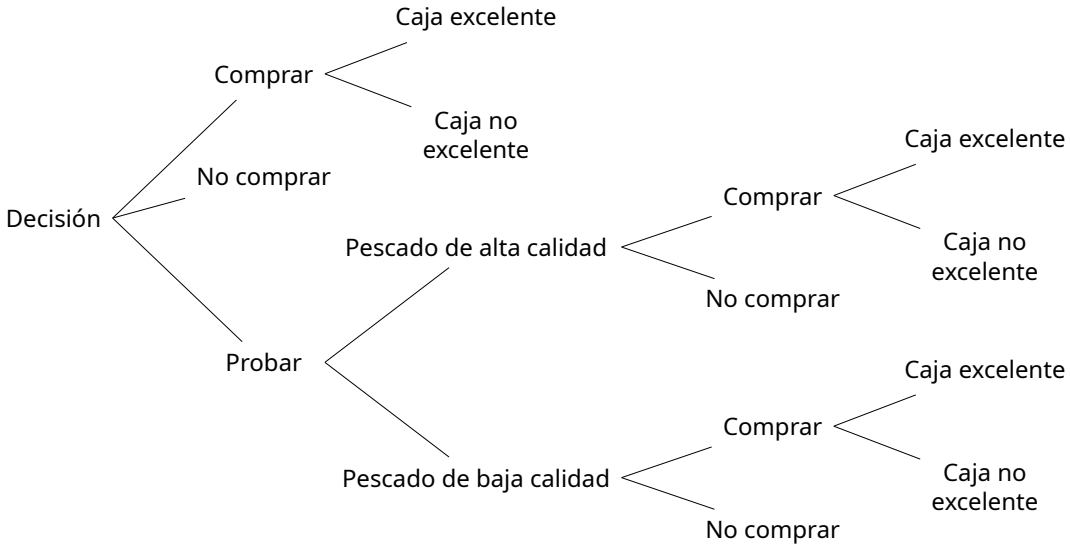
Paso 1: Enumere las diferentes alternativas de decisión:

- Comprar la caja de pescado
- NO comprar la caja de pescado
- Probar.

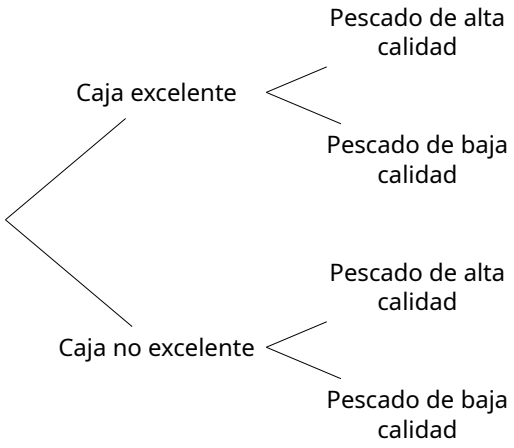
Paso 2: Enumere para cada una de las alternativas de decisión, los estados de la naturaleza asociados a la misma.

Alternativas	Estados de la naturaleza
Comprar	Caja de pescado excelente
	Caja de pescado no excelente
No comprar	–
Probar	Ejemplar de pescado de alta calidad
	Ejemplar de pescado de baja calidad

Paso 3: Dibuja el diagrama de decisión tambien conocido como árbol de decisión



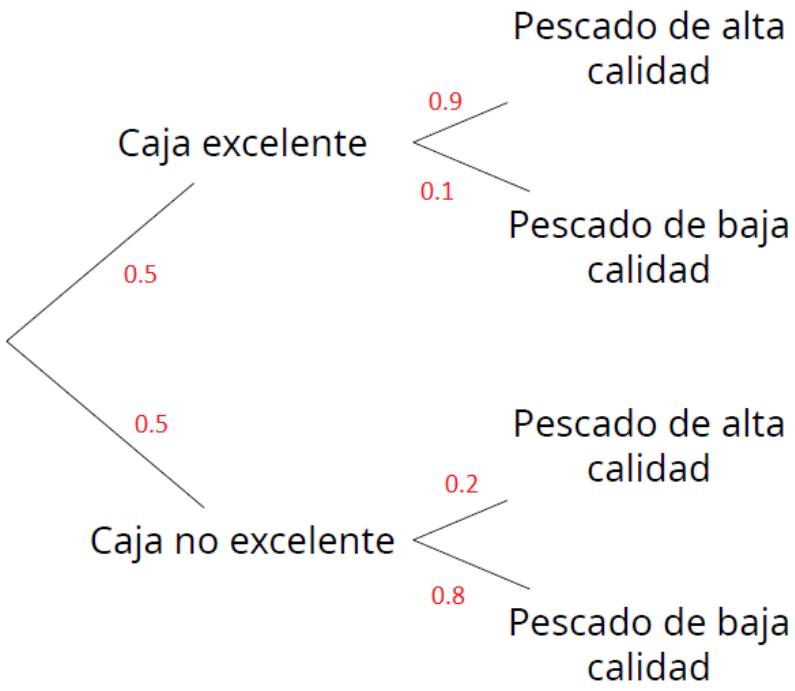
Paso 4: Asigne las probabilidades de cada uno de los estados de la naturaleza. En este caso se trata de probabilidades a posteriori, por lo que debe utilizar el teorema de Bayes para calcular dichas probabilidades. **Para la aplicación del teorema de Bayes** puede utilizar el árbol que se muestra a continuación. Los estados de la naturaleza son que la caja comprada sea o no de excelente calidad, y los acontecimientos, que el ejemplar de pescado verificado es de alta o baja calidad.



La probabilidad a priori de que una caja de pescado sea o no de excelente calidad es del 50%. Por su parte, las probabilidades condicionales vienen dadas, según se indica en el enunciado del ejercicio, por los siguientes valores:

$$
\begin{aligned}
P(Pescado\ sea\ de\ alta\ calidad \ / \ Caja\ es\ excelente) &= 0.9 \\
P(Pescado\ sea\ de\ baja\ calidad \ / \ Caja\ es\ excelente) &= 0.1 \\
P(Pescado\ sea\ de\ alta\ calidad \ / \ Caja\ no\ es\ excelente) &= 0.2 \\
P(Pescado\ sea\ de\ baja\ calidad \ / \ Caja\ no\ es\ excelente) &= 0.8
\end{aligned}
$$

Esto lo podemos ver en el árbol de la siguiente manera:



De donde, la probabilidad a priori de cada uno de los acontecimientos:

P(Pescado de alta calidad) = [P(Caja excelente) x P(Pescado de alta calidad / Caja excelente)] + [P(Caja no excelente) x P(Pescado de alta calidad / Caja no excelente)] = [0,5 x 0,9] + [0,5 x 0,2] = **0,55**

P(Pescado de baja calidad) = [P(Caja excelente) x P(Pescado de baja calidad / Caja excelente)] + [P(Caja no excelente) x P(Pescado de baja calidad / Caja no excelente)] = [0,5 x 0,1] + [0,5 x 0,8] = **0,45**

Ahora, mediante la aplicación del teorema de Bayes determine las probabilidades a posteriori de cada uno de los estados de la naturaleza.

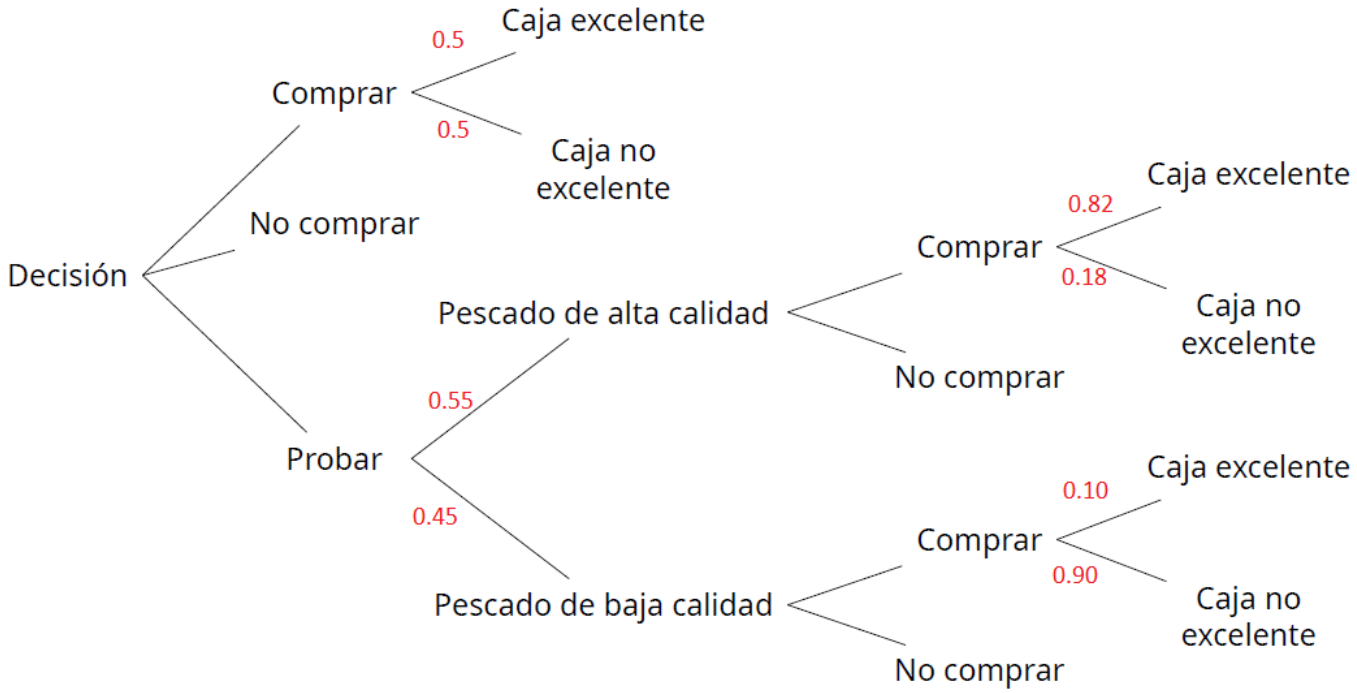
$$P(C.E/P.A.C.) = \frac{P(C.E) \times P(P.A.C/C.E)}{P(P.A.C)} = \frac{0.5 \times 0.9}{0.55} = 0.8$$

$$P(C.E/P.B.C.) = \frac{P(C.E) \times P(P.B.C/C.E)}{P(P.B.C)} = \frac{0.5 \times 0.1}{0.45} = 0.1$$

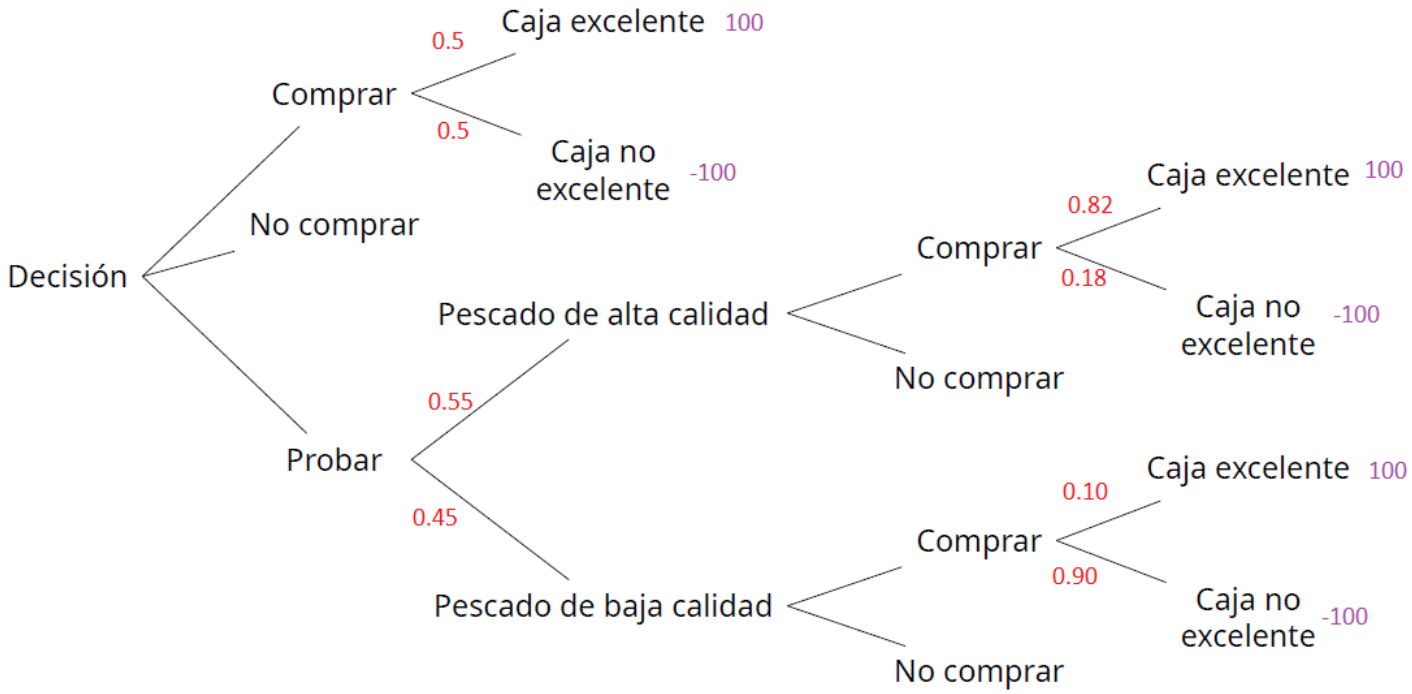
$$P(C.N.E/P.B.C.) = \frac{0.5 \times 0.8}{0.45} = 0.9$$

$$P(C.N.E/P.A.C.) = \frac{0.5 \times 0.2}{0.55} = 0.18$$

De donde, el árbol de decisión incluyendo las probabilidades, queda de la siguiente manera:



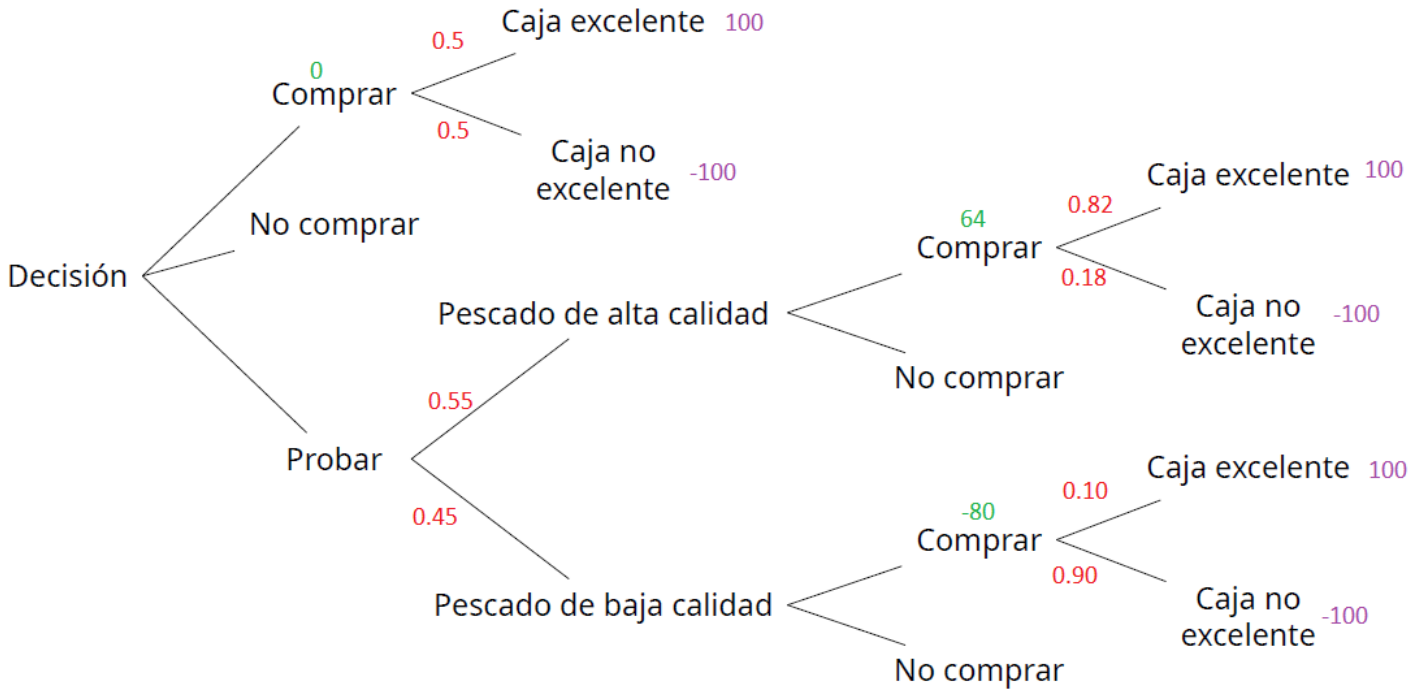
Paso 5: Ahora calculamos el beneficio de cada una de las ramas del arbol



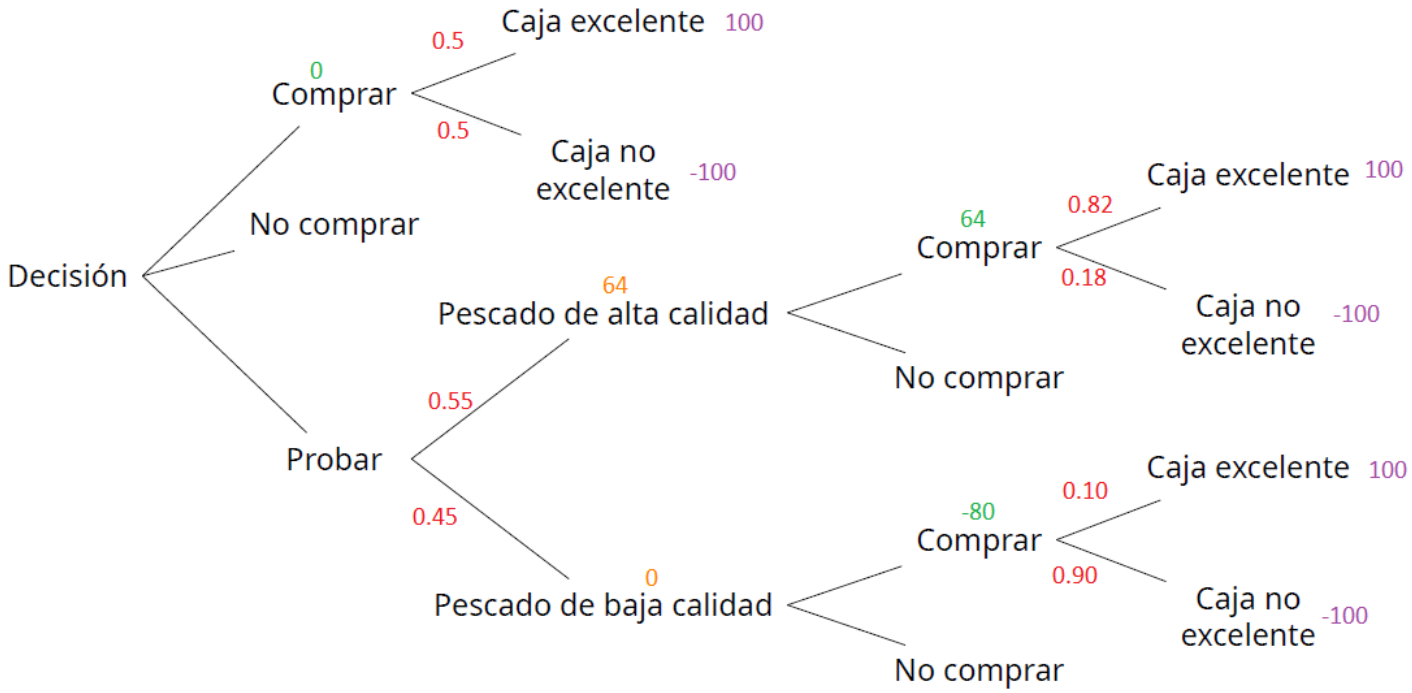
Paso 6: Resuelva el árbol de decisión de derecha a izquierda. Dado que la etapa final es probabilista debes aplicar el criterio de la esperanza matemática con el objetivo de determinar el beneficio esperado de cada alternativa de decisión.

$$
\begin{aligned}
(100x0,82) &+ ((-100)x0,18) = 64 \\
(100x0,10) &+ ((-100)x0,90) = -80 \\
(100x0,50) &+ ((-100)x0,50) = 0
\end{aligned}
$$

Y hay que ponerlas en el diagrama, de la siguiente manera:



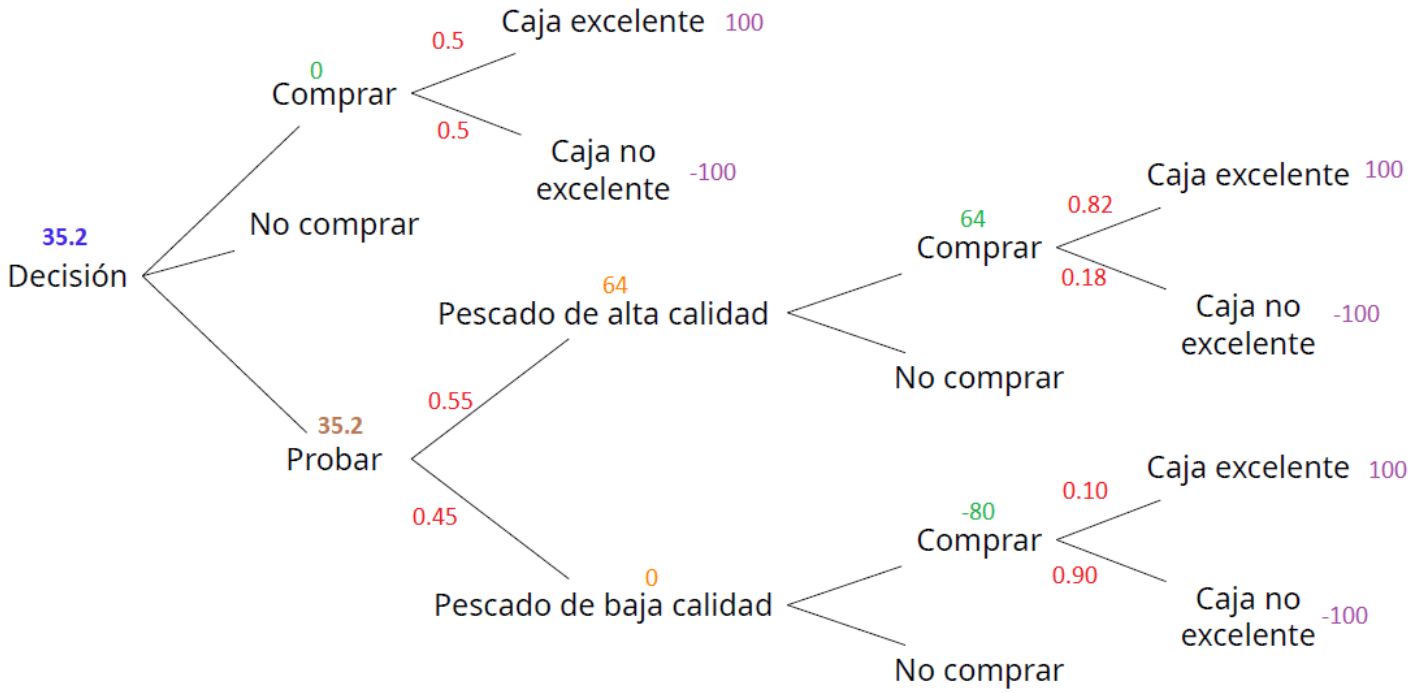
Paso 7: Resuelva la etapa anterior. Dado que dicha etapa es determinista y que los valores que ha calculado son beneficios, **debemos elegir la alternativa cuyo beneficio sea mayor** y colocar el resultado encima del nudo correspondiente.



Paso 8: Resuelva las dos últimas etapas. La penúltima etapa es probabilista por lo que debe aplicar el criterio de la esperanza matemática con el objetivo de determinar el beneficio esperado.

$$(64x0,55) + (0x0,45) = 35,2$$

La última etapa es determinista, debe pues elegir la alternativa cuyo beneficio sea mayor y colocar el resultado encima del nudo correspondiente.



Paso 9: Define la estrategia: La estrategia que debe seguir el empresario es la de extraer un ejemplar de pescado con el objetivo de verificar si se trata o no de pescado de alta calidad, **en el caso de que el pescado extraído sea de alta calidad, debe comprar la caja de pescado**, por el contrario, **si el pescado extraído es de baja calidad, no debe comprar la caja de pescado**. Con esta estrategia el beneficio esperado es de **35,2.**

Cual sería la estrategia si me cobraran 50 por sacar un ejemplar de pescado?

En este caso, **el valor de la información** = 0, que es el valor de la información que aporta la extracción de un ejemplar de pescado con el objetivo de verificar si se trata o no de pescado de alta calidad. Si por llevar a cabo este control de calidad le cobraran más de 35,2 , no interesa llevarlo a cabo.

Sin embargo, no es la unica forma de resolver el problema.

3.5 Otros criterios de decisión

Hasta ahora nos hemos centrado en un solo decisor, esto es que solo estamos estudiando las preferencias de una persona, pero no siempre es asi...

A menudo, una decisión debe ser tomada por un comité o una asamblea y aqui vienen los problemas por que claramente las preferencias de cada quien son diferentes y pueden ser contrarias.

La fuerza del criterio de maximización de la utilidad esperada reside en su fundamento axiomatico, ya que otros criterios podrian ser incoherentes.

Por ejemplo: Una persona puede afirmar que prefiere 100 pesos seguros a 101 con probailidad p y nada con probabilidad 1-p. Por que arriesgarme a perder 100 por la posibilidad de ganar un poco más?

En efecto, si:

$$100 > \{101|p, 0(1-p)\}$$

entonces,

$$101 > \{102|p, 0(1-p)\}$$

y por tanto

$$100 > \left\{102|p^2, 0(1-p^2)\right\}$$

Por lo tanto:

$$100 > \left\{1000|p^{900}, 0(1-p^{900})\right\}$$

Evidentemente, esta afirmación es incoherente ya que para un p tal que p^{900} sea suficientemente grande, cualquiera preferiria 1000 con esa probabilidad a 100 seguros.

En conclusión el criterio de maximización de la utilidad no toma en cuenta el riesgo...

Pensemos en inversiones, maximizar la utilidad esperada implicaria invertir todo el capital en el tipo de accion que se espere que sea más rentable. D:

La distinción entre el riesgo e incertidumbre fue establecida por F. Knight en 1921, quien en su obra *Risk, uncertainty and profit* se refería a la primera como aquella situación en la que no existe certeza sobre el resultado de la decisión, aunque se conoce al menos la probabilidad de los distintos resultados alternativos. Este sería el caso, por ejemplo, de la elección entre cara o cruz de una moneda: desconocemos de antemano el resultado (si la moneda no está trucada, claro está) pero conocemos la probabilidad objetiva de las dos alternativas. Las situaciones de incertidumbre se caracterizarían, en cambio, por el hecho de que no sólo desconocemos el resultado final, sino que no podemos predecirlo tampoco en términos de probabilidades objetivas.. Así pues, uno de los problemas centrales a los que se ha enfrentado la teoría de la decisión ha consistido en establecer algún criterio (o criterios) que nos permita optar por una acción u otra en situaciones de incertidumbre. Algunos de los criterios más conocidos se exponen a continuación:

3.5.1 Criterio Pesimista:

El criterio *minimax* consiste en tomar la decision que maximiza la utilidad garantizada: se determina lo peor que puede pasar en cada caao, y se elige aquella decisión para la que resulte menos mal.

La decisión optima sería la que maximiza la utilidad minima qe se pudiera tener:

	ϵ_1	ϵ_2	max	
a_1	1	0	0	
a_2	0.3	0.3	0.3	max

Por lo que la decisión optima seria a_2

3.5.2 Criterio Optimista:

El criterio *maximax* consiste en tomar la decisión qe maximixa el maximo de la utilidad .

En nuestro ejemplo seria:

	ϵ_1	ϵ_2	max	
a_1	1	0	1	max
a_2	0.3	0.3	0.3	

Por lo que la decisión optima seria a_1

3.5.3 Criterio de Laplace:

Este criterio considera que todos los estados son equiprobables y por ello debemos calcular la esperanza de la utilidad para cada a_i y elegir el maximo.

En nuestro ejemplo seria:

	ϵ_1	ϵ_2	max	
a_1	1	0	0.5	max
a_2	0.3	0.3	0.3	

Por lo que la decisión optima seria a_1

Capítulo 4

De la información a priori a la distribución *a priori*

4.1 Introducción

La distribución inicial, o distribución *a priori* (*prior distribution*), $p(\theta)$ describe la información que se obtiene sobre θ antes de los valores observados $x_1, \ldots, x_n$. Si se cuenta con algún tipo de información inicial acerca de θ, como juicios, creencias, o información histórica, ésta se traduce en la distribución inicial $p(\theta)$.

No hay una única forma de elegir una distribución a priori. Distintas distribuciones iniciales pueden dar lugar a inferencias distintas, lo cual tiene ventajas y desventajas.

La elección de una distribución a priori influye en la inferencia. Esta influencia sobre los resultados puede ser poca o mucha. Será necesario tomar en cuenta lo siguiente:

- Distribuciones a priori mal fundadas producen inferencias a posteriori No justificadas.
- No hay una distribución a priori que sea la única, excepto en casos muy especiales.

4.2 Información histórica

Para la determinación subjetiva y las aproximaciones de la distribución a priori, será necesario especificar la distribución a priori con base en la información inicial que se tiene:

- Información histórica.

- Experiencia del especialista.

Para obtener la distribución incial será necesario elicitar (*elicit*) una distribución inicial, es decir, traducir esa información inicial en una distribución inicial.

Ejemplo 4.1. Considera la familia conjugada tal que $X \sim \text{Binomial}(n, \theta)$, $\theta \sim \text{Beta}(a, b)$. Sabemos que la distribución final de θ es $\theta|x \sim \text{Beta}(a + x, b + n - x)$.

Una aplicación de este tipo de distribuciones es la siguiente. Se necesita calcular la prevalencia de una enfermedad[1]. Se toma una muestra de $n = 20$ sujetos, y se define como X al número de personas infectadas, y a θ la proporción de sujetos infectados.

Se tiene *información inicial* del experto, y se sabe que la tasa de infección se encuentra entre 0.05 y 0.20 con un 95% de probabilidad, y que la prevalencia promedio es de 0.10. Si nosotros queremos traducir (*elicitar*) esta información inicial en una distribución inicial de tipo $\theta \sim \text{Beta}(a, b)$, tendríamos que resolver el sistema de ecuaciones:

$$\begin{aligned} \mathbb{E}[\theta] = \frac{a}{a+b} &= 0.10 \\ \mathbb{P}[0.05 < \theta < 0.20] &= 0.95 \end{aligned}$$

Lo cual resulta aproximadamente en $a = 8$ y $b = 72$. ■

4.3 Distribuciones poco informativas y muy informativas

Para explicar las distribuciones muy informativas y poco informativas, utilizaremos un ejemplo.

Considera una muestra $x_1, \ldots, x_n$ de Binomial(k, θ), y la distribución inicial $\theta \sim \text{Beta}(a, b)$. Sabemos que la distribución final de θ es $\theta|x_1, \ldots, n_n \sim \text{Beta}\left(a + \sum_{i=1}^n x_i \ , \ b + kn - \sum_{i=1}^n x_i\right)$.

El estimador puntual de θ es $\hat{\theta} = \mathbb{E}[\theta|x] = \frac{a+\sum_{i=1}^n x_i}{a+b+nk}$.

Distribución a Priori Muy Informativa
Si la magnitud de a y b son 'muy grandes', en comparación a la magnitud de $\sum_{i=1}^n x_i$ y nk, entonces la información a priori tendrá más peso que la información muestral sobre las inferencias que se realicen con la distribución a posteriori.

Distribución a Priori Poco Informativa
Si la magnitud de a y b es 'cercana a cero' entonces la muestra tendrá mayor peso, la

[1]En epidemiología, se denomina prevalencia a la proporción de individuos de un grupo o una población (en medicina, persona), que presentan una característica o evento determinado (en medicina, enfermedades).

distribución a priori tendrá varianza muy grande, las inferencias que se hagan se basarán prácticamente en la información obtenida a partir de los datos muestrales.

4.3.1 Familia conjugada

Si $p(\theta)$ es conjugada a una familia paramétrica $f(x|\theta)$, entonces las distribuciones a priori y a posteriori, y las distribuciones predictivas a priori y a posteriori pertenecen a la familia de distribuciones. ¿Bajo qué condiciones es posible construir una familia conjugada? Usualmente dos distribuciones pretenecientes a la Familila Exponencial:

$$f(x|\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\}$$

donde $a(\theta)$ y $c(\theta)$ son funciones de θ, y $b(x)$ y $d(x)$ son funciones solo de x, tales que,

$$\int f(x|\theta)\mathrm{d}x = \int a(\theta)b(x)\exp\{c(\theta)d(x)\}\,\mathrm{d}x = 1$$

Familia Paramétrica	Familia Conjugada
Bernoulli(θ)	Beta($\theta|a,b$)
Binomial(k,θ)	Beta($\theta|a,b$)
Poisson(θ)	Gamma($\theta|a,b$)
Binomial-Negativa(α,θ)	Beta($\theta|a,b$)
Geométrica(θ)	Beta(θ)
Normal(μ)	Normal($\mu|m_0,s_0^2$)
Normal(μ,λ)	Normal-Gamma($\mu,\lambda|m_0,n_0,a,b$)
Exponencial(θ)	Gamma($\theta|a,b$)
Uniforme($0,\theta$)	Pareto($\theta|a,b$)

Ejemplo 4.2. Sea $x_1, \ldots, x_n$ una muestra de v.a.i.i.d. $f(x|\theta)$ perteneciente a la familia exponencial:

$$f(x|\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\}$$

La función de verosimilitud:

$$\begin{aligned} L(\theta|\underline{x}) &= f(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} a(\theta)b(x_i)\exp\{c(\theta)d(x_i)\} \\ &= [a(\theta)]^n \prod_{i=1}^{n} b(x_i)\exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)\right\} \\ &\propto [a(\theta)]^n \exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)\right\} \end{aligned}$$

Para la distribución inicial, se debe elegir $f(\theta)$ tal que pertenezca a la familia exponencial y sea de la forma:

$$f(\theta) \propto [a(\theta)]^{n_0}\exp\{c(\theta)d_0\}$$

tal que $a(\theta)$ y $c(\theta)$ sean iguales en $f(x|\theta)$ y $f(\theta)$.

La distribución final entonces es:

$$\begin{aligned} f(\theta|\underline{x}) &\propto L(\theta|\underline{x})f(\theta) \\ &\propto [a(\theta)]^n \exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)\right\} \times [a(\theta)]^{n_0}\exp\{c(\theta)d_0\} \\ &\propto [a(\theta)]^{n+n_0} \exp\left\{c(\theta)\left[\sum_{i=1}^{n} d(x_i) + d_0\right]\right\} \\ &\propto [a(\theta)]^{n_1}\exp\{c(\theta)d_1\} \end{aligned}$$

Por lo tanto, la distribución final de θ también pertenece a la familia exponencial.

Note que, una distribucón inicial conjugada con hiperparámetros n_0 y d_0 contiene la misma información que una muestra de tamaño n_0 que resulta en un valor de la estadística suficiente igual a d_0. ■

4.4 Distribuciones No informativas

Cuando No existe información inicial, es imposible justificar la elección de las distribuciones iniciales sobre una base subjetiva, y los hiperparámetros de las distribuciones iniciales No pueden determinarse.

Para las distribuciones a priori No informativas:

- No se espera que representen ignorancia total. Se toman como referencia o 'por defecto'.
- Algunos son más útiles o eficientes que otras, pero No necesariamente son 'menos informativas' que otras.

NOTA: Puede ocurrir que las observaciones No modifican la distribución de algunos prámetros. Esto ocurre cuando la distribución de X no depende de estos parámetros, como en escenarios No identificables.

Ejemplo 4.3. Sean la distribución de X y la distribución a priori:

$$\begin{aligned} X &\sim \text{Normal}\left(\frac{\theta_1+\theta_2}{2}, 1\right) \\ f(\theta_1,\theta_2) &= f_1(\theta_1+\theta_2) f_2(\theta_1-\theta_2) \end{aligned}$$

Haciendo cambio de variables:

$$\xi_1 = \frac{\theta_1+\theta_2}{2} \qquad \xi_2 = \frac{\theta_1-\theta_2}{2}$$

o equivalentemente

$$2\xi_1 = \theta_1+\theta_2 \qquad 2\xi_2 = \theta_1-\theta_2$$

La distribución final de ξ_2 es:

$$\begin{aligned} f(\xi_2|x) &\propto \int_{\mathbb{R}} \exp\left\{-\tfrac{1}{2}(x-\xi_1)^2\right\} 2f_1(2\xi_1) 2f_2(2\xi_2) \mathrm{d}\xi_1 \\ &\propto f_2(2\xi_2) \int_{\mathbb{R}} \exp\left\{-\tfrac{1}{2}(x-\xi_1)^2\right\} f_1(2\xi_1) \mathrm{d}\xi_1 \\ &\propto f_2(2\xi_2) \qquad \forall x \end{aligned}$$

donde $f_2(2\xi_2)$ No depende de x. Por lo tanto, la información de x No aporta información a la distribución de ξ_2. ■

4.4.1 Prior Laplace

Las distribuciones a priori de Laplace consisten en asignar distribuciones uniformes, basado sobre la equiprobabilidad de los eventos.

Existen varias críticas respecto a asignar distribuciones iniciales uniformes.

- Las distribuciones son impropias, es decir,

$$\int_{\Theta} f(\theta)\mathrm{d}\theta = \infty$$

  cuando el espacio paramétrico Θ No es compacto.
  Por ejemplo,

$$\begin{aligned} f(\theta) &\propto \frac{1}{\theta}, \quad \theta \in \Theta, \qquad \text{donde } \Theta = [0, \infty) \\ f(\theta) &\propto 1, \quad \theta \in \Theta, \qquad \text{donde } \Theta = (-\infty, \infty) \end{aligned}$$

  Esto a veces resulta en 'paradojas de marginalización'.

- El principio de eventos equiprobables No es coherente ante particiones.
  Por ejemplo,

$$\Theta = \{\theta_1, \theta_2\} \quad \text{entonces} \quad f(\theta_1) = f(\theta_2) = \frac{1}{2}$$

  Pero si se particiona $\theta_2 = \{\omega_1, \omega_2\}$, entonces, bajo el principio de eventos equiprobables,

$$\Theta = \{\theta_1, \omega_1, \omega_2\} \quad \text{entonces} \quad f(\theta_1) = f(\omega_1) = f(\omega_2) = \frac{1}{3}$$

  Lo cual resulta contradictorio para θ_1.

- Invarianza bajo reparametrizaciones.
  Sea $\theta \in \Theta$, reparametrizar $\eta = g(\theta)$, donde $g(\cdot)$ es una transformación uno-a-uno. Si $f(\theta) = 1$, la distribución de η es

$$\begin{aligned} f(\eta) &= \left| \frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta) \right| \times f\left(\theta(\eta)\right) \\ f(\eta) &= \left| \frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta) \right| \times 1 \end{aligned}$$

  por el Jacobiano, donde ahora $f(\eta)$ No necesariamente será constante.

Ejemplo, sea $\theta \sim \mathrm{U}(0,1)$, transformando por su momio (odds) $\eta = g(\theta) = \frac{\theta}{1-\theta}$, e

$$\begin{aligned} \eta &= g(\theta) = \frac{\theta}{1-\theta} \\ \theta &= g^{-1}(\eta) = \frac{\eta}{1+\eta} \\ \frac{\mathrm{d}}{\mathrm{d}\eta} g^{-1}(\eta) &= \frac{(1+\eta)-\eta}{(1+\eta)^2} = \frac{1}{(1+\eta)^2} \\ f(\eta) &= \frac{1}{(1+\eta)^2} \qquad \text{No constante} \end{aligned}$$

4.4.2 Familia Conjugada No Informativa o Poco Informativa

Dentro de las distribuciones pertenecientes a las familias conjugadas pueden considerarse casos 'poco informativos' o 'No informativos',

Ejemplos:

Considerando una prior $\theta \sim \text{Gamma}(a, b)$, el caso No informativo implicaría que $a \to 0$ y $b \to 0$, digamos, $\theta \sim \text{Gamma}(0.001, 0.001)$.

Considerando una prior $\mu \sim \text{Normal}(m_0, s_0^2)$, el caso No informativo implicaría que $m_0 \to 0$ y $s_0^2 \to \infty$ esto parece una línea horizontal en $\mathbb{R}$) digamos, $\mu \sim \text{Normal}(0, 10000)$.

Considerando una prior $\theta \sim \text{Beta}(a, b)$, el caso No informativo implicaría que $a \to 0$ y $b \to 0$, digamos, $\theta \sim \text{Beta}(0.001, 0.001)$.

4.4.3 Criterio de Jeffreys

La distribución No informativa de Jeffreys consiste en:

- Buscar simultáneamente invarianza antes transformaciones.
- Proveer menor información a priori en relación a la información muestral, vía la información de Fisher.
- Es posible para espacios paramétricos infinitos.

Caso unidimensional

Información de Fisher, considerando θ unidimensional,

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

Bajo condiciones de regularidad[2], se cumple que

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

La distribución a priori de Jeffreys se define como

$$f(\theta) \quad \propto \quad \sqrt{\mathbb{I}(\theta)}$$

La proporcionalidad está sujeta a un coeficiente de normalización cuando $f(\theta)$ es propia[3].

La información de Fisher es una forma de medir la cantidad de información que tiene una v.a. X acerca del parámetro θ, del cual depende X a través de la función de densidad $f(x|\theta)$.

[2]Sea $X_1, \ldots, X_n$ una muestra aleatoria de $f(x;\theta)$ y sea $\mathrm{T}(\underline{\mathrm{X}})$ un estimador insesgado de $\tau(\theta)$. Las siguientes se conocen como **condiciones de regularidad**:

- El soporte de $f(x;\theta)$ se define como $sop(f) = \{x : f(x) > 0\}$ y este es el mismo para toda θ.
- Para todo $x \in sop(f)$, $\frac{\partial}{\partial\theta}\ln f(x;\theta)$ existe.
- $\frac{\partial}{\partial\theta}\int\int\ldots\int T(\underline{x})f(\underline{x};\theta)dx_1\ldots dx_n = \int\int\ldots\int\frac{\partial}{\partial\theta}T(\underline{x})f(\underline{x};\theta)dx_1\ldots dx_n$.
- $\frac{\partial}{\partial\theta}\int\int\ldots\int f(\underline{x};\theta)dx_1\ldots dx_n = \int\int\ldots\int\frac{\partial}{\partial\theta}f(\underline{x};\theta)dx_1\ldots dx_n$.
- $0 < \mathbb{E}\left[\left(\frac{\partial\ \ln f(x;\theta)}{\partial\theta}\right)^2\right] < \infty$.

[3]Una función de densidad $f(\theta)$ se dice propia cuando $\int_\theta f(\theta)\mathrm{d}\theta = k < \infty$, con k el coeficiente de normalización. Y es impropia cuando $\int_\theta f(\theta)\mathrm{d}\theta = \infty$.

Bajo la distribución a priori de Jeffreys se satisface el requerimiento de invarianza ante reparametrizaciones, ya que si $\phi = \phi(\theta)$, y $\phi(\cdot)$ es una transformación uno-a-uno,

$$\begin{aligned} f(\phi) &= f(\theta)\left|\phi'(\theta)\right|^{-1} \\ \mathbb{I}(\phi) &= \mathbb{I}(\theta)\left|\frac{\mathrm{d}}{\mathrm{d}\phi}\theta\right|^2 \qquad \text{(explica la potencia 1/2)} \\ \sqrt{\mathbb{I}(\phi)} &= \sqrt{\mathbb{I}(\theta)}\left|\frac{\mathrm{d}}{\mathrm{d}\phi}\theta\right| \end{aligned}$$

La información de Fisher $\mathbb{I}(\theta)$ es un indicador del incremento de la información producida por el modelo (o las observaciones) acerca de θ.

Ejemplo 4.4. Sea $X \sim \text{Binomial}(n, \theta)$, entonces,

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ \log f(x|\theta) &= \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta) \\ \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta) &= x\frac{1}{\theta} + (n-x)\frac{(-1)}{1-\theta} = \frac{x - x\theta - n\theta + x\theta}{\theta(1-\theta)} = \frac{x-n\theta}{\theta(1-\theta)} \end{aligned}$$

$$\mathbb{I}(\theta) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{x-n\theta}{\theta(1-\theta)}\right)^2\right] = \frac{\mathbb{E}\left[(x-n\theta)^2\right]}{\left[\theta(1-\theta)\right]^2} = \frac{Var\,(x)}{\left[\theta(1-\theta)\right]^2}$$

$$\begin{aligned} \mathbb{I}(\theta) &= \frac{n\theta(1-\theta)}{\left[\theta(1-\theta)\right]^2} = \frac{n}{\theta(1-\theta)} \\ \sqrt{\mathbb{I}(\theta)} &\propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \text{Beta}(\tfrac{1}{2}, \tfrac{1}{2}) \end{aligned}$$

Entonces la distribución no informativa inicial de Jeffreys es equivalente a usar una distribución inicial conjugada Beta$(\frac{1}{2}, \frac{1}{2})$. ■

Ejemplo 4.5. Sea $X \sim \text{Poisson}(\theta)$, entonces,

$$\begin{aligned} f(x|\theta) &= \frac{e^{-\theta}\theta^x}{x!} \\ \log f(x|\theta) &= -\theta + x\log\theta - \log x! \\ \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta) &= -1 + x\frac{1}{\theta} = \frac{-\theta + x}{\theta} \end{aligned}$$

$$\mathbb{I}(\theta) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{-\theta+x}{\theta}\right)^2\right] = \frac{\mathbb{E}\left[(-\theta+x)^2\right]}{\theta^2} = \frac{Var\,(x)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

$$\sqrt{\mathbb{I}(\theta)} \propto \theta^{-1/2}$$

■

Ejemplo 4.6. Sea $X \sim \text{Normal}(\mu, \sigma^2)$, con σ^2 conocida.

$$\begin{aligned} f(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ \log f(x|\mu) &= -\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x-\mu)^2 \\ \frac{\mathrm{d}}{\mathrm{d}\mu}\log f(x|\mu) &= \frac{1}{2\sigma^2}2(x-\mu) \end{aligned}$$

$$\mathbb{I}(\mu) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\mu}\log f(x|\mu)\right)^2\right] = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right] = \frac{\mathbb{E}\left[(x-\mu)^2\right]}{\sigma^4} = \frac{Var\,(x)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} \propto 1$$

$$\sqrt{\mathbb{I}(\mu)} \propto 1$$

■

Caso multidimensional

La matriz de Información de Fisher es

$$\begin{aligned} \mathbb{I}(\theta) &= [\mathbb{I}_{ij}(\theta)] \\ \mathbb{I}_{ij}(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x|\theta)\right] \end{aligned}$$

donde $\mathbb{I}_{ij}(\theta)$ son las entradas (i,j), para $i,j = 1,\ldots,k$, de la matriz $\mathbb{I}(\theta)$, donde $\theta \in \Theta \subseteq \mathbb{R}^k$.

La distribución a priori de Jeffreys se define como

$$f(\theta) \propto \sqrt{\det\mathbb{I}(\theta)}$$

donde $\det\mathbb{I}(\theta)$ es el determinante de la matriz de información de Fisher.

Ejemplo 4.7. Sea $X \sim \text{Normal}(\mu, \sigma^2)$, con mu y σ^2 desconocidas.

$$
\begin{aligned}
f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\
\log f(x|\mu,\sigma^2) &= -\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x-\mu)^2 \\
\frac{\partial}{\partial\mu}\log f(x|\mu,\sigma^2) &= \frac{1}{2\sigma^2}2(x-\mu) \;=\; \frac{1}{\sigma^2}(x-\mu) \\
\frac{\partial}{\partial\sigma^2}\log f(x|\mu,\sigma^2) &= -\frac{1}{2}\frac{(\sigma^2)^{-1/2}}{(\sigma^2)^{1/2}} + \frac{1}{2\sigma^4}(x-\mu)^2 \;=\; -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \\
\frac{\partial^2}{\partial\mu^2}\log f(x|\mu,\sigma^2) &= -\frac{1}{\sigma^2} \\
\frac{\partial^2}{\partial(\sigma^2)^2}\log f(x|\mu,\sigma^2) &= +\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \\
\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(x|\mu,\sigma^2) &= -\frac{1}{\sigma^4}(x-\mu) \\
\frac{\partial^2}{\partial\sigma^2\partial\mu}\log f(x|\mu,\sigma^2) &= -\frac{1}{\sigma^4}(x-\mu) \\
\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\log f(x|\mu,\sigma^2)\right] &= -\frac{1}{\sigma^2} \\
\mathbb{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log f(x|\mu,\sigma^2)\right] &= +\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\mathbb{E}\left[(x-\mu)^2\right] \;=\; \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}Var(x) \;=\; \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \;=\; -\frac{1}{2\sigma^4} \\
\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(x|\mu,\sigma^2)\right] &= \mathbb{E}\left[\frac{\partial^2}{\partial\sigma^2\partial\mu}\log f(x|\mu,\sigma^2)\right] \;=\; +\frac{1}{\sigma^4}\mathbb{E}\left[(x-\mu)\right] \;=\; 0 \\
\mathbb{I}(\mu,\sigma^2) \;=\; \left[\mathbb{I}_{ij}(\mu,\sigma^2)\right] &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \\
\sqrt{\det\mathbb{I}(\mu,\sigma^2)} &\propto \frac{1}{\sigma^3}
\end{aligned}
$$

Proposición 4.1 (Invarianza de Jefffreys)**.** La distribución prior No informativa de Jeffreys $f(\theta) \propto \sqrt{\mathbb{I}(\theta)}$ es invariante ante transformaciones uno-a-uno, es decir, si $\eta = \eta(\theta)$ es una transformación uno-a-uno de θ, entonces la distribución a priori de η es $f(\eta) \propto \sqrt{\mathbb{I}(\eta)}$

Demostración. Sea $\eta = \eta(\theta)$ una transformación uno-a-uno de θ.

$$
\frac{\mathrm{d}}{\mathrm{d}\eta}\log f(x|\eta) \;=\; \frac{\mathrm{d}\theta}{\mathrm{d}\eta}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\eta(\theta))
$$

donde $\theta(\eta) = \theta$ es la inversa de $\eta = \eta(\theta)$. Entonces

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\log f(x|\eta) &= \frac{\mathrm{d}^2\theta}{\mathrm{d}\eta^2}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\eta(\theta)) + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\eta(\theta)) \\
\mathbb{I}(\eta) &= -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\log f(x|\eta)\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2\theta}{\mathrm{d}\eta^2}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\eta(\theta)) + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\eta(\theta))\right] \\
&= -\mathbb{E}\left[\frac{\mathrm{d}^2\theta}{\mathrm{d}\eta^2}\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\eta(\theta))\right] - \mathbb{E}\left[\left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\eta(\theta))\right] \\
&= -\frac{\mathrm{d}^2\theta}{\mathrm{d}\eta^2}\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\eta(\theta))\right] - \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x|\eta(\theta))\right] \\
&= \frac{\mathrm{d}^2\theta}{\mathrm{d}\eta^2} \times 0 + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \mathbb{I}(\theta) \\
&= \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2 \mathbb{I}(\theta)
\end{aligned}$$

porque

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta)\right] &= \mathbb{E}\left[\frac{1}{f(x|\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}f(x|\theta)\right] = \int \frac{1}{f(x|\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}f(x|\theta)f(x|\theta)\mathrm{d}x \\
&= \int \frac{\mathrm{d}}{\mathrm{d}\theta}f(x|\theta)\mathrm{d}x \qquad \text{por condiciones de regularidad} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\int f(x|\theta)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}1 = 0
\end{aligned}$$

Por lo tanto

$$\sqrt{\mathbb{I}(\eta)} \propto \sqrt{\mathbb{I}(\theta)}\sqrt{\left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right)^2} = \sqrt{\mathbb{I}(\theta)}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right|$$

donde $\left|\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right|$ es el valor absoluto del Jacobiano de la transformación inverse.

Por lo tanto, si la prior de Jeffreys para θ es

$$f(\theta) \propto \sqrt{\mathbb{I}(\theta)}$$

entonces, la prior de Jeffreys para η es

$$f(\eta) \propto \sqrt{\mathbb{I}(\theta(\eta))}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right| = \sqrt{\mathbb{I}(\eta)}$$

■ □

4.5 Análisis de Sensibilidad - Robustez

En el análisis de datos, será necesario tomar en consideración lo siguiente:

- Que el modelo sea robusto ante posibles faltas a los supuestos.
- Especificaciones en la distribución inicial.
- Cuando el tamaño de muestra es lo suficientemente grande, a veces las inferencias de la distribución final No se modifican sustancialmente ante cambios moderados en la distribución inicial.
- Identificar si cambios en la distribución inicial producen inferencias distintas.

Existen varias sugerencias al respecto:

- Comparar los resultdos considerando 3 distribuciones iniciales distintas:
  - No informativa.
  - Informativa.
  - Poco informativa.
- En caso de No exitir cambios, No será necesario preocuparse demasiado por la distribución inicial.
- En caso de Sí existir cambios significativos, entonces especificar una distribución inicial que refleje con mayor precisión la información inicial.

Bibliografía

Albert, J. (2009). *Bayesian Computation with R* (second ed.). New York: Springer.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Bernardo, J. M. (1981). *Bioestadística: Una Perspectiva Bayesiana*. Vicens-Vives.

Bernardo, J. M. and A. Smith (1994). *Bayesian Theory*. Wiley.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley.

Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim (2000). *Monte Carlo Methods in Bayesian Computation*. Series in Statistics. New York: Springer.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Chichester: John Wiley & Sons.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (2000). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (Second ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. Chapman & Hall-CRC.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gomez-Rubio, V. (2020). *Bayesian Inference with INLA* (1st ed.). Boca Raton, Florida: Chapman and Hall/CRC.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer.

Johnson, V. E. and J. H. Albert (1999). *Ordinal Data Modeling*. New York: Springer.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2012). *The BUGS Book - A Practical Introduction to Bayesian Analysis*. Boca Raton, Florida: CRC Press / Chapman & Hall.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing 10*(4), 325–337.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New Jersey: John Wiley & Sons.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: John Wiley & Sons.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robert, C. P. (1994). *The Bayesian Choice*. New York: Springer-Verlag.

Robert, C. P. (2007). *The Bayesian Choice* (Second ed.). New York: Springer.

Stan Development Team (2015). *Stan Modeling Language User's Guide and Reference Manual*. Version 2.9.0.

Tan, M. T., G.-L. Tian, and K. W. Ng (2010). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. Boca Raton, Florida: Chapman & Hall/CRC Biostatistics Series.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

Wang, X., Y. Yue, and J. Faraway (2018). *Bayesian Regression Modeling with INLA*. Chapman & Hall/CRC.