ESTADÍSTICA BAYESIANA

Iker Caballero Bragagnini

Tabla de contenido

La probabilidad y la inferencia

- La inferencia bayesiana es el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir el resultado con una distribución de probabilidad sobre los parámetros del modelo y sobre la cantidades no observadas (como predicciones para nuevas observaciones)

    - La característica esencial de los métodos bayesianos es el uso explícito de la probabilidad para cuantificar la incertidumbre en las inferencias basadas en el análisis estadístico de los datos. El proceso de análisis de datos bayesiano se puede dividir en los tres siguientes pasos:

        - La construcción de un modelo de probabilidad completo, el cual es una distribución de probabilidad conjunta para todas las cantidades observables e inobservables en un problema. El modelo debe ser consistente con el conocimiento del problema científico subyacente y el proceso de recolección de datos

        - El condicionamiento de los datos observados, que se basa en calcular e interpretar la distribución posterior apropiada, la cual es una distribución de probabilidad condicional para las cantidades no observadas de interés a partir de los datos observados

        - La evaluación del ajuste del modelo y las implicaciones de los resultados de la distribución posterior, que consiste en saber como de bien se ajusta el modelo a los datos, si la conclusiones son razonables y qué tan sensibles son los resultados a la suposiciones del modelo. En respuesta a esta evaluación, uno puede alterar o expandir el modelo y repetir estos tres pasos

    - Una motivación primaria para el pensamiento bayesiano es que facilita la interpretación lógica de las conclusiones estadísticas

        - Por ejemplo, un intervalo de probabilidad bayesiano para una cantidad que no se conoce puede ser interpretado directamente como un intervalo que tiene una alta probabilidad de contener aquella cantidad desconocida, en contraste a un intervalo frecuentista, el cual se interpreta solo en relación a la secuencia de inferencias similares que pueden hacerse en la práctica repetida

        - La característica central de la inferencia bayesiana, la cuantificación directa de la incertidumbre, significa que no hay impedimento para ajustar modelos de muchos parámetros y especificaciones de probabilidad complicadas con muchas capas

- Thomas Bayes demostró como las probabilidades inversas pueden usarse para calcular la probabilidad de eventos antecedentes a partir de la ocurrencia de un evento consecuente, y en el siglo XX se desarrolló un método de inferencia estadística completo basado en el teorema de Bayes

    - Como hay incertidumbre sobre los valores reales de los parámetros, se considera que estos son variables aleatorias

    - Las reglas de la probabilidad se usan directamente para hacer inferencias sobre los parámetros

    - Las proposiciones de probabilidad sobre los parámetros se interpretan como un grado de creencia. La distribución *a priori* tiene que ser subjetiva, de modo que cada persona tiene su propia distribución que contiene los pesos relativos que se da a cada valor posible de los parámetros (la plausibilidad de que ocurra cada valor antes de que se observen según cada persona)

    - Se revisan las creencias sobre los parámetros después de obtener los datos a través del teorema de Bayes. Esto permite obtener una distribución posterior, que da los pesos relativos que se dan a cada parámetro después de analizar los datos, por lo que la distribución posterior proviene de la *a priori* y de los datos

- Esto tiene un número de ventajas sobre el enfoque frecuentista convencional:

    - El teorema de Bayes es la única manera consistente de modificar las creencias sobre los parámetros dados los datos. Esto significa que la inferencia se basa en la ocurrencia real de los datos, no en todos los posibles conjuntos de datos que podrían ocurrir, pero no han ocurrido (como en el enfoque frecuentista)

    - Dejar que los parámetros sean aleatorios permite que se hagan proposiciones de probabilidad sobre estos, después de obtener los datos. Esto contrasta con lo convencional, que se basa en todos los conjuntos de datos que podrían haberse obtenidos dado un parámetro fijo (por lo que solo se pueden hacer proposiciones de confianza basado en lo que podría haber ocurrido)

    - La estadística bayesiana tiene una manera general de lidiar con parámetros molestos (a diferencia de la estadística frecuentista), que son aquellos para los cuales no se quiere inferenciar, pero

interfieren con las inferencias hechas sobre los parámetros de interés

- La estadística bayesiana es predictiva, a diferencia de la estadística convencional, de modo que se puede encontrar una distribución de probabilidad condicional para la siguiente observación de los datos muestrales

- La inferencia estadística se centra en extraer conclusiones de datos numéricos sobre cantidades que no han sido observadas. Para ello, uno normalmente se basa en muestras y se centra en la inferencia causal
  - Se distingue entre dos tipos de estimandos, los cuales son cantidades no observadas para el cual se hacen las inferencias estadísticas: las cantidades potencialmente observables y las cantidades no directamente observables
    - Las cantidades potencialmente observables son cantidades tales como observaciones futuras de un proceso o el resultado de un tratamiento no recibido en un individuo
    - Las cantidades no directamente observables son cantidades como los parámetros que gobiernan el proceso hipotético que causan los datos observados (por ejemplo, coeficientes de una regresión)
    - La distinción entre ambos tipos de estimandos no siempre es precisa, pero es generalmente útil como una manera de entender como un modelo estadístico para un problema particular se ajusta al mundo real
  - Como notación general, $\boldsymbol{\theta}$ se referirá al vector de cantidades no observables o de parámetros de interés, $\boldsymbol{y}$ se referirá al vector de datos observados y $\tilde{\boldsymbol{y}}$ se referirá al vector de datos potencialmente observables
    - En muchos estudios estadísticos, los datos se recogen en cada uno del conjunto de n objetos o unidades, y se puede escribir el vector $\boldsymbol{y}$ como $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$
    - Estas variables se interpretan como resultados y se consideran aleatorias, en el sentido que, cuando se hacen inferencias, se desea que haya la posibilidad de que los valores observados de las variables puedan tomar valores diferentes (debido al proceso de muestreo o a la variación de la población)

- A partir de estas nociones, es posible definir el concepto de modelo estadístico y las implicaciones que tendrá este para construir el modelo bayesiano

    - El modelo estadístico M es una lista de modelos de probabilidad indexado en el vector de parámetros $\boldsymbol{\theta}$ que pertenece a un espacio muestral $\boldsymbol{\Theta}$. Las distribuciones de probabilidad que pertenecen a la lista comparten el mismo soporte o espacio muestral $\tilde{\boldsymbol{y}} \in \boldsymbol{\Omega}$

    $$M = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

    - Como se puede observar, se especifica para $\tilde{\boldsymbol{y}}$ y no para $\boldsymbol{y}$ porque se asume que aún no se han observado los datos generados, solo que se distribuirán con una de las distribuciones incluidas en la lista

    - El diseño de los experimentos o de la recolección de datos determinarán M, ya que este modelo se determina por la naturaleza de los datos y por el tipo de muestreo

    - Una vez se recogen los datos $\boldsymbol{y}$, entonces se tiene que decidir si el modelo de probabilidad generado por los datos pertenece a M para ver si el modelo es correcto (validación del modelo) y, finalmente, realizar inferencia estadística para poder adivinar cuál es el verdadero vector $\boldsymbol{\theta}$

    $$p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) \Rightarrow p(\boldsymbol{y}|\boldsymbol{\theta}) \in M \Rightarrow Statistic\ Inference\ for\ \boldsymbol{\theta}^*$$

- El punto de partida común de un análisis estadístico es que los n valores y_i son intercambiables, es decir, que la incertidumbre se expresa con una densidad de probabilidad conjunta $p(y_1, \dots, y_n)$ que es invariante a permutaciones en los índices

    - Un modelo no intercambiable sería apropiado si la información relevante para el resultado residiera en los índices de las unidades y no en las variables explicativas

        - Normalmente se modelan los datos de una distribución intercambiable como independientes e idénticamente distribuidos (iid) dado un vector de parámetros desconocidos $\boldsymbol{\theta}$ con una distribución $p(\boldsymbol{\theta})$

    - Es común que haya observaciones en cada unidad que no se tengan que modelar como aleatorias, llamadas variables explicativas $\boldsymbol{x}$

        - Se utiliza $\boldsymbol{X}$ para denotar el conjunto entero de variables explicativas para todas las n unidades; si hay k variables explicativas, entonces la matriz $\boldsymbol{X}$ tiene n filas y k columnas

        - Tratando $\boldsymbol{X}$ como aleatorio, la noción de intercambiabilidad puede extenderse a requerir que la distribución de los n valores de $(x, y)_i$ no sean cambiados por permutaciones arbitrarias de los índices

        - Siempre es apropiado asumir intercambiabilidad en un modelo después de incorporar suficiente información relevante de $\boldsymbol{X}$ para que los índices se consideren asignados aleatoriamente

        - A partir de la suposición de intercambiabilidad se puede ver como la distribución de $\boldsymbol{y}$ dado $\boldsymbol{x}$ es la misma para todas las unidades del estudio, de modo que si dos unidades tienen el mismo valor de $\boldsymbol{x}$ sus distribuciones de $\boldsymbol{y}$ son iguales

    - Los modelos jerárquicos son modelos que se usan cuando la información está disponible en diferentes niveles de unidades observacionales

        - En un modelo jerárquico, es posible hablar de intercambiabilidad en cada nivel de unidades

- Las conclusiones estadísticas bayesianas sobre un vector de parámetros $\boldsymbol{\theta}$ o datos no observables $\tilde{\boldsymbol{y}}$ se hacen en términos de proposiciones de probabilidad. Estas proposiciones de probabilidad son condicionales al valor observado $\boldsymbol{y}$, usando entonces $p(\boldsymbol{\theta}|\boldsymbol{y})$ y $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ para denotar estas proposiciones

    - Además, se hace un condicionamiento implícito a los valores conocidos de cualquier variable explicativa $\boldsymbol{x}$

        - La inferencia bayesiana se diferencia del enfoque de la inferencia estadística por su nivel fundamental de condicionamiento de los datos observados, ya que normalmente la última se basa una evaluación retrospectiva del procedimiento usado para estimar $\boldsymbol{\theta}$ o $\tilde{\boldsymbol{y}}$ sobre la distribución de posibles valores de $\boldsymbol{y}$ condicional al verdadero valor desconocido de $\boldsymbol{\theta}$ o $\tilde{\boldsymbol{y}}$ (denotada por $p(\boldsymbol{y}|\boldsymbol{\theta})$ o $p(\boldsymbol{y}|\tilde{\boldsymbol{y}})$)

        - Aunque haya diferencias, se puede ver como en muchos análisis simples, se obtienen conclusiones superficialmente similares con ambos enfoques. No obstante, análisis obtenidos usando métodos bayesianos pueden ser normalmente extendidos a problemas más complejos

- La densidad de probabilidad condicional se representa como $p(\cdot \mid \cdot)$ y $p(\cdot)$ representa la densidad de probabilidad marginal (se usan los términos "distribución" y "densidad" de la misma manera), y esta notación se suele utilizar tanto para funciones de densidad (continuas) o de masa (discretas). Para denotar la probabilidad de un evento discreto también se suele utilizar $p(\cdot)$, aunque a veces se cambiará la notación por claridad

- Con tal de poder hacer proposiciones de probabilidad sobre $\boldsymbol{\theta}$ dado $\boldsymbol{y}$, se tiene que comenzar con un modelo para la distribución de probabilidad conjunta para $\boldsymbol{\theta}$ y $\boldsymbol{y}$

  - La función de densidad o de masa probabilidad conjunta puede escribirse como el producto de dos densidades: la distribución *a priori* $p(\boldsymbol{\theta})$ y la distribución muestral o verosimilitud $p(\boldsymbol{y}|\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$$

  - La distribución *a priori* da el peso que se da a cada posible valor de $\boldsymbol{\theta}$ a partir de el conocimiento *a priori*, y la distribución muestral o verosimilitud da el peso relativo que se da a cada posible valor de $\boldsymbol{\theta}$ cuando ocurren unos datos $\boldsymbol{y}$ (diferente a dar las probabilidades de un $\boldsymbol{\theta}$ concreto dado que $\boldsymbol{y}$ ocurre, expresadas por $p(\boldsymbol{\theta}|\boldsymbol{y})$)

- El condicionamiento sobre los valores de los datos $\boldsymbol{y}$ a través de la regla de Bayes permite obtener la densidad posterior, en donde $p(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$ si es una función de masa de probabilidad o $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$ si es una función de densidad de probabilidad

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

  - El cálculo más complicado siempre es el denominador, dado que hay que resolver al integral $\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$, la cual resultaría en una constante (porque se evalúa en los datos observados $\boldsymbol{y}$). No obstante, se puede identificar la familia de la distribución posterior solo con la forma del numerador

  - Aunque $p(\boldsymbol{y})$ sea una constante cuando se evalúa en $\boldsymbol{y}$, cuando se considera que $\boldsymbol{y}$ es una variable, se obtendría una función de densidad o una función de masa de probabilidad. Es importante distinguir si $p(\boldsymbol{\theta})$ es una densidad o una función de masa, porque dependiendo de ello se integrará (independientemente de si $p(\boldsymbol{y}|\boldsymbol{\theta})$ es continua o no) o se evaluará una suma como en el teorema de probabilidad total (independientemente de si $p(\boldsymbol{y}|\boldsymbol{\theta})$ es discreta o no)

- Una forma equivalente omite el factor $p(\boldsymbol{y})$ (el cual no depende de $\boldsymbol{\theta}$) y que, para una $\boldsymbol{y}$ fija, se puede considerar constante, de modo que se obtendría una densidad posterior no normalizada, que es la parte derecha de la siguiente relación:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$$

- El término $p(\boldsymbol{y}|\boldsymbol{\theta})$ en esta expresión se toma como una función de $\boldsymbol{\theta}$, no de $\boldsymbol{y}$ (al ser la verosimilitud), mientras que $p(\boldsymbol{\theta}|\boldsymbol{y})$ si es una función de los datos $\boldsymbol{y}$ (la distribución posterior)

- También es posible usar variables explicativas, de modo que estas condicionan las probabilidades

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) \propto p(\boldsymbol{\theta}|x)p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$$

- Estas fórmulas son el núcleo de la inferencia bayesiana, en donde la principal motivación es desarrollar un modelo $p(\boldsymbol{\theta}, \boldsymbol{y})$ y desarrollar cálculos para resumir $p(\boldsymbol{\theta}|\boldsymbol{y})$ apropiadamente. Estas permiten definir lo que se conoce como un modelo bayesiano

  - El modelo bayesiano M_B , en cambio, nace del modelo estadístico M pero considera que $\boldsymbol{\theta}$ es un vector o variable aleatoria, de modo que se puede escoger una distribución de probabilidad $p(\boldsymbol{\theta})$ sobre Ω basada en el conocimiento previo o creencias que se tenga sobre $\boldsymbol{\theta}$: la distribución *a priori*

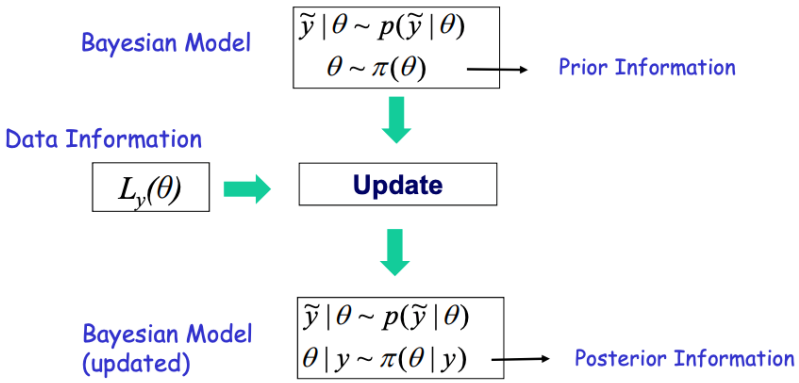
$$M_B = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}; p(\boldsymbol{\theta})\}$$

  - Comparando M y M_B se puede ver que el modelo estadístico no cambia, dado que es el modelo que se asume para una población concreta de estudio, y el único modelo que cambia es el modelo bayesiano, el cual cambia a través de cambiar $p(\boldsymbol{\theta})$

  - Definir el modelo bayesiano significa, por tanto, escoger un modelo estadístico y una distribución *a priori*. La manera más común es dibujando la distribución *a priori* en caso de que se tenga información sobre el espacio paramétrico, o dibujar la distribución predictiva *a priori* (la cual se introduce después) si se tiene información sobre el espacio muestral pero no sobre el paramétrico

  - Igual que con el modelo estadístico M, una vez se recogen los datos $\boldsymbol{y}$, se utilizan estos para poder obtener la función posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, de modo que la función de verosimilitud $p(\boldsymbol{y}|\boldsymbol{\theta})$ es la

que realmente se usa para los cálculos y no $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$ (ya se tienen observaciones)



- Hacer inferencias sobre una variable desconocida no observable, llamadas inferencias predictivas, sigue una lógica similar a lo visto anteriormente

  - Antes de considerar los datos $\boldsymbol{y}$, es posible obtener la distribución de la variable desconocida pero observable $\tilde{\boldsymbol{y}}$ integrando en $\boldsymbol{\theta}$, dado que no resulta en una constante al no tener valores para los datos (no han sido observados). Esta densidad será la siguiente:

$$p(\boldsymbol{y} = \tilde{\boldsymbol{y}}) = p(\tilde{\boldsymbol{y}}) = \int_{\Theta} p(\boldsymbol{\theta}, \tilde{\boldsymbol{y}})\, d\boldsymbol{\theta} = \int_{\Theta} p(\boldsymbol{\theta}) p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

  - Esta distribución se suele llamar distribución marginal de $\boldsymbol{y}$, aunque un nombre más informativo es la distribución predictiva *a priori*: es *a priori* porque no está condicionada a la observación anterior del proceso (no depende de las observaciones porque no se han obtenido, solo depende de $\tilde{\boldsymbol{y}}$), y es predictiva porque es la distribución de una cantidad $\tilde{\boldsymbol{y}}$ que es observable pero que se desconoce (permite hacer predicciones antes de observar los datos)

  - La distribución predictiva *a priori* se puede entender como una media ponderada de todos los posibles modelos de probabilidad en donde los pesos se determinan por la distribución *a priori*

$$p(\boldsymbol{y} = \tilde{\boldsymbol{y}}) = p(\tilde{\boldsymbol{y}}) = E_{p(\boldsymbol{\theta})}[p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})]$$

  - Después de que los datos $\boldsymbol{y}$ se hayan observado, se puede predecir una variable no observable $\tilde{\boldsymbol{y}}$ a partir del mismo proceso. La distribución de $\tilde{\boldsymbol{y}}$ se denomina distribución predictiva posterior (se obtiene la función de densidad): es posterior porque está condicionada a los valores observados $\boldsymbol{y}$

(no depende de ellos, solo depende de $\tilde{\boldsymbol{y}}$) y predictiva porque es una predicción para una $\tilde{\boldsymbol{y}}$ observable (permite hacer predicciones sobre valores futuros teniendo en cuenta los datos)

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{\theta}, \tilde{\boldsymbol{y}}|\boldsymbol{y})\, d\boldsymbol{\theta} = \int_{\Theta} p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = \\ = \int_{\Theta} p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$

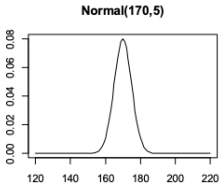
- La segunda y la tercera igualdad muestran la distribución predictiva posterior como la media de las predicciones condicionales sobre la distribución posterior de $\boldsymbol{\theta}$**.** La expresión final se debe al hecho de que se asume independencia condicional entre $\boldsymbol{y}$ e $\tilde{\boldsymbol{y}}$, por lo que $p(\tilde{\boldsymbol{y}}, \boldsymbol{y}|\boldsymbol{\theta}) = p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}, \tilde{\boldsymbol{y}}|\boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\tilde{\boldsymbol{y}}, \boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{y})$$

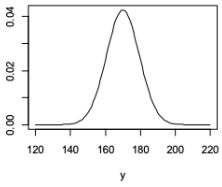
$$= p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})$$

- Las distribuciones predictivas traducen la información sobre el espacio paramétrico $\boldsymbol{\Theta}$ (y la de los datos observados $\boldsymbol{y}$ si se considera la posterior) al espacio muestral de $\tilde{\boldsymbol{y}}$, ya que, si se sabe algo de los parámetros, se sabe algo sobre como lucirán los datos y viceversa

$\pi(\mu)$



$p(\tilde{y})$



- Aunque las distribuciones predictivas *a priori* y posteriores se pueden obtener analíticamente integrando, estas se pueden aproximar utilizando métodos de simulación computacionales

  - Para poder aproximar la distribución predictiva *a priori*, se considera que se hacen $j = 1, \dots, M$ simulaciones. Primero se simulan valores $\boldsymbol{\theta}^{(j)}$ de $p(\boldsymbol{\theta})$, y después se simulan valores $\tilde{\boldsymbol{y}}^{(j)}$ de $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}^{(j)})$. Estos valores simulados $\tilde{\boldsymbol{y}}^{(1)}, \tilde{\boldsymbol{y}}^{(2)}, \dots, \tilde{\boldsymbol{y}}^{(M)}$

pertenecen a la distribución predictiva *a priori*, por lo que se puede hacer inferencia con la distribución aproximada resultante

- Para poder aproximar la distribución predictiva posterior, se considera que se hacen $j = 1, \dots, M$ simulaciones. Primero se simulan valores $\boldsymbol{\theta}^{(j)}$ de $p(\boldsymbol{\theta}|\boldsymbol{y})$, y después se simulan valores $\tilde{\boldsymbol{y}}^{(j)}$ de $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}^{(j)})$. Estos valores simulados $\tilde{\boldsymbol{y}}^{(1)}, \tilde{\boldsymbol{y}}^{(2)}, \dots, \tilde{\boldsymbol{y}}^{(M)}$ pertenecen a la distribución predictiva posterior, por lo que se puede hacer inferencia con la distribución aproximada resultante

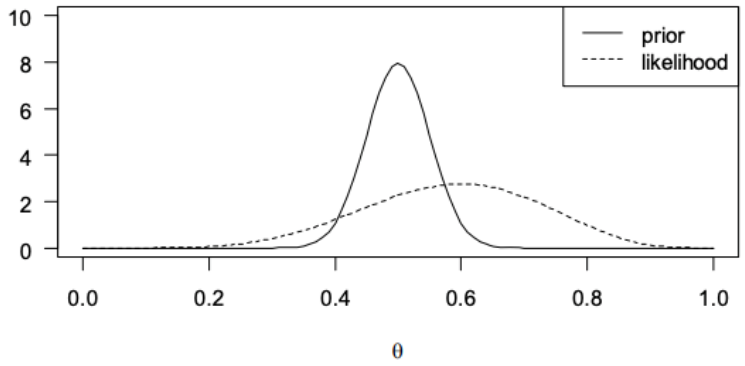
- Usando la regla de Bayes con un modelo estadístico escogido significa que los datos $\boldsymbol{y}$ afectan a la inferencia posterior solo a través de $p(\boldsymbol{y}|\boldsymbol{\theta})$ que, cuando se considera una función de $\boldsymbol{\theta}$ para un vector $\boldsymbol{y}$ fijo, se denomina función de verosimilitud

  - De este modo, la inferencia bayesiana obedece el principio de verosimilitud, que dice que, para unos datos muestrales dados, dos modelos de probabilidad cualesquiera que tienen la misma función de verosimilitud deben de resultar en la misma inferencia para $\boldsymbol{\theta}$

  - Este principio es razonable, pero solo dentro del marco de un modelo o familia de modelos adoptada para un análisis particular. En la práctica uno rara vez puede estar seguro que el modelo escogido es correcto

  - Al ser una función de verosimilitud y no una de probabilidad, esta función no tiene que integrar o sumar 1 para todo los valores posibles considerados de $\boldsymbol{\theta}$. Por lo tanto, para poder graficar la función se suele utilizar la verosimilitud estandarizada $p_{std}(\boldsymbol{y}|\boldsymbol{\theta})$, definida de la siguiente manera:

$$p_{std}(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$



- Esta estandarización permite que la función de verosimilitud $p_{std}(\boldsymbol{y}|\boldsymbol{\theta})$ integre o sume 1, lo cual se puede demostrar con su integral:

$$\int_{\Theta} p_{std}(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\Theta} \frac{p(\boldsymbol{y}|\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}\, d\boldsymbol{\theta} = \frac{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} = 1$$

  - La razón entre la densidad posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ evaluada en los puntos $\boldsymbol{\theta_1}$ y $\boldsymbol{\theta_2}$ bajo un modelo de probabilidad dado se conoce como las *odds* para $\boldsymbol{\theta_1}$ comparado con $\boldsymbol{\theta_2}$

    - La aplicación más familiar de este concepto es con parámetros discretos, con $\boldsymbol{\theta_2}$ tomado como el complemento de $\boldsymbol{\theta_1}$

    - Las *odds* proporcionan una representación alternativa de las probabilidades y tienen la propiedad de que la regla de Bayes toma una forma simple cuando se expresa con las *odds*

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

    - En otras palabras, las *odds* posteriores son iguales a las *odds a priori* multiplicadas por la razón de verosimilitud

- Como los usos de la probabilidad son más amplios en un marco bayesiano que en la estadística no bayesiana, es importante considerar fundamentos del concepto de probabilidad

  - En estadística bayesiana, la probabilidad se usa como medida fundamental de la incertidumbre

    - Dentro de este paradigma, es natural considerar tanto la probabilidad de que un estimando desconocido esté en un rango particular de valores como la probabilidad de que un estadístico de una muestra aleatoria de una población fija conocida esté dentro de un rango concreto

    - La primera de estas probabilidades es de mayor interés después de que los datos se recojan, mientras que la segunda es más relevante antes de eso

    - Los métodos bayesianos permiten hacer proposiciones con conocimiento parcial disponible (basado en los datos) en relación a una situación o estado de la naturaleza (inobservable o aún no observado) de una manera sistemática a través de la probabilidad. El principio más importante es que el estado del

conocimiento de cualquier cosa desconocida se describe con una distribución de probabilidad

- Las propiedades deseables que tienen las medidas de plausibilidad son las siguientes:

  - Los grados de plausibilidad se representan por números reales no negativos

  - Estas medidas concuerdan con el sentido común, de modo que números más grandes quieren decir más plausibilidad

  - Si una proposición puede representarse de más de una manera, entonces la plausibilidad de cada una debe ser la misma

  - Siempre se tiene que tener en cuenta toda la evidencia relevante

  - Los estados equivalentes de conocimiento tienen asignados la misma plausibilidad

  - El doctor R.T. Cox demostró como cualquier conjunto de plausibilidades que satisfacen las propiedades anteriores deben operar acorde a las mismas reglas de la probabilidad. Por lo tanto, la manera más inteligente para revisar las plausibilidades es usando las leyes de probabilidad, en la que se basa la estadística bayesiana

- Todos los métodos estadísticos que utilizan la probabilidad son subjetivos en el sentido de que se apoyan de idealizaciones matemáticas del mundo

  - Los métodos bayesianos suelen ser especialmente subjetivos porque se apoyan en la distribución *a priori*, aunque se necesite juicio científico para especificar la verosimilitud y la parte *a priori* del modelo

  - Cuando hay replicación, en el sentido que muchas unidades intercambiables se observan, entonces hay cabida a estimar las características de la distribución de probabilidad de los datos y hacer un análisis más objetivo

  - Si un experimento se replica varias veces, entonces los parámetros de la distribución *a priori* pueden ser estimados a partir de los datos. No obstante, ciertos elementos que requieren de juicio científico seguirán siendo necesarios (como

la selección de los datos, las formas paramétricas y las evaluaciones)

- Algunos resultados útiles para la manipulación de probabilidades y distribuciones de probabilidad son las siguientes:

  - Siendo $\mathcal{E}$ un experimento con un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, es posible que se tenga información incompleta del resultado real de $\mathcal{E}$ sin saber todo sobre este resultado

    - Si A y B son eventos y se dice que B ocurre, entonces, con esta información, la nueva probabilidad de que ocurra A puede no coincidir con la probabilidad $P(A)$

    - En esta nueva circunstancia, A ocurre si, y solo si, $A \cap B$ ocurre, sugiriendo que la nueva probabilidad de A puede ser proporcional a $P(A \cap B)$

  - Si $A, B \in \mathcal{F}$ y $P(B) > 0$, la probabilidad condicional de A dado B se denota por $P(A|B)$ y se define con la siguiente fórmula:

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

    - La constante de proporcionalidad de la definición se ha escogido de manera que la probabilidad $P(B|B)$ satisface $P(B|B) = 1$ (lo cual es lógico)

    - Si $A, B \in \mathcal{F}$ y $P(B) > 0$, entonces $(\Omega, \mathcal{F}, Q)$ es el espacio de probabilidad donde $Q: \mathcal{F} \to \mathbb{R}$ se define por $Q \equiv P(A|B)$

$$Q(A) \geq 0 \qquad Q(\Omega) = P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$A_1, A_2, \ldots \in \mathcal{F} \ \ such\ that \ \ A_i \cap A_j = \emptyset \ \ for\ i \neq j$$

$$\Rightarrow \ Q\left(\bigcup_i A_i\right) = \frac{1}{P(B)} P\left(\left(\bigcup_i A_i\right) \cap B\right) =$$

$$= \frac{1}{P(B)} P\left(\bigcup_i (A_i \cap B)\right) = \frac{1}{P(B)} \sum_i P(A_i \cap B) = \sum_i Q(A_i)$$

    - Si $A, B, C \in \mathcal{F}$ y $P(B \cap C), P(C) > 0$, entonces la probabilidad de la intersección de estos eventos es $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

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

$$= \frac{P(A \cap B \cap C)}{P(B \cap C)} P(B \cap C) = P(A \cap B \cap C)$$

- Si $A, B \in \mathcal{F}$ y $P(A), P(B) > 0$, entonces la probabilidad condicional $P(B|A) = P(A|B)P(B)/P(A)$

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

- Se llama independientes a dos eventos A y B si la ocurrencia de uno de los dos no afecta a la probabilidad (nueva) de que el otro ocurra. De manera más formal, los eventos A y B del espacio de probabilidad $(\Omega, \mathcal{F}, Q)$ son independientes si $P(A \cap B) = P(A)P(B)$, y son dependientes de otro modo

  - Esto quiere decir que, si $P(A), P(B) > 0$ y A y B son eventos independientes, entonces $P(A|B) = P(A)$ y $P(B|A) = P(B)$. No obstante, en la definición se permite que $P(A) = 0$ o $P(B) = 0$

  - Esta definición se puede generalizar aún más si se consideran más de dos eventos. Una familia $\mathcal{A} \equiv (A_i : i \in I)$ de eventos se llama independiente si, para todos los subconjuntos finitos J de I, se cumple la siguiente igualdad:

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

  - La familia $\mathcal{A}$ se denomina independiente a pares si la igualdad anterior se sostiene cuando $|J| = 2$ (para cada dos eventos)

  - Tres eventos A, B y C son independientes si, y solo si, se cumplen las siguientes igualdades:

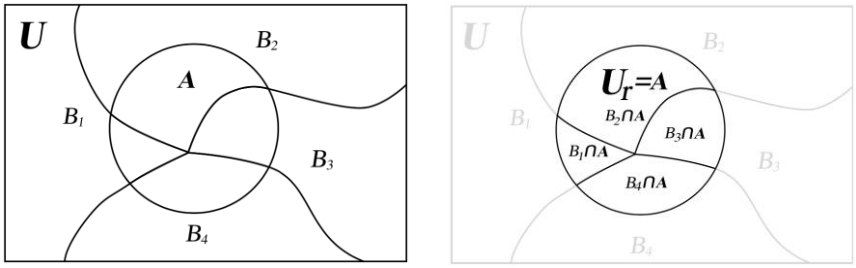
$$P(A \cap B \cap C) = P(A)P(B)P(C) \qquad P(A \cap B) = P(A)P(B)$$

$$P(B \cap C) = P(B)P(C) \qquad P(A \cap C) = P(A)P(C)$$

  - En consecuencia, hay familias de eventos que son independientes a pares pero que no son independientes (porque

hay alguna de las intersecciones que no es el producto de las probabilidades de cada evento)

- Siendo $(\Omega, \mathcal{F}, Q)$ un espacio de probabilidad, una partición de Ω es una colección $\{B_i : i \in I\}$ de eventos disjuntos (de modo que $B_i \in \mathcal{F}$ para toda i y $B_i \cap B_j = \emptyset$ si $i \neq j$) con union $\bigcup_i B_i = \Omega$. A partir de este concepto, se puede derivar el teorema de la partición (o de probabilidad total)



  - Si $\{B_1, B_2, \dots\}$ es una partición de Ω (de modo que $\bigcup_i B_i = \Omega$) con $P(B_i) > 0$ para toda i, entonces se da la siguiente igualdad:

$$P(A) = \sum_i P(A|B_i)P(B_i) \quad for \ A \in \mathcal{F}$$

  - La demostración de este teorema se basa en las expresiones equivalentes de A y la probabilidad para la unión de eventos disjuntos. Existe una demostración análoga para el caso continuo

$$P(A) = P(A \cap \Omega) = P\left(A \cap \left(\bigcup_i B_i\right)\right) = P\left(\bigcup_i (A \cap B_i)\right)$$

$$= \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

- Hay muchas situaciones en las que se quiere deducir algo a partir de una pieza de evidencia, de modo que A sería la evidencia y $B_1, B_2, \dots$ los posibles estados de la naturaleza. Entonces, se quiere saber $P(B_j|A)$ a partir de $P(A|B_j)$ y, para ello, se puede utilizar el teorema de Bayes

  - Siendo $\{B_1, B_2, \dots\}$ una partición de Ω (de modo que $\bigcup_i B_i = \Omega$) y $P(A), P(B_i) > 0$ para toda i, entonces se da la siguiente igualdad para cualquier evento A:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

- La demostración de este teorema se basa en el teorema de la partición anteriormente visto

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{P(A \cap \Omega)} = \frac{P(A|B_j)P(B_j)}{P\left(A \cap (\cup_i B_i)\right)} =$$

$$= \frac{P(A|B_j)P(B_j)}{P\left(A \cap (\cup_i B_i)\right)} = \frac{P(A|B_j)P(B_j)}{P(\cup_i (A \cap B_i))} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A \cap B_i)} =$$

$$= \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

- Debido a que el numerador de $P(B_j|A)$ es la probabilidad *a priori* de B_j por la verosimilitud $P(A|B_j)$ y el denominador es la suma de productos de probabilidad *a priori* de B_i por la verosimilitud $P(A|B_i)$, si se multiplicara una constante *c* por cada una de las verosimilitudes, esta constante se cancelaría en la división

$$\frac{cP(A|B_j)P(B_j)}{\sum_i cP(A|B_i)P(B_i)} = \frac{cP(A|B_j)P(B_j)}{c\sum_i P(A|B_i)P(B_i)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

  - A partir de este hecho, solo es necesario saber la verosimilitud dentro de una constante de proporcionalidad: los pesos relativos para todas las posibilidades (la verosimilitud) es todo lo que se necesita

  - De manera similar, si se multiplica cada probabilidad *a priori* por una constante, de modo que ocurriría lo mismo que antes: lo único que sería necesario son los pesos relativos que se daría a cada probabilidad *a priori*

  - De este modo, se suele expresar el teorema de Bayes en su forma proporcional, dado que se multiplican todos los valores por $1/\sum_i P(A|B_i)P(B_i)$ (los pesos relativos)

$$P(B_j|A) \propto P(B_j)P(A|B_j)$$

- Se puede, por tanto, resumir el uso del teorema de Bayes para eventos con lo siguientes tres pasos:

  - El primer paso es la multiplicación de $P(B_i)P(A|B_i)$ para cada una de las B_i. Esto encuentra la probabilidad de $A \cap B_i$ por la regla de la multiplicación

- El segundo paso consiste en sumar para $i = 1,2, \dots, n$ todas las probabilidades $P(B_i)P(A|B_i)$. Esto encuentra la probabilidad de A a partir del teorema de probabilidad total

- El tercer paso es la división de cada $P(B_i)P(A|B_i)$ entre la suma anterior. Esto encuentra la probabilidad condicional $P(B_i|A)$ para cada B_i posible

- Otra manera de poder lidiar con los eventos inciertos que se modelan como aleatorios es formar las *odds* de lo eventos. Las *odds* para un evento C es igual a la probabilidad del evento que ocurre dividida por la probabilidad de que el evento no ocurra

$$odds(C) = \frac{P(C)}{P(C^c)} = \frac{P(C)}{1 - P(C)}$$

  - Como la probabilidad de que un evento no ocurra es $P(C^c) = 1 - P(C)$, se puede ver como hay una correspondencia uno a uno entre la probabilidad $P(C)$ y las $odds(C)$. Aislando para $P(C)$, se puede obtener la siguiente relación:

$$P(C) = \frac{odds(C)}{1 - odds(C)}$$

  - El factor de Bayes, denotado por B, contiene la evidencia en los datos relevantes para el evento C, denotados por D, que han ocurrido. Este es el factor por el cual las *odds a priori* se han cambiado a las *odds* posteriores

$$prior\ odds(C) \times B = posterior\ odds(C)$$

$$\Rightarrow\ B = \frac{posterior\ odds(C)}{prior\ odds(C)}$$

  - Se puede sustituir en la *ratio* de probabilidades para las *odds* posteriores y *a priori* para poder encontrar la siguiente fórmula:

$$B = \frac{P(D|C)}{P(D|C^c)}$$

  - Por lo tanto, el factor de Bayes se puede interpretar como la *ratio* entre las probabilidades de obtener los datos cuando ha ocurrido un evento dado entre las probabilidades de obtener los datos cuando no ha ocurrido tal evento. Cuando este factor es mayor a uno, entonces es más probable que el evento haya ocurrido a partir de los datos, y si es menor, lo más probable es que no haya pasado

- Siempre hay que tener cuidado con hacer una indicación apropiada del condicionamiento de la probabilidad. Con tal de ser conciso, se suele

omitir el condicionamiento a las hipótesis que se mantienen a lo largo del desarrollo

- No obstante, estas se podrían representar a través de la siguiente notación, en donde H representa el conjunto de hipótesis o suposiciones usadas para definir el modelo

$$P(A, B|H) = P(A|B, H)p(B|H)$$

- A veces también se omite el condicionamiento explícito de las variables explicativas conocidas

Los modelos de un solo parámetro

- En el modelo binomial simple, el objetivo es estimar la proporción poblacional desconocida θ de una secuencia de experimentos de Bernoulli (datos $y_1, y_2, \dots, y_n$ que pueden tomar valores cero o uno). Este problema representa un buen comienzo para entender la inferencia bayesiana
    - La distribución binomial proporciona un modelo natural para los datos que nacen de una secuencia de n experimentos intercambiables de una gran población en donde cada experimento puede resultar en dos únicos resultados (éxitos y fallos)
        - Debido a la intercambiabilidad de los datos, estos se pueden resumir con el número total de éxitos en los n experimentos, denotado por y
        - Convertir una formulación en términos de experimentos intercambiables a una usando variables aleatorias independientes e idénticamente distribuidas se consigue a través de dejar que el parámetro θ represente la proporción de éxitos en la población o la probabilidad de éxito en cada experimento
        - El modelo muestral binomial, representado por $Bin(y|\theta, n)$, es el siguiente:

$$p(y|\theta) = Bin(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

        - En $p(y|\theta)$ se elimina la dependencia de n porque se asume que esta depende del diseño experimental que se considera fijo, por lo que todas las probabilidades utilizadas para discutir este problema se asumen como condicionadas a n

- Para poder realizar inferencia bayesiana en el modelo de binomial, es necesario especificar una distribución *a priori* para θ, de modo que se puede asumir cualquier distribución. En este caso, se asume una uniforme estándar, en donde la probabilidad de que θ tome un valor concreto en $[0,1]$ es igual para todos los valores

    - Una aplicación elemental del teorema de Bayes da una función posterior para θ como la siguiente, en donde como n e y son fijas (se saben una vez se han realizado los experimentos), el factor $\binom{n}{y}$ se puede omitir al ser una constante y 1, que sería la probabilidad de la distribución uniforme para cada valor de θ, también se omite

$$p(\theta|y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \propto \theta^y(1-\theta)^{n-y}$$

    - En muchos problemas se podrá escribir la distribución posterior con una forma cerrada (multiplicada por una constante de proporcionalidad)

    - En este caso, se puede reconocer que la función posterior $p(\theta|y)$ tiene la forma de una distribución beta no normalizada $Beta(y+1, n-y+1)$

$$\theta|y \sim Beta(y+1, n-y+1)$$

- En el caso de la binomial con una distribución *a priori* uniforme, la distribución *a priori* predictiva puede evaluarse de manera explícita

    - Bajo el modelo bayesiano obtenido, todos los valores de y son igualmente probables *a priori*

$$p(y) = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\, d\theta = \frac{1}{n+1}$$

    - Para la predicción posterior de este modelo, no obstante, se está más interesado en el resultado de un nuevo experimento más que en otro conjunto de n experimentos nuevos, así que denotando $\tilde{y}$ como el resultado de un nuevo experimento de Bernoulli (intercambiable con los primeros n experimentos), se obtiene la siguiente identidad a través de las propiedades de la distribución beta:

$$P(\tilde{y} = 1|y) = \int_0^1 P(\tilde{y} = 1|y, \theta)p(\theta|y)\, d\theta = \int_0^1 \theta p(\theta|y)\, d\theta =$$

$$= E(\theta|y) = \frac{y+1}{n+2}$$

- Este resultado obtenido, basado en la distribución uniforme estándar como distribución *a priori*, se denomina ley de sucesión de Laplace. En las observaciones extremas $y = 0$ e $y = n$, la ley de sucesión predice que las probabilidades serán $1/(n+2)$ y $(n+1)/(n+2)$

- El proceso de inferencia bayesiana involucra pasar de una distribución *a priori* $p(\theta)$ a una distribución posterior $p(\theta|\boldsymbol{y})$, y es natural esperar que algunas relaciones generales se mantengan entre estas dos distribuciones

  - Algo razonable es esperar que la distribución posterior tenga menos variabilidad o varianza que la distribución *a priori* $p(\theta)$ debido a que la distribución posterior $p(\theta|\boldsymbol{y})$ incorpora la información de los datos (la varianza de la verosimilitud $p(\boldsymbol{y}|\theta)$ suele ser más pequeña o igual)

    - Esta notación se formaliza a través de las siguientes expresiones:

$$E(\theta) = E_{\boldsymbol{y}}[E(\theta|\boldsymbol{y})]$$

$$Var(\theta) = E_{\boldsymbol{y}}[Var(\theta|\boldsymbol{y})] + Var_{\boldsymbol{y}}[E(\theta|\boldsymbol{y})]$$

    - La segunda identidad muestra como la varianza posterior $E_{\boldsymbol{y}}[Var(\theta|\boldsymbol{y})]$ es, de media, más pequeña que la varianza *a priori* $Var(\theta)$ por una cantidad que depende de la variación de las medias posteriores sobre la distribución de los datos posibles $Var_{\boldsymbol{y}}[E(\theta|\boldsymbol{y})]$

  - En consecuencia, cuanto mayor es la variación $Var_{\boldsymbol{y}}[E(\theta|\boldsymbol{y})]$, mayor es el potencial de reducir la incertidumbre con respecto a θ, dado que, si la diferencia es mayor, quiere decir que hay menos certeza en la distribución *a priori* que en el posterior

    - Las relaciones de la media y la varianza solo describen las expectativas, y en situaciones particulares la varianza posterior puede ser similar o incluso mayor que la varianza *a priori* (aunque esto puede indicar conflicto o inconsistencia entre el modelo muestral o verosímil del modelo de probabilidad y las distribuciones *a priori* del modelo bayesiano)

  - En el ejemplo binomial, le media de la distribución *a priori* es $1/2$, mientras que la media de la distribución posterior es $(y+1)/(n+2)$, la cual es un compromiso entre la media *a priori* y la proporción muestral y/n, en donde la media *a priori* tiene un rol cuanto mayor es el tamaño muestral

        - Esto es una característica general de la inferencia bayesiana: la distribución posterior está centrada en un punto que representa un compromiso entre la información *a priori* y los datos, y el compromiso es controlado principalmente por los datos cuanto mayor es la muestra

        - De este modo, la distribución posterior viene determinada intrínsecamente por la información *a priori* y los datos

- La distribución de probabilidad posterior contiene toda la información actual sobre el parámetro θ. Idealmente, uno puede representar gráficamente la distribución posterior, pero una ventaja del enfoque bayesiano es que las inferencias posteriores se pueden resumir fácilmente y de manera flexible

    - Para muchos propósitos prácticos, los resúmenes numéricos de la distribución son deseables

        - Los resúmenes más comunes de localización son la media, la mediana y la moda de la distribución

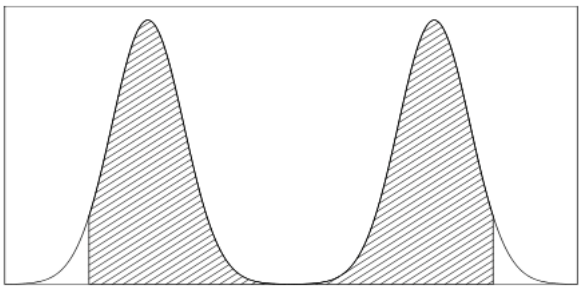
        - La variabilidad normalmente se resume a través de la desviación estándar, el rango intercuartílico y otros cuantiles

        - Mucha parte de la inferencia práctica se apoya en el uso de aproximaciones normales, normalmente mejorados aplicando transformaciones a θ

        - Cuando la distribución posterior tiene una forma cerrada, normalmente los resúmenes como la media, la mediana o la desviación estándar de la distribución posterior tienen también una forma cerrada

    - Además de los resúmenes puntuales, también es imprescindible reportar la incertidumbre. El enfoque más usual es presentar cuartiles de la distribución posterior de los estimandos de interés o un intervalo central de probabilidad posterior

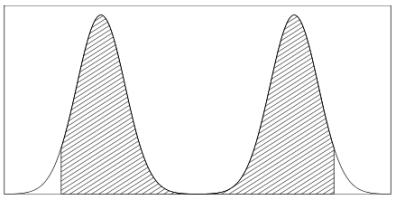
        - El intervalo central de probabilidad posterior corresponde al rango de valores por encima y por debajo de $100(\alpha/2)\%$ en el caso de un intervalo del $100(1-\alpha)\%$, y este tipo de intervalo se denomina intervalo posterior

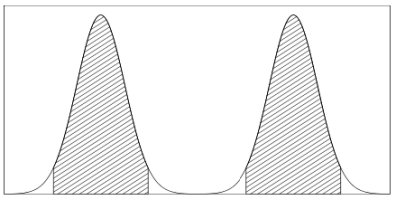


- Para modelos simples como el binomial o el normal, el intervalo se puede calcular directamente utilizando las funciones integradas en el *software*, pero se pueden calcular los intervalos a través de simulaciones de la distribución posterior

- Otro resumen de la incertidumbre posterior es la región de mayor densidad posterior, que es el conjunto de valores que contiene el $100(1-\alpha)\%$ de la probabilidad posterior y que tiene la característica que la densidad en la región nunca es menor que la de fuera de esta (da los intervalos más estrechos posibles)

- Esta región es idéntica al intervalo posterior central si la distribución posterior es unimodal y simétrica. En el caso en que no sea unimodal, la región se compondrá de intervalos disjuntos, y en el caso de asimetría, es muy probable que el intervalo posterior central sea diferente a la región





- Se consideran dos interpretaciones básicas que pueden darse de las distribuciones *a priori*: la poblacional y la de estado del conocimiento. Ambas interpretaciones tienen implicaciones a la hora de asignar una distribución *a priori*, pero ambas son distribuciones *a priori* informativas

    - En la interpretación poblacional, la distribución *a priori* representa una población de posibles valores paramétricos del cual θ se ha obtenido, mientras que en la interpretación de estado del conocimiento el principio es que se debe expresar el conocimiento o incertidumbre sobre θ como si el valor fuera una realización de la distribución *a priori*

        - Para muchos problemas, como el de estimar probabilidades de eventos, no hay una población perfectamente relevante de $\boldsymbol{\theta}$ de la cual se pueda obtener un valor (excepto en contemplaciones hipotéticas)

- Normalmente la distribución *a priori* debería incluir todos los valores plausibles de θ, pero la distribución no necesita estar realísticamente concentrada alrededor del valor verdadero, ya que la mayoría de veces la información sobre θ contenida en los datos pesará mucho más que cualquier especificación de probabilidad *a priori* razonable

- Para poder escoger una distribución *a priori*, normalmente se dibuja la distribución *a priori* o se resuelve un sistema de ecuaciones basado en momentos o cuartiles de una distribución *a priori* predeterminada. El segundo método es útil cuando se tiene información sobre la distribución de los parámetros, ya que se podrán fijar unos valores para los momentos o cuartiles

$$\begin{cases} E(\theta) = \int_{-\infty}^{\infty} \theta \; p(\theta) \; d\theta = k_1 \\ \dots \\ E(\theta^n) = \int_{-\infty}^{\infty} \theta^n \; p(\theta) \; d\theta = k_n \end{cases}$$

- Debido a que la varianza de la distribución *a priori* $p(\theta)$ se entiende como la incertidumbre presente sobre θ, aquellas distribuciones *a priori* que tengan menor varianza (que sean menos planas y más concentradas) serán más informativas respecto a los valores que puede tomar θ

  - Por lo tanto, aquellas con más varianza lo serán menos (más planas y con menos concentración), y eso dará lugar al concepto de distribuciones *a priori* no informativas o débilmente informativas

- Considerando el ejemplo binomial anterior, se puede ver como, si se escoge una distribución *a priori* diferente para poder estudiar algunas propiedades útiles

  - Considerada como una función de θ, la función de verosimilitud tendrá una verosmilitud de forma $p(\boldsymbol{y}|\theta) \propto \theta^a(1-\theta)^b$, por lo que, si se escoge una distribución con la misma forma, con sus propios valores a y b, entonces la distribución posterior de probabilidad también tendrá esta forma

  - Si se escoge una forma de la distribución *a priori* como $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, que es una distribución beta con parámetros α y β. Comparando $p(\boldsymbol{y}|\theta)$ con $p(\theta)$, se puede ver como la *a priori* mostraría que hay $\alpha - 1$ éxitos y $\beta - 1$ fallos si se interpreta como la binomial

- Los parámetros de $p(\theta)$ se suelen llamar hiperparámetros, y estos se pueden modificar para poder especificar una distribución *a priori* concreta fijando parámetros como la media y la varianz. La densidad posterior en este caso sería la siguiente:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta) = \theta^{a}(1-\theta)^{b}\theta^{\alpha-1}(1-\theta)^{\beta-1} =$$

$$= \theta^{a+\alpha-1}(1-\theta)^{b+\beta-1} \sim Beta(\theta|a+\alpha, b+\beta)$$

- A partir de la distribución posterior, es posible obtener la media y la varianza posterior con los nuevos parámetros:

$$E(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

$$Var(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} =$$

$$= \frac{E(\theta|y)[1 - E(\theta|y)]}{(\alpha + \beta + n + 1)}$$

- Cuanto más grande es y y $n - y$ con unas α y β fijas, se puede demostrar que $E(\theta|y) \approx y/n$ y $Var(\theta|y) \approx (y/n^2)(1 - y/n)$ que se aproxima a cero a una tasa de $1/n$. En el límite, los parámetros de la distribución *a priori* no tienen influencia en la distribución posterior

- La conjugación se define formalmente de la siguiente manera: si $\mathcal{F}$ es una clase de distribuciones de muestreo o de verosimilitud $p(\boldsymbol{y}|\theta)$ y $\mathcal{P}$ es una clase de distribuciones *a priori* para θ, entonces la clase $\mathcal{P}$ es conjugada para $\mathcal{F}$ si se cumple la siguiente condición:

$$p(\theta|\boldsymbol{y}) \in \mathcal{P} \quad for\ all \quad p(\cdot\,|\theta) \in \mathcal{F} \ \ \& \ \ p(\theta) \in \mathcal{P}$$

- Uno está interesado en las familias conjugadas *a priori* naturales, las cuales nacen tomando $\mathcal{P}$ como el conjunto de todas las densidades que tienen la misma forma funcional como la función de verosimilitud

- La justificación básica para el uso de distribuciones *a priori* conjugadas es similar a la de usar modelos estándar para la verosimilitud: es fácil entender los resultados, se pueden poner de forma analítica, y suelen ser una buena aproximación y simplifican cálculos

- Estos modelos son útiles para modelos más complicados, en las que se incluyen nuevas dimensiones, en donde normalmente la conjugación es imposible. Por estas razones, las mezclas de

familias conjugadas pueden ser útiles a veces cuando las distribuciones conjugadas simples no son razonables

- Aunque las distribuciones *a priori* no conjugadas puedan hacer las interpretaciones de las inferencias posteriores menos transparentes y hagan que los cálculos sean más difíciles, estas no crean problemas conceptuales

- En la práctica, distribuciones *a priori* conjugadas pueden no ser ni siquiera posibles

- Es posible relacionar las familias de distribuciones conjugadas con los conceptos clásicos de familias exponenciales y los estadísticos suficientes

  - Las distribuciones de probabilidad que pertenecen a la familia exponencial tienen distribuciones *a priori* conjugadas naturales. La verosimilitud que corresponde al vector $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ de observaciones independientes e idénticamente distribuidas sería la siguiente:

$$p(\boldsymbol{y}|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) g(\theta)^n \exp\left(\phi(\theta)' \sum_{i=1}^{n} u(y_i)\right)$$

  - Para toda n e y, esta tiene una forma fija (como una función de θ) como la siguiente, en donde $t(\boldsymbol{y})$ es el estadístico suficiente para θ (la verosimilitud depende de los datos solo a través de $t(\boldsymbol{y})$)

$$p(\boldsymbol{y}|\theta) \propto g(\theta)^n e^{\phi(\theta)' t(\boldsymbol{y})} \quad where \quad t(\boldsymbol{y}) = \sum_{i=1}^{n} u(y_i)$$

  - Los estadísticos suficientes son útiles para manipulaciones algebraicas de verosimilitudes y distribuciones posteriores, y si la densidad *a priori* se especifica como $p(\theta) \propto g(\theta)^\eta e^{\phi(\theta)'\nu}$, la densidad posterior será la siguiente:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)\, p(\theta) = g(\theta)^n e^{\phi(\theta)' t(\boldsymbol{y})} g(\theta)^\eta e^{\phi(\theta)'\nu} =$$

$$= g(\theta)^{n+\eta} e^{\phi(\theta)'[t(\boldsymbol{y})+\nu]}$$

  - Se ha demostrado que, en general, la familia exponencial es la única familia de distribuciones que tienen distribuciones *a priori* conjugadas dado que, aparte de casos irregulares, las únicas distribuciones teniendo un número fijo de estadísticos suficientes para toda n son del tipo exponencial

- La distribución normal es fundamental para la mayoría de modelos estadísticos, gracias al uso del teorema del límite central y las aproximaciones que permite hacer. Por lo tanto, se puede estudiar el caso de estimar la media cuando la varianza es conocida

    - Como un primer caso simple, se considera una la función de verosimilitud $p(y|\theta)$ de una observación escalar singular y de una distribución normal parametrizada por una media θ, en donde se que σ^2 es conocida

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$$

        - Considerando la función de verosimilitud como una función μ, la función de verosimilitud es una función exponencial de una forma cuadrática en θ, de modo que la familia de densidades *a priori* conjugadas se parecerá a la siguiente:

$$p(\theta) = e^{A\theta^2 + B\theta + C}$$

        - Se puede parametrizar esta familia de densidades de la siguiente manera, en donde $\theta \sim N(\mu_0, \tau_0^2)$ con hiperparámetros μ_0 y τ_0^2 (los cuales se asumen que son conocidos):

$$p(\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

    - La densidad conjugada *a priori* implica que la distribución posterior para θ es la exponencial de una forma cuadrática, y por tanto normal (aunque se necesita hacer transformaciones algebraicas para poder revelar la forma exacta)

        - En la densidad posterior, todas las variables excepto θ se asumen como constantes, la cual da la siguiente densidad condicional:

$$p(\theta|y) \propto e^{-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]}$$

        - Expandiendo los exponentes, juntando términos y completando el cuadrado en θ da la siguiente función de densidad posterior:

$$p(\theta|y) \propto e^{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2}$$

$$where \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma_0^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma_0^2}} \quad \& \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- Al manipular las distribuciones normales, la inversa de la varianza juega un rol prominente y se denomina la precisión

  - El álgebra anterior demuestra que para los datos normales y para una distribución *a priori* (cada una con su precisión conocida), la precisión posterior iguala la precisión *a priori* sumada a la precisión de los datos

  - Hay varias maneras diferentes de interpretar el la forma de la media posterior μ_1. En la expresión anterior, la media posterior se expresa como una media ponderada de la media *a priori* μ_0 y el valor observado y, en donde las ponderaciones son proporcionales a las precisiones

  - Alternativamente, se puede expresar μ_1 como la media *a priori* μ_0 ajustada hacia la observación y o como los datos contraídos hacia la media *a priori* μ_0. Cada una de las formulaciones representa la media posterior como un *trade-off* entre la media *a priori* y el valor observado

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

$$\mu_1 = y + (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

  - Si $\tau_0^2 = 0$, entonces la distribución *a priori* es infinitamente más precisa que los datos, de modo que la distribución posterior y la *a priori* son idénticas y se concentran en el valor μ_0. Además, si $\mu_0 = y$, lógicamente se obtiene que $\mu_1 = \mu_0 = y$

$$\mu_1 = \mu_0 \quad if \quad y = \mu_0 \quad or \quad \tau_0^2 = 0$$

  - Si $\sigma^2 = 0$, entonces los datos son perfectamente precisos y la distribución posterior está concentrada en el valor observado y. Además, si $\mu_0 = y$, lógicamente se obtiene que $\mu_1 = \mu_0 = y$

$$\mu_1 = y \quad if \quad y = \mu_0 \quad or \quad \sigma^2 = 0$$

- La distribución posterior predictiva $p(\tilde{y}|y)$ de una observación futura $\tilde{y}$ se puede calcular directamente integrando, obteniendo el siguiente resultado:

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)\, d\theta \;\propto \int_{\Theta} e^{-\frac{(\tilde{y}-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu_1)^2}{2\tau_1^2}}\, d\theta$$

$$\propto \int_{\Theta} e^{-\frac{\tau_1^2(\tilde{y}-\theta)^2+\sigma^2(\theta-\mu_1)^2}{2\sigma^2\tau_1^2}}\, d\theta$$

- A partir de las propiedades de la distribución normal multivariante, se puede determinar más fácilmente la distribución de $\tilde{y}$. El producto en el integrando es la exponencial de una función cuadrática de $(\tilde{y}, \theta)$, por lo que $\tilde{y}$ y θ siguen una distribución normal conjunta posterior, por lo que la distribución marginal de $\tilde{y}$ es normal

- Se puede determinar la media y la varianza de la distribución posterior predictiva usando el conocimiento de la distribución posterior sobre que $E(\tilde{y}|\theta) = \theta$ y $Var(\tilde{y}|\theta) = \sigma^2$

$$E(\tilde{y}) = E[E(\tilde{y}|\theta, y)|y] = E(\theta|\tilde{y}) = \mu_1$$

$$Var(\tilde{y}) = E[Var(\tilde{y}|\theta, y)|y] + Var[E(\tilde{y}|\theta, y)|y] =$$

$$= E(\sigma^2|y) + \; Var(\theta|y) = \sigma^2 + \tau_1^2$$

- Por lo tanto, la distribución posterior predictiva de $\tilde{y}$ tiene una media igual a la media posterior de θ y dos componentes de la varianza: la varianza predictiva σ^2 del modelo y la varianza τ_1^2 debido a la incertidumbre posterior en θ

- Este desarrollo del modelo normal con una sola observación puede extenderse de manera simple a una situación más realista en la que se tiene una muestra iid de observaciones $\boldsymbol{y} = (y_1, \ldots, y_n)$

  - Procediendo formalmente, la densidad posterior es la siguiente:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta) = p(\theta)\prod_{i=1}^{n} p(y_i|\theta)$$

$$\propto \; e^{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}} \prod_{i=1}^{n} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} \;\propto\; e^{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{\sigma^2}\right]}$$

  - La simplificación algebraica de esta expresión muestra que la distribución posterior depende de y solo a través de $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ ($\bar{y}$ es un estadístico suficiente). Debido a que $\bar{y}|\theta, \sigma^2 \sim N(\theta, \sigma^2/n)$, los resultados derivados de la

observación normal única (considerando $\bar{y}$ como observación única) aplican inmediatamente y permite obtener los siguientes resultados:

$$p(\theta|y_1, y_2, \dots, y_n) = p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2)$$

$$where\ \ \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \& \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- Este resultado se puede obtener de manera equivalente a partir de añadir una observación cada vez, usando la distribución posterior en cada paso como la distribución *a priori* para la siguiente

- En las expresiones para la media y la varianza posterior, la precisión *a priori* $1/\tau_0^2$ y la precisión de los datos n/σ^2 tienen papeles equivalentes, de modo que si n es grande, la distribución posterior se determina en gran parte por σ^2 y por el valor muestral $\bar{y}$. Si $\tau_0^2 = \sigma^2$, entonces la distribución *a priori* tiene el mismo peso que una observación extra con valor μ_0.

- Más específicamente, cuando $\tau_0 \to \infty$ con n fija o cuando $n \to \infty$ con τ_0^2 fija, se ve que la normal es una buena aproximación cuando las creencias *a priori* son relativamente difusas sobre el rango de θ en donde la verosimilitud es sustancial

$$p(\theta|y) \overset{D}{\to} N\left(\theta \middle| \bar{y}, \frac{\sigma^2}{n}\right)$$

- En general, la densidad posterior $p(\theta|y)$ no tiene una forma cerrada y la constante normalizadora $p(y)$ es especialmente difícil de calcular debido al integrando $p(y|\theta)p(\theta)$
    - Mucha de la estadística bayesiana formal se concentra en situaciones en donde las formas cerradas están disponibles
        - Esos modelos a veces no son realistas, pero su análisis normalmente proporciona un buen punto inicial para construir modelos más realistas
    - Las distribuciones estándar (la binomial, la normal, la Poisson y la exponencial) tienen derivaciones naturales de modelos de probabilidad simples

- La distribución binomial se motiva de el conteo de resultados intercambiables y la distribución normal aplica a una variable aleatoria que es la suma de muchos términos intercambiables o independientes

- También se podrá aplicar la distribución normal para el logaritmo de datos positivos, que aplicarían naturalmente a las observaciones que se modelan como el producto de muchos factores multiplicativos independientes

- La distribución de Poisson y la exponencial surgen como el número de veces contadas y los tiempos de espera, respectivamente, con una tasa de ocurrencia constante

- Normalmente se construirán modelos de probabilidad realísticos para resultados más complicados al combinar estas distribuciones básicas

- Todos estos modelos tienen asociados una familia de distribuciones *a priori* conjugadas, las cuales se discuten a continuación

- El modelo normal con media conocida θ y varianza desconocida es un ejemplo importante debido a que funciona como una introducción para modelos mucho más complicados y útiles como el modelo normal con media y varianza desconocida

    - Además, este ejemplo proporciona un ejemplo de como es la estimación de un parámetro de escala

    - Para $p(\boldsymbol{y}|\theta,\sigma^2) = N(\boldsymbol{y}|\theta,\sigma^2)$ con θ conocida y σ^2 desconocida, la verosimilitud para un vector $\boldsymbol{y}$ de n observaciones independientes e idénticamente distribuidas sería la siguiente:

$$p(\boldsymbol{y}|\theta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2}}$$

$$\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{nv}{2\sigma^2}} \quad where \quad v = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$$

    - En este caso, v es el estadístico suficiente y la densidad *a priori* conjugada correspondiente es la densidad de la distribución gamma inversa, la cual tiene hiperparámetros (α,β)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

- Una parametrización conveniente es una distribución chi cuadrada inversa escalada, con escala σ_0^2 y v_0 grados de libertad, de modo que la distribución de σ^2 ahora es la distribución de $\sigma_0^2 v_0/X$, donde $X \sim \chi_{v_0}^2$

$$\sigma^2 \sim Inv.\chi^2(v_0, \sigma_0^2)$$

- La densidad posterior resultante para σ^2 es la siguiente:

$$p(\sigma^2|\boldsymbol{y}) \propto p(\sigma^2)p(\boldsymbol{y}|\sigma^2) \propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\frac{v_0}{2}+1} e^{-\frac{v_0\sigma_0^2}{2\sigma^2}}(\sigma^2)^{-\frac{n}{2}}e^{-\frac{nv}{2\sigma^2}}$$

$$\propto (\sigma^2)^{-\left[\frac{(n+v_0)}{2}+1\right]}e^{-\frac{v_0\sigma_0^2+nv}{2\sigma^2}}$$

- Por lo tanto, $\sigma^2|\boldsymbol{y}$ se distribuye como una distribución chi cuadrada inversa con escala igual a la media ponderada de los grados de libertad de la *a priori* y de los datos. La distribución *a priori* puede interpretarse como una que proporciona información equivalente a v_0 observaciones con una desviación cuadrada media σ_0^2

$$\sigma^2|\boldsymbol{y} \sim Inv.\chi^2\left(n + v_0, \frac{v_0\sigma_0^2 + nv}{v_0 + n}\right)$$

- La distribución de Poisson surge naturalmente en el estudio de los datos que toman la forma de conteo. Si un dato y sigue una distribución de Poisson con tasa θ, entonces la distribución de probabilidad de una sola observación y es la siguiente:

$$p(y|\theta) = \frac{e^{-\theta}\theta^y}{y!} \quad for\ \ y = 0,1,2,\ldots$$

- Para un vector $\boldsymbol{y}$ de observaciones independientes e idénticamente distribuidas, la verosimilitud es la siguiente:

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n}\frac{1}{y_i!}\theta^{y_i}e^{-\theta} = \frac{1}{(y_i!)^n}\theta^{\Sigma_{i=1}^n y_i}e^{-n\theta} \propto \theta^{t(y)}e^{-n\theta}$$

$$where\ \ t(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$$

- La función de verosimilitud se puede reescribir en una forma perteneciente a la familia exponencial, lo cual permite ver que el parámetro natural $\phi(\theta) = \log(\theta)$ y la distribución conjugada *a*

priori natural sería la siguiente, indexada por los hiperparámetros (η, v):

$$p(\boldsymbol{y}|\theta) \propto e^{t(y)\theta \log(\theta)} e^{-n\theta} \;\Rightarrow\; p(\theta) \propto e^{v \log(\theta)} \left(e^{-\theta}\right)^{\eta}$$

- En otras palabras, la verosimilitud es de la forma $\theta^{a} e^{-b\theta}$, por lo que la densidad conjugada *a priori* debe ser $p(\theta) \propto \theta^{A} e^{-B\theta}$. Si se considera la *a priori* como una densidad de una distribución gamma con parámetros (α, β), entonces se puede interpretar que $p(\theta)$ es equivalente a un conteo total de $\alpha - 1$ con β observaciones

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \quad for \quad Gamma(\alpha, \beta)$$

- Con esta distribución conjugada *a priori*, la distribución posterior será una gamma de la siguiente forma:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)\, p(\theta) \propto \theta^{t(y)} e^{-n\theta} \theta^{\alpha-1} e^{-\beta\theta}$$

$$\propto \theta^{\alpha+t(\boldsymbol{y})-1} e^{-(\beta+n)\theta} \propto \theta^{\alpha+n\bar{y}-1} e^{-(\beta+n)\theta}$$

$$where \;\; \bar{y} = t(\boldsymbol{y})/n$$

$$\Rightarrow \; \theta|\boldsymbol{y} \sim Gamma(\alpha + n\bar{y}, \beta + n)$$

- Con familias conjugadas, la forma conocida de las densidades *a priori* y posterior se pueden usar para encontrar la distribución marginal $p(y)$ (la distribución predictiva *a priori*) usando la siguiente fórmula:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

- La demostración de esta fórmula es intuitiva dadas las definiciones de cada término:

$$\frac{p(y|\theta)p(\theta)}{p(\theta|y)} = \frac{\frac{p(\theta|y)p(y)}{p(\theta)}p(\theta)}{p(\theta|y)} = \frac{p(\theta|y)p(y)}{p(\theta|y)} = p(y)$$

- Por ejemplo, el modelo de Poisson para una sola observación y tiene una distribución predictiva *a priori* que se reduce a la densidad de una distribución negativa binomial:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)} = \frac{Pois(y|\theta)Gamma(\theta|\alpha,\beta)}{Gamma(\alpha+y, \beta+1)} =$$

$$= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!\,(1+\beta)^{\alpha+y}}$$

$$\Rightarrow\ p(y) = \binom{\alpha + y - 1}{y}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{\alpha}$$

$$\Rightarrow\ \ y \sim Neg.bin(\alpha, \beta)$$

- La derivación muestra como la distribución negativa binomial es una mezcla de distribuciones de Poisson con tasas θ que siguen una distribución gamma

$$Neg.bin(\alpha, \beta) = \int_{\Theta} Pois(y|\theta)\ Gamma(\theta|\alpha, \beta)\ d\theta$$

- En muchas aplicaciones, es conveniente extender el modelo de Poisson para los datos $y_1, y_2, \ldots, y_n$ de la siguiente forma, en donde los valores de x_i son valores positivos conocidos de una variable explicativa x y θ es el parámetro desconocido de interés:

$$y_i \sim Pois(x_i\theta)$$

- En epidemiología, el parámetro θ normalmente se denomina tasa, mientras que x_i se denomina la exposición de la unidad i. Este modelo no es intercambiable en las y_i pero es intercambiable en los pares (x_i, y_i)

- La verosimilitud de este modelo sería la siguiente, la cual tendría como distribución *a priori* la distribución Gamma vista anteriormente:

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-x_i\theta} = \frac{1}{(y_i!)^n}\theta^{\sum_{i=1}^{n} y_i}e^{-\sum_{i=1}^{n} x_i\theta}$$

$$\propto \theta^{\sum_{i=1}^{n} y_i}e^{-\sum_{i=1}^{n} x_i\theta}$$

$$\Rightarrow\ Gamma(\alpha, \beta)\ \ as\ prior$$

- Esto hace que la distribución posterior resultante también sea una Gamma de la siguiente forma:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)\ p(\theta) \propto \theta^{\sum_{i=1}^{n} y_i}e^{-\sum_{i=1}^{n} x_i\theta}\theta^{\alpha-1}e^{-\beta\theta}$$

$$\propto \theta^{(\alpha+\sum_{i=1}^{n} y_i)-1}e^{-(\beta+\sum_{i=1}^{n} x_i)\theta}$$

$$\Rightarrow\ \theta|\boldsymbol{y} \sim Gamma\left(\alpha + \sum_{i=1}^{n} y_i\,, \beta + \sum_{i=1}^{n} x_i\right)$$

- La distribución exponencial se usa comúnmente para modelar los tiempos de espera y otras variables reales continuas con valores positivos, normalmente medidos en una escala de tiempo

    - La distribución muestral de un resultado y, dado el parámetro θ, es la siguiente, y $1/E(y|\theta)$ se denomina tasa

$$p(y|\theta) = \theta e^{-\theta y} \quad for \;\; y > 0$$

    - Matemáticamente, el caso exponencial es el caso especial en el que la distribución gamma tiene parámetros $(\alpha, \beta) = (1, \theta)$. En este caso, sin embargo, se usa como distribución muestral y no como una distribución *a priori* para θ

    - La distribución exponencial tiene la propiedad de pérdida de memoria, lo cual hace que sea un modelo natural para datos de supervivencia (la probabilidad de que un objeto sobreviva una longitud de tiempo adicional t es independiente del tiempo transcurrido hasta este punto

$$P(y > t + s|y > s, \theta) = P(y > t|\theta) \quad for \;\; \forall s, t$$

    - La distribución conjugada *a priori* para el parámetro exponencial θ será la gamma con parámetros (α, β), tal como en el caso de la Poisson, y su distribución posterior será una gamma de la siguiente forma:

$$p(y|\theta) \propto p(y|\theta)p(\theta) \propto \theta e^{-\theta y} \theta^{\alpha-1} e^{-\beta\theta} \propto \theta^{\alpha+1-1} e^{-(\beta+y)\theta}$$

$$\Rightarrow \; \theta|y \sim Gamma(\alpha + 1, \beta + y)$$

    - La distribución muestral para n observaciones exponenciales independientes con tasa constante θ será la siguiente, la cual es proporcional a la densidad de una $Gamma(n + 1, n\bar{y})$:

$$p(\mathbf{y}|\theta) = \theta^n e^{-n\bar{y}\theta} \quad for \;\; \bar{y} \geq 0$$

- Cuando las distribuciones *a priori* no tienen una base poblacional, estas pueden ser difíciles de construir, y siempre ha habido un deseo de distribuciones *a priori* que garanticen un papel mínimo en la distribución posterior, las cuales se denominan distribuciones *a priori* de referencia y su densidad se describe como no informativa

    - El racional para usar distribuciones *a priori* no informativas es hacer que los datos sean quienes dicten la distribución del parámetro, de modo que las inferencias no se vean afectadas por la información externa a los datos actuales

- Una idea relacionada son las distribuciones *a priori* débilmente informativas, que contiene alguna información (la suficiente para regularizar la distribución posterior o mantenerla bajo unos límites razonables) sin intentar capturar completamente el conocimiento científico sobre el parámetro subyacente

- Si uno vuelve al problema de estimar la media θ de un modelo normal con varianza conocida σ^2 y con una distribución *a priori* $N(\mu_0, \tau_0^2)$ sobre θ, si la precisión *a priori* $1/\tau_0^2$ es pequeña en comparación a la precisión de los datos n/σ^2 , entonces la distribución posterior es aproximadamente como si $\tau_0^2 = \infty$

$$p(\theta|y) \approx N(\theta|\bar{y}, \sigma^2/n)$$

  - La distribución posterior es aproximadamente la que resultaría de asumir que $p(\theta)$ es proporcional a una constante c para $\theta \in (-\infty, \infty)$. Esta distribución no es estrictamente posible, dado que la integral de la $p(\theta)$ es infinito, lo cual viola la suposición de que las probabilidades sumen 1

$$p(\theta) \propto c \;\Rightarrow \int_\Theta p(\theta)\, d\theta = \int_\Theta c\, d\theta$$

  - En general, se dice que una distribución *a priori* es propia o *proper* si esta no depende de los datos e integra a 1. Si $p(\theta)$ integra a cualquier valor positivo finito, entonces se denomina densidad no normalizada y se puede renormalizar multiplicando por esa constante para que integre 1

$$\frac{p(\theta)}{\int_\Theta p(\theta)\, d\theta} \;\Rightarrow \int_\Theta \frac{p(\theta)}{\int_\Theta p(\theta)\, d\theta}\, d\theta = 1$$

- Como un segundo ejemplo de distribución *a priori* no informativa, se considera el modelo normal de media conocida, pero varianza desconocida, con la distribución *a priori* conjugada escalada chi cuadrada inversa

  - Si los grados de libertad v_0 son pequeños relativamente a los grados de libertad de los datos n, entonces la distribución posterior es aproximadamente como si $v_0 = 0$

$$p(\sigma^2|y) \approx Inv.\chi^2(\sigma^2|n, v)$$

  - Esta forma límite de la distribución posterior también se puede derivar al definir la densidad posterior para σ^2 como $p(\sigma^2) \propto$

$1/\sigma^2$, que es impropia, teniendo una integral que no converge sobre el rango de $(0, \infty)$

- En ninguno de los dos ejemplos anteriores la distribución *a priori* se combina con la función de verosimilitud para definir un modelo de probabilidad conjunta $p(y, \theta)$ propia

  - No obstante, se puede proceder como se ha hecho anteriormente para definir una función de densidad posterior no normalizada a través de $p(\theta|y) \propto p(y|\theta)p(\theta)$

  - En los ejemplos anteriores, la densidad posterior es propia, de modo que se pueden obtener distribuciones posteriores propias a partir de una distribución *a priori* impropia. No obstante, este no siempre es el caso

  - Las distribuciones posteriores de distribuciones *a priori* impropias se tienen que interpretar con mucho cuidado, de modo que uno siempre debe comprobar que la distribución posterior tiene una integral finita y una forma sensible. Su interpretación más razonable es como unas aproximaciones en situaciones en donde la verosimilitud domina la densidad *a priori*

- Una manera común de definir distribuciones *a priori* no informativas fue introducida por Jeffreys, basada en considerar una transformación uno a uno de los parámetros (por lo que $\phi = h(\theta)$)

  - Por la transformación de variables, la densidad *a priori* $p(\theta)$ es equivalente, en términos de expresar las mismas creencias, a la siguiente densidad *a priori* sobre ϕ:

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

  - El principio general de Jeffreys es que cualquier regla para determinar la densidad *a priori* $p(\theta)$ debería dar un resultado equivalente si se aplica al parámetro transformado

  - La densidad $p(\phi)$ calculada determinando $p(\theta)$ y aplicando el principio anterior debería concordar con la distribución que se obtiene al determinar $p(\phi)$ directamente utilizando el modelo transformado $p(y, \phi) = p(y|\phi)p(\phi)$

  - El principio de Jeffrey comporta definir la densidad *a priori* no informativo como $p(\theta) \propto [I(\theta)]^{1/2}$, donde $I(\theta)$ es la información de Fisher para θ:

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

- Para ver que el modelo invariante a la parametrización, se evalúa $I(\theta)$ en $\theta = h^{-1}(\phi)$

$$I(\phi) = -E\left(\frac{d^2\log[p(y|\theta)]}{d^2\theta}\right) =$$

$$= -E\left(\frac{d^2\log\left[p\left(y\middle|\theta = h^{-1}(\phi)\right)\right]}{d^2\theta}\left|\frac{d\theta}{d\phi}\right|^2\right) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

$$\Rightarrow\ [I(\phi)]^{1/2} = [I(\theta)]^{1/2}\left|\frac{d\theta}{d\phi}\right|$$

- El principio de Jeffreys se puede extender a modelos de múltiples parámetros, pero los resultados son más controversiales

  - Enfoques más simples basados en asumir distribuciones *a priori* no informativas independientes para los componentes de un vector de parámetros $\boldsymbol{\theta}$ pueden dar diferentes resultados de los que se obtienen con el principio de Jeffreys

  - Cuando el número de parámetros en el problema es grande, se abandonan las distribuciones no informativas puras para utilizar modelos jerárquicos

- La búsqueda de distribuciones *a priori* no informativas tiene varios problemas

  - Buscar una distribución *a priori* que siempre sea no informativa no parece ser una buena guía: si la verosimilitud del problema es muy dominante en un problema, entonces la selección de las posibles densidades *a priori* planas no debería importar. Por lo tanto, seleccionar una como la referencia parece indicar que no se está usando apropiadamente

  - Para muchos problemas no hay una clara elección de distribución *a priori* no informativa, dado que la densidad es plana o uniforme en una parametrización no estará en otra parametrización

  - También se dan otras dificultades cuando se hace la media sobre un conjunto de posibles modelos que tienen distribuciones *a priori* propias

        - No obstante, las densidades *a priori* de referencia o no informativas normalmente son útiles cuando no vale la pena cuantificar el conocimiento real propio como una distribución de probabilidad, siempre que uno compruebe que la densidad posterior es propia y se determine la sensibilidad de las inferencias posteriores para modelar las suposiciones de conveniencia

- Una distribución *a priori* se caracteriza como débilmente informativa si es propia o *proper*, pero está configurada de manera que la información que proporcione sea intencionalmente menor que cualquier conocimiento (distribución concreta) *a priori* disponible

    - Más que modelar una completa ignorancia del problema, en la mayoría de problemas se prefiere utilizar distribuciones *a priori* débilmente informativas que incluyan una pequeña cantidad de información del mundo real

        - En este caso, la cantidad es tal que asegura que la distribución posterior tiene sentido (no sale una forma rara)

    - En la mayoría de problemas reales, el analista tendrá más información de la que se querría incluir convenientemente en el modelo estadístico

        - Esto es un problema con la verosimilitud y con la distribución *a priori*

        - En la práctica hay varios problemas: describir el modelo convenientemente, es difícil expresar el conocimiento precisamente en forma probabilística, hay que simplificar los cálculos o se quiere evitar utilizar información no fiable

    - Excepto por la última razón, estos son argumentos para la conveniencia y se justifican diciendo que la respuesta no habría variado mucho si se hubiera sido más preciso

        - De ese modo, si tan pocos datos disponibles hay que la elección de la distribución *a priori* no informativa marca la diferencia, entonces se debe poner más énfasis en poner información relevante en la *a priori*, usando igual un modelo jerárquico

    - Uno podría decir que todos los modelos estadísticos son débilmente informativos: un modelo siempre da algo de información (ya sea por los parámetros y la forma funcional) pero no es posible codificar todas las creencias *a priori* en un conjunto de distribuciones de probabilidad

        - Con eso en mente, se pueden seguir dos principios para crear distribuciones *a priori* débilmente informativas, yendo al problema en dos direcciones diferentes

        - Se puede comenzar por una versión de una distribución *a priori* débilmente informativa y añadir información de modo que las inferencias estén restringidas para ser razonables

        - Se puede comenzar por una distribución muy informativa y ampliarla (aumentar la variabilidad sobre los valores posibles) para tener en cuenta la incertidumbre sobre las creencias *a priori* y en la aplicabidad a nuevos datos

    - Ninguno de los dos enfoques anteriores es totalmente puro o no tiene problemas

        - En verdad puede pasar que la distribución inicial sea demasiado fuerte, ya que puede darse una distribución *a priori* tal que la posterior se concentre en un rango no deseado. También puede pasar que la distribución inicial sea demasiado débil

        - Las distribuciones *a priori* se deberían hacer más precisas cuando las inferencias posteriores sean vagas

    - Hay situaciones en las que no se recomienda utilizar todo el conocimiento relevante, aunque este pueda mejorar las inferencias posteriores

        - La preocupación principal es que la distribución *a priori* no debería dirigir las inferencias a ninguna dirección predeterminada

        - Estas preocupaciones pueden y deberían estar incluidas en el análisis de decisión y en algún tipo de modelo del proceso científico entero, de modo que se compensen las ganancias de la identificación temprana de efectos grandes y reales contra las pérdidas de sobreestimar las magnitudes de efectos y reacionar exageradamente a patrones que pueden ser aleatorios

        - No obstante, se sabe que las inferencias estadísticas se toman como evidencias de los efectos, y guía la toma de decisiones futuras, y por esta razón tiene sentido requerir que los modelos tengan ciertas restricciones como simetría en 0 para la *a priori* de un efecto de tratamiento individual

Los métodos de inferencia bayesianos

- Una vez se tiene la distribución posterior del parámetro, esta se puede utilizar para actualizar el modelo bayesiano y hacer inferencias a través del cálculo de estimaciones puntuales, interválicas, predicciones y contrastes de hipótesis

    - Después de recolectar los datos y evaluar el modelo, la inferencia estadística intenta adivinar cual es el verdadero valor del vector de parámetros $\boldsymbol{\theta}$ a partir de la distribución posterior del parámetro

        - En otras palabras, se intenta saber cuál es el modelo que ha generado los datos

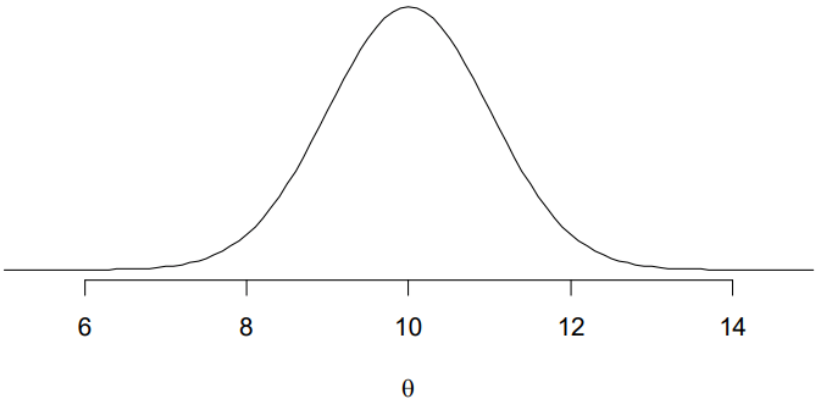
        - La inferencia estadística principalmente consiste de estimación puntual, estimación interválica, predicción y contrastes de hipótesis

        - Obviamente, es posible utilizar la distribución *a priori* para poder hacer estimaciones puntuales, interválicas y otras, pero solo reflejará el conocimiento *a priori* sobre el parámetro y no tendrá en cuenta los datos

    - La distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ tiene toda la información sobre el parámetro una vez se han observado los datos, dado que comprende la información de la distribución *a priori* y de la verosimilitud

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

        - Esta distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ es el estimator bayesiano natural para $\boldsymbol{\theta}$, dado que a partir de los datos y del conocimiento *a priori* se puede obtener una distribución que permite ver los posibles valores de $\boldsymbol{\theta}$



        - Teniendo la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, es posible hacer inferencia de varios tipos utilizando estadísticos y la misma distribución

- También es posible realizar inferencia sobre una función del parámetro $\phi = h(\boldsymbol{\theta})$ en vez de sobre $\boldsymbol{\theta}$

    - Una manera de poder realizar esta inferencia es a través de aplicar el principio de Jeffreys visto anteriormente para poder obtener la densidad posterior para $\phi = h(\theta)$, en donde θ es un solo parámetro

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

    - Como se discutió anteriormente, se asume un solo parámetro porque el caso múltiple es más controversial

    - Otra manera de poder obtener $p(\phi|\boldsymbol{y})$ es a través de una simulación, de modo que primero se fija un número de simulaciones M y, de $j = 1$ a M se simulan $\theta^{(j)}$ parámetros con $p(\theta|\boldsymbol{y})$ y se calculan los parámetros $\phi^{(j)} = h\left(\theta^{(j)}\right)$

    - Estas $\phi^{(1)}, \phi^{(2)}, \dots, \phi^{(M)}$ simuladas son simulaciones de $p(\phi|\boldsymbol{y})$. Por lo tanto, usando estas simulaciones se puede calcular todo lo que se necesite para hacer inferencia

- Una vez se ha entendido que la inferencia bayesiana se basará en la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$, se puede explicar de manera general cómo se obtienen estas inferencias dependiendo del tipo de estimador que se quiere utilizar o cómo hacer predicciones

    - Un estimador puntual $\widehat{\boldsymbol{\theta}}$ es un estadístico calculado a partir de los datos que se usa como estimación de los parámetros $\boldsymbol{\theta}$. Las estimaciones puntuales bayesianas son valores singulares como medidas de localización calculadas a partir de la distribución posterior

        - Las medidas de localización más utilizadas son la media, la mediana y el estimador máximo posterior MAP (equivalente a la moda de la densidad posterior) , las cuales se pueden calcular generalmente de la siguiente manera:

$$\widehat{\boldsymbol{\theta}}_{mean} = E(\boldsymbol{\theta}|\boldsymbol{y}) = \int_{\Theta} \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$

$$\widehat{\boldsymbol{\theta}}_{med} \quad such\ that \quad \int_{-\infty}^{\widehat{\boldsymbol{\theta}}_{med}} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = 0.5$$

$$\widehat{\boldsymbol{\theta}}_{MAP} = \widehat{\boldsymbol{\theta}}_{mode} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$$

- Escoger uno u otro estimador es equivalente a escoger una función de pérdida específica para escoger un estimador que minimice esta pérdida. La función de pérdida más utilizada es la de media cuadrática, la cual es minimizada por la media

$$PMSE(\widehat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} =$$

$$= \int_{-\infty}^{\infty} (\boldsymbol{\theta} - m + m - \widehat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} =$$

$$= \int_{-\infty}^{\infty} \left[(\boldsymbol{\theta} - m)^2 - 2(\boldsymbol{\theta} - m)(m - \widehat{\boldsymbol{\theta}}) + (m - \widehat{\boldsymbol{\theta}})^2\right] p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$

$$= Var(\boldsymbol{\theta}|\boldsymbol{y}) + (m - \widehat{\boldsymbol{\theta}})^2$$

$$m = E(\boldsymbol{\theta}|\boldsymbol{y}) \;\Rightarrow\; m = \arg\min_{\widehat{\boldsymbol{\theta}}} PMSE(\boldsymbol{\theta})$$

- Estos estimadores se pueden obtener de manera analítica (a través de la integración) o por simulación, de modo que se pueden obtener M simulaciones de $\boldsymbol{\theta}^{(j)}$ de $p(\boldsymbol{\theta}|\boldsymbol{y})$ y, a partir del vector de simulaciones, obtener los estadísticos deseados

- El valor esperado posterior se puede escribir como una media ponderada entre el estimador de máxima verosimilitud y el valor esperado *a priori*

$$E(\theta|\boldsymbol{y}) = \lambda \widehat{\boldsymbol{\theta}}_{MLE} + (1 - \lambda) E(\boldsymbol{\theta}) \quad for\ \ \lambda \in [0,1]$$

- Cuando más grande sea λ mayor peso tendrá la verosimilitud, lo cual ocurrirá cuando el tamaño muestral sea grande. En cambio, cuanto menor sea λ, mayor peso se pondrá en la *a priori* y esto ocurrirá cuando esta sea más informativa (tenga menor varianza)

- Un ejemplo de esta forma para la esperanza es considerando una distribución binomial $Bin(n, \theta)$ con una *a priori* $Beta(a, b)$, la cual permite obtener una posterior $Beta(a + \sum_{i=1}^{n} y_i\,, b + n)$

$$E(\theta|\boldsymbol{y}) = \frac{a + \sum_{i=1}^{n} y_i}{a + b + n} = \frac{a}{a + b + n} + \frac{\sum_{i=1}^{n} y_i}{a + b + n} =$$

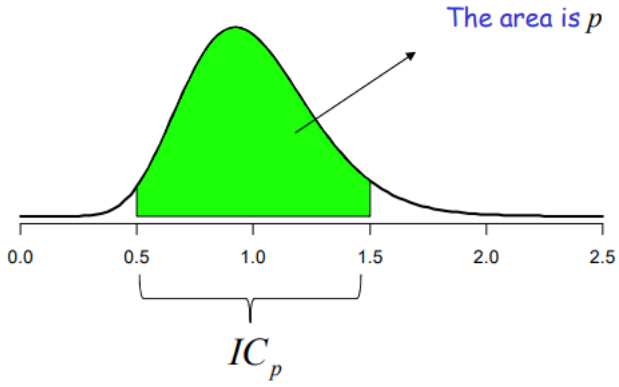
$$= \left(\frac{a + b}{a + b + n}\right)\left(\frac{a}{a + b}\right) + \left(\frac{n}{a + b + n}\right)\frac{\sum_{i=1}^{n} y_i}{n} =$$

$$= \lambda \bar{y} + (1 - \lambda) E(\theta) \quad for\ \ \lambda = \frac{n}{a + b + n}$$

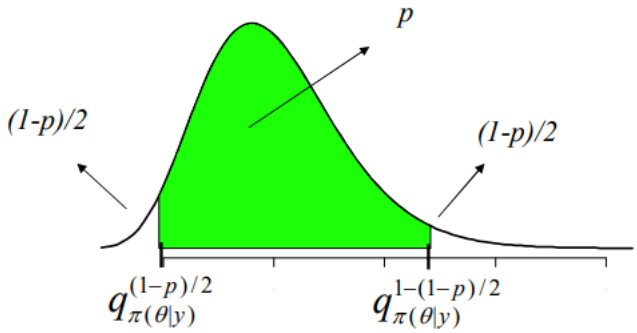
- Además de los resúmenes puntuales, también es imprescindible reportar la incertidumbre. El enfoque más usual es presentar cuartiles de la distribución posterior o un intervalo de probabilidad o de credibilidad posterior

    - Un intervalo de probabilidad o de credibilidad de p para θ es cualquier región en el espacio paramétrico Θ de modo que se cumpla la siguiente identidad:

$$P(q_\alpha < \theta < q_{1-\alpha}|\boldsymbol{y}) = \int_{q_\alpha}^{q_{1-\alpha}} p(\theta|\boldsymbol{y})\, d\theta = p$$



    - El intervalo central de probabilidad corresponde al rango de valores por encima y por debajo de $100(\alpha/2)\%$ en el caso de un intervalo del $100(1-\alpha)\%$, y este tipo de intervalo se denomina intervalo posterior



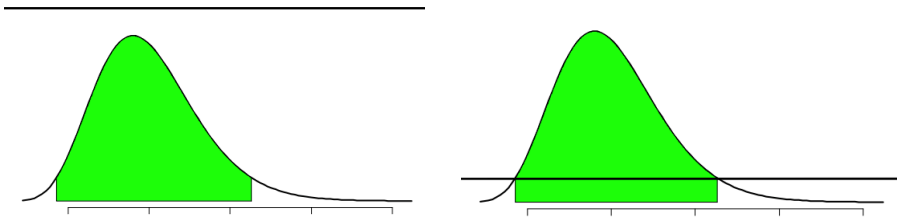
    - Estos intervalos se pueden estimar de manera analítica a través de obtener los cuartiles para los cuales la probabilidad a la izquierda de q y a la derecha de q son 0.025

$$\int_{-\infty}^{q_{0.025}} p(\theta|\boldsymbol{y})\, d\theta = 0.025 \qquad \int_{q_{0.975}}^{\infty} p(\theta|\boldsymbol{y})\, d\theta = 0.025$$

- Otra manera de estimarlos es a través de simulaciones, de modo que se obtienen $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(M)}$ simulaciones de $p(\theta|\boldsymbol{y})$ y se calculan los cuartiles que dejan un 0.025 del total de simulaciones a la izquierda y a la derecha

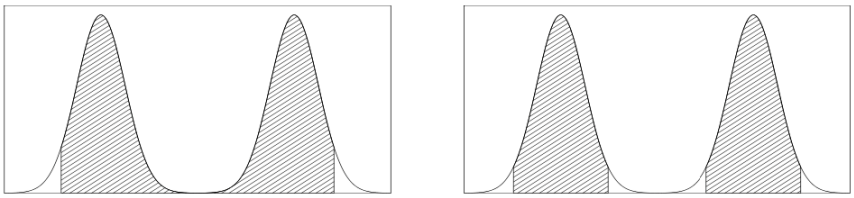
$$\frac{\sum_{j=1}^{q_{0.025}} \theta^{(j)}}{M} = 0.025 \qquad \frac{\sum_{j=q_{0.975}}^{M} \theta^{(j)}}{M} = 0.025$$

- Otro resumen de la incertidumbre posterior es la región de mayor densidad posterior, que es el conjunto de valores que contiene el $100(1-\alpha)\%$ de la probabilidad posterior y que tiene la característica de que la densidad en la región nunca es menor que la de fuera de esta, dando así el intervalo más estrecho posible en longitud



  - Esta región es idéntica al intervalo posterior central si la distribución posterior es unimodal y simétrica

  - En el caso en que no sea unimodal, la región se compondrá de intervalos disjuntos, y en el caso de asimetría, es muy probable que el intervalo posterior central sea diferente a la región



  - Si esta no es simétrica, entonces los cuartiles de la distribución posterior que se usan en el intervalo también pueden variar

- Como se ha comentado anteriormente, es posible hacer predicciones a través de la distribución posterior predictiva $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$, dado que esta representa toda la información que se tiene sobre los valores futuros de $\tilde{\boldsymbol{y}}$

  - La distribución, como se ha visto anteriormente, se puede calcular de manera analítica integrando sobre los valores de θ

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

- Otra manera de poder calcular esta distribución es a través de simulaciones, en donde se simulan M valores $\boldsymbol{\theta}^{(j)}$ simulaciones de $p(\boldsymbol{\theta}|\boldsymbol{y})$ y, con estas, se simulan M valores $\tilde{\boldsymbol{y}}^{(j)}$ de $p\left(\tilde{\boldsymbol{y}}\middle|\boldsymbol{\theta}^{(j)}\right)$. Estos $\tilde{\boldsymbol{y}}^{(1)}, \tilde{\boldsymbol{y}}^{(2)}, \ldots, \tilde{\boldsymbol{y}}^{(M)}$ pueden ser vectores y provienen de la distribución predictiva posterior

- Lo que esto permite es poder calcular estimaciones puntuales e intervalos de probabilidad o credibilidad para esta distribución de predicciones. Asumiendo que $\tilde{\boldsymbol{y}} = \tilde{y}$, se pueden obtener los siguientes resultados

$$E(\tilde{y}|\boldsymbol{y}) = \int_{\Omega} \tilde{y}\, p(\tilde{y}|\boldsymbol{y})\, d\tilde{y} = \int_{\Omega} \tilde{y} \left(\int_{\Theta} p(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \right) d\tilde{y}$$

$$or\ \ E(\tilde{y}|\boldsymbol{y}) = \sum_{j=1}^{M} \tilde{y}^{(j)}$$

$$\int_{-\infty}^{q_{0.025}} p(\tilde{y}|\boldsymbol{y})\, d\tilde{y} = 0.025 \qquad \int_{q_{0.975}}^{\infty} p(\tilde{y}|\boldsymbol{y})\, d\tilde{y} = 0.025$$

$$or \qquad \frac{\sum_{j=1}^{q_{0.025}} \tilde{y}^{(j)}}{M} = 0.025 \qquad \frac{\sum_{j=q_{0.975}}^{M} \tilde{y}^{(j)}}{M} = 0.025$$

- También es posible realizar contrastes de hipótesis de manera mucho más flexible pero diferente a cómo se realiza en estadística frecuentista, todo a través de utilizar la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$

    - Hay dos situaciones que se estudiarán: el contraste de hipótesis de un solo lado y el contraste de hipótesis a dos lados

        - En el primer tipo, solo interesa detectar el efecto en una sola dirección. En este caso, el contraste de hipótesis bayesiano funciona extremadamente bien (sin las contradicciones que se dan en el enfoque frecuentista) y se realiza con la probabilidad posterior de la hipótesis nula (usando la distribución posterior)

        - En el segundo tipo se encuentran dos casos: uno en el que se contrasta una hipótesis conjunta frente a dos alternativas y otra en la que se contrasta una hipótesis singular (no conjunta) frente a dos alternativas

        - Si se contrasta la pertenencia del valor del parámetro a un conjunto concreto contra su no pertenencia, entonces se puede

utilizar la probabilidad posterior de la hipótesis nula (usando la distribución posterior)

- Cuando solo se contrasta para un valor concreto, su probabilidad *a priori* y posterior deberían ser nulas debido a que se tiene una densidad *a priori* y posterior (un punto exacto tiene probabilidad nula), por lo que no se puede contrastar la hipótesis nula usando la distribución posterior y se tendrán que usar otros métodos

- Comenzando desde el modelo bayesiano $M_B = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta; p(\boldsymbol{\theta})\}$, se divide el espacio paramétrico Θ en dos o más subespacios, correspondientes a aquellos valores a contrastar

  - El objetivo con esta subdivisión es poder saber a qué subespacio pertenece el valor real de los parámetros $\boldsymbol{\theta}$

- Dado un espacio paramétrico Θ, se pueden definir dos subespacios disjuntos Θ_0 y Θ_1 (tales que $\Theta = \Theta_0 \cup \Theta_1$) para poder decidir a qué subespacio pertenece el parámetro

$$H_0: \boldsymbol{\theta} \in \Theta_0 \quad or \quad H_1: \boldsymbol{\theta} \in \Theta_1$$

  - Después de observar los datos, se puede actualizar el modelo bayesiano para obtener la probabilidad posterior para cada hipótesis a partir de la distribución posterior

$$p(H_0|\boldsymbol{y}) = p(\boldsymbol{\theta} \in \Theta_0|\boldsymbol{y}) = \int_{\Theta_0} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$

$$p(H_1|\boldsymbol{y}) = p(\boldsymbol{\theta} \in \Theta_1|\boldsymbol{y}) = \int_{\Theta_1} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = 1 - p(H_0|\boldsymbol{y})$$

  - A partir de las probabilidades que se obtengan, se decidirá favorecer la hipótesis con más probabilidad, por lo que se rechaza la hipótesis nula si $p(H_0|\boldsymbol{y}) < p(H_1|\boldsymbol{y})$ y no se rechaza si $p(H_0|\boldsymbol{y}) > p(H_1|\boldsymbol{y})$

  - Dividir el espacio paramétrico en diferentes subespacios es equivalente a dividir el modelo bayesiano en diferentes submodelos, por lo que realizar un contraste de hipótesis se puede entender como una manera de seleccionar diferentes modelos

$$M_B = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta; p(\boldsymbol{\theta})\} \;\Rightarrow\; \begin{cases} M_0' = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0; p(\boldsymbol{\theta})\} \\ M_1' = \{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_1; p(\boldsymbol{\theta})\} \end{cases}$$

$$\Rightarrow \begin{cases} p(M_0'|\boldsymbol{y}) = p(H_0|\boldsymbol{y}) = p(\boldsymbol{\theta} \in \Theta_0|\boldsymbol{y}) = \int_{\Theta_0} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \\ p(M_1'|\boldsymbol{y}) = p(H_1|\boldsymbol{y}) = p(\boldsymbol{\theta} \in \Theta_1|\boldsymbol{y}) = \int_{\Theta_1} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \end{cases}$$

- Con esta metodología, se pueden realizar contrastes de hipótesis de una cola y contrastes de hipótesis de dos colas cuando la hipótesis a contrastar es conjunta

- En el caso en el que el espacio paramétrico Θ sea discreto y solo tenga dos posibilidades $\Theta = \{\theta_0, \theta_1\}$, entonces se puede calcular la distribución posterior de cada hipótesis a través del teorema de Bayes y el teorema de la probabilidad total

$$M' = \{p(\tilde{\boldsymbol{y}}|\theta), \theta \in \{\theta_0, \theta_1\}; p(\theta)\} \quad where \quad p(\theta) = \begin{cases} p & if \ \ \theta = \theta_0 \\ 1-p & if \ \ \theta = \theta_1 \end{cases}$$

$$H_0: \theta = \theta_0 \quad or \quad H_1: \theta = \theta_1$$

  - Debido a que la probabilidad de que $\theta = \theta_0$ es p, la probabilidad incondicional de H_0 es $p(H_0) = p$ y la de H_1 es $p(H_1) = 1 - p$

  - A través de los siguientes cálculos, se puede obtener la probabilidad posterior a través de la probabilidad incondicional de cada hipótesis y la verosimilitud

$$p(H_0|\boldsymbol{y}) = \frac{p(H_0)p(\boldsymbol{y}|H_0)}{p(\boldsymbol{y})} = \frac{p(H_0)p(\boldsymbol{y}|H_0)}{p(H_0)p(\boldsymbol{y}|H_0) + p(H_1)p(\boldsymbol{y}|H_1)}$$

$$p(H_1|\boldsymbol{y}) = 1 - p(H_0|\boldsymbol{y}) = \frac{p(H_1)p(\boldsymbol{y}|H_1)}{p(H_0)p(\boldsymbol{y}|H_0) + p(H_1)p(\boldsymbol{y}|H_1)}$$

  - Como se puede observar, la distribución posterior no se utiliza para calcular las probabilidades, si no que son estas mismas probabilidades las que determinarán la posterior. A partir de esta, se escoge la hipótesis con la probabilidad más alta

$$p(\theta|\boldsymbol{y}) = \begin{cases} p(H_0|\boldsymbol{y}) & if \ \ \theta = \theta_0 \\ p(H_1|\boldsymbol{y}) & if \ \ \theta = \theta_1 \end{cases}$$

- Dadas dos hipótesis H_i y H_j, es posible obtener las *odds* posteriores a través de la función de verosimilitud para cada hipótesis de la siguiente manera:

$$\frac{p(H_i|\boldsymbol{y})}{p(H_j|\boldsymbol{y})} = \frac{\frac{p(H_i)p(\boldsymbol{y}|H_i)}{p(\boldsymbol{y})}}{\frac{p(H_i)p(\boldsymbol{y}|H_j)}{p(\boldsymbol{y})}} = \frac{p(H_i)}{p(H_j)}\frac{p(\boldsymbol{y}|H_i)}{p(\boldsymbol{y}|H_j)} = \frac{p(H_i)}{p(H_j)}FB_{ij}$$

- La *ratio* $p(\boldsymbol{y}|H_0)/p(\boldsymbol{y}|H_1)$ es el factor de Bayes, de modo que las *odds* posteriores se pueden entender como el producto entre las *odds a priori* y el factor de Bayes

- Debido a esta igualdad anterior, el factor de Bayes se puede expresar a través de las probabilidades posteriores de las hipótesis y las *a priori*

$$FB_{ij} = \frac{p(\boldsymbol{y}|H_i)}{p(\boldsymbol{y}|H_j)} = \frac{p(H_j)}{p(H_i)}\frac{p(H_i|\boldsymbol{y})}{p(H_j|\boldsymbol{y})}$$

- Dadas dos hipótesis H_i y H_j, si $FB_{ij} > 1$, entonces $p(\boldsymbol{y}|H_i) > p(\boldsymbol{y}|H_j)$ y se escogería H_i, mientras que si $FB_{ij} < 1$, entonces $p(\boldsymbol{y}|H_i) < p(\boldsymbol{y}|H_j)$

- Kass y Raftery (1995) propone una escala para poder interpretar mejor el factor de Bayes dependiendo de su valor y la fuerza de la evidencia que este proporciona

$\log_{10}$ FB	FB	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

- En el caso en que se contraste una hipótesis singular contra una hipótesis conjunta (un contraste de dos colas en el que la hipótesis nula es un solo valor) y la distribución de los datos es discreta, se puede utilizar el factor de Bayes para decidir

  - Se puede definir el contraste de hipótesis como una elección entre dos modelos, en donde la distribución *a priori* para el modelo nulo sería una función delta de Dirac

$$H_0: \theta = \theta_0 \quad or \quad H_1: \theta \neq \theta_0$$

$$\begin{cases} M_0' = \{p(\tilde{\boldsymbol{y}}|\theta), \theta = \theta_0; p_0(\theta)\} & where \quad p_0(\theta) = \delta(x - \theta_0) \\ M_1' = \{p(\tilde{\boldsymbol{y}}|\theta), \theta \neq \theta_0; p_1(\theta)\} & where \quad p_1(\theta) \ is \ other \ prior \end{cases}$$

$$\Rightarrow \ p(H_i) = p_i(\theta) \quad for \ \ i = 0{,}1$$

- A través del teorema de Bayes se puede sacar la probabilidad posterior de que se de uno u otro modelo (equivalente a la hipótesis) y así utilizar el factor de Bayes para decidir sobre la hipótesis con la probabilidad más alta

$$p(H_0|\boldsymbol{y}) = \frac{p(H_0)p(\boldsymbol{y}|H_0)}{p(H_0)p(\boldsymbol{y}|H_0) + p(H_1)p(\boldsymbol{y}|H_1)}$$

$$p(H_1|\boldsymbol{y}) = 1 - p(H_0|\boldsymbol{y})$$

$$\Rightarrow\ FB_{01} = \frac{p(\boldsymbol{y}|H_0)}{p(\boldsymbol{y}|H_1)} = \frac{p(H_1)}{p(H_0)}\frac{p(H_0|\boldsymbol{y})}{p(H_1|\boldsymbol{y})}$$

- Si, en cambio, se está en el caso anterior pero la distribución de probabilidad es continua, entonces no se puede aplicar el método anterior y se recurre a los intervalos de credibilidad o probabilidad

  - Como la distribución es discreta, es posible calcular $p(\boldsymbol{y}|H_i)$ en el método anterior. De otro modo, este método no se podría aplicar debido a la definición de densidad de probabilidad

  - Calculando un intervalo de probabilidad $100(1-\alpha)\%$ para θ, si θ_0 no pertenece a este intervalo, entonces se rechaza la hipótesis nula, mientras que si θ_0 pertenece, entonces no se puede rechazar (se mantiene como un valor creíble)

- Los contrastes de hipótesis vistos anteriormente son un caso particular de un enfoque más general, en donde se pueden plantear más de dos hipótesis. Para cada hipótesis, se podría plantear un modelo bayesiano, de modo que se pueden tener k modelos con sus correspondientes k hipótesis

$$H_0 :\ M_0' = \{p_0(\tilde{\boldsymbol{y}}|\theta_0), \boldsymbol{\theta_0} \in \Theta_0; p_0(\theta_0)\}\ \ where\ \ p(H_0) = p(M_0)$$

$$H_1 :\ M_1' = \{p_1(\tilde{\boldsymbol{y}}|\theta_1), \boldsymbol{\theta_1} \in \Theta_1; p_1(\theta_1)\}\ \ where\ \ p(H_1) = p(M_1)$$

$$\dots$$

$$H_k :\ M_k' = \{p_k(\tilde{\boldsymbol{y}}|\theta_k), \boldsymbol{\theta_k} \in \Theta_k; p_k(\theta_k)\}\ \ where\ \ p(H_k) = p(M_k)$$

  - Cada modelo es una lista de distribuciones de probabilidad con sus propios parámetros y espacio paramétrico

$$M_i' = \{p_i(\tilde{\boldsymbol{y}}|\boldsymbol{\theta_i}), \boldsymbol{\theta_i} \in \Theta_i\}$$

- La distribución *a priori* en cada modelo se define en su propio espacio paramétrico

$$p_i(\boldsymbol{\theta_i}) = \begin{cases} p_i(\boldsymbol{\theta_i}) & if \ \boldsymbol{\theta_i} \in \Theta_i \\ 0 & otherwise \end{cases}$$

- La probabilidad *a priori* del modelo se define para cada modelo, y esta es equivalente a la probabilidad *a priori* de la hipótesis

$$p(H_i) = p(M_i') = \int_{\Theta_i} p_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}$$

- Cada modelo tiene su propia distribución predictiva *a priori*

$$p(\tilde{\boldsymbol{y}}|M_i') = \int_{\Theta_i} p_i(\tilde{\boldsymbol{y}}|\boldsymbol{\theta_i}) p_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}$$

- Una manera de poder calcular las probabilidades posteriores y decidir entre las hipótesis es a través de teorema de Bayes, lo cual permite utilizar el factor de Bayes y poder hacer predicciones

  - La probabilidad posterior para cada modelo o hipótesis se puede obtener aplicando el teorema de Bayes y el de probabilidad total a la de los modelos (dado que es equivalente a las hipótesis)

$$(1) \quad p(M_i') = \int_{\Theta_i} p_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}$$

$$(2) \quad p(\boldsymbol{y}|M_i') = \int_{\Theta_i} p_i(\boldsymbol{y}|\boldsymbol{\theta_i}) p_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}$$

$$(3) \quad p(M_i'|\boldsymbol{y}) = \frac{p(M_i')p(\boldsymbol{y}|M_i')}{\sum_{j=0}^{k} p(M_j')p(\boldsymbol{y}|M_j')}$$

  - A partir de las probabilidades posteriores para cada modelo, se puede calcular el factor de Bayes para cada par de modelos y obtener una tabla. Sin embargo, también se puede obtener el factor de Bayes de una hipótesis H_i (o modelo M_i') contra todas las otras hipótesis

i \ j	1	2	3
1	-	0.48	0.71
2	2.08	-	1.48
3	1.4	0.67	-

$$FB_{ij} = \frac{p(\boldsymbol{y}|M_i')}{p(\boldsymbol{y}|M_j')} = \frac{p(M_j')}{p(M_i')}\frac{p(M_i'|\boldsymbol{y})}{p(M_j'|\boldsymbol{y})}$$

$$\Rightarrow FB_{i.} = \frac{p(\boldsymbol{y}|M_i')}{1-p(\boldsymbol{y}|M_i')} = \frac{1-p(M_i')}{p(M_i')}\frac{p(M_i'|\boldsymbol{y})}{1-p(M_i'|\boldsymbol{y})}$$

- Para poder hacer predicciones, normalmente se utiliza el método de promediar los modelos, lo cual consiste en utilizar una media ponderada de las distribuciones predictivas posteriores de todos los modelos y así poder hacer predicciones

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{i=0}^{k} p_i(M_i'|\boldsymbol{y})p_j(\tilde{\boldsymbol{y}}|\boldsymbol{y})$$

$$where\ \ p_j(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \int_{\Theta_j} p_j(\tilde{\boldsymbol{y}}|\boldsymbol{\theta_j})p_j(\boldsymbol{\theta_j}|\boldsymbol{y})\, d\boldsymbol{\theta_j}$$

- El contraste de hipótesis también se puede llevar a cabo a través de métodos computacionales con la distribución posterior y con la distribución posterior predictiva. Para ello, se tienen que seguir los siguientes pasos:

  - El primer paso es obtener la distribución posterior o la distribución posterior predictiva a través de los métodos vistos anteriormente, obteniendo así una distribución con la que trabajar

  - A partir de esta, se pueden utilizar las funciones integradas en el *software* para obtener probabilidades de diversos rangos (como el porcentaje de observaciones que están en un rango) que conforman las hipótesis y así aceptar o rechazar en base a esta probabilidad

Los modelos jerárquicos

- Muchas aplicaciones estadísticas requieren muchos parámetros que se pueden interpretar como relacionados o conectados de alguna manera por la estructura del problema, implicando que un modelo de probabilidad conjunto para estos parámetros debería reflejar su dependencia

  - Hay contextos en donde se puede esperar que las estimaciones de los parámetros θ_i estén relacionadas con las otras. Esto se puede conseguir modelar de manera natural si se usa una distribución *a priori* en la que los parámetros θ_i se consideran una muestra de una distribución poblacional común

- Una importante característica de estas aplicaciones es que los datos observados y_{ij} (de un individuo i en un grupo j) se pueden usar para estimar aspectos de la distribución poblacional de las θ_i aunque los valores de θ_i no se observen

- Es muy común modelar estos problemas de manera jerárquica, con resultados observables modelados condicionalmente en ciertos parámetros, a los cuales se da una especificación probabilística en términos de otros parámetros llamados hiperparámetros

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}) \quad (1st\ level)$$

$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \quad (2nd\ level)$$

$$p(\boldsymbol{\gamma}) \quad (3rd\ level)$$

- Este pensamiento jerárquico ayuda a entender los problemas multiparámetro y juega un papel importante al desarrollar estrategias computacionales

  - Lo que más importante es en la práctica es que los modelos no jerárquicos simples suelen ser inapropiados para datos jerárquicos: con pocos parámetros, estos tienden a sobreajustarse a los datos

    - En cambio, los modelos jerárquicos pueden tener los parámetros suficientes para ajustarse bien a los datos, mientras que usan una distribución poblacional para estructurar la dependencia en los parámetros, evitando asó los problemas de sobreajuste

- Considerando un conjunto de experimentos $i = 1,2, \dots, I$ en la que el experimento i tiene un vector de datos $\boldsymbol{y}_i$, un vector de parámetros $\boldsymbol{\theta_i}$ y función de verosimilitud $p(\boldsymbol{y}_i|\boldsymbol{\theta_i})$, se puede configurar un modelo jerárquico a través de la idea de intercambiabilidad

  - Si no hay más información (a parte de los datos) disponible para distinguir cualquiera de las $\boldsymbol{\theta_i}$ de otras, y no hay un orden o agrupación de los parámetros, uno debe asumir simetría de los parámetros en su distribución *a priori*

    - Esta simetría es representada probabilísticamente por la intercambiabilidad: los parámetros $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \dots, \boldsymbol{\theta_I})$ son intercambiables en su distribución conjunta si $p(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \dots, \boldsymbol{\theta_I})$ es invariante a permutaciones de los índices $(1, \dots, I)$

- El concepto de intercambiabilidad ya se ha encontrado construyendo modelos independientes e idénticamente distribuidos para datos directos

- En la práctica, la ignorancia implica intercambiabilidad: cuanto menos se sepa de un problema, lo más confiado que uno puede suponer intercambiabilidad

- La forma más simple de una distribución intercambiable tiene todos los parámetros $\boldsymbol{\theta_i}$ como una muestra independiente de una distribución *a priori* (o poblacional) gobernada por un vector de parámetros desconocido

$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \prod_{i=1}^{I} p(\boldsymbol{\theta_i}|\boldsymbol{\gamma})$$

- En general, $\boldsymbol{\gamma}$ es desconocido, de modo que la distribución para $\boldsymbol{\theta}$ debe promediar sobre la incertidumbre de $\boldsymbol{\gamma}$. Esta forma de mezcla de distribuciones iid es todo lo que se necesita para capturar la intercambiabilidad en la práctica

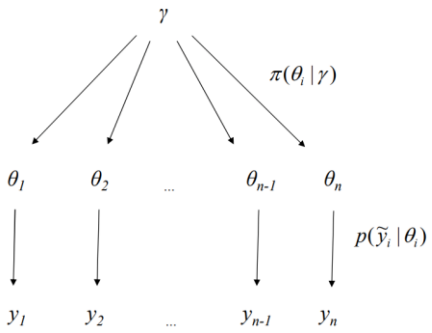
$$p(\boldsymbol{\theta}) = \int_{\Theta_\gamma} \left[\prod_{i=1}^{I} p(\boldsymbol{\theta_i}|\boldsymbol{\gamma}) \right] p(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}$$

- Un resultado teórico relacionado es el teorema de Finetti, que expresa que en límite $I \to \infty$, cualquier distribución intercambiable de buen comportamiento sobre $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \dots, \boldsymbol{\theta_I})$ puede expresarse como una mezcla de distribuciones iid. Este teorema no se mantiene cuando I es finita

- Estadísticamente, el modelo de mezcla caracteriza los parámetros $\boldsymbol{\theta}$ como observaciones sacadas de una "superpoblación" común que es determinada por unos hiperparámetros desconocidos $\boldsymbol{\gamma}$

- Es común que las observaciones no sean completamente intercambiables, sino que lo sean parcialmente o condicionalmente

- Si las observaciones pueden ser agrupadas, se puede hacer un modelo jerárquico en donde cada grupo tenga su propio submodelo, pero las propiedades del grupo son desconocidas. Si se asume que las propiedades del grupo son intercambiables, se puede usar una distribución *a priori* común para las propiedades del grupo



- Si $\boldsymbol{y_i}$ tiene información adicional $\boldsymbol{x_i}$ de modo que $\boldsymbol{y_i}$ no es intercambiable, pero $(\boldsymbol{y_i}, \boldsymbol{x_i})$ si lo es, entonces se puede hacer un modelo conjunto para $(\boldsymbol{y_i}, \boldsymbol{x_i})$ o un modelo condicional para $\boldsymbol{y_i}|\boldsymbol{x_i}$

- En general, la manera usual de modelar la intercambialidad con variables explicativas es a través de asumir independencia condicional. De este modo, modelos intercambiables se vuelven aplicables universalmente, dado que cualquier información disponible para distinguir diferentes unidades debería estar en las variables $\boldsymbol{y}$ y $\boldsymbol{x}$

$$p(\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_I}|\boldsymbol{x_1}, \dots, \boldsymbol{x_I}) = \int_{\Theta_\gamma} \left[\prod_{i=1}^{I} p(\boldsymbol{\theta_i}|\boldsymbol{\gamma}, \boldsymbol{x_i})\right] p(\boldsymbol{\gamma}|\boldsymbol{x})\, d\boldsymbol{\gamma}$$

- En cualquier aplicación estadística, es natural poner objeciones a la intercambiabilidad debido a la diferencia entre unidades, pero esta información no invalida la intercambiabilidad

  - Que los experimentos difieran implica que las $\boldsymbol{\theta_i}$ difieren, pero puede ser perfectamente aceptable considerarlas como si fueran una muestra de una distribución común

  - Si no hay información para distinguirlas, no se tiene otra elección lógica más que modelar las $\boldsymbol{\theta_i}$ de modo intercambiable. Por lo tanto, objetar esto es lo mismo que objetar un modelo iid, los modelos de regresión u otros

- Volviendo al problema de la inferencia, la parte jerárquica clave de estos modelos es que $\boldsymbol{\gamma}$ no es conocido y, por tanto, tiene su propia distribución $p(\boldsymbol{\gamma})$

  - La distribución posterior bayesiana apropiada es del vector $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, por lo que la distribución *a priori* conjunta es la siguiente:

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$$

- La distribución posterior conjunta es la siguiente, con la simplificación de que $p(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ debido a que $p(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\theta})$ depende solo de $\boldsymbol{\theta}$ porque los hiperparámetros solo afectan a través de $\boldsymbol{\theta}$ ($\boldsymbol{y}$ no depende de $\boldsymbol{\gamma}$ condicionalmente):

$$p(\boldsymbol{\gamma},\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\gamma},\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\gamma},\boldsymbol{\theta}) =$$

$$= \frac{p(\boldsymbol{y},\boldsymbol{\theta})p(\boldsymbol{\gamma})}{p(\boldsymbol{\theta},\boldsymbol{\gamma})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{p(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})} = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma},\boldsymbol{\theta})$$

- Para crear una distribución conjunta para $(\boldsymbol{\gamma},\boldsymbol{\theta})$, se tiene que asignar una distribución *a priori* para $\boldsymbol{\gamma}$

  - Si se sabe poco sobre $\boldsymbol{\gamma}$, se puede asignar una distribución *a priori* difusa, pero se tiene que tener cuidado si se utiliza una distribución impropia porque se debe comprobar que la posterior resultante sí es propia y si las conclusiones son muy sensibles a esta simplificación

  - En la mayoría de problemas reales, uno debe tener suficiente información sustantiva sobre los parámetros en $\boldsymbol{\gamma}$ para al menos restringir estos a una región finita o asignar una distribución concreta

  - Igual que en los modelos no jerárquicos, normalmente es práctico comenzar con una distribución simple y no muy informativa e ir añadiendo información *a priori* si hay demasiada variabilidad

- En consecuencia, tomando una muestra aleatoria $\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_I}$ de tamaño I, el modelo bayesiano jerárquico se puede escribir de la siguiente manera:

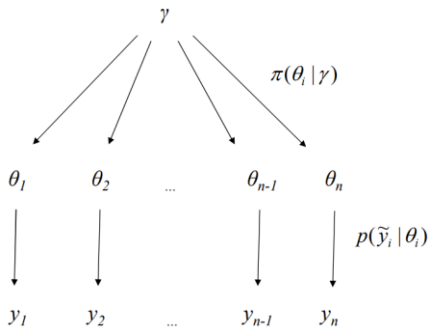
$$\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{i}} \sim p(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{i}}) \quad \boldsymbol{\theta}_{\boldsymbol{i}}|\boldsymbol{\gamma} \sim p(\boldsymbol{\theta}_{\boldsymbol{i}}|\boldsymbol{\gamma}) \quad \boldsymbol{\gamma} \sim p(\boldsymbol{\gamma}) \quad for\ i = 1,2,\dots,I$$

  - Es posible pasar de un modelo jerárquico a uno no jerárquico, aunque eso provoca perder la opción de realizar inferencias predictivas en el segundo nivel (para $\boldsymbol{\theta}$). Solo hace falta integrar para todos los posibles valores de $\boldsymbol{\theta_i}$, de modo que solo dependería de $\boldsymbol{\gamma}$

$$\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\gamma} \sim p(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\gamma}) = \prod_{i=1}^{I} \int_{\Theta_i} p(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{i}}) p(\boldsymbol{\theta}_{\boldsymbol{i}}|\boldsymbol{\gamma})\, d\boldsymbol{\theta}_{\boldsymbol{i}} \qquad \boldsymbol{\gamma} \sim p(\boldsymbol{\gamma})$$

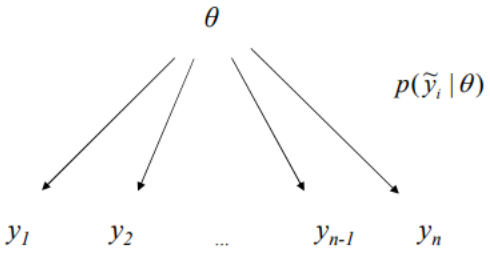
$$for\ i = 1,2,\dots,n$$

- En este caso, las $\boldsymbol{y_i}$ son independientes, pero no se distribuyen idénticamente (su distribución $p(\tilde{\boldsymbol{y}}_{\boldsymbol{i}}|\boldsymbol{\theta_i})$ depende de parámetros diferentes $\boldsymbol{\theta_i}$), mientras que las θ_i son independientes e idénticamente distribuidas (su distribución $p(\boldsymbol{\theta_i}|\boldsymbol{\gamma})$ depende de los mismos parámetros $\boldsymbol{\gamma}$)



- En un modelo no jerárquico, en cambio, las y_i son independientes e idénticamente distribuidas, ya que depende de la misma distribución *a priori* $p(\boldsymbol{\gamma})$

$$\tilde{y}_i|\boldsymbol{\theta} \sim p(\tilde{y}_i|\boldsymbol{\theta}) \quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \quad for\ i = 1,2,\dots,n$$



- Los modelos jerárquicos se caracterizan por hiperparámetros $\boldsymbol{\gamma}$ y parámetros $\boldsymbol{\theta}$, por lo que hay dos distribuciones posteriores predictivas que pueden ser de interés para el analista

  - La primera es la distribución de las futuras observaciones de $\tilde{\boldsymbol{y}}$ correspondientes a una $\boldsymbol{\theta_i}$ existente, lo que sería la distribución predictiva posterior del primer nivel $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta_i})$

  - La segunda es la distribución de observaciones $\tilde{\boldsymbol{y}}$ correspondientes a futuras $\boldsymbol{\theta_i}$ sacadas de la misma superpoblación ($\boldsymbol{\theta_i}$ no está definida, es un valor futuro de otro experimento), denotada por $p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}})$

  - En el primer caso, las observaciones $\tilde{\boldsymbol{y}}$ obtenidas de la posterior predictiva se basan en observaciones posteriores $\boldsymbol{\theta_i}$ para el experimento existente. En cambio, en el segundo caso, uno

primero debe obtener las observaciones $\widetilde{\boldsymbol{\theta}}$ para el nuevo experimento de la distribución de población, dadas las observaciones posteriores de $\boldsymbol{\gamma}$, y entonces sacar observaciones de $\widetilde{\boldsymbol{y}}$ dadas las obtenidas $\widetilde{\boldsymbol{\theta}}$

- Ambos tipos de replicaciones se pueden usar para evaluar la adecuación del modelo

- La estrategia inferencial para los modelos jerárquicos sigue el mismo enfoque general que para los modelos multiparamétricos, pero es más difícil en la práctica por el gran número de parámetros de los modelos jerárquicos

    - Para realizar la derivación analítica de las distribuciones condicionales y marginales se tienen que seguir los siguientes pasos:

        - Se escribe la densidad posterior conjunta $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})$ en forma no normalizada como el producto de la distribución *hiperpriori* $p(\boldsymbol{\gamma})$, la distribución poblacional $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ y la verosimilitud $p(\boldsymbol{y}|\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma}, \boldsymbol{\theta})$$

        - Se determina analíticamente la densidad posterior condicional de $\boldsymbol{\theta}$ dados los hiperparámetros $\boldsymbol{\gamma}$, de modo que para una $\boldsymbol{y}$ observada fija, esta es una función $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$

$$p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y}) = \int_{\Theta_\gamma} p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})\, d\boldsymbol{\gamma} = \int_{\Theta_\gamma} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma}, \boldsymbol{\theta})\, d\boldsymbol{\gamma}$$

        - Se estima $\boldsymbol{\gamma}$ utilizando el paradigma bayesiano, es decir, obteniendo la distribución marginal posterior $p(\boldsymbol{\gamma}|\boldsymbol{y})$

$$p(\boldsymbol{\gamma}|\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})\, d\boldsymbol{\theta} = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma}, \boldsymbol{\theta})\, d\boldsymbol{\gamma}$$

        - Para muchas distribuciones estándar, sin embargo, se puede obtener la distribución $p(\boldsymbol{\gamma}|\boldsymbol{y})$ algebraicamente usando la probabilidad condicional

$$p(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma}, \boldsymbol{\theta})}{\int_{\Theta_\gamma} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\gamma}, \boldsymbol{\theta})\, d\boldsymbol{\gamma}}$$

        - Lo malo es que el denominador de $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$ tiene un factor normalizador (la constante de Bayes) que depende de $\boldsymbol{\gamma}$ y de $\boldsymbol{y}$, por lo que se tiene que comprobar que esta constante es, en verdad, una constante

- Para poder simular observaciones de la distribución posterior conjunta $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})$ de modelos jerárquicos simples, se pueden realizar los siguientes pasos:

  - Se saca el vector de hiperparámetros $\boldsymbol{\gamma}$ de su distribución posterior marginal $p(\boldsymbol{\gamma}|\boldsymbol{y})$. Si $\boldsymbol{\gamma}$ tiene pocas dimensiones, entonces se pueden utilizar métodos simples, pero si no, se necesitan métodos más sofisticados

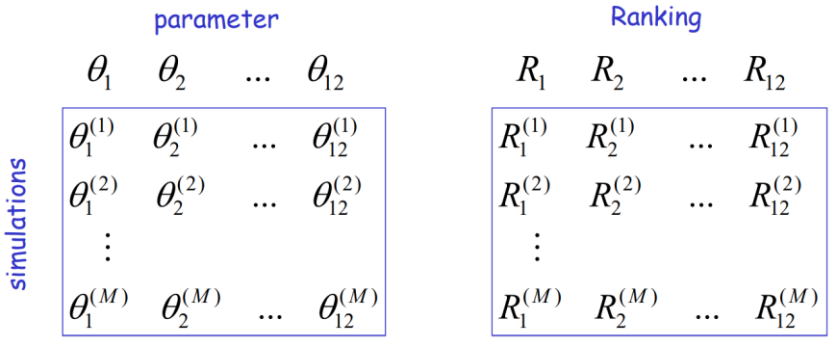
  - Se saca el vector de parámetros $\boldsymbol{\theta}$ de su distribución condicional posterior $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$ dado el valor del vector $\boldsymbol{\gamma}$. En ejemplos sencillos, la factorización $p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y}) = \prod_i p(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$ se cumple, por lo que los componentes θ_i se pueden obtener uno a uno de manera independiente y una a una

  - Si se desea, se pueden sacar los valores predictivos $\tilde{\boldsymbol{y}}$ de la distribución posterior predictiva dados los valores simulados de $\boldsymbol{\theta}$. Dependiendo del problema, primero será necesario obtener los valores de $\tilde{\boldsymbol{\theta}}$ dados los de $\boldsymbol{\gamma}$ para obtener la predictiva para un nuevo experimento

  - Estos pasos se realizan L veces con tal de obtener L observaciones simuladas. Con las simulaciones posteriores conjuntas de $\boldsymbol{\theta}$ e $\tilde{\boldsymbol{y}}$, uno puede calcular la distribución posterior de cualquier estimando o cantidad predictiva de interés

- Existen veces en las que se quiere realizar una clasificación o establecer un orden dependiendo de los parámetros θ_i, lo cual es posible a través de los métodos de simulación

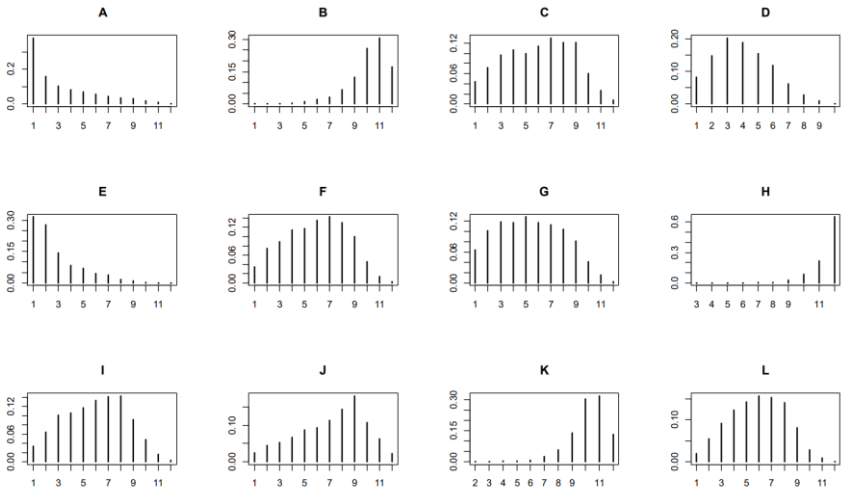
  - Para poder realizarlo, se tiene que crear una variable de clasificación R para cada indiviudo i, de modo que para cada simulación $\theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_I^{(j)}$, se asigna un valor del 1 al I a cada individuo $R_1^{(j)}, R_2^{(j)}, \dots, R_I^{(j)}$ dependiendo del orden cuantitativo que ocurre en j



$$\theta_1^{(j)} < \theta_7^{(j)} < \theta_3^{(j)} < \cdots < \theta_I^{(j)} < \theta_9^{(j)}$$

$$\Rightarrow R_1^{(j)} = 1, R_2^{(j)} = 5, R_3^{(j)} = 3, \ldots, R_I^{(j)} = I - 1$$

- Como se realizan varias iteraciones, se puede establecer una distribución de la variable R para cada individuo. Además, a partir de estas distribuciones se puede calcular un intervalo de credibilidad para cada uno de los individuos y así saber la oscilación de puestos



80% CI	A	B	C	D	E	F	G	H	I	J	K	L
10%	1	8	2	2	1	2	2	10	3	3	9	3
90%	7	12	9	7	6	9	9	12	9	10	12	9

La comprobación de modelos

- Una vez que se han realizado los dos primeros pasos del análisis bayesiano (construir un modelo de probabilidad y calcular la distribución posterior para todos los estimandos), se debe comprobar el ajuste del modelo a los datos y al conocimiento subjetivo

    - La comprobación del modelo es crucial para el análisis estadístico: las inferencias bayesianas de *a priori* a posterior asumen la estructura completa de un modelo de probabilidad y puede dar inferencias engañosas cuando el modelo es pobre

        - Un buen análisis bayesiano, por tanto, debería incluir al menos alguna comprobación de la adecuación del ajuste del modelo a los datos y la plausibilidad del modelo para los propósitos para los que el modelo se utilizará

        - Esto a veces se discute como un problema de sensibilidad de la distribuión *a priori*, pero en la práctica la verosimilitud del modelo es típicamente solo una sospecha

        - Se utiliza “modelo” como término que comprende la distribución muestral, la distribución *a priori*, cualquier estructura jerárquica y temas como qué variables explicativas se han incluido en la regresión

- Normalmente pasa que más de un modelo de probabilidad puede proporcionar un ajuste adecuado a los datos en el problema científico. La pregunta básica del análisis de sensibilidad es cuánto cambian las inferencias posteriores al utilizar diferentes modelos de probabilidad razonables

    - Otros modelos razonables pueden diferir sustancialmente del modelo actual en la especificación *a priori*, la distribución muestral o en que información se incluye (variables predictivas en una regresión, por ejemplo)

        - Es posible que el modelo actual proporcione un ajuste adecuado a los datos, pero que las inferencias posteriores difieran bajo modelos alternativos posibles

    - En teoría, tanto la comprobación del modelo y el análisis de sensibilidad se pueden incorporar en el análisis *a priori* a posterior. Bajo esta perspectiva, la comprobación del modelo se realiza al construir una distribución conjunta comprensiva, tal que cualquier dato que se pueda observar estén en el espacio muestral de la distribución

        - Por lo tanto, la distribución conjunta es una mezcla de todos los posibles modelos verdaderos o realidades, incorporando toda la información substantiva conocida. La distribución *a priori* en este caso incorpora las creencias *a priori* sobre la verosimilitud de las diferentes realidades y sobre los parámetros de los modelos

        - La distribución posterior de ese modelo exhaustivo de probabilidad automáticamente incorpora todo el análisis de sensibilidad, pero aún se apoya en la verdad de algún miembro dentro una clase de modelos más grande

    - En la práctica, no obstante, construir un supermodelo así para incluir todas las posibilidades y todo el conocimiento substantivo es conceptualmente imposible e inviable computacionalmente (excepto para los problemas más sencillos)

        - Por lo tanto, es necesario examinar los modelos de otras maneras y ver como fallan al ajustarse a la realidad y que tan sensibles son las distribuciones posteriores a especificaciones arbitrarias

- Otro elemento que se tiene que comprobar es si las inferencias posteriores obtenidas tienen sentido o no

    - En cualquier problema aplicado, habrá conocimiento que no se incluye formalmente en la distribución *a priori* o la verosimilitud, por razones de conveniencia u objetividad

        - Si la información adicional sugiere que las inferencias posteriores de interés son falsas, entonces hay potencial para crear un modelo de probabilidad más preciso

    - De manera más formal, se puede comprobar el modelo por validación externa, usando el modelo para hacer predicciones sobre los datos futuros y entonces recoger estos datos y compararlos con las predicciones

        - Las medias posteriores deberían ser correctas de media, los intervalos de credibilidad del 50% deberían contener los valores reales el 50% del tiempo, etc.

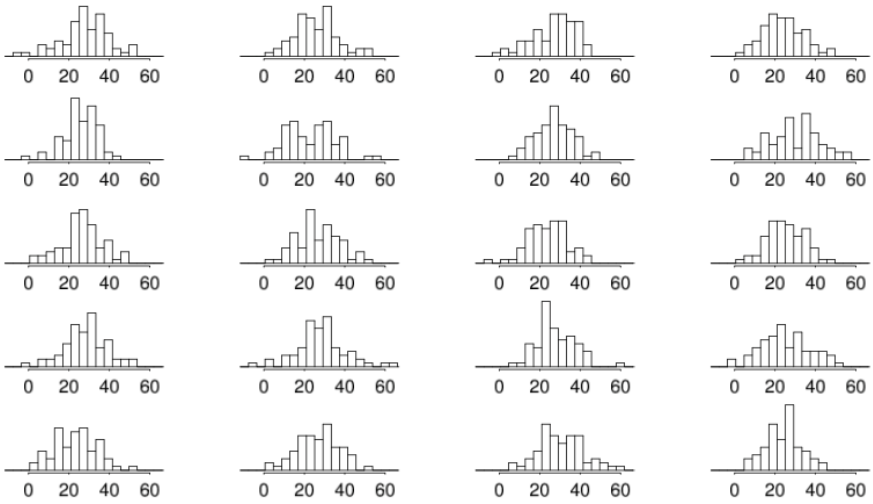
        - Normalmente se necesitará comprobar el modelo antes de obtener nuevos datos o esperar a que pase el futuro. Existen métodos que pueden aproximar la validación externa usando datos disponibles

    - Un solo modelo se puede usar para hacer diferentes predicciones, y hay contextos en los que se tienen diferentes elecciones para definir el foco de las predicciones

        - Para ellos se utiliza la comprobación de la distribución posterior predictiva, que utiliza resúmenes globales para comprobar la distribución posterior predictiva conjunta $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$

- Si el modelo ajusta, entonces réplicas de datos generadas bajo el modelo tendrían que ser similares a los datos observados (los datos deberían verse plausibles bajo la distribución posterior predictiva), lo cual es una comprobación de consistencia propia: una discrepancia se puede deber a un mal ajuste o por azar

    - La técnica básica para comprobar el ajuste del modelo a los datos es sacar valores simulados de la distribución posterior predictiva conjunta de datos replicados y comparar estas muestras con los datos observados



- Cualquier diferencia sistemática entre las simulaciones y los datos indica fallos potenciales del modelo

- Siendo $\boldsymbol{y}$ los datos observados y $\boldsymbol{\theta}$ el vector de parámetros (incluyendo todos los hiperparámetros si el modelo es jerárquico), se definen las replicas $\boldsymbol{y}^{rep}$ como los datos que podrían haberse observado o que se observarían en una réplica del experimento que ha producido $\boldsymbol{y}$ pero en el futuro

  - Se distingue entre $\boldsymbol{y}^{rep}$ y $\tilde{\boldsymbol{y}}$ porque $\tilde{\boldsymbol{y}}$ es cualquier vector futuro observable, mientras que $\boldsymbol{y}^{rep}$ es específicamente una réplica como $\boldsymbol{y}$ (realizaciones). Por ejemplo, si el modelo tiene variables explicativas $\boldsymbol{x}$, estas serán idénticas para $\boldsymbol{y}^{rep}$ y para $\boldsymbol{y}$, pero $\tilde{\boldsymbol{y}}$ podría tener unas variables explicativas diferentes $\tilde{\boldsymbol{x}}$

  - Se trabajará con la distribución de $\boldsymbol{y}^{rep}$ dado el estado actual del conocimiento, lo que quiere decir que se trabaja con la siguiente distribución predictiva posterior:

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta}$$

- La discrepancia entre el modelo y los datos se mide al definir las cantidades de contraste, que son los aspectos de los datos que se quieren comprobar

  - Una cantidad de contraste o medida de discrepancia $T(\boldsymbol{y}, \boldsymbol{\theta})$ es un resumen escalar de parámetros y datos que se usa como un estándar cuando se comparan los datos a las simulaciones predictivas

- Las cantidades de contraste juegan el mismo rol que los estadísticos de contraste frecuentistas pero en la comprobación del modelo bayesiano

- Se utiliza la notación $T(\boldsymbol{y})$ para un estadístico de contraste, que es una cantidad de contraste que depende solo de los datos. En el contexto bayesiano, se pueden generalizar los estadísticos de contraste para permitir dependencia a los parámetros del modelo bajo la distribución posterior

- Las cantidades de contraste en esta sección suelen ser funciones de los datos $\boldsymbol{y}$ o de los datos replicados $\boldsymbol{y}^{rep}$, aunque se pueden calibrar funciones de $\boldsymbol{y}$ y de $\boldsymbol{y}^{rep}$ a la vez

- La falta de ajuste de los datos con respecto a la distribución posterior predictiva puede medirse por la probabilidad del área de cola, o *p-value*, de la cantidad de contraste, y calculado utilizando las simulaciones posteriores $(\boldsymbol{\theta}, \boldsymbol{y}^{rep})$

  - El *p-value* clásico para el estadístico de contraste $T(\boldsymbol{y})$ se define de la siguiente manera, en donde la probabilidad se toma sobre la distribución de $\boldsymbol{y}^{rep}$ con un $\boldsymbol{\theta}$ fijo (la distribución de $\boldsymbol{y}^{rep}$ dado $\boldsymbol{y}$ y $\boldsymbol{\theta}$ es lo mismo que su distribución dado $\boldsymbol{\theta}$):

$$p_C = P(T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})|\boldsymbol{\theta}, \boldsymbol{y}) = P(T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})|\boldsymbol{\theta})$$

  - Los estadísticos de contraste normalmente se derivan de muchas maneras, pero generalmente representan una medida resumen de discrepancia entre los datos observados y lo que sería esperado bajo un modelo con un valor particular de $\boldsymbol{\theta}$. Este valor puede corresponder al de la hipótesis nula o a un estimador puntual máximo verosímil (en este caso, un estimador puntual se tiene que sustituir por $\boldsymbol{\theta}$ para calcular un *p-value*)

- Para evaluar el ajuste de la distribución posterior en el modelo bayesiano, se comparan los datos observados con la distribución posterior predictiva

  - En el enfoque bayesiano, las cantidades de contraste pueden ser funciones de los parámetros desconocidos o pueden ser datos, ya que la cantidad de contraste se evalúa sobre observaciones sacadas de la distribución posterior de los parámetros desconocidos

  - El *p-value* bayesiano se define como la probabilidad de que los datos replicados sean más extremos que los datos observados, medidos por la cantidad de contraste. Este tiene la siguiente

forma, donde la probabilidad se toma sobre la distribución posterior de $\boldsymbol{\theta}$ y la distribución posterior predictiva de $\boldsymbol{y}^{rep}$ (la distribución conjunta $p(\boldsymbol{y}^{rep}, \boldsymbol{\theta}|\boldsymbol{y})$):

$$p_B = P(T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}) =$$

$$= \int_{\Theta} \left[\int_{\Omega} I_{\{T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})\}} p(\boldsymbol{y}^{rep}|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{y}^{rep} \right] d\boldsymbol{\theta}$$

$$= \int_{\Theta} \left[\int_{\Omega} I_{\{T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})\}} p(\boldsymbol{y}^{rep}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{y}^{rep} \right] d\boldsymbol{\theta}$$

- En la práctica, normalmente se calcula la distribución posterior predictiva utilizando simulación

  - Si ya se tienen M simulaciones de la densidad posterior de $\boldsymbol{\theta}$, solo se necesita sacar una $\boldsymbol{y}^{rep}$ de la distribución predictiva para cada $\boldsymbol{\theta}^{(j)}$ simulada, teniendo así M réplicas de la distribución posterior conjunta $p(\boldsymbol{y}^{rep}, \boldsymbol{\theta}|\boldsymbol{y})$

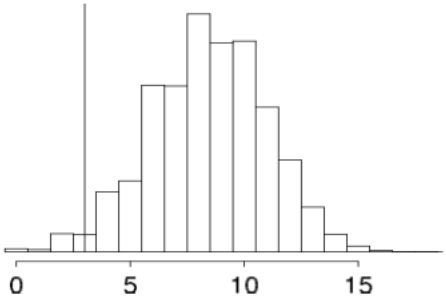
  - La comprobación posterior predictiva es la comparaxión de las cantidades de comprobación $T\left(\boldsymbol{y}, \boldsymbol{\theta}^{(j)}\right)$ y las cantidades de comprobación predictivas $T\left(\boldsymbol{y}^{rep}, \boldsymbol{\theta}^{(j)}\right)$. El *p-value* estimado será la proporción de las M simulaciones para las cuales la cantidad de comprobación iguala o excede la de su valor realizado $T\left(\boldsymbol{y}^{rep}, \boldsymbol{\theta}^{(j)}\right) \geq T\left(\boldsymbol{y}, \boldsymbol{\theta}^{(j)}\right)$ para $j = 1,2, \dots, M$

$$p_B = \frac{\sum_{j=1}^{M} I_{\{T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})\}}}{M}$$

  - A diferencia del enfoque clásico, la comprobación de la modelo bayesiana no requiere métodos especiales para manejar los parámetros molestos: usando simulaciones posteriores, implícitamente se promedia sobre todos los parámetros del modelo

- El procedimiento para llevar a cabo una comprobación del modelo posterior predictivo requiere especificar una cantidad $T(\boldsymbol{y})$ o $T(\boldsymbol{y}, \boldsymbol{\theta})$ y una distribución predictiva apropiada para las réplicas $\boldsymbol{y}^{rep}$

  - Si $T(\boldsymbol{y})$ no parece ser consistente con el conjunto de valores $T(\boldsymbol{y}^{rep\,1}), T(\boldsymbol{y}^{rep\,2}), \dots, T(\boldsymbol{y}^{rep\,M})$, entonces el modelo está haciendo predicciones que no se ajustan a los datos. La discrepancia entre $T(\boldsymbol{y})$ y la distribución de $T(\boldsymbol{y}^{rep})$ puede estar resumida por un *p-value*, pero se prefiere mirar la magnitud de la discrepancia y su *p-value* a la vez (en qué lugar cae dentro de la distribución)



- Para muchos problemas, una función de los datos y de los parámetros pueden enfocarse en un problema particular de un modelo de una manera que sería difícil usando una función únicamente de los datos. Si la cantidad de comprobación depende de $\boldsymbol{\theta}$ y de $\boldsymbol{y}$, entonces $T(\boldsymbol{y}, \boldsymbol{\theta})$ y su réplica $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ son desconocidas y se representan por M simulaciones, y la comparación se puede representar con un diagrama de dispersión de los valores $T(\boldsymbol{y}, \boldsymbol{\theta})$ y $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$, que debería ser simétrico en la línea de 45º, o con un histograma de las diferencias $T(\boldsymbol{y}, \boldsymbol{\theta}) - T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$, el cual debe incluir el cero

- Debido a que un modelo de probabilidad puede fallar en reflejar el proceso que generó los datos de varias maneras, los *p-values* posteriores predictivos se pueden calcular para una variedad de cantidades de comprobación para evaluar más de un fallo posible del modelo

- Idealmente, las cantidades T se escogerán para reflejar aspectos que son relevantes para los propósitos científicos. Estas cantidades suelen reflejar aspectos que no cubre el modelo estadístico (por ejemplo, los rangos de la muestra o las correlaciones de los residuos entre algunas variables explicativas)

  - La comprobación posterior predictiva es una manera directa útil de evaluar el ajuste del modelo a estos aspectos de los datos. El objetivo aquí es explorar más maneras en que cualquiera de los modelos flaquea

    - Las cantidades de comprobación numéricas pueden construirse de patrones visuales detectados. Esto puede ser útil para cuantificar el patrón de potencial interés o para resumir la comprobación del modelo que se va a realizar repetidamente

La computación bayesiana

- La computación bayesiana se basa en dos pasos: el calculo de la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ y el cálculo de la distribución predictiva posterior $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$.

Antes se han visto ejemplos que se pueden calcular analíticamente, pero esto no siempre es posible

- Para modelos complicados, inusuales o de altas dimensiones, es necesario utilizar algoritmos más elaborados para aproximar la distribución posterior
  - Para las distribuciones estándar, las simulaciones se realizan usando una combinación de rutinas preprogramadas y tablas de cálculo numérico
  - Normalmente, la computación más eficiente se puede obtener combinando diferentes algoritmos
- Uno se refiere a la distribución que se quiere simular como a la distribución objetivo y se denota como $p(\boldsymbol{\theta}|\boldsymbol{y})$
  - Se asume que $p(\boldsymbol{\theta}|\boldsymbol{y})$ puede calcularse fácilmente para cualquier valor $\boldsymbol{\theta}$, hasta un factor involucrando solo los datos $\boldsymbol{y}$, por lo que se asume que hay una función fácilmente calculable $q(\boldsymbol{\theta}|\boldsymbol{y})$ (que es una densidad no normalizada) para la cual $q(\boldsymbol{\theta}|\boldsymbol{y})/p(\boldsymbol{\theta}|\boldsymbol{y})$ es una constante que depende solo de $\boldsymbol{y}$
- Para poder evitar desbordamientos computacionales, uno debe calcular con los logaritmos de las densidad posteriores cuando sea posible
  - La exponenciación solo se debe realizar cuando sea necesario y tan tarde como sea posible

- La integración numérica, también llamada cuadratura, se refiere a los métodos en los que la integral sobre una función continua es evaluada calculando el valor de la función en un número finito de puntos
  - Incrementando el número de puntos en donde la función es evaluada, se puede obtener una precisión deseada para los cálculos
    - Los métodos de integración numéricos se pueden dividir métodos de simulación, como el método de Monte Carlo, y en métodos deterministas, como las reglas de cuadratura
  - La esperanza posterior de cualquier función $h(\boldsymbol{\theta})$ se define como $E[h(\boldsymbol{\theta})|\boldsymbol{y}] = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$, en donde la integral tiene tantas dimensiones como $\boldsymbol{\theta}$
    - De manera conversa, se puede expresar cualquier integral sobre el espacio de $\boldsymbol{\theta}$ como una esperanza posterior al definir $h(\boldsymbol{\theta})$ apropiadamente

- Si se tiene una muestra de M valores de $\boldsymbol{\theta}^{(j)}$ de $p(\boldsymbol{\theta}|\boldsymbol{y})$, se puede estimar la integral a través de la media muestral $\sum_{j=1}^{M} h\left(\boldsymbol{\theta}^{(j)}\right)/M$

- Para cualquier número finito de $\boldsymbol{\theta}^{(j)}$, la precisión del estimador se puede medir aproximadamente a través de la desviación estándar

- Si no es fácil simular valores de la distribución posterior o si $h\left(\boldsymbol{\theta}^{(j)}\right)$ es muy variable, entonces se necesitan más métodos de muestreo

- Los métodos de simulación o estocásticos están basados en obtener muestras aleatorias $\boldsymbol{\theta}^{(j)}$ de la distribución deseada $p(\boldsymbol{\theta})$ y estimar la esperanza de cualquier función $h(\boldsymbol{\theta})$

$$E[h(\boldsymbol{\theta})|\boldsymbol{y}] = \int_{\Theta} h(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \approx \frac{1}{M}\sum_{j=1}^{M} h\left(\boldsymbol{\theta}^{(j)}\right)$$

  - La estimación es estocástica dependiendo de números generados aleatoriamente, pero la precisión de la simulación se puede mejorar obteniendo más muestras

  - Los métodos básicos de Monte Carlo, que producen muestras independientes, y los métodos de Monte Carlo con cadenas de Markov, que pueden adaptarse mejor a distribuciones complejas con muchas dimensiones (pero producen muestras dependientes)

  - Los métodos de Monte Carlo con cadenas de Markov han sido importantes para hacer que la inferencia bayesiana sea práctica para modelos jerárquicos

  - Los métodos de simulaciones se pueden usar para distribuciones de muchas dimensiones, y hay algoritmos generales para trabajar con una gran variedad de modelos. Cuando es necesario, un cálculo más eficiente se puede obtener combinando ideas generales con métodos de simulación, métodos deterministas y aproximaciones distribucionales

- Los métodos de integración numérica deterministas se basan en evaluar el integrando $h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})$ en puntos seleccionados $\boldsymbol{\theta}^{(j)}$ basada en una versión ponderada del promedio muestral

$$E[h(\boldsymbol{\theta})|\boldsymbol{y}] = \int_{\Theta} h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \approx \frac{1}{M}\sum_{j=1}^{M} w_j h\left(\boldsymbol{\theta}^{(j)}\right)p\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{y}\right)$$

- La ponderación w_j corresponde al volumen de espacio representado por el punto $\boldsymbol{\theta}^{(j)}$. Reglas más elaboradas, tales como las de Simpson, usan polinomios locales para más precisión

- Las reglas de integración numéricas deterministas típicamente tienen menor varianza que los métodos de simulación, pero la selección de localizaciones es más difícil en altas dimensiones

- Los métodos de tablas pueden hacerse adaptables empezando la formación de la tabla a partir de la moda posterior

- Los métodos de cuadratura existen para regiones acotadas y no acotadas

La simulación de cadenas de Markov

- La simulación de cadenas de Markov o método de Monte Carlo con cadenas de Markov (MCMC) es un método general basado en obtener valores de $\boldsymbol{\theta}$ de distribuciones aproximadas y corregir estas para obtener una mejor aproximación de la distribución posterior objetivo $p(\boldsymbol{\theta}|\boldsymbol{y})$

    - La simulación de cadenas de Markov se usa cuando no es posible (o no es computacionalmente eficiente) muestrear $\boldsymbol{\theta}$ directamente de $p(\boldsymbol{\theta}|\boldsymbol{y})$

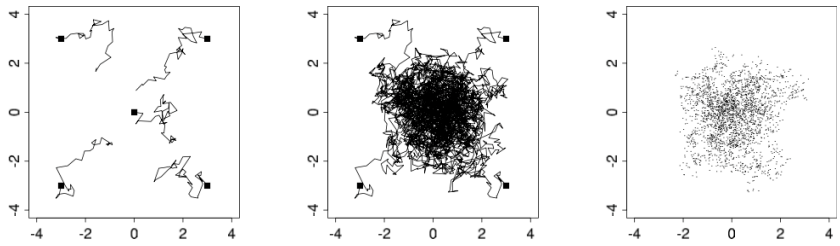
        - Por lo tanto, se muestrean iterativamente de manera que en cada paso del proceso se espera obtener una simulación de una distribución que se vuelve cada vez más cercana a $p(\boldsymbol{\theta}|\boldsymbol{y})$

        - Para una amplia clase de problemas (incluidos los modelos jerárquicos) esta es la manera más fácil de obtener resultados fiables

    - El muestreo se realiza secuencialmente, con la distribución de los valores simulados dependiendo del último valor obtenido (las muestras de una cadena de Markov, ya que dependen de su último valor)

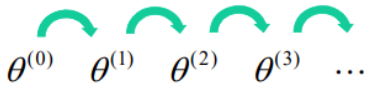
        - La clave para el éxito de estos métodos no es la propiedad de Markov, sino que las distribuciones aproximadas se pueden mejorar en cada paso en la simulación, en el sentido de que pueden converger a la distribución objetivo

- De este modo, se puede visualizar este método como un método que crea un número de cadenas de Markov y que, a lo largo del camino de cada cadena, se puede ver como estas convergen a una distribución estacionaria (la objetivo). Una vez convergen, se usan solo las observaciones de aquella mitad que converge a esta distribución y se realizan las inferencias simulando puntos



- Un usuario solo se interesa por las simulaciones obtenidas una vez las cadenas han convergido a la distribución objetivo, dado que son las de la distribución posterior con la que se hará la inferencia

- En las aplicaciones de la simulación de cadenas de Markov, se crean varias secuencias independientes, en donde cada secuencia $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ se produce empezando por un punto inicial $\boldsymbol{\theta}^{(0)}$



  - Entonces, para cualquier j, se saca $\boldsymbol{\theta}^{(j)}$ de una distribución de transición $T_j\left(\boldsymbol{\theta}^{(j)}\middle|\boldsymbol{\theta}^{(j-1)}\right)$ que depende solo del valor simulado anterior $\boldsymbol{\theta}^{(j-1)}$

  - Normalmente es conveniente dejar que la distribución de transición depende del número de iteración (j)

  - Las distribuciones de probabilidad de transición se deben de construir tal que la cadena de Markov converja a una única distribución estacionaria (la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$)

- La clave para la simulación de cadenas de Markov es crear un proceso de Markov cuya distribución estacionaria se $p(\boldsymbol{\theta}|\boldsymbol{y})$ y ejecutar la simulación durante mucho tiempo para que la distribución de las simulaciones actuales sea lo más cercana posible a esta distribución estacionaria

- Para cualquier $p(\boldsymbol{\theta}|\boldsymbol{y})$ específica o densidad no normalizada $q(\boldsymbol{\theta}|\boldsymbol{y})$ hay una variedad de cadenas de Markov que se pueden construir

- Cuando el algoritmo de simulación se ha implementado y las simulaciones se han obtenido, es absolutamente necesario comprobar la convergencia de las secuencias simuladas

- Un algoritmo particularmente útil en muchos problemas multidimensionales es el muestreador de Gibbs o *Gibbs sampler*, que se define en términos de subvectores de $\boldsymbol{\theta}$

  - Suponiendo que el vector de parámetros $\boldsymbol{\theta}$ se ha dividio en d compoenntes o subvectores $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$, cada iteración del muestreador de Gibbs hace un ciclo a través de los subvectores de $\boldsymbol{\theta}$, obteniendo una simulación de cada subconjunto condicional al valor de todos los otros

    - Por lo tanto, hay d pasos en cada iteración j (en cada observación extraída de la simulación) después de escoger los valores iniciales $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)}$

    - En cada iteración j, una ordenación de los d subvectores de $\boldsymbol{\theta}$ se escoge y, por lo tanto, en cada $\theta_i^{(j)}$ para $i = 1,2, \dots, d$ se muestrea de la distribución condicional dados todos los otros componentes de $\boldsymbol{\theta}$

$$p\left(\theta_i \middle| \boldsymbol{\theta}_{-i}^{(j-1)}, \boldsymbol{y}\right)$$

$$where\ \boldsymbol{\theta}_{-i}^{(j-1)} = \left(\theta_1^{(j)}, \dots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \dots, \theta_d^{(j-1)}\right)$$

  - Por lo tanto, cada subvector θ_i se actualiza condicionado a los últimos valores de los otros componentes de $\boldsymbol{\theta}$, que son los valores en la iteración j para los componentes que ya están actualizados y los valores en la iteración $j - 1$ para los otros

$$Simulate\ \theta_1^{(j)}\ from\ p\left(\theta_1 \middle| \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, \boldsymbol{y}\right)$$

$$Simulate\ \theta_2^{(j)}\ from\ p\left(\theta_2 \middle| \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}, \boldsymbol{y}\right)$$

$$\dots$$

$$Simulate\ \theta_d^{(j)}\ from\ p\left(\theta_d \middle| \theta_1^{(j)}, \theta_3^{(j)}, \dots, \theta_{d-1}^{(j)}, \boldsymbol{y}\right)$$

        - Los componentes se van actualizando secuencialmente debido a que los d subvectores están ordenados, por lo que una vez se actualiza el primero, el siguiente tiene que tener en cuenta esta actualización

    - Para muchos problemas que involucran modelos estadísticos estándar, es posible muestrear directamente de la mayoría de distribuciones condicionales posteriores de los parámetros

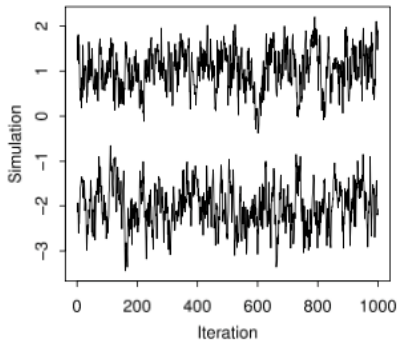
        - Normalmente se construyen modelos usando una secuencia de distribuciones condicionales, igual que con los modelos jerárquicos

        - Suele pasar que las distribuciones condicionales en estos modelos son distribuciones conjugadas que proporcionan una simulación fácil

- El método básico de inferencia de una simulación iterativa es el mismo que para una simulación bayesiana en general: usar la colección de valores simulados para hacer inferencia. No obstante, hay que vigilar el uso de la simulación iterativa y comprobar los resultados

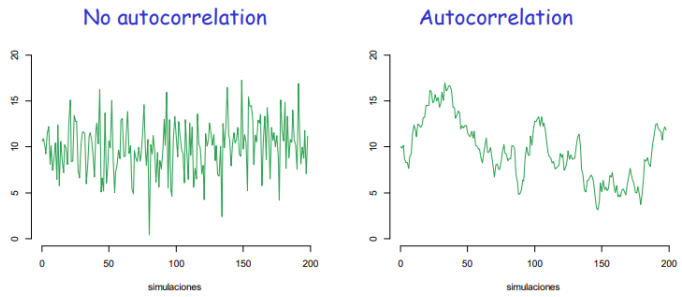
    - La simulación iterativa añade dos problemas a la inferencia de las simulaciones: las iteraciones pueden no haber procedido lo suficiente y las simulaciones están correlacionadas

        - Si las iteraciones no han procedido lo suficiente, estas pueden no ser representativas de la distribución objetivo. Aunque las simulaciones hayan conseguido una convergencia aproximada, las iteraciones anteriores aún reflejarán la aproximación inicial y no la de la distribución objetivo



        - Aparte de problemas de convergencia, la inferencia de simulaciones correlacionadas es generalmente menos precisa que si fueran independientes. La correlación serial en las simulaciones no es necesariamente un problema porque, cuando convergen, se distribuyen idénticamente y son

independientes y se pueden ignorar estos problemas al inferenciar



- No obstante, esta correlación puede causar ineficiencias en la simulación, dado que se necesitarán simular más valores para aproximar adecuadamente la distribución posterior

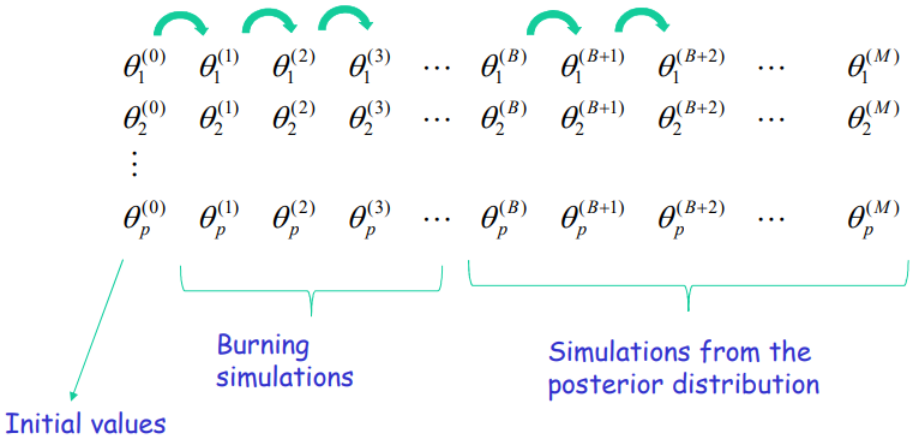
- Estos problemas de la simulación iterativa se pueden gestionar de tres maneras diferentes

  - La primera es intentar diseñar las ejecuciones de la simulación para permitir una monitorización efectiva de la convergencia, normalmente simulando múltiples secuencias con puntos iniciales esparcidos por el espacio paramétrico

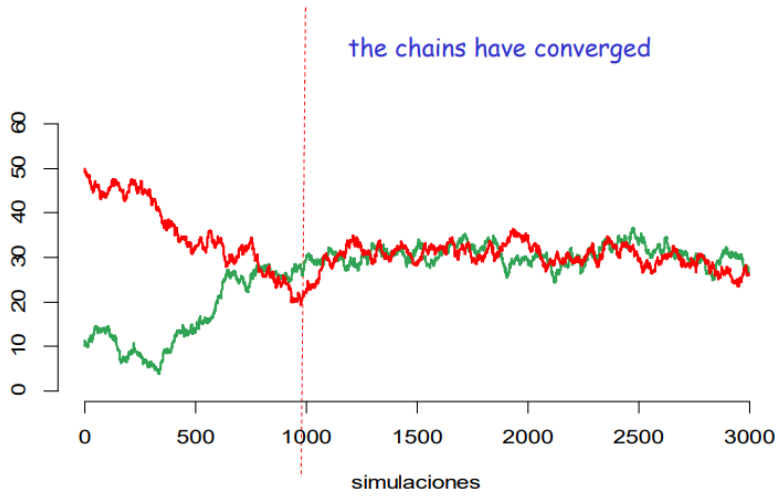
  - La segunda es monitorizar la convergencia de todas las cantidades de interés comparando la variación entre y dentro de las secuencias simuladas hasta que la variación dentro de cada secuencia sea más o menos la misma que entre secuencias. Solo se puede decir que se aproxima la distribución objetivo cuando la distribución de cada secuencia simulada está cerca a la de todas las secuencias mezcladas

  - La tercera es alterar el algoritmo para mejorar la eficiencia de la simulación

- Para disminuir la influencia de los valores iniciales, normalmente se descarta la primera mitad de cada secuencia y se presta atención a la segunda mitad



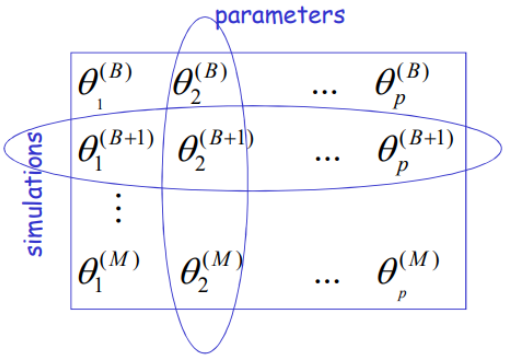
- Las inferencias se basarán en la suposición de que las distribuciones de los valores simulados $\boldsymbol{\theta}^{(j)}$ para una j lo bastante grande (denotada por B normalmente) se acercan a la distribución objetivo



B

- A estas simulaciones que se eliminan se les llama simulaciones de quema o *burning simulations* (antes de la convergencia), y la práctica de descartar estas se llama *warm-up* o calentamiento. Dependiendo del contexto, diferentes fracciones de calentamiento pueden ser apropiadas

- Cuando se simulan vectores de parámetros, se puede crear una tabla orientativa en la que cada fila es una simulación de la distribución posterior conjunta y cada columna representa una de las cadenas de Markov simuladas



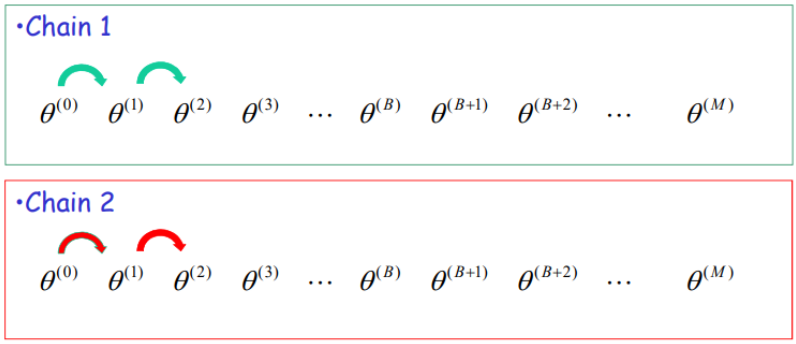
The rows are simulations from the posterior joint distribution

$$\pi(\theta_1, \theta_2, ..., \theta_p \mid y)$$

The columns are simulations from the posterior marginal distribution.

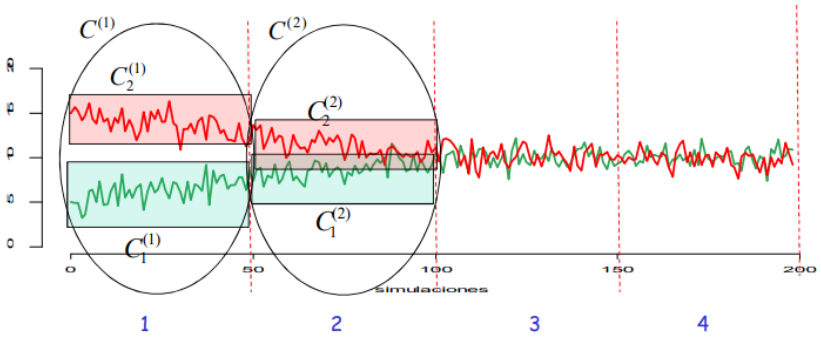
$$\pi(\theta_2 \mid y) = \int_{\Omega_p} ... \int_{\Omega_3 \Omega_1} \pi(\theta_1, ..., \theta_p \mid y) d\theta_1 d\theta_3 ... d\theta_p$$

- Otro problema que a veces surge, una vez la convergencia aproximada se ha obtenido, es si afinar las secuencias manteniendo fuera cada simulación número k de cada secuencia y descartando el resto
  - En las aplicaciones, se ha encontrado útil pasar de las iteraciones en problemas con grandes números de parámetros en donde la capacidad de almacenamiento del ordenador es un problema
  - Aunque las secuencias sean o no afinadas, si las secuencias han conseguido una convergencia aproximada, se pueden usar directamente para inferenciar
- El enfoque más recomendado para hacer inferencia de la simulación iterativa es basándose en la comparación de diferentes secuencias simuladas



  - Para poder ver esas disparidades, es necesario tener más de una secuencia independiente. Por lo tanto, el plan es simular independientemente al menos dos secuencias, con puntos iniciales obtenidos de una distribución sobredispersa

- Se diagnostica convergencia al comprobar la mezcla y la estacionariedad. Hay varias maneras de realizar esto, pero una manera simple es dividiendo cada cadena por la mitad y comprobando que todos las secuencias se han mezclado



  - Esto comprueba simultáneamente la mezcla (si todas las cadenas se han mezclado bien, las partes separadas de las cadenas deberían mezclarse también) y la estacionariedad (tanto la primera como la segunda mitad deberían aproximar la misma distribución)

  - Se comienza por un número de secuencias simuladas en las que el periodo de calentamiento ya se ha descartado, y después se toma cada una de estas cadenas y se divide en dos mitades (después de descartar las simulaciones de quema)

  - Siendo N el número de cadenas y M el número de simulaciones de cada cadena, N siempre tiene que ser al menos cuatro (porque se tienen que mezclar dos secuencias)

- Para cada estimando escalar ψ, se etiqeutan las simulaciones como ψ_{ij} para $i = 1,2,\dots,M$ y $j = 1,2,\dots,N$ y se calcula la varianza entre secuencias B y la varianza dentro de las secuencias W, definidas de la siguiente manera:

$$B = \frac{M}{N-1}\sum_{j=1}^{N}\left(\bar{\psi}_j - \bar{\bar{\psi}}\right)^2 \quad where \quad \bar{\psi}_j = \frac{1}{M}\sum_{i=1}^{M}\psi_{ij} \quad \& \quad \bar{\bar{\psi}} = \frac{1}{N}\sum_{j=1}^{N}\bar{\psi}_j$$

$$W = \frac{1}{N}\sum_{j=1}^{N}s_j^2 \quad where \quad s_j^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(\psi_{ij} - \bar{\psi}_j\right)^2$$

  - Se puede estimar la varianza posterior marginal del estimando $Var(\psi|\boldsymbol{y})$ como una media ponderada entre B y W

$$\widehat{Var}^{+}(\psi|\boldsymbol{y}) = \frac{M-1}{M}W + \frac{1}{M}B$$

- Esta cantidad sobreestima la varianza posterior marginal asumiendo que la distribución inicial esté apropiadamente sobredispersa, pero no está sesgado bajo estacionariedad (cuando la distribución inicial equivale a la posterior) o en el límite $M \rightarrow \infty$

- Para cualquier cantidad finita M, la varianza dentro de cada secuencia W debería subestimar la varianza $Var(\psi|\boldsymbol{y})$ porque las secuencias individuales no han tenido tiempo para cubrir toda la distribución objetivo y, en consecuencia, tienen menos variabilidad. Cuando $M \rightarrow \infty$, la esperanza de W se aproxima a $Var(\psi|\boldsymbol{y})$

- Se monitoriza la convergencia de la simulación iterativa estimando un factor por el cual la escala de la distribución corriente para ψ puede ser reducida si las simulaciones se continuaran para $M \rightarrow \infty$

  - Esta reducción de escala es estimada por el índice $\hat{R}$ o *r-hat*, el cual se inspira en el análisis de la varianza y disminuye a 1 cuando $M \rightarrow \infty$

$$\hat{R} = \sqrt{\frac{\widehat{Var}^{+}(\psi|\boldsymbol{y})}{W}}$$

  - Si la reducción de escala potencial es alta, entonces se tienen razones para creer que proceder con más simulaciones puede mejorar la inferencia sobre la distribución objetivo del estimando escalar asociado

- El método de monitorización de convergencia presentado tiene la ventaja de que no necesita examinar gráficos de series temporales de las secuencias simuladas

  - La inspección de esos gráficos es un método poco fiable para evaluar la convergencia y no es factible si se quiere monitorizar un gran número de cantidades de interés, como en los modelos jerárquicos

  - Debido a que está basado en medias y varianzas, el método es más efectivo para cantidades cuyas distribuciones marginales posteriores sean aproximadamente normales

  - Cuando se realiza inferencia para los cuartiles extremos o para parámetros con distribuciones marginales posteriores

multimodales, uno debe también monitorizar los cuartiles extremos de las secuencias entre y dentro de

La computación en R y Stan

- La computación bayesiana moderna normalmente utiliza lenguajes de programación especializados y lenguajes estadísticos como R
    - El lenguaje Stan es un lenguaje de programación de alto nivel en el que el usuario especifica el modelo y tiene la opción de proporcionar valores iniciales, y después se implementa automáticamente una simulación de cadenas de Markov para la distribución posterior resultante
        - Es posible construir y ajustar modelos enteramente dentro de Stan, pero en la práctica es casi siempre necesario procesar los datos antes de introducirlos en el modelo, y procesar las inferencias después de que el modelo se haya ajustado
        - Por lo tanto, se suele ejecutar Stan llamándolo directamente desde R a través de la función $stan()$
    - Cuando se trabaja en R y Stan, es recomendable utilizar RStudio, dado que permite trabajar con cuatro ventanas de manera flexible
        - Se pueden utilizar otras alternativas mientras se pueda ver la consola de R, los gráficos, un editor de texto para los *scripts* de R y un editor de texto para los *scripts* de Stan
- En esta sección se describen los pasos que se tienen que seguir con tal de poder realizar inferencias con la distribución posterior a través de funciones básicas o de paquetes de R (sin usar Stan)
    - El paquete *dplyr* es un paquete de *tidyverse* el cual permite trabajar con datos de manera más sencilla. Los elementos principales con los que se trabaja son los *tibbles*, los verbos de *dplyr* y el operador “>%>”
        - Un *tibble* es un tipo de *data frame* que tiene propiedades más útiles a la hora de trabajar con *dplyr*. Para poder hacer una selección de variable, se debe usar $[[x]]$ o $

```
                     Company Revenue Activity     City
1     ACCIONA CONSTRUCCION SA.    1527     4121   Madrid
2          ACERINOX EUROPA SAU    1483     2410    Cadiz
3              AENA S.M.E. SA.    3755     5223   Madrid
4  AIR EUROPA LINEAS AEREAS SA    1934     5110 Baleares
5 AIRBUS DEFENCE AND SPACE SA.    2988     3030   Madrid
6           AIRBUS MILITARY SL    1681     3030   Madrid
```

```
df <- as_tibble(df)
df
```

```
# A tibble: 100 x 4
   Company                           Revenue Activity City
   <fct>                               <int>    <int> <fct>
 1 "ACCIONA CONSTRUCCION SA."           1527     4121 Madrid
 2 "ACERINOX EUROPA SAU"                1483     2410 Cadiz
 3 "AENA S.M.E. SA."                    3755     5223 Madrid
 4 "AIR EUROPA LINEAS AEREAS SA"        1934     5110 Baleares
 5 "AIRBUS DEFENCE AND SPACE SA."       2988     3030 Madrid
 6 "AIRBUS MILITARY SL"                 1681     3030 Madrid
 7 "AIRBUS OPERATIONS SL"               1749     3030 Madrid
 8 "ALCAMPO SA"                         3294     4711 Madrid
 9 "ALUMINIO ESPA\x84OL SL"             2761     2442 Madrid
10 "ALVEAN SUGAR SOCIEDAD LIMITADA."    3831     4636 Bizkaia
# … with 90 more rows
```

- La función $filter()$ permite extraer filas que cumplan una o más condiciones lógicas, las cuales se introducen como argumentos a la derecha

```
filter(df, City == "Barcelona", Revenue > 3000)
```

```
# A tibble: 2 x 4
  Company                                      Revenue Activity City
  <fct>                                          <int>    <int> <fct>
1 "SEAT SA"                                       9552     2910 Barcelona
2 "VOLKSWAGEN GROUP ESPA\x84A DISTRIBUCION SA."   3628     4511 Barcelona
```

- La función $arrange()$ permite ordenar filas de valores más pequeños a más altos, escogiendo la variable de ordenación deseada a la derecha como argumento. Para poder invertir el orden de mayor a menor, se utiliza la función $desc()$ sobre la variable

```
arrange(df, Revenue)
```

```
# A tibble: 100 x 4
   Company                                        Revenue Activity City
   <fct>                                            <int>    <int> <fct>
 1 "SAGANE SA"                                       1450     3523 Madrid
 2 "TECH DATA ESPA\x84A SL"                          1450     4651 Madrid
 3 "PRIMARK TIENDAS SLU"                             1472     4751 Madrid
 4 "ACERINOX EUROPA SAU"                             1483     2410 Cadiz
 5 "IBERDROLA COMERCIALIZACION DE ULTIMO RECURSO SOCI…  1516     3514 Bizkaia
 6 "ACCIONA CONSTRUCCION SA."                        1527     4121 Madrid
 7 "TELEFONICA AUDIOVISUAL DIGITAL SL."              1556     6020 Madrid
 8 "PUNTO FA SL"                                     1577     4771 Barcelo…
 9 "IBERDROLA ESPA\x84A SOCIEDAD ANONIMA."           1611     6420 Bizkaia
10 "EDP ESPA\x84A SA."                               1616     3516 Asturias
# … with 90 more rows
```

- La función $select()$ permite extraer columnas por el nombre de las variables, las cuales se introducen como argumentos a la derecha. Para escoger un rango, se utiliza : entre el nombre de dos variables, mientras que para escoger todas menos unas variables concretas, se usa un vector $-c(\dots)$

```
select(df, Company, City)
```

```
# A tibble: 100 x 2
   Company                           City
   <fct>                             <fct>
 1 "ACCIONA CONSTRUCCION SA."        Madrid
 2 "ACERINOX EUROPA SAU"             Cadiz
 3 "AENA S.M.E. SA."                 Madrid
 4 "AIR EUROPA LINEAS AEREAS SA"     Baleares
 5 "AIRBUS DEFENCE AND SPACE SA."    Madrid
 6 "AIRBUS MILITARY SL"              Madrid
 7 "AIRBUS OPERATIONS SL"            Madrid
 8 "ALCAMPO SA"                      Madrid
 9 "ALUMINIO ESPA\x84OL SL"          Madrid
10 "ALVEAN SUGAR SOCIEDAD LIMITADA." Bizkaia
# … with 90 more rows
```

- También es posible especificar con que letra empieza o acaba la columna o las columnas a seleccionar a través de las funciones $starts\_with()$ y $ends\_with()$. Adicionalmente, se pueden seleccionar columnas cuyo nombre contenga unos caracteres a través de la función $contains()$

```
select(df, starts_with("C"))
```

```
# A tibble: 100 x 2
   Company                           City
   <fct>                             <fct>
 1 "ACCIONA CONSTRUCCION SA."        Madrid
 2 "ACERINOX EUROPA SAU"             Cadiz
 3 "AENA S.M.E. SA."                 Madrid
 4 "AIR EUROPA LINEAS AEREAS SA"     Baleares
 5 "AIRBUS DEFENCE AND SPACE SA."    Madrid
 6 "AIRBUS MILITARY SL"              Madrid
 7 "AIRBUS OPERATIONS SL"            Madrid
 8 "ALCAMPO SA"                      Madrid
 9 "ALUMINIO ESPA\x84OL SL"          Madrid
10 "ALVEAN SUGAR SOCIEDAD LIMITADA." Bizkaia
# … with 90 more rows
```

- La función $mutate()$ crea nuevas columnas, de modo que se añade el nombre de la columna y después del símbolo $=$ se introduce una fórmula. Para crear columnas eliminando las otras, se utiliza la función $transmute()$ de la misma manera

```
mutate(df,
  Rev_per_month = Revenue / 12,
  Rev_per_week = Rev_per_month / 4)
```

```
# A tibble: 100 x 6
   Company                  Revenue Activity City    Rev_per_month Rev_per_week
   <fct>                      <int>    <int> <fct>           <dbl>        <dbl>
 1 "ACCIONA CONSTRUCCION SA…   1527     4121 Madrid           127.         31.8
 2 "ACERINOX EUROPA SAU"       1483     2410 Cadiz            124.         30.9
 3 "AENA S.M.E. SA."           3755     5223 Madrid           313.         78.2
 4 "AIR EUROPA LINEAS AEREA…   1934     5110 Balear…          161.         40.3
 5 "AIRBUS DEFENCE AND SPAC…   2988     3030 Madrid           249          62.2
 6 "AIRBUS MILITARY SL"        1681     3030 Madrid           140.         35.0
 7 "AIRBUS OPERATIONS SL"      1749     3030 Madrid           146.         36.4
 8 "ALCAMPO SA"                3294     4711 Madrid           274.         68.6
 9 "ALUMINIO ESPA\x84OL SL"    2761     2442 Madrid           230.         57.5
10 "ALVEAN SUGAR SOCIEDAD L…   3831     4636 Bizkaia          319.         79.8
# … with 90 more rows
```

- La función $summarise()$ toma una base de datos como primer argumento y crea una tabla con estadísticos (nombres y fórmulas se tienen que especificar)

```
summarise(df, n = n(),
  Rev_total = sum(Revenue),
  Rev_mean = mean(Revenue),
  Tot_act = n_distinct(Activity))
```

```
# A tibble: 1 x 4
      n Rev_total Rev_mean Tot_act
  <int>     <int>    <dbl>   <int>
1   100    414449    4144.      55
```

- El operador %>% pasa el resultado en la izquierda como el primer argumento de la función a la derecha

```
df %>%
    summarise(n = n(),
      Rev_total = sum(Revenue),
      Rev_mean = mean(Revenue),
      Tot_act = n_distinct(Activity))
```

```
# A tibble: 1 x 4
      n Rev_total Rev_mean Tot_act
  <int>     <int>    <dbl>   <int>
1   100    414449    4144.      55
```

- La función $group\_by()$ agrupa los casos por valores comunes, de modo que necesita una base de datos como primer argumento y las variables por las que se agrupa como otros argumentos

```
df %>%
    group_by(City) %>%
    summarise(n = n(),
      Rev_total = sum(Revenue),
      Rev_mean = mean(Revenue),
      Tot_act = n_distinct(Activity))
```

```
# A tibble: 18 x 5
   City             n Rev_total Rev_mean Tot_act
   <fct>        <int>     <int>    <dbl>   <int>
 1 "Asturias"       2      5020    2510        2
 2 "Baleares"       2      3668    1834        2
 3 "Barcelona"      7     23096    3299.       6
 4 "Bizkaia"        9     32758    3640.       6
 5 "Cadiz"          1      1483    1483        1
 6 "Coru\x96a"      2     12137    6068.       2
 7 "Granada"        1      2238    2238        1
 8 "Huelva"         1      1789    1789        1
 9 "Lerida"         1      1674    1674        1
10 "Madrid"        63    274168    4352.      44
11 "Malaga"         1      2175    2175        1
12 "Navarra"        1      2525    2525        1
13 "Pontevedra"     1      5047    5047        1
14 "Sevilla"        1      4530    4530        1
15 "Tenerife"       1      1736    1736        1
16 "Valencia"       2     23365   11682.       1
17 "Valladolid"     2     10810    5405        2
18 "Zaragoza"       2      6230    3115        2
```

- Existen muchas más funciones y comandos útiles en el paquete *dplyr*, pero solo se han enumerado las principales con las que se suele trabajar

- Una vez se saben usar los paquetes de R más importantes para manipular datos y utilizar gráficos, se puede comenzar a programar con Stan utilizando R

- Lo primero que es necesario es crear un archivo de texto Stan que tenga como sufijo $.stan$. Dentro de un *script* básico de Stan, se encuentran tres bloques básicos: el bloque de datos, el de parámetros y el del modelo

    - El bloque de datos se usa para declarar todos los datos que se pasarán a Stan, lo que permitirá estimar el modelo. Se debe declarar el tipo de datos o parámetros que se usarán en el modelo, y este tipo no se puede cambiar después

```
data {
  real Y[10]; // heights for 10 people
}
```

    - Algunos ejemplos de tipos de datos son los siguientes:

`real Y[10]` : array with 10 elements
`real<lower=0, upper=1> Z` : continuous variable bounded between 0 and 1
`int<lower=0> Z` : a discrete variable that takes integer values with a minimum value of 0
`vector[N] Z` : a vector of continuous variables of length `N`
`matrix[3, 3] Z` : a 3 x 3 matrix of continuous variables
`matrix[3, 3] Z[5, 2]` : a 5 x 2 array of 3 x 3 matrices

    - En el bloque de parámetros, se declaran todos los parámetros que se inferirán en el modelo (los parámetros de todos los niveles del modelo). Stan no soporta parámetros enteros, pero es posible incluirlos indirectamente al marginalizarlos

```
parameters {
  real mu; // mean height in population
  real<lower=0> sigma; // sd for height distribution
}
```

    - Igual que con el bloque de datos, se pueden utilizar diferentes tipos de datos para los parámetros. Algunos ejemplos son los siguientes:

`simplex[K] Z` : a vector of K non-negative continuous variables whose sum is 1
`corr_matrix[K] Z` : a K x K dimensional correlation matrix
`ordered[K] Z` : a vector of K continuous ordered elements

    - El bloque del modelo, finalmente, se utiliza para especificar la función de verosimilitud y las distribuciones *a priori*

```
model {
  for (i in 1:10) {
    Y[i] ~ normal(mu, sigma); // likelihood
  }

  mu ~ normal(1.5, 0.1);  // prior for mu
  sigma ~ gamma(1, 1);    // prior for sigma
}
```

- Una manera más eficiente y compacta de escribir un *for loop* para las observaciones o los parámetros es utilizar vectores directamente

```
Y ~ normal(mu, sigma); // likelihood
```

- Stan tiene un conjunto de distribuciones de probabilidad útil que soporta. Algunas de las distribuciones más populares son las siguientes:

Discrete: Bernoulli, binomial, Poisson, beta-binomial, negative-binomial, multinomial
Continuous unbounded: normal, Student-t, Cauchy
Continuous bounded: uniform, beta, log-normal, exponential, gamma, chi-squared, Weibull, Pareto
Multivariate continuous: normal, Student-t

- Para ejecutar el programa de Stan desde R, se tienen que seguir los siguientes pasos:

  - El primer paso consiste en seleccionar un directorio de trabajo que contenga tanto el *script* de R como una carpeta con los archivos de Stan

  - Después se tiene que cargar el paquete de *RStan* en R y después configurarlo para que pueda ejecutar las cadenas de Markov en múltiples procesadores en paralelo. Para problemas o más información sobre la instalación de Stan y los compiladores que se usan para el programa, se tiene que visitar la web https://github.com/stan-dev

```
library(rstan)
```

```
options(mc.cores = parallel::detectCores())
```

  - El tercer paso es crear una base de datos y crear una lista de estos datos en la que se use la notación que se pondrá en el bloque de datos y se incluyan los datos de la base

```
# y[i]: the number of deaths performing a specific cardiac surgery in hospital i-th
# n[i]: number of operations in hospital i-th

N <- 12
df <- tibble(
        hosp = LETTERS[1:N],
        y = c(0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24),
        n = c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360)
    )

## Stan model
data_list <- list(
  N = N,
  y = df$y,
  n = df$n,
  a = 1,
  b = 1
  )
```

- El último paso es compilar y ejecutar la simulación de Monte Carlo con cadenas de Markov en el programa Stan a través de la siguiente función:

```
fit <- stan("stan_models/simple.stan", iter = 200, chains = 4,
  data = list=(Y = Y))
```

- Una vez se ha ejecutado la simulación, es posible interpretar los resultados

  - Si se llama al ajuste de la función, se puede obtener la siguiente información

    - Number of chains considered, and how many iterations each
    - Warm-up discarded iterations
    - Total post-warmup draws
    - Summary statistics for the parameters, and for the log probability of the model
    - Convergence diagnostics for each parameter:
    - `n_eff` : number of effective samples
    - `Rhat` : potential scale reduction factor on split chains (at convergence, Rhat=1)

  - A partir del modelo de Stan, se pueden obtener diversas cantidades como cuartiles, las distribuciones posteriores (en una sola matriz), gráficos de autocorrelación y otros gráficos

```
# Interpreting the results  ------------------------------------------------

# Model Summary
print(fit, probs = c(0.1, 0.25, 0.5, 0.75, 0.9, 0.95))

# Posterior predictions
posterior <- as.matrix(fit)
colnames(posterior)

# Autocorrelation plot
acf(posterior[, "mu"])
acf(posterior[, "sigma"])

# Samples vs. iteration plot
traceplot(fit)

# Posterior of prob1
plot(density(posterior[, "mu"]))
plot(density(posterior[, "sigma"]))
```

  - Además, a través del paquete *bayesplot*, es posible dibujar áreas y densidades de las distribuciones posteriores (con la matriz de distribuciones) y gráficos sobre la simulación de cadenas

```r
library(bayesplot)

plot_title <- ggtitle("Posterior distributions of mu and sigma",
                      "with medians and 80% intervals")
mcmc_areas(posterior,
  pars = c("mu", "sigma"),
  prob = 0.8) + plot_title

posterior
mcmc_trace(posterior,
  pars = c("mu", "sigma"))

mcmc_trace(posterior,
           pars = c("mu", "sigma"),
           facet_args = list(nrow = 2))

save(posterior, file=paste("results/simple.RData", sep=""))
```

- También es posible la ejecución de otros bloques a parte de los tres principales

  - El bloque de cantidades generadas sirve principalmente para obtener muestras de la distribución posterior predictiva y hacer comprobaciones de predicción del ajuste del modelo

```
generated quantities {
  vector[10] lSimData;
  int aMax_indicator;
  int aMin_indicator;

  // Generate posterior predictive samples
  for (i in 1:10) {
    lSimData[i] = normal_rng(mu, sigma);
  }

  // Compare with real data
  aMax_indicator = max(lSimData) > max(Y);
  aMin_indicator = min(lSimData) > min(Y);
}
```

  - No obstante, también es muy útil para obtener predicciones a través del modelo estimado por Stan

```
model {
  for (j in 1:J)
    a[j] ~ normal(g_0 + g_1 * u[j], sigma_a);
  for (n in 1:N)
    y[n] ~ normal(a[county[n]] + b * x[n], sigma_y);
}
generated quantities {
  real u_tilde;
  real a_tilde;
  real y_tilde;
  u_tilde = mean(u);
  a_tilde = normal_rng(g_0 + g_1 * u_tilde, sigma_a);
  y_tilde = normal_rng(a_tilde + b * 1, sigma_y);
}
```

  - Los bloques soportan todo lo definido en bloques anteriores. Por lo tanto, el orden de los bloques importa, y el orden típico es el siguiente:

functions: definition - once at the beginning

data: once at the beginning

transformed data: once after the *data block* is executed

parameters: each time the log probability is evaluated

transformed parameters: each time the log probability is evaluated

model: each time the log probability is evaluated

generated quantities: once per sample

- También es posible ejecutar *for loops* y otras proposiciones condicionales como *while loops* en Stan

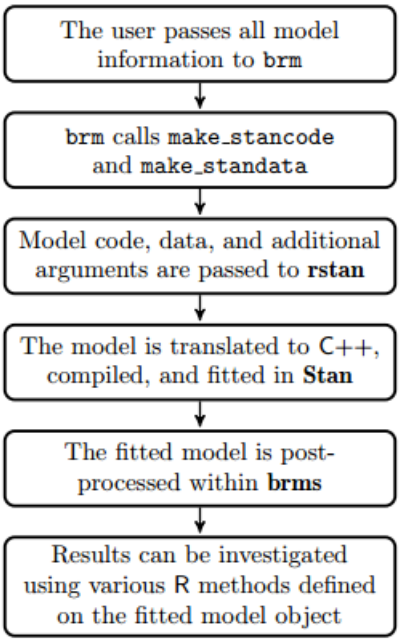
```
for (i in 1:10) {
  execute this;
}
```

```
int i = 0;
while (i < 10) {
  execute this;
  i = i + 1;
}
```

```
if (i < 2) {
  execute this;
} else if (i == 2) {
  execute that;
} else {
  execute other thing;
}
```

  - Todos se definen como se harían en R, solo que necesitarán un punto y coma después de cada línea dentro del *loop*, tal como se ha visto que se necesita para Stan

- El paquete *brms* es un paquete que implementa modelos bayesianos multinivel en R utilizando Stan, y es muy útil para implementar modelos de regresión, aunque también se puede usar para otros modelos más simples



- La función más importante es la función $brm()$, la cual tiene la siguiente sintaxis:

```
R> fit1 <- brm(formula =
+    time | cens(censored) ~ age * sex + disease + (1 + age|patient),
+    data = kidney, family = lognormal(),
+    prior = c(set_prior("normal(0,5)", class = "b"),
+      set_prior("cauchy(0,2)", class = "sd"),
+      set_prior("lkj(2)", class = "cor")), warmup = 1000,
+    iter = 2000, chains = 4, control = list(adapt_delta = 0.95))
```

- En este caso, la fórmula será lo que especifique el modelo de regresión para los datos, y se conforma de la variable respuesta, una función de una variable $fun(var)$ (que haga una función en el modelo con una variable concreta, como datos censurados o el número de intentos), las variables explicativas (separadas por +) y $(coefs \mid groups)$, que son variables cuyos efectos pueden variar según los niveles de los grupos de una variable factor. Si se quiere que los efectos no tengan correlación entre sí, se usa $(coefs||groups)$

```
response | addition ~ fixed + (random | group)
```

- La familia indica la distribución de la variable respuesta, y pueden haber de diferentes tipos

```
Families:
gaussian, student, cauchy, binomial, bernoulli, beta, categorical, poisson,
negbinomial, geometric, gamma, inverse.gaussian, exponential, weibull,
cumulative, cratio, sratio, acat, hurdle_poisson, hurdle_negbinomial,
hurdle_gamma, zero_inflated_poisson, and zero_inflated_negbinomial
```

- Las distribuciones *a priori* se pueden fijar con un vector que incluya funciones $set\_prior()$ con la distribución (en formato *string*), la clase de parámetro (b indica parámetros, sd indica desviación estándar, y cor indica correlación) seguido de $\_coef\_group$ (si se puede especificar)

- Después, se especifican características de la simulación de cadenas, tales como las observaciones *burning* o *warmup*, las iteraciones y el número de cadenas

- Finalmente, también se puede fijar un parámetro de control del muestreador NUTS que usa la función $brm()$, de modo que se utiliza $control = list(adapt\_delta = \ldots)$ y se fija un número entre 0.8 y 1. Este número, al aumentarse, reduce la velocidad, pero hace que haya menos divergencia de las cadenas

- Los resultados que se obtienen al ejecutar la función $summary()$ sobre el modelo ajustado son los siguientes:

```
R> summary(fit1, waic = TRUE)

 Family: lognormal (identity)
Formula: time | cens(censored) ~ age * sex + disease + (1 + age | patient)
   Data: kidney (Number of observations: 76)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: 673.51

Group-Level Effects:
~patient (Number of levels: 38)
                  Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)         0.40      0.28     0.01     1.01       1731    1
sd(age)               0.01      0.01     0.00     0.02       1137    1
cor(Intercept,age)   -0.13      0.46    -0.88     0.76       3159    1

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept        2.73      0.96     0.82     4.68       2139    1
age              0.01      0.02    -0.03     0.06       1614    1
sexfemale        2.42      1.13     0.15     4.64       2065    1
diseaseGN       -0.40      0.53    -1.45     0.64       2664    1
diseaseAN       -0.52      0.50    -1.48     0.48       2713    1
diseasePKD       0.60      0.74    -0.86     2.02       2968    1
age:sexfemale   -0.02      0.03    -0.07     0.03       1956    1

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     1.15      0.13     0.91     1.44       4000    1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Las aproximaciones modales y distribucionales

Los modelos de regresión

- Muchos estudios se enfocan en relaciones entre dos o más observable, enfocándose en cómo varía una cantidad y varía en función de otra cantidad o vector de cantidades $\boldsymbol{x}$

    - En general, uno se interesa por la distribución condicional de y dado $\boldsymbol{x}$, parametrizada como $p(y|\boldsymbol{\theta}, \boldsymbol{x})$ bajo el modelo en la que n observaciones $(x, y)_i$ son intercambiables

        - La cantidad primaria de interés, y, se llama respuesta o variable dependiente, y se asume que es continua

        - Las variables $\boldsymbol{x} = (x_1, \dots, x_k)$ se denominan variables explicativas y pueden ser discretas o continuas. A veces se escoge una sola variable x_j de interés primario y se le denomina variable de tratamiento, llamando así a las otras variables como variables de control

        - La distribución de y dado $\boldsymbol{x}$ es típicamente estudiada en el contexto de un conjunto de unidades o sujetos experimentales $i = 1, \dots, n$ sobre los que y_i y $x_{i1}, x_{i2}, \dots, x_{ik}$ se miden. Se suele utilizar el índice i para indicar unidades y j para indicar los componentes de $\boldsymbol{x}$

        - Se usa $\boldsymbol{y}$ para denotar el vector de resultados para los n sujetos y $\boldsymbol{X}$ para denotar la matriz $n \times k$ de predictores

    - La versión más simple y la más usada del modelo es el modelo lineal normal, en la que la distribución de $\boldsymbol{y}$ dado $\boldsymbol{X}$ es una normal con media que es una función lineal de $\boldsymbol{X}$

    $$E(y_i|\boldsymbol{\beta}, \boldsymbol{X}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad for \;\; i = 1,2, \dots, n$$

        - Para muchas aplicaciones, x_{1i} se fija en uno, de modo que $\beta_1 x_{1i}$ es una constante β_1 para toda i

        - En el marco del modelo lineal normal, los problemas de modelaje más importantes son la definición de las variables $\boldsymbol{x}$ e y (posiblemente usando transformaciones), de modo que la esperanza condicional de $\boldsymbol{y}$ sea razonablemente lineal como una función de las columnas de $\boldsymbol{X}$ con errores normales aproximadamente, y configurar una distribución *a priori* en los parámetros del modelo que reflejen precisamente el conocimiento sustantivo

        - El problema de inferencia estadística es estimar los parámetros $\boldsymbol{\theta}$ condicionales a $\boldsymbol{X}$ e $\boldsymbol{y}$

- Debido a que se pueden coger tantas variables $\boldsymbol{X}$ como uno quiere y transformar $\boldsymbol{X}$ e $\boldsymbol{y}$ de una manera conveniente, el modelo lineal normal es una herramienta muy flexible para establecer relaciones cuantitativas entre variables

- Los datos numéricos en el problema de regresión incluyen tanto $\boldsymbol{X}$ como $\boldsymbol{y}$. Por lo tanto, un modelo bayesiano completo incluye una distribución de $\boldsymbol{X}$, $p(\boldsymbol{X}|\boldsymbol{\psi})$, indexado por un vector de parámetros ψ, por lo que involucra una verosimilitud conjunta $p(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\theta})$ junto con una distribución *a priori* $p(\boldsymbol{\psi}, \boldsymbol{\theta})$

  - En el contexto de regresión estándar, se asume que la distribución de $\boldsymbol{X}$ no proporciona información sobre la distribución condicional de $\boldsymbol{y}$ dado $\boldsymbol{X}$ (se asume independencia de los parámetros $\boldsymbol{\theta}$ determinando $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ y de los parámetros $\boldsymbol{\psi}$ determinando $p(\boldsymbol{X}|\boldsymbol{\psi})$)

  - Por lo tanto, desde una perspectiva bayesiana, la característica que define el modelo de regresión es que ignora la información que proporciona $\boldsymbol{X}$ sobre $(\boldsymbol{\psi}, \boldsymbol{\theta})$. Suponiendo que $\boldsymbol{\psi}$ y $\boldsymbol{\theta}$ son independientes en su distribución *a priori*, los factores de la distribución posterior se pueden analizar de la siguiente manera sin perder información:

$$p(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\boldsymbol{\psi})p(\boldsymbol{\theta})$$

$$\Rightarrow\ p(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = p(\boldsymbol{\psi}|\boldsymbol{X}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = p(\boldsymbol{\psi}|\boldsymbol{X})p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$$

$$\Rightarrow\ p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{y})} \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$$

  - Cuando las variables explicativas $\boldsymbol{X}$ se escogen, la proabbilidad $p(\boldsymbol{X})$ se conoce, y no hay parámetros $\boldsymbol{\psi}$

  - La ventaja práctica de utilizar este modelo de regresión es que es más fácil especificar una distribución condicional realista de una variable dado k variables que especificar una distribución conjunta de $k + 1$ variables

- Una gran parte del análisis estadístico aplicado se basa en técnicas de regresión lineal que se pueden interpretar como una inferencia posterior bayesiana basada en una distribución *a priori* no predictiva para los parámetros del modelo normal lineal

  - En el caso más simple, a veces llamado regresión lineal ordinaria, los errores de las observaciones son independientes y tienen la misma varianza

$$\boldsymbol{y}|\boldsymbol{\beta},\sigma,\boldsymbol{X} \sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_{k\times k})$$

- Se discuten extensiones y alternativas a estas suposiciones, sobre todo en cuanto a la varianza constante y a las correlaciones condicionales se refiere

- En el modelo de regresión normal, una distribución no informativa *a priori* conveniente es uniforme en $(\boldsymbol{\beta}, \log\sigma)$

$$p(\boldsymbol{\beta},\sigma^2|\boldsymbol{X}) \propto \sigma^{-2}$$

  - Cuando hay muchos datos y solo hay pocos parámetros, la distribución *a priori* no informativa es útil porque da resultados aceptables y requiere menos esfuerzo al especificar el conocimiento *a priori* de forma probabilística

  - Para un tamaño pequeño de muestra o un gran número de parámetros, la verosimilitud no tiene un pico tan puntiagudo, por lo que las distribuciones *a priori* y los modelos jerárquicos son más importantes

- Igual que con la distribución normal con media y varianza desconocida, se determina primero la distribución posterior para $\boldsymbol{\beta}$, condicional en σ, y después la distribución posterior para σ^2. Esto quiere decir que se factoriza la distribución posterior conjunta para $\boldsymbol{\beta}$ y σ^2 de la siguiente manera:

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

  - La distribución condicional posterior de $\boldsymbol{\beta}$ dado σ es la exponencial de una forma cuadrática en $\boldsymbol{\beta}$, por lo que es normal

$$\boldsymbol{\beta}|\sigma,\boldsymbol{y},\boldsymbol{X} \sim N\left(\widehat{\boldsymbol{\beta}},\boldsymbol{V}_{\boldsymbol{\beta}}\sigma^2\right) \quad where$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \& \quad \boldsymbol{V}_{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  - La distribución marginal posterior de σ^2 se puede escribir de la siguiente manera, la cual se puede interpretar como una distribución con forma de χ^2 inversa escalada

$$p(\sigma^2|\boldsymbol{y},\boldsymbol{X}) = \frac{p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})}{p(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})} \Rightarrow \sigma^2|\boldsymbol{y},\boldsymbol{X} \sim Inv.\chi^2(n-k,s^2)$$

$$where \quad s^2 = \frac{1}{n-k}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

        - La distribución marginal posterior de $\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}$, promediando sobre σ, es una distribución multivariante t con $n-k$ grados de libertad, pero raramente se usa el hecho en la práctica cuando se sacan inferencias por simulación, dado que para caracterizar la distribución posterior conjunta se pueden sacar simulaciones de σ y después de $\boldsymbol{\beta}|\sigma$

        - Las estimaciones estándar no bayesianas de $\boldsymbol{\beta}$ y c son $\widehat{\boldsymbol{\beta}}$ y s, respectivamente, como se ha definido. La estimación del error estándar clásico para $\boldsymbol{\beta}$ se obtiene a través de fijar $\sigma = s$

    - Para el análisis basado en una distribución *a priori* impropia, es importante comprobar que la distribución posterior es propia

        - Resulta que $p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})$ es propia mientras $n > k$ y el rango de $\boldsymbol{X}$ sea k

        - Estadísticamente, en la ausencia de información *a priori*, la primera condición requiere que haya al menos tantas observaciones como parámetros, mientras que la segunda condición requiere que las columnas de $\boldsymbol{X}$ sean linealmente independientes para que los k coeficientes β estén únicamente determinados

- Es fácil sacar muestras de distribución posterior $p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})$ calculando $\widehat{\boldsymbol{\beta}}$ y V_β con las fórmulas, calculando s^2, sacando σ^2 de la distribución $Inv.\chi^2$ y sacando β de la distribución normal multivariante

    - Para ser computacionalmente eficiente, la simulación se puede llevar a cabo de la siguiente manera:

        - Se calcula la factorización $\boldsymbol{QR}$ ($\boldsymbol{X} = \boldsymbol{QR}$), en donde $\boldsymbol{Q}$ es una matriz $n \times k$ con columnas ortonormales y $\boldsymbol{R}$ es una matriz triangular superior de tamaño $k \times k$

        - Se calcula $\boldsymbol{R}^{-1}$, lo cual es fácil debido a que $\boldsymbol{R}$ es una matriz superior triangular. La matriz $\boldsymbol{R}^{-1}$ es un factor de Cholesky (la raíz cuadrada de una matriz) de la matriz de covarianzas $V_{\boldsymbol{\beta}}$, dado que $\boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})' = (\boldsymbol{X}'\boldsymbol{X})^{-1} = V_{\boldsymbol{\beta}}$

        - Se calcula $\widehat{\boldsymbol{\beta}}$ resolviendo el sistema de ecuaciones lineales $\boldsymbol{R}\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}'\boldsymbol{y}$, usando el hecho de que $\boldsymbol{R}$ es una matriz triangular superior

    - Una vez que se ha simulado σ^2 (usando una simulación de la χ^2). $\boldsymbol{\beta}$ se puede simular fácilmente de la distribución normal multivariante apropiada usando la factorización de Cholesky y un programa para generar normales estándar independientes

- La factorización $\boldsymbol{QR}$ de $\boldsymbol{X}$ es útil para calcular la media de la distribución posterior y para simular el componente aleatorio en la distribución posterior de $\boldsymbol{\beta}$

- Para algunos problemas grandes involucrando miles de datos y cientos de variables explicativas, hasta la descomposición $\boldsymbol{QR}$ puede requerir mucho espacio en la computadora y tiempo, por lo que hay métodos más efectivos

- También es posible realizar un desarrollo formal para la distribución posterior predictiva para nuevos datos

  - Ahora el objetivo es aplicar el modelo de regresión a un nuevo conjunto de datos para los que se ha observado la matriz $\tilde{\boldsymbol{X}}$ de variables explicativas, y se desea predecir los resultados $\tilde{\boldsymbol{y}}$

    - Si $\boldsymbol{\beta}$ y σ^2 se conocieran exactamente, el vector $\tilde{\boldsymbol{y}}$ tendría una distribución normal con medio $\tilde{\boldsymbol{X}}\boldsymbol{\beta}$ y con matriz de varianzas y covarianzas $\sigma^2\boldsymbol{I}$. En cambio, el conocimiento actual $\boldsymbol{\beta}$ y σ se resume por su distribución posterior

  - La distribución posterior predictiva de datos no observados $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ tiene dos componentes de incertidumbre

    - La primera es la variabilidad fundamental del modelo, representado por la varianza σ^2 en $\boldsymbol{y}$ y $\boldsymbol{X}$ que no se tiene en cuenta $\boldsymbol{X\beta}$

    - La segunda es la incertidumbre posterior en $\boldsymbol{\beta}$ y σ por el tamaño finito de la muestra de $\boldsymbol{y}$ y $\boldsymbol{X}$

    - Mientras el tamaño de la muestra $n \to \infty$, la varianza debido a la incertidumbre posterior $(\boldsymbol{\beta}, \sigma^2)$ decrece hacia cero, pero la incertidumbre predictiva se mantiene. Para sacar la muestra aleatoria $\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{X}}$ para su distribución posterior predictiva, primero se saca $(\boldsymbol{\beta}, \sigma^2)$ de la distribución posterior conjunta, y entonces simular $\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{X}} \sim N\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$

  - El modelo lineal normal es lo suficientemente simple que también se puede determinar la distribución posterior predictiva analíticamente

    - Derivar la forma analítica no es necesario, uno puede simular valores $(\boldsymbol{\beta}, \sigma^2)$ y después simular $\tilde{\boldsymbol{y}}$, como anteriormente se ha descrito. No obstante, de esta manera se puede obtener una perspectiva útil para estudiar la incertidumbre predictiva analíticamente

- Primero se considera la distribución condicional posterior predictiva $p(\tilde{\boldsymbol{y}}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}})$ y después se promedia sobre la incertidumbre posterior en $\sigma|\boldsymbol{y}, \tilde{\boldsymbol{X}}$. Dada σ, la observación futura $\tilde{\boldsymbol{y}}$ tiene una distribución normal, y se deriva su media al promediar sobre $\boldsymbol{\beta}$

$$E(\tilde{\boldsymbol{y}}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}) = E_{\boldsymbol{\beta}}\left[E_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}})|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] =$$

$$= E_{\boldsymbol{\beta}}\left[\tilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] = \tilde{\boldsymbol{X}}\boldsymbol{\beta}$$

- Después, es posible derivar la varianza condicional de $\tilde{\boldsymbol{y}}$. Este resultado tiene sentido: condicional a σ, la varianza posterior predictiva tiene dos términos, $\sigma^2\boldsymbol{I}$, representando la variación muestral, y $\tilde{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{\beta}}\tilde{\boldsymbol{X}}'\sigma^2$, debido a la incertidumbre sobre $\boldsymbol{\beta}$

$$Var(\tilde{\boldsymbol{y}}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}) =$$

$$= E\left[Var(\tilde{\boldsymbol{y}}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}})|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] + Var\left[E(\tilde{\boldsymbol{y}}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}})|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] =$$

$$= E\left[\sigma^2\boldsymbol{I}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] + Var\left[\tilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right] =$$

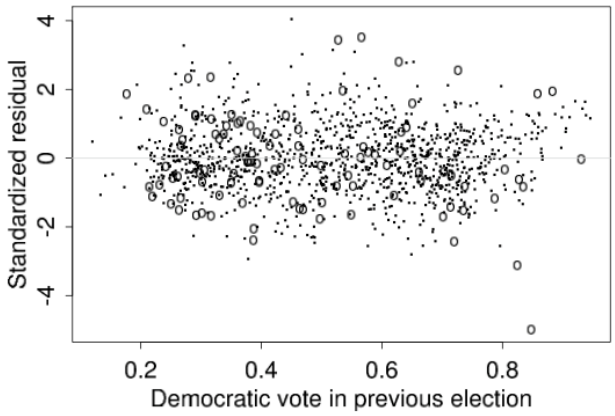
$$= \sigma^2\boldsymbol{I} + \tilde{\boldsymbol{X}}Var\left[\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \tilde{\boldsymbol{X}}\right]\tilde{\boldsymbol{X}}' = \sigma^2\left(\boldsymbol{I} + \tilde{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{\beta}}\tilde{\boldsymbol{X}}'\right)$$

- Dado σ, las observaciones futuras tienen una distribución normal con media $\tilde{\boldsymbol{X}}\boldsymbol{\beta}$, que no depende de σ, y la varianza es proporcional a σ^2. Para completar la determinación de la distribución posterior predictiva, uno tiene que promediar sobre la distribución marginal posterior de σ^2 en $Inv.\chi^2(n-k, s^2)$

- Es más difícil predecir $\tilde{\boldsymbol{y}}$ si no todas las variables explicativas $\tilde{\boldsymbol{X}}$ se conocen, porque entonces las variables explicativas tienen que modelarse una distribución de probabilidad

- Comprobando el ajuste y la robustez del modelo de regresión lineal es un tema bien desarrollado en la estadística

  - Los modelos estándar tales como examinar gráficos de residuos contra las variables explicativas son útiles y pueden interpretarse directamente como comprobaciones posteriores predictivas



- Una ventaja del enfoque bayesiano es que se puede calcular, por simulación, la distribución posterior predictiva para cualquier resumen de los datos, de modo que no se necesita tanto esfuerzo para estimar las distribuciones muestrales de los estadísticos de contraste

- Por ejemplo, para evaluar la significancia estadística y práctica de los patrones en un gráfico de residuos, se puede obtener una distribución posterior predictiva de un estadístico apropiado (por ejemplo, la correlación entre los residuos cuadrados y los valores ajustados)

	Observed proportion of outliers	Posterior predictive dist. of proportion of outliers		
		2.5%	median	97.5%
Open seats	$41/1596 = 0.0257$	0.0013	0.0038	0.0069
Incumbent running	$84/10303 = 0.0082$	0.0028	0.0041	0.0054

- Los modelos de regresión satisfacen al menos tres objetivos: entender el comportamiento de y dado $\boldsymbol{x}$, predecir y dado $\boldsymbol{x}$ para observaciones futuras, y la inferencia causal (o predecir como y cambiará si $\boldsymbol{x}$ se cambia de una manera específica

  - Una vez sus parámetros se han estimado, una regresión puede usarse para predecir las observaciones futuras de las unidades en las que las variables explicativas $\tilde{\boldsymbol{X}}$, pero no el resultado $\tilde{\boldsymbol{y}}$, se han observado

    - Cuando se hacen predicciones, se asume que las viejas y las nuevas observaciones son intercambiables dados los mismos valores de $\boldsymbol{x}$, de modo que el vector contiene toda la información que se tiene para distinguir las nuevas observaciones de las viejas (esto incluye, por ejemplo, la suposición de que el tiempo de observación es irrelevante si no se incluye en $\boldsymbol{x}$)

        - Igual que con la intercambiabilidad en general, no se requiere que los individuos sean idénticos o similares, solo que todo el conocimiento relevante sobre ellos esté incluido en $\boldsymbol{x}$. Cuanto mayor es la similitud, menor es la varianza, pero eso es una cuestión de precisión, no de validez

    - Cuando las viejas y las nuevas observaciones no son intercambiables, la información relevante debería estar introducida en el vector $\boldsymbol{x}$

        - Por ejemplo, si se quieren estudiar dos individuos de dos poblaciones diferentes, se necesita utilizar una variable indicadora para la población

        - La manera más simple es reemplazar el término constante en $\boldsymbol{x}$ por dos variables indicadoras (reemplazar la columna de unos en $\boldsymbol{X}$ por dos columnas de las variables indicadoras)

    - Los objetivos de describir la relación entre y y $\boldsymbol{x}$ y usando el modelo resultante para predicción son aplicaciones directas de estimar $p(y|\boldsymbol{x})$

        - La inferencia causal es más sutil. Cuando se piensa en la inferencia causal, se piensa en una variable de interés conocida como variable de tratamiento y las otras variables explicativas son variables de control

        - Los tratamientos representan atributos que son manipulados o al menos potencialmente manipulables por el investigador, mientras que las variables de control miden otras características de la unidad experimental o el contexto experimental, medidas antes del tratamiento

        - Se tiene que tener cuidado al considerar las variables de control para la inferencia causal, ya que la relación entre las variables de control y la de tratamiento no pueden estar muy correlacionadas, ya que afectaría el análisis *Ceteris paribus*

- La elección de que variables incluir en un modelo de regresión depende en el propósito del estudio. Desde un punto de vista bayesiano, se pueden discutir temas que nacen en la regresión clásica

    - Los parámetros en una regresión clásica no pueden estimarse únicamente si hay más parámetros que datos o, más generalmente, si las columnas de la matriz $\boldsymbol{X}$ de variables explicativas no son linealmente independientes

- En estos casos, se dicen que los datos son colineares, y $\boldsymbol{\beta}$ no puede estimar únicamente solo de los datos, sin importar que tan grande sea el tamaño muestral

- Esto ocurre porque si una o más variables (columnas) es combinación lineal de otra, se podría utilizar el mismo coeficiente para estas y eso hace que los posibles valores de las β de cada variable sean infinitos (problema de identificación)

- Pensando en un gráfico de dispersión k-dimensional de los n puntos, si los n puntos caen cerca de un hiperplano de menos dimensiones (así como un plano en un espacio tridimensional), entonces los datos son colineares o coplanares y proporcionan poca información sobre algunas combinaciones lineales de las β

- Una vez las variables se han seleccionado, normalmente tiene sentido transformarlas de modo que la relación entre $\boldsymbol{x}$ e y sea aproximadamente lineal

  - Las transformaciones como los logaritmos y los *logits* son útiles en varias situaciones. Sin embargo, una transformación cambia la interpretación del coeficiente de regresión al cambio en la y transformada por un cambio unitario en la variable x transformada

  - Si se piensa que una variable x_j tiene un efecto no lineal en y, es posible incluir más de una transformación de la x_j en la regresión (incluir, por ejemplo, x_j y x_j^2 en una misma regresión)

- Para incluir una variable categórica en una regresión, un enfoque natural es construir una variable indicador para cada categoría

  - Esto permite separar el efecto para cada nivel de la categoría, sin asumir ningún orden u otra estructura en las categorías

  - Cuando hay dos categorías, un indicador binario 0/1 funciona, y cuando hay k categorías, se requieren $k-1$ indicadores más el término constante

  - Si hay una ordenación natural de las categorías de una variable discreta, suele ser útil tratar la variable como si fuera continua (codificándolas con números)

- En el modelo lineal, un cambio en una unidad en x_j se asocia a un cambio constante en la respuesta media de y_i, dado cualquier valor fijo de los otros predictores

- Si las respuesta a un cambio unitario en x_j depende del valor fijo que toma otro predictor x_r , entonces es necesario incluir una interacción en los términos del modelo

- Generalmente se puede permitir la interacción añadiendo el término de producto cruzado $(x_j - \bar{x}_j)(x_r - \bar{x}_r)$ como un predictor adicional, aunque estos términos no puedan ser directamente interpretables si ambas variables son continuas (en este caso, lo mejor es categorizar al menos una de las variables)

- Estas interacciones se tratan como cualquier otra variable explicativa: creando una nueva columan en $\boldsymbol{X}$ y estimando un nuevo elemento de β

- Normalmente solo se desea incluir variables que tienen una conexión sustantiva razonable con el problema que se está estudiando

  - Es común que en una regresión haya un gran número de variables de control potencial, algunas de las que pueden parecer tener valor predictivo

  - No obstante, lo importante en esta situación es entender por qué estas variables pueden tener poder predictivo y qué relación tienen con otras variables de interés, de modo que se resuma toda la información sustancial de esos aspectos en el menor número de variables

- Idealmente, un modelo estadístico debería incluir toda la información relevante. En una regresión, por tanto, $\boldsymbol{x}$ tendría que incluir todas las variables explicativas que posiblemente predigan y

  - El intento de incluir predictores relevantes es difícil en la práctica, pero es generalmente útil. La posible pérdida de precisión cuando se incluyen predictores no importantes suele verse como un riesgo pequeño para la validez general de las predicciones y las inferencias sobre los estimandos de interés

  - En la regresión clásica, hay desventajas directas a incrementar el número de variables explicativas. Por una parte, está la restricción de que $k < n$, y por otra parte, usar un gran número de variables explicativas deja poca información disponible para obtener estimaciones precisas de la varianza

  - Estos problemas, normalmente conocidos como problemas de sobreajuste, son de mucho menos importancia con distribuciones *a priori* razonables

- Con tal de poder seleccionar los predictores que se usarán en la regresión, se pueden utilizar métodos de regularización y de reducción dimensional para múltiples predictores

    - Enfoques tales como la regresión a pasos o *stepwise regression* y la selección se subconjuntos son métodos no bayesianos tradicionales para escoger un conjunto de variables explicativas que incluir en una regresión

        - Matemáticamente, no incluir una variable es equivalente a fijar el coeficiente de esta en cero

    - La estimación de la regresión clásica basada en un procedimiento de selección es equivalente a obtener la moda posterior correspondiente a una distribución *a priori* que tiene una probabilidad positiva en varios hiperplanos de pocas dimensiones del espacio de $\boldsymbol{\beta}$

        - El procedimiento de selección se ve muy influenciado por la cantidad de datos disponibles, de modo que las variables importantes pueden ser omitidas porque la variación aleatoria no se puede descartar como una explicación alternativa para su poder predictivo

        - Geométricamente, si el espacio de $\boldsymbol{\beta}$ se interpreta como un hipercubo, el modelo obtenido de la selección clásica del modelo expresa que el vector $\boldsymbol{\beta}$ real tiene unas ciertas probabilidades *a priori* de estar en cualquier zona de ese hipercubo

    - En un enfoque bayesiano, es más atractivo incluir información *a priori* de manera más continua

        - Con muchas variables explicativas x_j , cada una con una probabilidad de ser irrelevante para modelar y, uno puede dar a cada coeficiente una distribución *a priori* $p(\boldsymbol{\beta})$ con un pico en cero (una distribución t-Student centrada en cero con moda concentrada cerca de cero y una cola larga). Esto dice que cada variable es posiblemente no importante, pero que, si tiene poder predictivo, este podría ser grande

        - Por ejemplo, cuando las coeficientes de un conjunto de predictores se modelan, puede ser importante aplicar primero transformaciones lineales u otras para poner los predictores en una escala común (aproximadamente)

- En la estimación por máxima verosimilitud o mínimos cuadrados, o si la regresión lineal se realiza con una distribución *a priori* uniforme no informativa en los coeficientes β_j, transformaciones lineales de los predictores no tienen efecto en las inferencias de la predicción. En un enfoque bayesiano, sin embargo, estas pueden ser importantes

    - En la inferencia bayesiana práctica, uno de los criterios para escoger modelos es la conveniencia, y como algo predeterminado, es conveniente usar una sola distribución *a priori* para todos los coeficientes en un modelo, o posiblemente categorizar los coeficientes en dos o tres categorías

    - En cualquier caso, no se asignaría cuidadosamente una distribución *a priori* separada para cada coeficiente, sino que se utilizaría una clase de modelos flexibles pero convenientes que se espera que rindan bien en la mayoría de casos

- La regularización es un término general utilizado para procedimientos estadísticos que proporcionan estimaciones más estables, y en la regularización bayesiana, se suelen considerar tres aspectos

    - Las estimaciones de mínimos cuadrados con una gran número de predictores puede tener mucho ruido, por lo que distribuciones *a priori* pueden regularizar estas estimaciones

    - Un aspecto es la distribución *a priori* de localización y escala. Por ejemplo, una *a priori* más concentrada y realiza más regularización

    - Otro aspecto es la forma analítica de la distribución *a priori*. Por ejemplo, la normal empuja las estimaciones a la media de la *a priori* en una proporción constante; la Laplaciana desplaza las estimaciones en una cantidad constante; y la distribución de Cauchy hace más regularización cerca de la media y poca si se está muy lejos de esta

    - El último aspecto es como se resumen las inferencias posteriores. Por ejemplo, la moda posterior, que omite alguna variabilidad, puede verse más suave o regularizada que la posterior completa

- Estos tres puntos se pueden reflejar en el lasso, una forma de regularización popular no bayesiana que corresponde a estimar los coeficientes por su moda posterior, después de asignar una distribución *a priori* de Laplace centrada en cero

$$p(\boldsymbol{\beta}) \propto \prod_j e^{-\lambda|\beta_j|}$$

- Combinada con la verosimilitud normal de un modelo de regresión, esto resulta en una distribución posterior que está parcialmente agrupada hacia cero, con la cantidad de agrupación determinada por λ (un hiperparámetro que se puede fijar basándose en información externa o estimado de los datos, de manera bayesiana o no bayesiana)

- La clave del lasso es la combinación de una distribución *a priori* con un pico puntiagudo en cero y la decisión de resumir la distribución posterior (la función de verosimilitud penalizada) por su moda. Poniendo ambas cosas juntas permite obtener estimaciones para coeficientes que se vuelven exactamente cero en contextos con muchos predictores, ruido en los datos, o muestras pequeñas o moderadas

- Para las regresiones que se suelen considerar, no se cree que cualquier coeficiente sea puramente nulo y no se considera generalmente como una ventaja conceptual

  - Sin embargo, las estimaciones regularizadas tales como las obtenidas con el lasso pueden ser mucho mejores que aquellas que resultan de un ajuste de mínimos cuadrados con distribuciones *a priori* planas que, en la práctica, pueden requerir implícitamente que los usuarios restrinjan masivamente el conjunto de posibles predictores para obtener estimaciones estables

- Desde esta perspectiva, la regularización siempre ocurrirá de una manera u otra, y métodos simples tales como el lasso tiene dos ventajas sobre otros métodos tradicionales de seleccionar predictores

  - La primera es que el lasso se define claramente y es algorítmico, de modo que sus elecciones son transparentes, lo cual no es el caso con métodos informales

  - La segunda es el que el lasso permite incluir más información en el ajuste del modelo y usa un enfoque basado en datos para decidir que predictores escoger

- Desde la perspectiva bayesiana, hay posibilidad de mejorar el lasso en varias direcciones

  - Un primer paso sería permitir incertidumbre en qué variables se seleccionan, volviendo posiblemente al conjunto de simulaciones representando los diferentes subconjuntos para incluir en el modelo

- Otro enfoque puede ser tomar la distribución lasso *a priori* y hacer inferencia bayesiana completa, en el que los coeficientes se agrupan parcialmente. Esto tiene sentido menos en el caso en que no haya una ventaja particular a la familia doble-exponencial y se tenga que usar algo de la familia t-Student

- Además, en cualquier aplicación, uno puede tener información sugiriendo que algunos coeficientes deberían agruparse más que otros (no hay necesidad de usar la misma distribución *a priori* para todos ellos)

- Como siempre, una vez se piensa seriamente en el modelo, hay muchas aplicaciones posibles. Para todas sus simplificaciones, el lasso es un paso útil en la dirección correcta, permitiendo el uso automático de muchos más predictores de los que sería posible bajo mínimos cuadrados

Los modelos de regresión generalizados

- Como se ha visto, un modelo estocástico basado en un predictor lineal $\boldsymbol{X\beta}$ es fácil de entender y puede funcionar en una variedad de problemas. El propósito del modelo lineal generalizado es extender la idea del modelaje lineal a casos en los que la relación lineal entre $\boldsymbol{X}$ y $E(\boldsymbol{y}|\boldsymbol{X})$ o la distribución normal no es apropiada

    - En algunos casos, es razonable aplicar un modelo lineal a un resultado transformado utilizando variables explicativas transformadas

        - No obstante, la relación entre $\boldsymbol{X}$ y $E(\boldsymbol{y}|\boldsymbol{X})$ no puede ser completamente modelada siempre como normal o lineal, aún después de las transformaciones

        - Por ejemplo, si y solo puede tomar valores positivos o cero, no se puede analizar $\log(y)$ solamente, aunque la relación entre $E(y)$ a $\boldsymbol{X}$ es generalmente multiplicativa. Si y toma valores discretos, en cambio, la media de y puede estar relacionada linealmente $\boldsymbol{X}$, pero el término de variación no puede ser descrito por la distribución normal

    - La clase de modelos generalizados unifica los enfoques necesitados para analizar los datos para los cuales cualquier suposición de una relación lineal entre $\boldsymbol{x}$ e y o la suposición de variación normal no es apropiada. Un modelo lineal generalizado está especificado en tres niveles:

        - Primero de todo se especifica un predictor lineal, $\boldsymbol{\eta = X\beta}$

- Después, se especifica una función de vínculo o *link function* $g(\cdot)$ que relaciona el predictor lineal a la media de la variable resultado

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X\beta})$$

- Finalmente, se especifica el componente aleatorio especificando la distribución de la variable resultado $\boldsymbol{y}$ con media $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\mu}$. La distribución también puede depender de un parámetro de dispersión ϕ

- Por lo tanto, la media de la distribución de $\boldsymbol{y}$, dado $\boldsymbol{X}$, se determina por $\boldsymbol{X\beta}$ a través de $E(\boldsymbol{y}|\boldsymbol{X}) = g^{-1}(\boldsymbol{X\beta}) = \boldsymbol{\mu}$

  - Se usa la misma notación como en una regresión lineal cuando es posible, de modo que $\boldsymbol{X}$ es la matriz $n \times p$ de variables explicativas y $\boldsymbol{\eta} = \boldsymbol{X\beta}$ es el vector de n valores del predictor lineal

  - Si se denota el predictor lineal para el caso i por $\boldsymbol{X_i\beta}$ y el parámetro de dispersión o varianza (si está presente) por ϕ, entonces la distribución de los datos toma la siguiente forma:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} p(y_i|\boldsymbol{X_i\beta}, \phi)$$

  - Los modelos lineales generalizados más comunes son el modelo de Poisson y de la distribución binomial, que no requieren ningún parámetro de dispersión (se fijan en 1). No obstante, en la práctica, el exceso de dispersión suele ser la regla más que la excepción

- La regresión lineal es un caso especial del modelo lineal generalizado, con una función identidad como función de vínculo $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$

  - Para datos continuas que son todos positivos, se pueden usar un modelo normal en la escala logarítmica. Cuando la familia distribucional no ajusta los datos, la distribución gamma o la Weibull se consideran como alternativas

- Los datos de conteo se suelen modelar utilizando un modelo de Poisson. El modelo lineal generalizado de Poisson, normalmente llamado modelo de regresión de Poisson, asume que $\boldsymbol{y}$ es Poisson con media y varianza μ

- La función de vínculo escogida suele ser el logaritmo, de modo que $\log(\boldsymbol{\mu}) = \boldsymbol{X\beta}$, haciendo que $\boldsymbol{\mu} = e^{\boldsymbol{X\beta}}$. La distribución para los datos $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ es la siguiente:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-\exp(\eta_i)} [\exp(\eta_i)]^{y_i}$$

- Cuando se considera la distribución posterior bayesiana, se condiciona a $\boldsymbol{y}$, y los factores $1/y_i!$ Pueden ser absorbidos en una constante arbitraria

- Uno de los modelos lineales generalizados más usados son los de datos binarios o binomiales. Suponiendo que $y_i \sim Bin(n_i, \mu_i)$ con n_i conocida, es común especificar el modelo en términos de la media de las proporciones y_i/n_i, mas que la media de y_i

  - Escogiendo la transformación *logit* de la probabilidad de éxito $g(\mu_i) = \log[\mu_i/(1-\mu_i)]$ como la función de vínculo, esto lleva al modelo de regresión logística, con la siguiente función de distribución:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{\eta_i - y_i}$$

  - Otras funciones de vínculo que se usan normalmente, el vínculo *probit* $g(\boldsymbol{\mu}) = \Phi^{-1}(\boldsymbol{\mu})$, son comúnmente utilizadas en econometría. La distribución de los datos en un modelo *probit* es la siguiente:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} \big(\Phi(\eta_i)\big)^{y_i} \big(1 - \Phi(\eta_i)\big)^{\eta_i - y_i}$$

  - El vínculo probit se obtiene reteniendo el proceso de variación normal en el modelo lineal normal, mientras que se asume que todos los resultados están dicotomizados

  - En la práctica, los modelos *probit* y *logit* son similares, difiriendo principalmente en los extremos de las colas. En cualquier caso, los factores $\binom{n_i}{y_i}$ dependen solo de las cantidades observadas y se pueden incluir en un factor constante de la densidad posterior

  - La distribución t-Student puede ser usada como una alternativa robusta a los modelos *probit* y *logit*, y otra función de vínculo estándar es la de log-log complementario $g(\boldsymbol{\mu}) =$

$\log(-log(\boldsymbol{\mu}))$, que es asimétrica en μ (por lo que $g(\boldsymbol{\mu}) \neq -g(1-\boldsymbol{\mu})$

- Los análisis clásicos de modelos lineales generalizados permiten la posibilidad de variación más allá de la que asume la distribución muestral, normalmente llamada sobredispersión

  - Los datos pueden indicar mucha variación, más que la esperada bajo una distribución concreta

  - Esta variación puede ser incorporada en un modelo jerárquico usando un indicador para cada individuo, los cuales tienen una distribución propia

- Una vez se han explicado algunos ejemplos de modelos lineales generalizados, se puede analizar los diferentes aspectos de trabajar con estos modelos lineales generalizados

  - La descripción de los modelos estándar en la sección anterior utiliza funciones de vínculo canónicas para cada familia

    - El vínculo canónico es la función del parámetro de la media que aparece en el exponente de la familia de la forma de la familia exponencial de la densidad de probabilidad

    - Normalmente se utilizan funciones canónicas, pero hay casos en los que no (por ejemplo, en el modelo binomial con el *probit*)

  - A veces es conveniente expresar el modelo lineal generalizado de manera que uno de los predictores tenga un coeficiente conocido

    - Un predictor de este tipo se denomina *offset* y comúnmente surgen en los modelos de Poisson, en donde la tasa de ocurrencia por unidad de tiempo es μ y, por tanto, con una exposición $\boldsymbol{T}$, el número esperado de incidentes es $\boldsymbol{\mu T}$

    - Uno puede querer tomar $log(\boldsymbol{\mu}) = \boldsymbol{X\beta}$ como en el modelo usual de Poisson, pero los modelos lineales generalizados están parametrizados a través de la media de y, que es $\boldsymbol{\mu T}$, en donde $\boldsymbol{T}$ ahora representa el vector de exposiciones para las unidades en la regresión

    - Se puede aplicar el modelo lineal generalizado de Poisson aumentando la matriz de variables explicativas con una columna conteniendo los valores $\log(\boldsymbol{T})$ (esta columna de la matriz corresponderá a un coeficiente con valor conocido igual a 1)

- La elección y parametrización de las variables explicativas $\boldsymbol{x}$ requiere las mismas consideraciones que las discutidas anteriormente para modelos lineales clásicos

    - El predictor lineal es usado para predecir la función de vínculo $g(\mu)$ y no $\mu = E(y)$, por lo que el efecto de cambiar la variable explicativa x_j por una cantidad fija depende del valor actual de $\boldsymbol{x}$

    - Una manera de traducir los efectos en la escala de y es medir los cambios en comparación al caso estándar, con el vector de predictores $\boldsymbol{x_0}$ y el resultado estándar $y_0 = g^{-1}(\boldsymbol{x_0\beta})$. Entonces, añadiendo o restando el vector $\Delta\boldsymbol{x}$ lleva a un cambio en el resultado estándar de y_0 a $g^{-1}(g(y_0) \pm (\Delta\boldsymbol{x})\boldsymbol{\beta})$

- Una idea importante, tanto en el entendimiento como en el cálculo de regresiones de datos discretos, es la expresión de términos no observados (latentes) en los datos continuos

    - El modelo *probit* para datos binarios es equivalente al modelo siguiente con datos latentes u_i, los cuales a veces pueden tener una interpretación útil:

$$u_i \sim N(\boldsymbol{X_i\beta}, 1)$$

$$y_i = \begin{cases} 1 & if \;\; u_i > 0 \\ 0 & if \;\; u_i < 0 \end{cases}$$

    - Otra ventaja de la parametrización latente para el modelo *probit* es que permite hacer un cálculo conveniente a través del muestreador de Gibbs. Condicional en las u_i latentes, el modelo es una regresión lineal simple, por lo que, condicional en los parámetros del modelo y de los datos, las u_i han truncado la distribución normal, con cada u_i truncada para ser negativa o positiva dependiendo de si $y_i = 0$ o $y_i = 1$

$$u_i | y_i = 0, \boldsymbol{X_i\beta} \sim \begin{cases} N(\boldsymbol{X_i\beta}, 1) & with \;\; u_i < 0 \\ 0 & with \;\; u_i > 0 \end{cases}$$

$$u_i | y_i = 1, \boldsymbol{X_i\beta} \sim \begin{cases} 0 & with \;\; u_i < 0 \\ N(\boldsymbol{X_i\beta}, 1) & with \;\; u_i > 0 \end{cases}$$

    - La misma idea se puede aplicar para la regresión logística, solo que se tiene que cambiar la distribución de u_i a $u_i \sim logistic(\boldsymbol{X_i\beta}, 1)$, la cual tiene como función de densidad $1/(e^{x/2} + e^{-x/2})$ (la derivada de la función *logit* inversa). El modelo es menos conveniente para su computación, pero puede ser útil para el modelaje

- Interpretaciones similares se pueden dar para regresiones ordenadas multinomiales del tipo descrito anteriormente. Por ejemplo, si los datos y toman unos ciertos valores discretos $0,1,2,\ldots$, entonces el modelo multinomial se puede definir en términos de los puntos de corte $c_0, c_1, c_2, \ldots$, de modo que la respuesta y_i iguala a 0 si $u_i < c_0$, 1 si $u_i \in (c_0, c_1)$, etc.

- Hay más de una manera de parametrizar los modelos de puntos de corte, y la selección de la parametrización tiene implicaciones cuando el modelo se pone en una estructura jerárquica. Por ejemplo, la parametrización anterior no se cambia si una constante se añade a todos los puntos de corte, por lo que el modelo es potencialmente no identificable (se suele gestionar esto fijando $c_0 = 0$, de modo que los otros puntos se estiman con los datos)

- Se han considerado modelos lineales generalizados con distribuciones *a priori* no informativas en $\boldsymbol{\beta}$, distribuciones *a priori* informativas en $\boldsymbol{\beta}$ y modelos jerárquicos para los cuales la distribución *a priori* en $\boldsymbol{\beta}$ depende de parámetros desconocidos

  - Se intenta tratar los modelos lineales generalizados de la misma manera, lo cual causa algunas dificultades. Por ejemplo, los modelos lineales generalizados con parámetro de dispersión ϕ además de los coeficientes de regresión (σ en el caso normal)

  - Se puede considerar una distribución *a priori* en el parámetro de dispersión, y cualquier información *a priori* sobre $\boldsymbol{\beta}$ puede describirse condicionada al parámetro de dispersión ϕ, de modo que $p(\boldsymbol{\beta}, \phi) = p(\phi)p(\boldsymbol{\beta}|\phi)$

  - El análisis clásico de modelos lineales generalizados se obtiene si se asume una distribución *a priori* no informativa para $\boldsymbol{\beta}$. La moda posterior que corresponde a una *a priori* no informativa es el estimador de máxima verosimilitud para el parámetro $\boldsymbol{\beta}$, que se puede obtener utilizando una regresión lineal iterativa

  - Un enfoque que es a veces útil para especificar una distribución *a priori* para $\boldsymbol{\beta}$ es hacerlo en términos de los datos hipotéticos obtenidos bajo el mismo modelo, lo cual quiere decir especificarla en términos de un vector $\boldsymbol{y_0}$ de n_0 individuos hipotéticos y de una matriz $\boldsymbol{X_0}$ de tamaño $n_0 \times k$ de variables explicativas. Esto resulta en una distribución posterior para el vector aumentado $(\boldsymbol{y}, \boldsymbol{y_0})'$ y la matriz aumentada $[\boldsymbol{X}|\boldsymbol{X_0}]$

- Suele ser más natural expresar la información *a priori* directamente en términos de los parámetros $\boldsymbol{\beta}$, por ejemplo, usando un modelo normal $\boldsymbol{\beta} \sim N(\boldsymbol{\beta_0}, \boldsymbol{\Sigma_\beta})$ con valores específicos $\boldsymbol{\beta_0}$ y $\boldsymbol{\Sigma_\beta}$. Una *a priori* normal en $\boldsymbol{\beta}$ es particularmente conveniente para métodos computacionales basados en la aproximación normal para la verosimilitud

- Igual que en la regresión lineal normal, las distribuciones *a priori* jerárquicas para modelos lineales generalizados son una manera natural de ajustar estructuras de datos complejas y permiten incluir más variables explicativas sin encontrarse con problemas de sobreajuste. Normalmente se escoge una distribución normal (como *a priori* debido a que se puede usar el modelaje visto para modelos de regresión jerárquicos normales usando la aproximación normal para la verosimilitud

- La inferencia posterior en modelos lineales generalizados normalmente requiere la aproximación y los métodos de muestreo vistos anteriormente. Para cálculos simples, puede ser conveniente aproximar la verosimilitud con una distribución normal para $\boldsymbol{\beta}$, condicionada al parámetro de dispersión y, si es necesario, para cualquier parámetro jerárquico

    - El método básico para aproximar el modelo lineal generalizado por un modelo lineal, de modo que para cada punto y_i se construye un pseudo-dato z_i y una pseudo-varianza σ_i^2 de modo que la verosimilitud del modelo lineal generalizado $p(y_i|\boldsymbol{X_i\beta}, \phi)$ se aproxima a través de la verosimilitud normal $N(z_i|\boldsymbol{X_i\beta}, \sigma_i^2)$. Después, se combinan los n pseudo-datos y se aproxima la la verosimilitud entera por un modelo de regresión lineal del vector $\boldsymbol{z} = (z_1, \dots, z_n)$ en la matriz de variables explicativas $\boldsymbol{X}$ con matriz de varianzas y covarianzas $diag(\sigma_1^2, \dots, \sigma_n^2)$, lo cual permite analizar los datos como si provinieran de un modelo normal y utilizar los algoritmos desarrollados anteriormente

    - En general, la aproximación normal dependerá del valor de $\boldsymbol{\beta}$ (y de ϕ si el modelo tiene un parámetro de dispersión) en el que se centra. En el desarrollo siguiente, se utiliza la notación $(\widehat{\boldsymbol{\beta}}, \hat{\phi})$ para el punto en el que la aproximación se centra y $\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ para el vector correspondiente de predictores lineales. Cuando se intenta encontrar la moda en la computación, se altera iterativamente el centro de la aproximación normal, y cuando se consigue aproximadamente llegar a la moda, se usa la aproximación normal en aquel valor fijo $(\widehat{\boldsymbol{\beta}}, \hat{\phi})$

- Se puede escribir la verosimilitud logarítmica de la siguiente manera, en donde L es la función de verosimilitud logarítmica para las observaciones individuales:

$$p(y_1, y_2, \dots, y_n|\boldsymbol{\eta}, \phi) = \prod_{i=1}^{n} p(y_i|\eta_i, \phi) = \prod_{i=1}^{n} e^{L(y_i|\eta_i, \phi)}$$

  - Se aproxima cada factor en el productorio superior con una densidad normal para η_i, por lo tanto, aproximando cada $L(y_i|\eta_i, \phi)$ por una función cuadrática en η_i, en donde la z_i, σ_i^2, y la constante dependen de y, de $\hat{\eta}_i = \boldsymbol{X_i}\widehat{\boldsymbol{\beta}}$ y $\hat{\phi}$. Esto quiere decir que un dato es aproximadamente equivalente a una observación z_i, normalmente distribuida con media η_i y con varianza σ_i^2

$$L(y_i|\eta_i, \phi) \approx -\frac{1}{2\sigma_i^2}(z_i - \eta_i)^2 + constant$$

  - Una manera estándar de determinar z_i y σ_i^2 para la aproximación es hacer que coincidan con el primer y el segundo término de la expansión de Taylor de $L(y_i|\eta_i, \phi)$ centrada alrededor de $\hat{\eta}_i = \boldsymbol{X_i}\widehat{\boldsymbol{\beta}}$, respectivamente

$$z_i = \hat{\eta}_i - \frac{L'(y_i|\hat{\eta}_i, \hat{\phi})}{L''(y_i|\hat{\eta}_i, \hat{\phi})} \qquad \sigma_i^2 = -\frac{1}{L''(y_i|\hat{\eta}_i, \hat{\phi})}$$

- La moda posterior se puede encontrar utilizando una regresión lineal ponderada iterativa: en cada paso, uno calcula la aproximación normal a la verosimilitud basada en la hipótesis actual de $(\boldsymbol{\beta}, \phi)$ y encuentra la moda de la distribución posterior aproximada resultante por una regresión lineal ponderada

  - Si hay cualquier información *a priori* sobre $\boldsymbol{\beta}$, esta debería incluirse como filas de datos y variables explicativas en la regresión

  - Iterando este proceso es equivalente a resolver un sistema de k ecuaciones no lineales $\partial p(\boldsymbol{\beta}|\boldsymbol{y})/\partial \boldsymbol{\beta} = \mathbf{0}$ usando el método de Newton-Raphson, y converge hacia la oda de manera rápida para modelos lineales generalizados estándar. Una posible dificultad es que los estimadores de los coeficientes tiendan a infinito, pero información *a priori* sustantiva suele eliminar este problema

  - Si un parámetro de dispersión ϕ está presente, uno puede actualizar ϕ en cada paso de la iteración al maximizar su

densidad posterior condicional (que es unidimensional), dado el valor incumbente para $\boldsymbol{\beta}$. Similarmente, uno puede incluir parámetros de varianza jerárquica cualesquiera que necesitan ser estimados y necesitan actualizarse los valores en cada paso

- Una vez se alcanza la moda $(\widehat{\boldsymbol{\beta}}, \hat{\phi})$, uno puede aproximar la distribución posterior condicional de $\boldsymbol{\beta}$ dado $\hat{\phi}$ con el resultado de la regresión lineal ponderada más reciente, donde $\boldsymbol{V_\beta} = \left[\boldsymbol{X}' diag\left(-L''(y_i, \hat{\eta}_i, \hat{\phi}_i)\right)\boldsymbol{X}\right]^{-1}$

$$p(\boldsymbol{\beta}|\hat{\phi}, \boldsymbol{y}) \approx N_p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \boldsymbol{V_\beta})$$

  - En general, uno solo necesita calcular el factor de Cholesky de $\boldsymbol{V_\beta}$ para poder trabajar

  - Si el tamaño de la muestra n es grande y $\hat{\phi}$ no es parte del modelo (como en las distribuciones binomiales y de Poisson), uno puede parar y resumir la distribución posterior por esta normal

  - Si un parámetro de dispersión está presente, entonces se puede aproximar la distribución marginal de ϕ usando el método visto anteriormente, pero aplicado a la moda condicional $\widehat{\boldsymbol{\beta}}(\phi)$

$$p_{approx}(\phi|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \phi|, \boldsymbol{y})}{p_{approx}(\boldsymbol{\beta}|\phi, \boldsymbol{y})} \propto p(\widehat{\boldsymbol{\beta}}(\phi), \phi|\boldsymbol{y})|\boldsymbol{V_\beta}|^{1/2}$$