# Inferencia Bayesiana

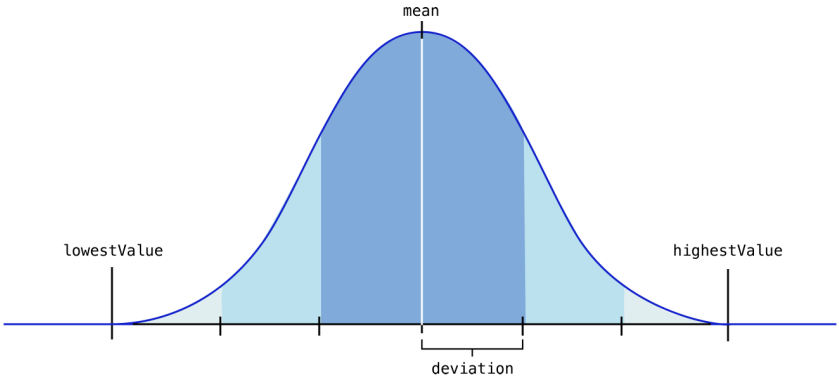
## Datos Moleculares II

Nelson R. Salinas

Septiembre 26, 2020

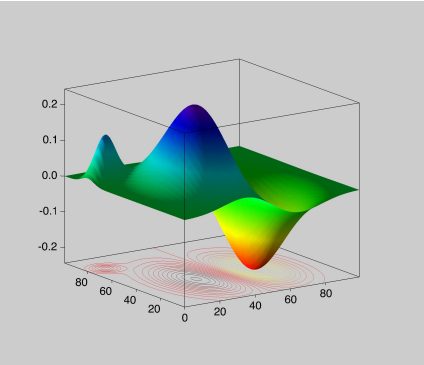
## Distribución estadística

Resumen de probabilidades de todos los posibles valores de un parámetro o un conjunto de parámetros.



## Distribución estadística

Resumen de probabilidades de todos los posibles valores de un parámetro o un conjunto de parámetros.



- Técnica estadística de estimación de parámetros.
- Resultados práctico $\rightarrow$ rango de distribución por parámetro.
- Cambio de paradigma estadístico.

Thomas Bayes (1701-1761)

**Teorema de Bayes**

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

donde

- $P(H|D)$: posterior o distribución *a posteriori*.
- $P(D|H)$: verosimilitud (ML).
- $P(H)$: prior o distribución *a priori*.
- $P(D)$: distribución marginal.

**Teorema de Bayes**

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

donde

- $P(H|D)$: probabilidad de que la hipótesis sea verdadera dados los datos.
- $P(D|H)$: verosimilitud de los datos dada la hipótesis (ML).
- $P(H)$: probabilidad de que la hipótesis sea verdadera.
- $P(D)$: probabilidad de todas las posibles combinaciones de datos.

**Teorema de Bayes**

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Problemas:

- Prior es una distribución subjetiva.
- Si verosimilitud domina sobre el prior, subjetividad del prior no es problemátiva. No siempre es el caso.
- Distribución marginal es computacionalmente intratable para problemas complejos (e.g., filogenética).

**Teorema de Bayes**

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Problemas:

- Prior es una distribución subjetiva.
- Si verosimilitud domina sobre el prior, subjetividad del prior no es problemátiva. No siempre es el caso.
- Distribución marginal es computacionalmente intratable para problemas complejos (e.g., filogenética).

**Markov chain Monte Carlo**

- Técnica para realizar inferencia bayesiana y evitar una estimación exhaustiva de la distribución marginal.

**Markov chain Monte Carlo**

- Técnica para realizar inferencia bayesiana y evitar una estimación exhaustiva de la distribución marginal.
- Se simulan secuencialmente parametros del modelo durante n iteraciones.

**Markov chain Monte Carlo**

- Técnica para realizar inferencia bayesiana y evitar una estimación exhaustiva de la distribución marginal.
- Se simulan secuencialmente parametros del modelo durante n iteraciones.
- En cada paso se estudia la probabilidad de que el nuevo set de parámetros se ajuste más a los datos que el anterior:

**Markov chain Monte Carlo**

- Técnica para realizar inferencia bayesiana y evitar una estimación exhaustiva de la distribución marginal.
- Se simulan secuencialmente parametros del modelo durante n iteraciones.
- En cada paso se estudia la probabilidad de que el nuevo set de parámetros se ajuste más a los datos que el anterior:

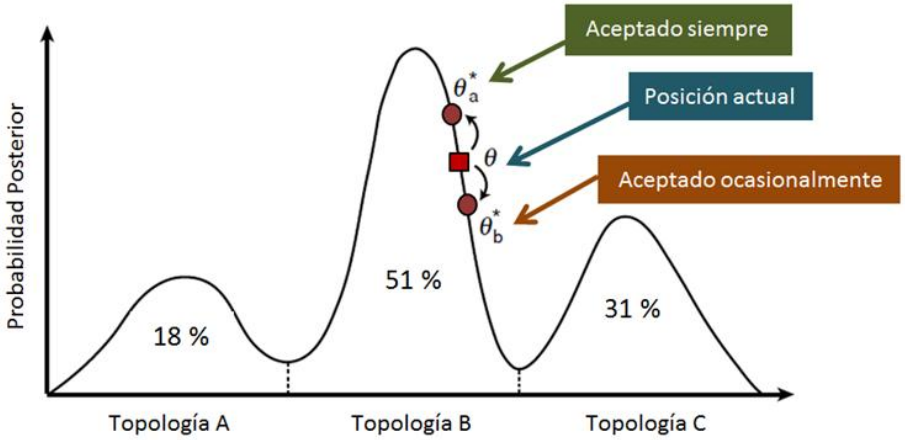
$$\frac{P(D|H_1)P(H_1)}{P(D)} \Big/ \frac{P(D|H_0)P(H_0)}{P(D)}$$

**Markov chain Monte Carlo**

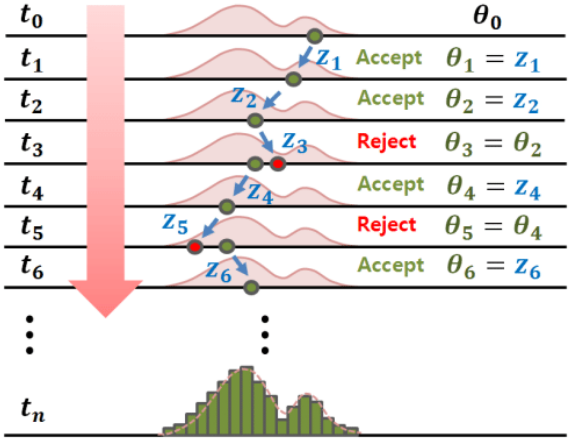
- Técnica para realizar inferencia bayesiana y evitar una estimación exhaustiva de la distribución marginal.
- Se simulan secuencialmente parametros del modelo durante n iteraciones.
- En cada paso se estudia la probabilidad de que el nuevo set de parámetros se ajuste más a los datos que el anterior:

$$\frac{P(D|H_1)P(H_1)}{\cancel{P(D)}} \Big/ \frac{P(D|H_0)P(H_0)}{\cancel{P(D)}}$$

**Markov chain Monte Carlo**



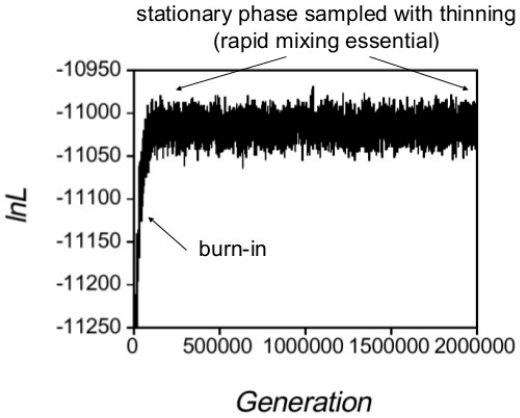
**Markov chain Monte Carlo**



**Markov chain Monte Carlo**

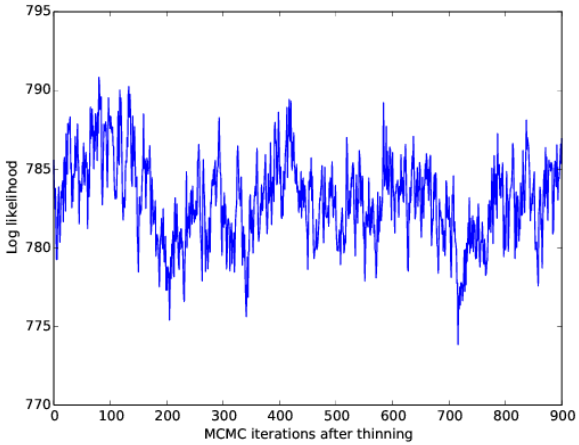
- El principal resultado es una tabla donde las columnas son los parámetros del modelo y las filas las iteraciones MCMC *muestreadas*.
- Se debe asegurar que MCMC haya explorado eficientemente el paisaje de posterior.
- Varias técnica para asegurar la calidad de MCMC:
  - Convergencia de varias réplicas.
  - Tamaño estimado de la población (*Estimated population size*) $> 200$.
  - Comportamiento de la cadena en figuras de muestreo.

- Análisis del reporte de parametros muestreados por MCMC.
- Eliminar 20% inicial de las cadenas (*burnin*).
- Resto de la cadena debe variar alrededor de la media, la varianza homogénea.
- Patrón "oruga peluda".

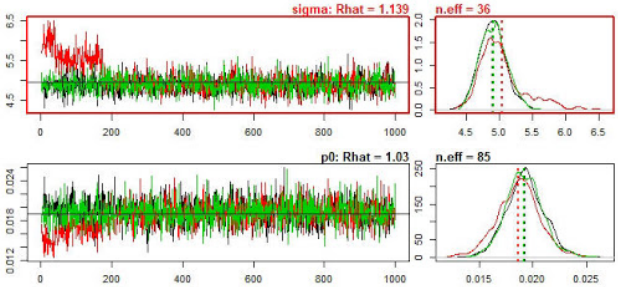


- Análisis del reporte de parametros muestreados por MCMC.
- Eliminar 20% inicial de las cadenas (*burnin*).
- Resto de la cadena debe variar alrededor de la media, la varianza homogénea.
- Patrón "oruga peluda".

## Mala convergencia de las cadenas



**Convergencia de multiples MCMC**




Los trazos y distribuciones de mcmc independientes deben estar sobrepuestos.

- Las MCMCs son aceptables, ¿donde está mi filogenia?
- Los resultados se consolidan en un solo arbol.
- Árbol de consenso: contiene los clados presentes en el 90% de las muestras MCMC.
- Árbol de máxima credibilidad: mayor probabilidad posterior.
- Cada nodo anotado con su respectiva frecuencia en el muestreo MCMC $\rightarrow$ probabilidad posterior del clado.