

---

# Inferencia Bayesiana

---

Modelos Bayesianos con aplicaciones ecológicas

Dr. Cole Monnahan

University of Concepción, Chile

Enero, 2019

# Resumen

- Para distribuciones continuas hay que integrar para obtener probabilidades
- Integración analítica es normalmente demasiado difícil
- Entonces se puede usar Monte Carlo como una flexible opción, pero a menudo (nunca?) la forma no es conocida
- Modelos Bayesianos resultan en distribuciones muy complejas que necesitan ser integrados

# Conceptos importantes

- La inferencia Bayesiana es un paradigma diferente que la frecuentista
- Las probabilidades son grados de creencia
- Se actualiza la creencia *a priori* con los datos
- La incertidumbre se cuantifica mediante probabilidades
- **Calculo de las probabilidades se requiere integración**

📌 Pinned Tweet



**\mathfrak{Michael Betancourt}** @betanalpha · 5 Jan 2017

Remember that using Bayes' Theorem doesn't make you a Bayesian. Quantifying uncertainty with probability makes you a Bayesian.

# La regla de Bayes

- Si  $\theta$  son los parámetros y  $y$  los datos (ambos son v.a)
- Entonces,  $P(\theta, y) = P(\theta)P(y|\theta)$
- Y  $P(\theta|y) = P(\theta, y)/P(y)$
- Combinando los...

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

- Posterior = (constant)(prior)(likelihood)

# Componentes de la regla de Bayes

- $P(\theta)$  = “Prior”: *la incertidumbre antes de experimento o conocimiento de un experto*
- $P(y|\theta)$  = “Likelihood”: *la verosimilitud de los datos dado los parámetros – lo mismo como clásica*
- $P(y)$  = *Una constante que no se puede calcular*
- $P(\theta|y)$  = “Posterior”: *la creencia que resulta de la combinación de dos fuentes da información: prior y datos.*
  - ❑ Es una distribución de probabilidad
  - ❑ La usamos para hacer inferencia

# Resumen de las diferencias de los paradigmas de inferencia.

	Frequentist	Bayesiana
Que es estimado?	$P(Y H)$ Datos dado el hipótesis	$P(H Y)$ Hipótesis dado los datos
La definición de probabilidad	frecuencias (infinitas) relativas de eventos	Grado de creencia
Fuentes de la información	Solo los datos	Los datos y información a priori
La definición de los parámetros	Estimaciones de cantidades “verdaderas”	Variables aleatorias estadísticas
Método de inferencia	Máximo verosimilitud	Integración (de posteriori)

# Las ventajas de inferencia Bayesiana

- Hay respuestas intuitivas: los parámetros son distribuciones de probabilidad.
- Poder formalmente incorporar conocimiento antes del experimento
- Las suposiciones asintóticas no son necesarios
- La estimación de los modelos jerárquicos es natural y fácil
- Análisis de decisión: Poder calcular probabilidades de las consecuencias de varias acciones. (Punt and Hilborn 1997)

# Desventajas

- Toma mas tiempo para estimar
- En general, la especificación de los priors
  - ❑ Poder ser sensitivo para la transformación de los parámetros. (e.g., Thorson and Cope 2017, Maunder 2003)
  - ❑ Poder ser difícil determinar apropiados “priors”
  - ❑ P.ej., no hay “uninformative priors”



# Priors

## ■ Cual es el rol de los priors?

Gelman et al. (2014):

1. Una población de valores posibles de los parámetros (perspectiva de la población)
2. Una declaración del conocimiento y incertidumbre de los parámetros (perspectiva del estado del conocimiento)

En ambos casos, la prior debe incluir todos los valores posibles – “*not in the prior not in the posterior*” .. Pero **ellas no pueden depender de los datos**

# La polémica

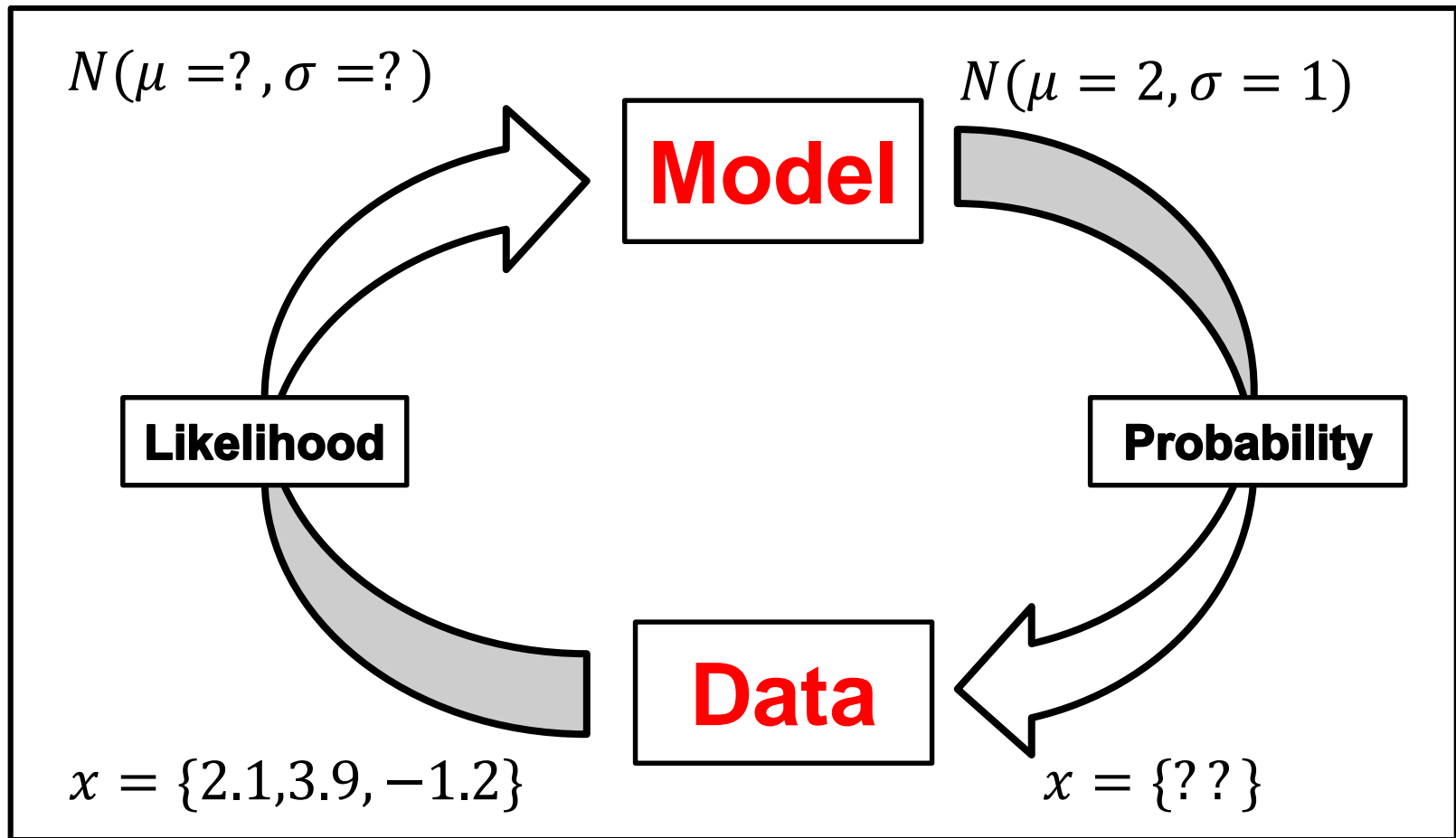
- Hay debates en la comunidad de los estadísticos por décadas
- Hay objeciones de priors “subjetivas” por el contrario de las decisiones “objetivas” con inferencia frecuentista
- Pero recientemente ha disminuido.. “*prior distributions are not necessarily any more subjective than other aspects of a statistical model*” (Gelman and Hennig 2017)
- Vamos a ignorarla y enfocar en aplicaciones

# Noninformative priors

- La idea es elegir una prior que resulta en un efecto pequeño de la posterior (*also reference, vague, or flat*)
- Puede permitir solo los datos guiar la inferencia a través de la verosimilitud
- “Improper priors” tiene una probabilidad infinita (e.g.,  $X \sim U(-\text{Inf}, \text{Inf})$ )
- Por otro lado, una “proper prior”

# Probabilities vs likelihoods

- La diferencia puede ser confusa



# Example: Normal likelihood

Probability (density):

$$Pr(\mathbf{x}|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}}$$

Likelihood:

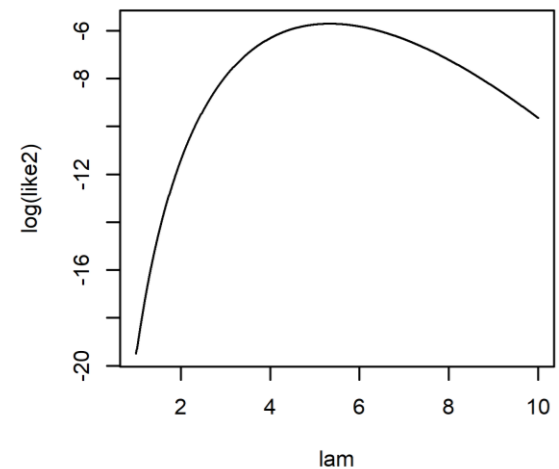
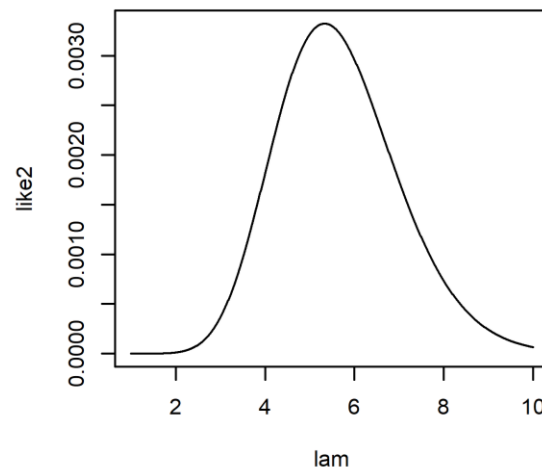
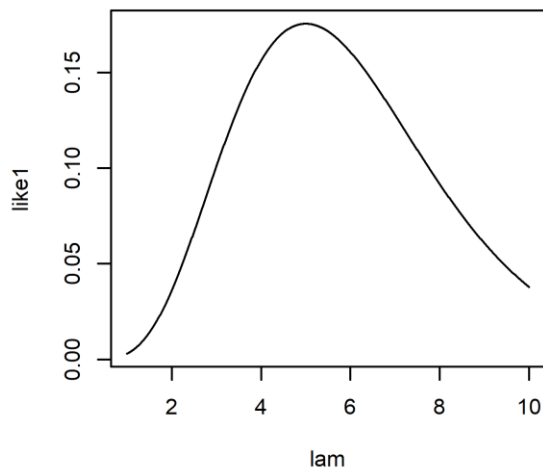
$$Pr(x; \mu, \sigma) = L(\mu, \sigma|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

But calculated the same in R!

```
dnorm(x=2, mean=3, sd=1)
```

# Exercise

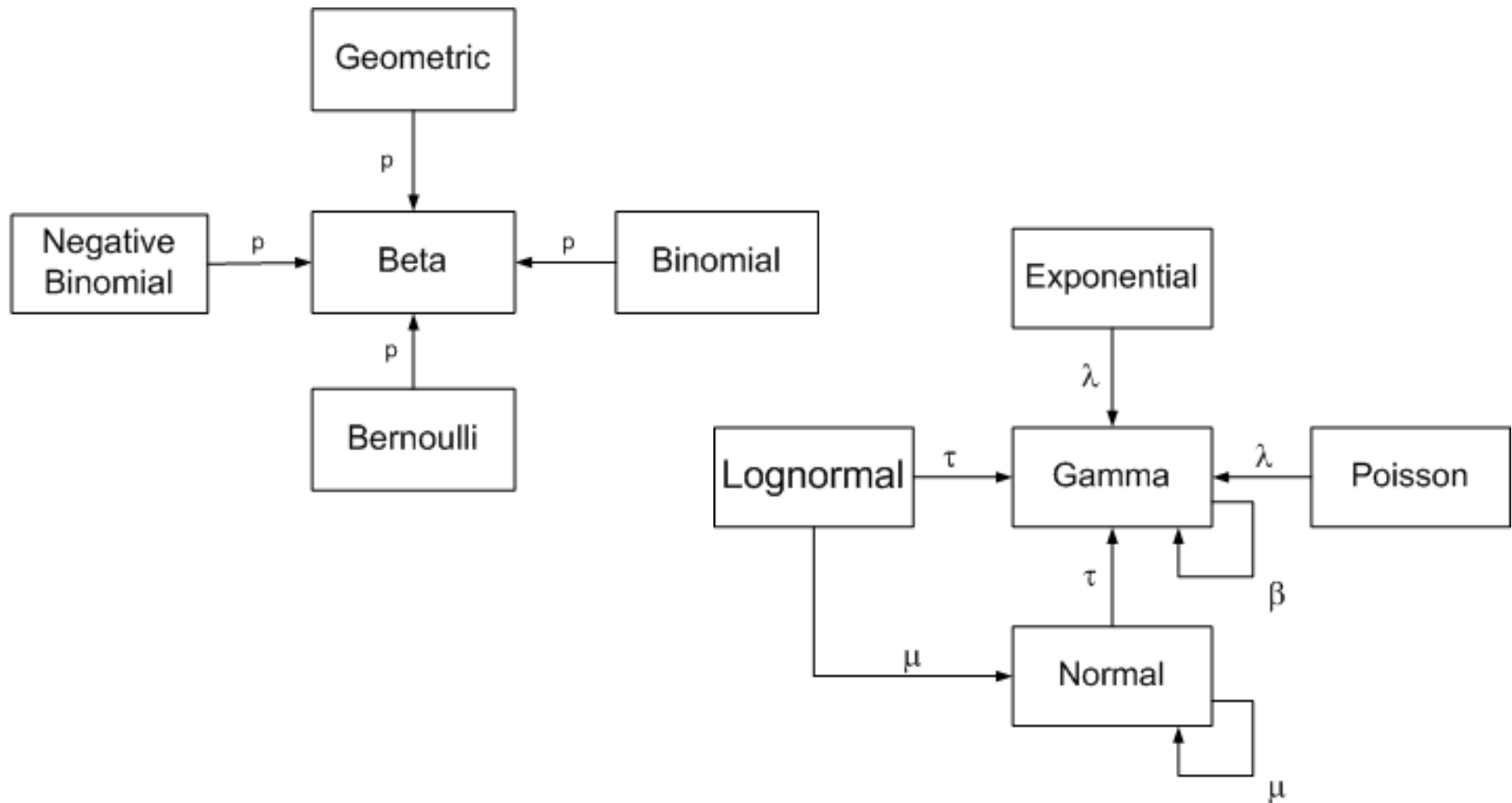
- Deja  $X \sim \text{Poisson}(\lambda)$  y una sola observación  $y=\{5\}$ . Crea una figura de la verosimilitud de  $\lambda$  de 1 a 10
- Repetirlo con  $y=\{5,4,7\}$  independiente y finalmente crea otra versión de la segunda que es un logaritmo de la verosimilitud.



# Conjugacy

- En casos muy específicos se puede reconocer la forma da la posterior
- Normal prior + normal likelihood = normal posterior
- Eso es “conjugacy” o “conjugate prior” (see [here](#))
- En tan casos la inferencia es fácil (p.ej.  $p_{\text{norm}}$ ) pero es muy raro
- Sin *conjugacy* se necesita otra manera de integración

# Conjugate examples





# Ejemplo I

- Suponga que usamos un solo dato ( $y$ ) de una distribución normal donde la media ( $\theta$ ) no es conocida pero la varianza sí:  $p(y|\theta) \sim N(\theta, \sigma)$ . La prior  $= p(\theta) \sim N(\mu_0, \tau_0)$

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

$$p(\theta | y) \propto \exp\left(-\frac{(y - \theta)^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{(\theta - \mu_0)^2}{2\tau_0^2}\right)$$

$$= \exp\left(-\frac{1}{2} \left[ \frac{(y - \theta)^2}{\sigma^2} + \frac{(\theta - \mu_0)^2}{\tau_0^2} \right]\right)$$

$$= \exp\left(-\frac{(\theta - \mu_1)^2}{2\tau_1^2}\right)$$

$$\mu_1 = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{1}{\sigma^2} y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

# Ejemplo I

- Suponga que usamos un solo dato ( $y$ ) de una distribución normal donde la media ( $\theta$ ) no es conocida pero la varianza sí:  $p(y|\theta) \sim N(\theta, \sigma)$ . La prior  $= p(\theta) \sim N(\mu_0, \tau_0)$

$$p(\theta | y) \propto \exp\left(-\frac{(\theta - \mu_1)^2}{2\tau_1^2}\right)$$

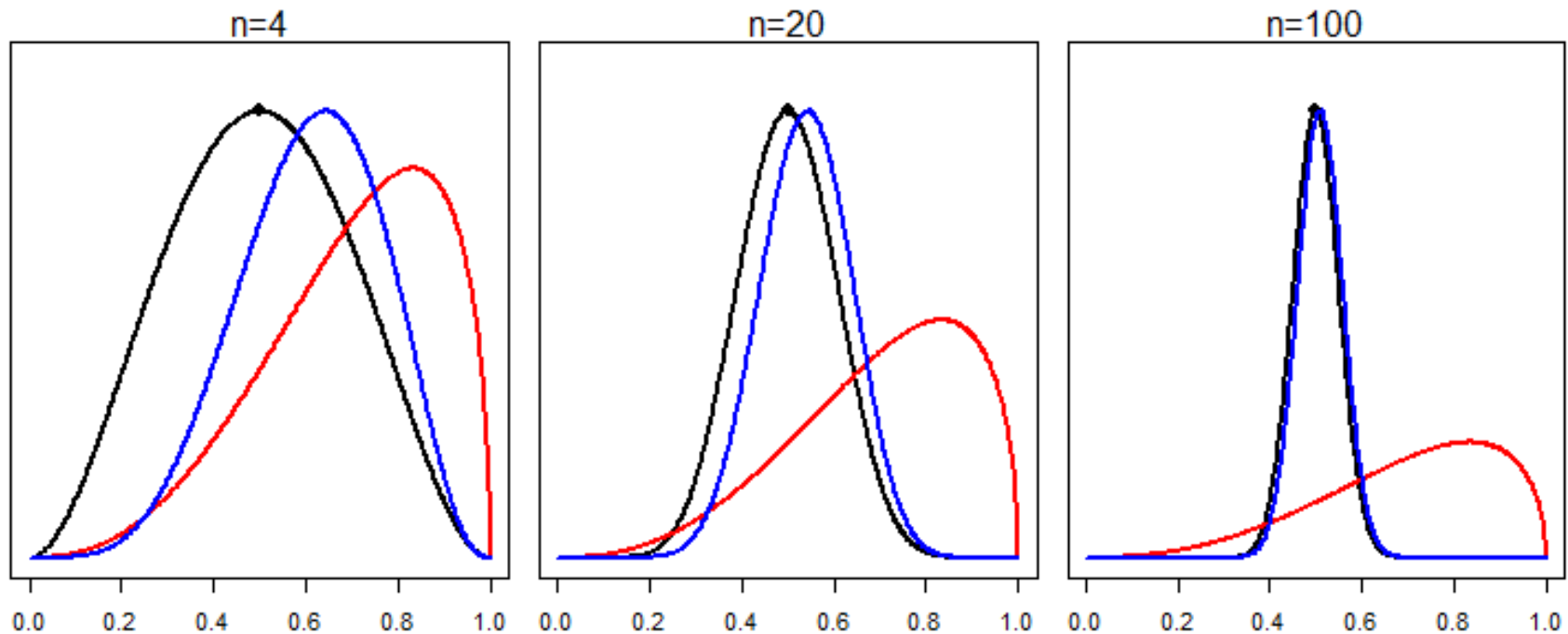
$$\mu_1 = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{1}{\sigma^2} y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

- Qué representa esta ecuación?
- Una distribución normal!  $N(\mu_1, \tau_1)$
- La media de la posterior “*es el promedio ponderado de la media de la distribución a priori y el dato*”
- [mostrar en R: prior, verosimilitud y posterior]

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

# Ejemplo II

- Posterior Beta-binomial=beta prior + binomial likelihood
- Supongamos que la mitad de los animales marcados mueren (queremos estimar sobrevivencia)
- Qué pasa al aumentar los datos sin cambiar la prior?



# Review of key concepts

- Actualizamos el conocimiento prior con los datos para formar la posterior
- Como todas las distribuciones, hay que integrarlas por inferencia (medianas, medias, cuantíales, etc.)
- Pero raramente tienen formas conocidas entonces no se puede usar Monte Carlo integración
- Entonces, como se puede integrarlas?

# Método 3: **Markov chain** Monte Carlo

- La idea principal es generar muestras aleatorias **correlacionadas** y calcular proporciones para aproximar probabilidades
- Usamos cadena de Márkov (Markov chains)
- Es un tipo especial de proceso estocástico en que cada evento depende *solamente* del evento inmediatamente anterior
- ..... Qué??

# Método 3: **Markov chain** Monte Carlo

- Un ejemplo:

$$X_{t+1} = X_t + U_t, \quad U_t \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$$

- X es una cadena de Markov



**Probability Fact**

@ProbFact

Following



One way to think of the Markov assumption:  
The future is independent of the past, given  
the present.

9:02 AM - 26 Dec 2018

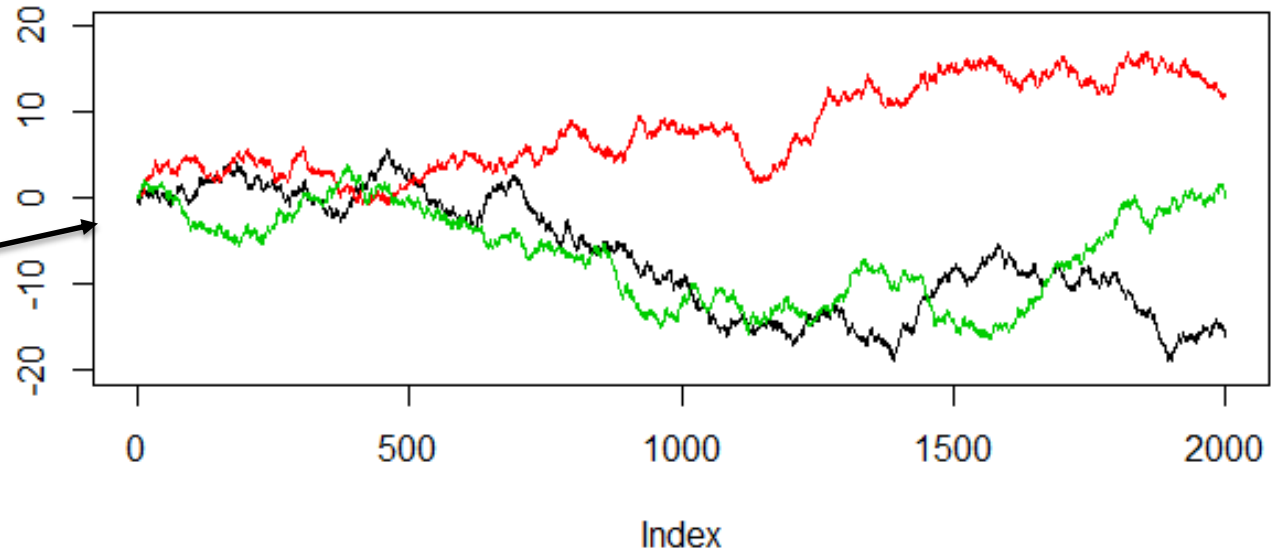
# Un ejemplo de cadena de Márkov simple

```
mc <- function(Niter) {  
  x <- rep(NA, Niter)  
  x[1] <- 0  
  for(i in 2:Niter) {  
    x[i] <- x[i-1] + runif(1, -.5, .5)  
  }  
  return(x)  
}
```

Estado inicial

Este estado  
depende *solamente* del  
evento anterior

“Random walk”  
comportamiento



# Método 3: **Markov chain** Monte Carlo

- Un ejemplo:

$$X_{t+1} = X_t + U_t, \quad U_t \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$$

- Esa cadena no es tan útil. No se puede usarla para hacer inferencia
- No es “Monte Carlo” en el sentido de inferencia Bayesiana
- Hay que cambiar la cadena un poco para usarla



# Una cadena de Márkov especial: MCMC

```
mcmc <- function(Niter, f, x0=0, U=1){
```

```
  x <- rep(NA,Niter)
```

```
  x[1] <- x0
```

Estado inicial

```
  for(i in 2:Niter){
```

```
    new <- x[i-1]+runif(1,-U,U)
```

El próximo estado depende *solamente* del evento anterior

```
    if( f(new)/f(x[i-1]) > runif(1)){
```

```
      ## Update or "accept" this new point
```

```
      x[i] <- new
```

```
    } else {
```

```
      ## Stay at previous point
```

```
      x[i] <- x[i-1]
```

Se acepta el estado nuevo depende de una condición aleatoria

```
    }
```

```
  }
```

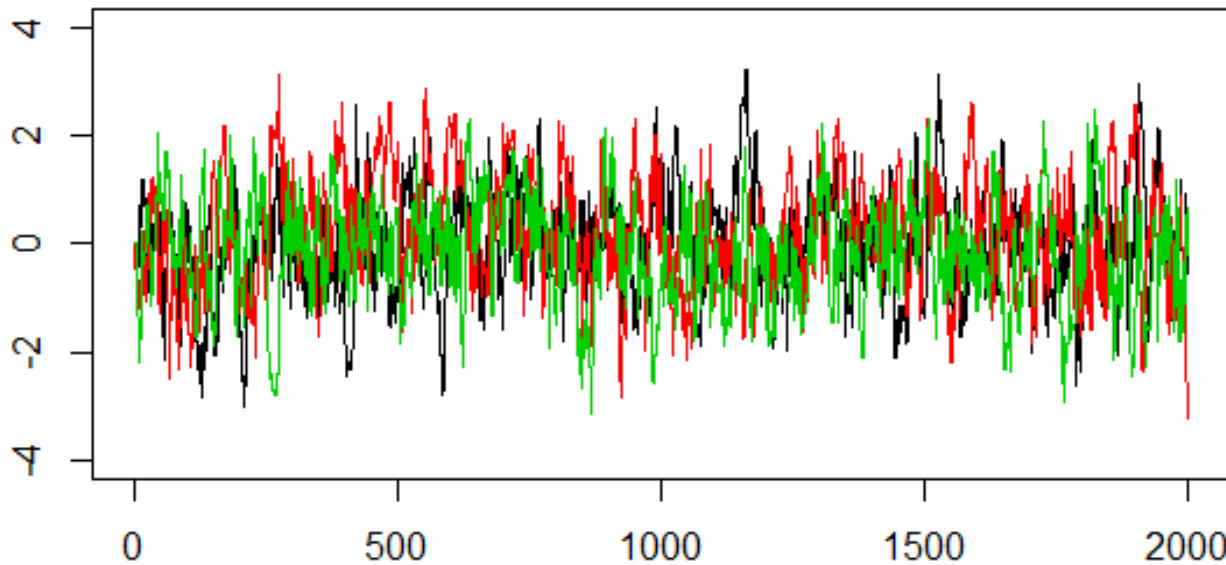
```
  return(x)
```

```
  f <- function(x) dnorm(x,0,1)
```

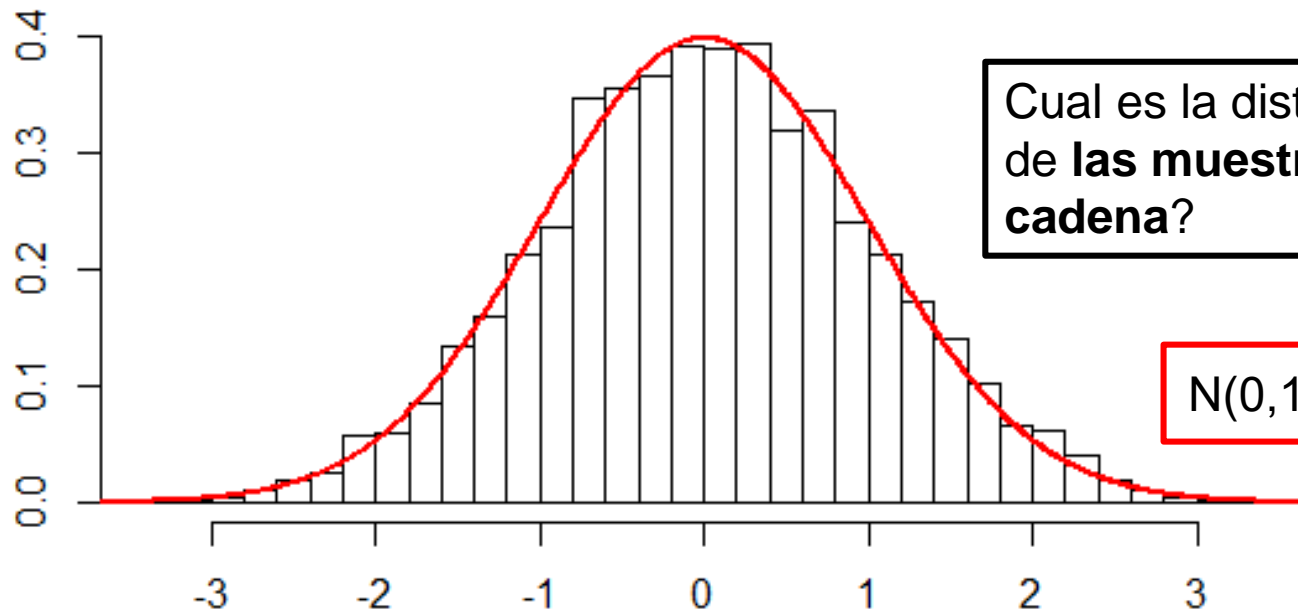
$f$  es la función de la densidad (PDF)

```
}
```

# Una cadena Márkov especial



Comportamiento  
diferente que  
anterior



Cual es la distribución  
de **las muestras de la  
cadena?**

$N(0,1)$ ... es  $\hat{f}$ !!

# Una cadena Márkov especial

- Notar que en la ecuación

$$c * f(\text{new}) / (c * f(x[i-1]))$$

- ... la constante (c) se cancelaría
- Significa que **no es necesario** conocer la constante para usar este método.
- Por eso podemos usarlo para aproximar las distribuciones a posteriori
- Este algoritmo se llama Metropolis-Hastings
- [Demonstrar con el ejemplo anterior]

## Método 3: **Markov chain** Monte Carlo

- Thus the strong law of large numbers applies to MCMC just as it does for Monte Carlo
- For integrable function  $h()$ , the average converges on its expectation

$$\frac{1}{T} \sum_{t=1}^T h(X^{(t)}) \longrightarrow \mathbb{E}_f[h(X)]$$

- So we can use the MCMC chain to approximate integrals just like with Monte Carlo

# Markov chain Monte Carlo

- Hay **muchos** tipos de cadenas de Márkov Monte Carlo:
  - Metropolis-Hastings, Gibbs, NUTS, slice sampling, etc.
- La idea es la misma: generar muestras para estimar probabilidades
- MCMC es lento, y hay algunas dificultades
- Las discutiremos durante del curso
- Pero MCMC es flexible y por eso es usado ampliamente en estadística Bayesiana

---

# Questions?

# Exercise

$$p(y | \theta) \sim N(\theta, \sigma = 1)$$

$$p(\theta) \sim N(\mu_0 = -2, \tau_0 = 0.5)$$

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

- Usa ejemplo 1:
- Usa mcmc función para generar muestras de la posterior:

```
samples <- mcmc(Niter=5000, f=posterior, x0=0, U=1)
```

- Contrasta a la solución analítica:

$$\mu_1 = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{1}{\sigma^2} y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

$$p(\theta | y) \sim N(\mu_1, \tau_1)$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- Estima 95% *credible interval* usando ambas maneras de la integración

---

# References

- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. Bayesian data analysis. Taylor & Francis.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21:1087-1092**.
- Ellison, A. M. (2004). "Bayesian inference in ecology." *Ecology Letters* 7(6): 509-520.