

# Introducción al Análisis Bayesiano con Aplicaciones en STAN

Ignacio Alvarez-Castro

conectaR

-

Universidad de Costa Rica

-

23 al 26 de Enero 2019

Introducir el paradigma Bayesiano para hacer inferencia con una perspectiva aplicada, que los participantes obtengan una guía a seguir para implementar un análisis Bayesiano en sus problemas de interés.

- Presentar la naturaleza de la inferencia Bayesiana y sus bases teóricas
- Introducción al cómputo Bayesiano utilizando STAN
- Introducción a la modelización Bayesiana: Tasa de Mortalidad Infantil

Ignacio Alvarez-Castro

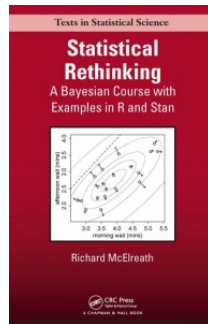
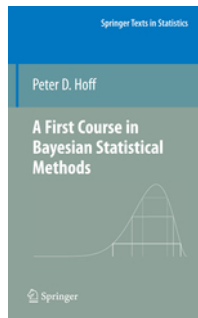
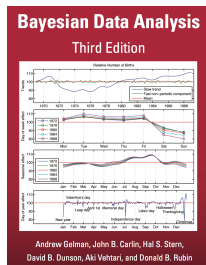
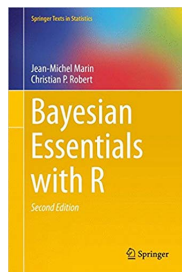
- Instituto de Estadística (IESTA), Universidad de la República, Uruguay
- `nachalca@iesta.edu.uy`
- `@nachalca`
- [https://github.com/nachalca/conectar\\_introBayes](https://github.com/nachalca/conectar_introBayes)



FACULTAD DE  
CIENCIAS ECONÓMICAS  
Y DE ADMINISTRACIÓN



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY





**COBAL 2019**

@cobal2019

Siguiendo



VI Congreso Bayesiano de América Latina (VI COBAL)

Lima, Perú, 19 al 21 Junio 2019.

[sites.google.com/site/cobal2019/](https://sites.google.com/site/cobal2019/)

@ISBA\_events #Bayesian #LatinAmerica  
#Peru #Statistics

1:58 - 18 sept. 2018

4 Retweets 6 Me gusta



6



- 1 Inferencia Bayesiana
- 2 Cómputo Bayesiano
- 3 Introducción a STAN
- 4 Modelos jerárquicos
- 5 Tasa de Mortalidad Infantil en Uruguay

La inferencia estadística se ocupa de elaborar métodos para obtener conclusiones basadas en datos observados

Si  $y = (y_1, y_2, \dots, y_n)$  corresponden a observaciones del fenómeno de interés:

- Suponemos que  $y$  corresponde a la realización de la *variable aleatoria*  $Y$  con función de densidad o masa  $p(y)$ .
- Objetivos:
  - *Explicar* características relevantes de  $Y$
  - *Predecir* el valor de observaciones futuras
  - *Comparar* modelizaciones alternativas

Hoy hablaremos únicamente sobre modelos *paramétricos*.

El término inferencia Bayesiana se refiere al paradigma teórico utilizado para realizar inferencia estadística (explicar o predecir fenómenos estocásticos).

- $y$ : los datos observados
- $M$ : aspectos de la realidad que se asumen conocidos
- $\theta$ : lo que no conocemos del problema



El término inferencia Bayesiana se refiere al paradigma teórico utilizado para realizar inferencia estadística (explicar o predecir fenómenos estocásticos).

- $y$ : los datos observados
- $M$ : aspectos de la realidad que se asumen conocidos
- $\theta$ : lo que no conocemos del problema

Basar nuestras decisiones sobre las cantidades desconocidas en la distribución de probabilidad condicional a los datos observados y otros aspectos conocidos del problema.

$$p(\theta|y, M)$$

Un modelo estadístico Bayesiano está formado por dos componentes:

Modelo para los datos  $p(y|\theta)$

Previa  $p(\theta)$

La estimación de  $\theta$  consiste en *hallar la distribución posterior*  $p(\theta|y)$

Hallar  $p(\theta|y)$  mediante la regla de Bayes:

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

- analíticamente, en problemas simples
- numéricamente: aproximando en una grilla de valores de  $\theta$
- numéricamente: mediante simulaciones

(tomado de McElreath (2016))

**Objetivo:** Estimar la proporción de superficie cubierta por AGUA en el planeta.

**Experimento:** tirar el globo hacia arriba y cuando lo agarramos registramos si el dedo índice de la mano derecha quedó sobre agua o sobre tierra.

Datos de 9 tiradas:

```
obs <- c('AA', 'TT', 'AA', 'AA', 'AA', 'TT', 'AA', 'TT', 'AA')  
y <- as.numeric(obs == 'AA')
```

- $Y_i$  vale 1 si el resultado de la tirada  $i$  es AGUA y 0 si es TIERRA.
- $\theta$  representa la probabilidad de observar AGUA en una tirada

**Modelo para los datos:**  $Y_i \stackrel{ind}{\sim} \text{Ber}(\theta)$ ,  $n = 9$ ,  $\sum_{i=1}^n y_i = n_1$

$$p(y|\theta) = \prod_{i=1}^n p(y_i|\theta) = \theta^{n_1} (1 - \theta)^{n - n_1}$$

**Previo:** ¿Cómo representamos nuestra incertidumbre sobre  $\theta$  ?

- $Y_i$  vale 1 si el resultado de la tirada  $i$  es AGUA y 0 si es TIERRA.
- $\theta$  representa la probabilidad de observar AGUA en una tirada

**Modelo para los datos:**  $Y_i \stackrel{ind}{\sim} Ber(\theta)$ ,  $n = 9$ ,  $\sum_{i=1}^n y_i = n_1$

$$p(y|\theta) = \prod_{i=1}^n p(y_i|\theta) = \theta^{n_1} (1 - \theta)^{n - n_1}$$

**Previa:** ¿Cómo representamos nuestra incertidumbre sobre  $\theta$  ?

Por ahora asignemos “igual probabilidad” a todos sus posibles valores:

$$\theta \sim Unif(0, 1)$$

Mediante regla de Bayes:

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

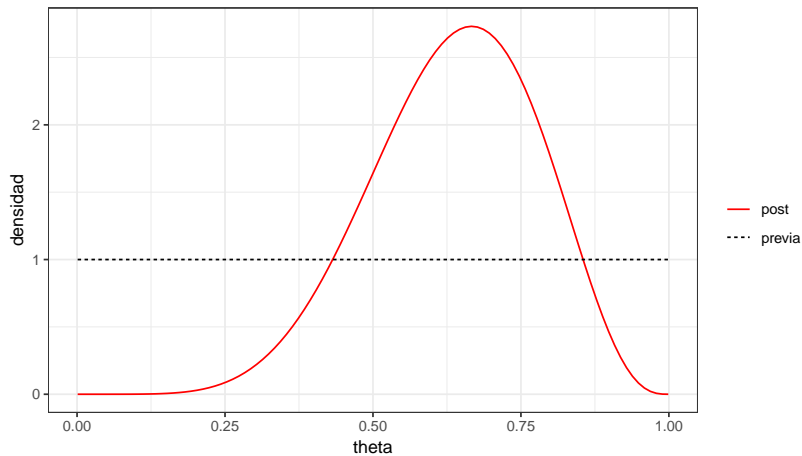
Notar que  $p(y)$  es constante respecto a  $\theta$ , al obtener  $p(\theta|y)$  podemos evitar calcular dicha constante,

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

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

$$\propto \theta^{n_1} (1 - \theta)^{n - n_1} \times I_{\theta \in (0,1)}$$

$$\theta|y \sim \text{Beta}(n_1 + 1, n - n_1 + 1)$$



- representa la incertidumbre sobre  $\theta$  antes de observar  $y$
- permite incorporar información previa en la modelización

## Seleccionar la previa

- 1 Hacer uso de información “genuina”: en experimentos secuenciales, conocimiento de expertos
- 2 Automáticas: incorporan la menor información posible (Jeffreys, objetivas, referencia)
- 3 Débilmente informativas: regularizar inferencias extremas
- 4 Convenientes: seleccionadas ad hoc, computacionalmente convenientes o referenciadas en la literatura. (ejemplo: Previas conjugadas)



Estimar  $\theta$  significa hallar su distribución posterior  $p(\theta|y)$ . Sin embargo, muchas veces conviene resumir la distribución

- 1 Estimación puntual,  $\hat{\theta}_{bayes}$ :

$$\hat{\theta}_{bayes} = \operatorname{argmin}_{\theta} \left\{ \int L(\theta, \hat{\theta}) p(\theta|y) d\theta \right\}$$

$L(\theta, \hat{\theta})$  función de pérdida que determina  $\hat{\theta}_{bayes}$ . Medidas de posición central son muy usadas (esperanza, mediana, moda)

- 2 Intervalo de credibilidad:  $P(\theta \in (a, b)|y) = \int_a^b p(\theta|y) d\theta = \alpha$ 
  - Definido con percentiles (  $(\theta|y)_{\alpha/2}, (\theta|y)_{1-\alpha/2}$  )
  - Región de máxima posterior:  $\{\theta : p(\theta|y) \geq k\}$
- 3 Probabilidad de eventos de interés:  $Pr(\theta \in A|y) = \int_A p(\theta|y) d\theta$

Aparte del objetivo de estimar  $\theta$  (hallar  $p(\theta|y)$ ), podemos estar interesados en predecir futuras observaciones (como objetivo en sí mismo o para evaluar si el modelo es adecuado)

Distribución predictiva posterior:

- $\tilde{y}$ : una nueva observación del fenómeno de estudio.
- $\tilde{y}|\theta$  es independiente de los datos observados  $y_i$
- el modelo de los datos es el mismo,  $\tilde{y} \sim p(y|\theta)$

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

- predicciones puntuales o por intervalos en base a  $p(\tilde{y}|y)$
- comparando  $\tilde{y}$  con  $y$  se puede evaluar el ajuste del modelo

Las inferencias basadas en la posterior dependen del modelo:  $p(y|\theta)$  y  $p(\theta)$ . Es necesario evaluar que tan razonables son esos supuestos.

## Checks posterior predictiva

- Sirve para evaluar que tan bien un modelo se ajusta a los datos observados.
- comparaciones en base a gráficos
- *p-valores* posterior predictivos
- otras comparaciones: comparar simulaciones de la posterior con la previa

## Comparación de modelos

- Análisis de sensibilidad
- Medir el poder predictivo del modelo

Considera el modelo:

$$\begin{aligned} y_i &\overset{\text{indep}}{\sim} \text{Poisson}(n\lambda) \\ \lambda &\overset{\text{indep}}{\sim} \text{Gamma}(a, b) \end{aligned}$$

¿Puedes mostrar que  $E(\lambda|y) = \frac{a+n\bar{y}}{b+n}$  ?

(tomado de Gelman et al. (2013) )

- 1 Proponer un modelo de probabilidad *completo*: para todas las cantidades de interés (observadas y no observadas)
- 2 Obtener la *distribución posterior* de interés. Es decir, la distribución de probabilidad de cantidades no observadas de interés condicional a las cantidades observadas.
- 3 Evaluar los resultados del modelo contenidos en la distribución posterior, evaluar los supuestos del modelo. Eventualmente modificar o expandir el modelo y repetir el proceso.

- 1 Inferencia Bayesiana
- 2 Cómputo Bayesiano**
- 3 Introducción a STAN
- 4 Modelos jerárquicos
- 5 Tasa de Mortalidad Infantil en Uruguay

Gran cantidad de objetivos de inferencia pueden representarse como

$$E_{\theta|y}[h(\theta)|y] = \int h(\theta)p(\theta|y)d\theta$$

donde  $h(\theta)$  es la cantidad de interes.

se obtiene

- analíticamente, en problemas simples
- numéricamente: grilla de valores, mediante simulaciones

**Monte Carlo:** usar valores simulados de la posterior:

$$\{\theta^1, \theta^2, \dots, \theta^S\}$$

donde  $\theta^i \sim p(\theta|y)$

Monte Carlo se basa en la Ley Fuerte de los grandes números

$$\frac{1}{S} \sum h(\theta^i) \longrightarrow \int h(\theta) p(\theta|y) d\theta = E(\theta|y)$$

Infinidad de algoritmos para obtener las simulaciones.

Dos grandes formas:

- Monte Carlo, simulaciones/realizaciones **independientes**
- Monte Carlo, simulaciones/realizaciones **dependientes**



Monte Carlo basado en cadenas de Markov **MCMC**:  
los valores simulados

$$\{\theta^1, \theta^2, \dots, \theta^n\}$$

forman una cadena de Markov, cuya distribución estacionaria es la posterior de interés  $p(\theta|y)$

Una cadena de Markov es una secuencia de variables aleatorias dependientes

$$X^1, X^2, \dots, X^t, \dots$$

tal que la distribución de  $X^t$  dada las variables pasadas sólo depende de  $X^{t-1}$ ,

$$p(X^t|X^1, \dots, X^{t-1}) = p(X^t|X^{t-1})$$

Si la cadena es **ergódica** entonces se cumple que

$$\frac{1}{S} \sum h(X^t) \longrightarrow \int h(x)p(x)dx$$

donde  $p(x)$  es la distribución estacionaria de la cadena.

Esto implica que podemos trabajar con las simulaciones de la cadena de igual forma que con simulaciones independientes (Monte Carlo).

$$\{\theta^1, \theta^2, \dots, \theta^S\}$$

Debemos definir:

- $m$ : número de cadenas independientes
- $\theta^0$ : valores iniciales para cada cadena
- $S$ : cantidad de iteraciones en cada cadena

Una estrategia

- Obtener simulaciones de varias cadenas,  $m = 3$  o  $4$
- Usar valores **dispersos** y descartar las iteraciones iniciales (warm up)
- Determinar si las iteraciones son suficientes mediante  $\hat{R}$  y  $n_{eff}$

Reduccion potencial en la varianza:

$$\hat{R} = \sqrt{\frac{\text{var}^+(\theta|y)}{W}}$$

representa cuanto se puede reducir la varianza si aumentamos las iteraciones.

Número de muestras efectivo

$$n_{\text{eff}} = \frac{n \times m}{\left(1 + 2 \sum_{k=1}^{\infty} \rho_k\right)}$$

representa la cantidad de muestras equivalentes si fueran independientes  
( $\rho_k = \text{corr}(\theta^t, \theta^{t-k})$ ).

- 1 Inferencia Bayesiana
- 2 Cómputo Bayesiano
- 3 Introducción a STAN**
- 4 Modelos jerárquicos
- 5 Tasa de Mortalidad Infantil en Uruguay

Obtener muestras de  $p(\theta|y)$  en cualquier modelos

$$\{\theta^1, \theta^2, \dots, \theta^S\}$$

- lenguaje de programación probabilística
- Utiliza Hamiltonian Monte Carlo
- Principalmente para inferencia Bayesiana
- interface con varios programas

No es necesario salir de R

```
install.packages("rstan",  
                 repos = "https://cloud.r-project.org/",  
                 dependencies = TRUE)
```

<https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started>

Hay muchas otras opciones y software para realizar inferencia Bayesiana:

- Obtener MCMC (BUGS, JAGS, NIMBLE, Greta)
- Manipular MCMC (coda, bayesplot, loo, shinystan)
- Modelos particulares (ARM, INLA, rstanarm, brms)
- Task View <https://cran.r-project.org/web/views/Bayesian.html>

Modelo:  $Y_i \overset{iid}{\sim} N(\mu, \sigma^2) \quad \mu \sim t_3(0, 1) \quad \sigma \sim Ca^+(0, 1)$

3 pasos:

- 1 escribir el modelo en lenguaje STAN, `inicial.stan`
- 2 compilar el modelo, para obtener un programa en C++
- 3 ejecutar el programa, para obtener muestras de la distribución posterior



Modelo:  $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$   $\mu \sim t_3(0, 1)$   $\sigma \sim Ca^+(0, 1)$

```
data {  
  // bloque de datos  
  int<lower=1> n;  
  real y[n];  
}  
parameters{  
  // bloque para definir parametros y su espacio  
  real mu;  
  real<lower=0> sigma;  
}  
model {  
  // Bloque del modelo: previas y modelo para datos  
  sigma ~ cauchy(0, 1);  
  mu ~ student_t(3, 0, 1);  
  
  y ~ normal(mu, sigma);  
}
```

Modelo:  $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$   $\mu \sim t_3(0, 1)$   $\sigma \sim Ca^+(0, 1)$

```
# cargamos la libreria y simulamos datos
library(tidyverse)
library(rstan)
rstan_options(auto_write = TRUE)

simulados <- data_frame(ys = rnorm(500, mean = 20, sd = 3))

# compilamos el modelo: se genera un objeto con el codigo
m = stan_model(file = 'inicial.stan')

# ponemos los datos en una lista con nombres
dt.ls <- with(simulados, list(n = length(ys), y = ys) )

# Obtenemos simulaciones de la posterior
res = sampling(m, data = dt.ls )
```

Modelo:  $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2) \quad \mu \sim t_3(0, 1) \quad \sigma \sim Ca^+(0, 1)$

```
# resumen de resultados
res
summary(res)

# varios dibujos posibles
plot(res, plotfun = 'trace') # iteraciones
plot(res, plotfun = 'rhat') # Reduccion varianza potencial
plot(res, plotfun = 'ess')  # Muestras efectivas

plot(res) # intervalos de credibilidad
plot(res, plotfun = 'hist') # histograma posterior

# Obtener los valores simulados de p(mu,sigma | y)
rstan::extract(res)
```

¿Puedes escribir el modelo que utilizamos para los datos del experimento con el globo terraqueo?

$$\text{Modelo: } Y_i \stackrel{iid}{\sim} \text{Bin}(9, \theta) \quad \theta \sim \text{Unif}(0, 1)$$

3 pasos:

- 1 escribir `globo.stan`
- 2 compilar el modelo con `stan_model()`
- 3 obtener posterior con `sampling()`

(2 y 3 se pueden combinar usando `stan()` )

- 1 Inferencia Bayesiana
- 2 Cómputo Bayesiano
- 3 Introducción a STAN
- 4 Modelos jerárquicos**
- 5 Tasa de Mortalidad Infantil en Uruguay

Llamamos **modelo jerárquico** a modelos en los que agregamos previas para los parámetros de las distribuciones que **NO** aparecen en el modelo para los datos.

$$\begin{array}{ll} y_j = (y_{j,1}, \dots, y_{j,n_j}) & \overset{ind}{\sim} p(y|\theta_j) \\ \theta_j & \overset{ind}{\sim} p(\theta|\phi) \\ \phi & \sim p(\phi) \end{array}$$

- $y_j = (y_{j,1}, \dots, y_{j,n_j})$  observaciones para el grupo  $j$
- $n_j$  es la cantidad de observaciones en el grupo  $j$

Llamamos **modelo jerárquico** a modelos en los que agregamos previas para los parámetros de las distribuciones que **NO** aparecen en el modelo para los datos.

$$\begin{array}{ll} y_j = (y_{j,1}, \dots, y_{j,n_j}) & \overset{\text{ind}}{\sim} p(y|\theta_j) \\ \theta_j & \overset{\text{ind}}{\sim} p(\theta|\phi) \\ \phi & \sim p(\phi) \end{array}$$

- $y_j = (y_{j,1}, \dots, y_{j,n_j})$  observaciones para el grupo  $j$
- $n_j$  es la cantidad de observaciones en el grupo  $j$
- $p(y|\theta_j)$  controla la variabilidad *al interior* de cada grupo
- $p(\theta|\phi)$  controla la variabilidad *entre* grupos
- $p(\phi)$  representa información previa sobre  $\phi$

Aparecen cuando los datos tienen un *agrupamiento* natural. Por ejemplo, datos con estructura espacial y/o temporal, o experimentos con varias medidas por unidad experimental, etc.

Si nos interesa un parámetro *para cada grupo*, nos conviene aprovechar la estructura de los datos.

- $y_{ij}$  observación para la unidad  $i$  en el grupo  $j$
- $y_{ij} \sim p(y|\theta_j)$ ,  $\theta_j$  parámetro de interés en el grupo  $j$
- $\theta_j$  son intercambiables

Es natural un modelo jerárquico en esta situación



La estimación (clásica) de modelos jerárquicos puede ser difícil

- teoría asintótica requiere que  $n_j$  y  $J$  sean grandes
- la cantidad de parámetros a estimar crece con los datos
- no hay p-valor en `lmer`, Douglas Bates dice  
<https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html>

Inferencia Bayesiana tiene ventajas aquí

- es válido para muestras finitas
- aprovecha estructura de dependencia entre parámetros

Usando MCMC el método de inferencia (Bayesiana) no cambia, pero puede ser computacionalmente costoso

- 1 Inferencia Bayesiana
- 2 Cómputo Bayesiano
- 3 Introducción a STAN
- 4 Modelos jerárquicos
- 5 Tasa de Mortalidad Infantil en Uruguay**

TMI es la proporción de niños que muere con menos de 1 año de edad.

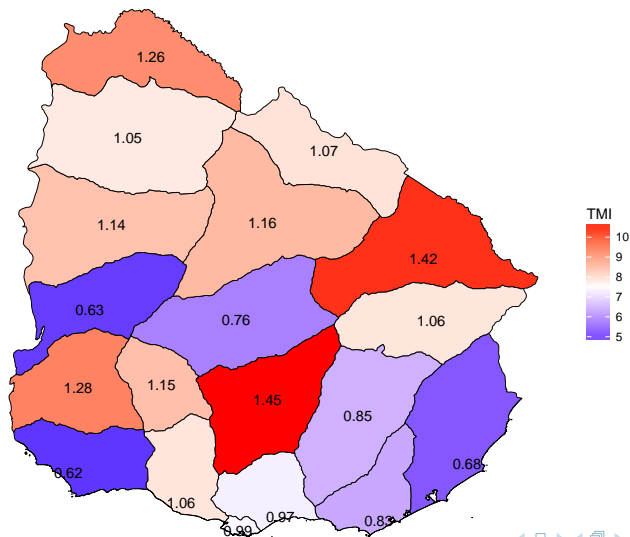
En 2015 en Uruguay hubo 48926 nacimientos, 367 de esos niños fallecieron antes de cumplir un año:  $TMI = 0,0075$  o 7,5 por mil.

Queremos modelar TMI en cada departamento (región) de Uruguay

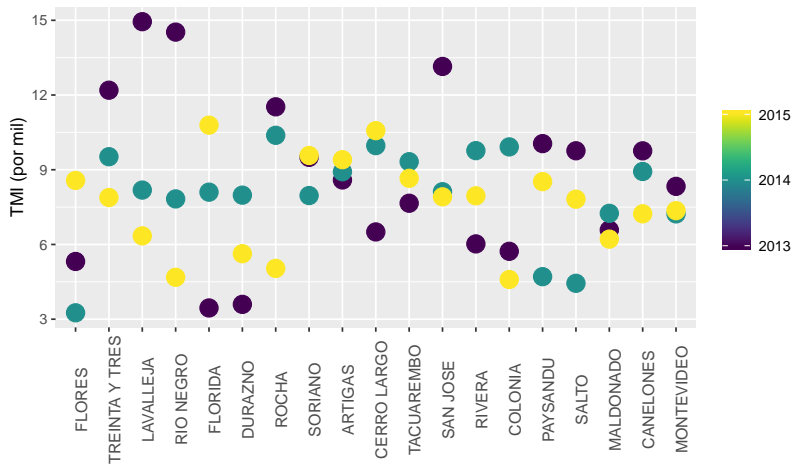
- $y_{it}$ : defunciones en la región  $i$  para el año  $t$
- $N_{it}$ : nacimientos en la región  $i$  para el año  $t$
- Hay 19 regiones,  $i = 1, \dots, 19$  y 3 años,  $t = 2013, 2014, 2015$

# Mapa de TMI en Uruguay - 2015

En cada departamento se presenta el correspondiente SMR



# TMI para 3 años



$$\begin{aligned}y_{it} &\overset{\text{indep}}{\sim} \text{Poisson}(N_{it} * \frac{TMI_i}{1000}) \\TMI_i &\overset{\text{indep}}{\sim} \text{Gamma}(\alpha, \beta) \\ \alpha &\sim \text{Exponencial}(1) \\ \beta &\sim \text{Exponencial}(1)\end{aligned}\tag{1}$$

- James-Stein:  $TMI_i$  son estimadas con menor *riesgo*.
- Flexibilidad: “estimamos” la previa
- Varias limitaciones (ej: no hay efecto en el tiempo)
- Ver Ugarte et al. (2009)

¿Puedes sugerir alguna modelización alternativa para este problema?

¿Puedes sugerir alguna modelización alternativa para este problema?

NOS VAMOS A R !

- `mdjer_poigam.stan` tiene el código del modelo (1)
- `tasas.R` tiene el código para analizar los datos



Muchas gracias

- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian data analysis*, Chapman and Hall/CRC.
- Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, Springer Science & Business Media.
- Marin, J.-M. and Robert, C. P. (2014), *Bayesian essentials with R*, vol. 48, Springer.
- McElreath, R. (2016), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, vol. 122, CRC Press.
- Ugarte, M. D., Goicoa, T., and Militino, A. F. (2009), "Empirical Bayes and Fully Bayes procedures to detect high-risk areas in disease mapping," *Computational Statistics & Data Analysis*, 53, 2938–2949.