Sesion 03 - Inferencia en distribuciones tipo mixtas - Parte 1

Juan Carlos Martinez Ovando

February 5, 2019

Objetivo

- Estudiar las implicaciones de la `propiedad de separacion` en el proceso inferencia tipo mixto.
- Incorporar `datos` en el proceso inferencial.
- Incorporar `datos` e `informacion complementaria` en el proceso inferencial.

Preambulo

Seguimos considerando la clase de modelos tipo mixto con soporte en $\mathcal{X} = \{0\} \cup (0, \infty)$, y "densidad" dada por

$$f(x|\theta_0, \theta_c) = \theta_0 \mathbb{I}(x = 0) + (1 - \theta_0)\theta_c \exp\{-\theta_c x\}\mathbb{I}(x > 0).$$

Como vimos, esta clase de modelos puede verse como una composicion considerando la variable auxiliar

$$Z = \begin{cases} 1, & \text{si } x = 0, \text{ con probabilidad } \theta_0, \\ 0, & \text{si } x > 0, \text{ con probabilidad } (1 - \theta_0). \end{cases}$$

Problema inferencial

Si consideramos un conjunto de `datos` dado por una coleccion de J observaciones/polizas que toman valores en $\mathcal{X}$, digamos

$$\text{datos} = \{x_1, \ldots, x_J\},$$

tendremos el **objetivo** de encontrar los valores de

$$\theta_0 \text{ y } \theta_c,$$

que **sean mas compatibles con los datos**.

> Recordemos que como $x_i \in \mathcal{X}$, esperaremos que varios x_is compartan el valor 0.

Datos

Consideremos los datos reales de la empresa **AllState** de un portafolio de seguros de autos.

Cargamos los datos desde un repositorio en GitHUb (en este caso, el repositorio de datos de nuestro curso, referido como `JCMO-ITAM/Data4Analysis`). Para esto, empleamos en RStudio el paquete `repmis`. El diccionario de datos se encuentra en el mismo repositorio.

```
if(!require("repmis")){install.packages("repmis")}
library("repmis")

data <- source_data("https://github.com/JCMO-ITAM/Data4Analysis/blob/master/d4a_allstateclaim_data.csv?
```

Resumen

Los `datos` incluyen las siguientes variables:

```
## Loading required package: repmis
```

```
## Downloading data from: https://github.com/JCMO-ITAM/Data4Analysis/blob/master/d4a_allstateclaim_data
```

```
## SHA-1 hash of the downloaded data file is:
## f99c63d65351dd1ff9e67aa3c66c94f5d9139f22
```

```
## [1] "Household_ID"  "Vehicle"       "Calendar_Year" "Model_Year"
## [5] "Blind_Make"    "Claim_Amount"
```

La variable `Claim_Amount` representa el monto de reclamo individual, nuestros valores x_i.

Los casos `Claim_Amount==0` representan no siniestro en la poliza correspondiente.

J es el numero de polizas en los `datos`, y n es el numero de polizas con no siniestro.

```
## [1] 330065
```

```
## [1] 327064
```

Verosimilitud

La funcion de verosimilitud para (θ_0, θ_c) con base en los datos es (bajo el supuesto `iid`),

$$
\begin{aligned}
lik(\theta_0, \theta_c|\text{datos}) &= \prod_{i=1}^{J} f(x_i|\theta_0, \theta_c) \\
&= \prod_{x_i:x_i=1}^{J} \theta_0 \times \prod_{x_i:x_i>0} (1-\theta_0)\theta_c \exp\{-\theta_c x_i\} \\
&= \theta_0^{n_0} \times (1-\theta_0)^{J-n_0} \theta_c^{J-n_0} \exp\{-\theta_c \sum_{x_i:x_i>0} x_i\},
\end{aligned}
$$

donde

$$n_0 = \#\{x_i : x_i = 0\}.$$

Comentarios

1. Noten que el componente
$$\theta_0^{n_0} \times (1-\theta_0)^{J-n_0},$$
resume la informacion en los datos para θ_0.
2. Noten que el componente
$$\theta_c^{J-n_0} \exp\left\{-\theta_c \sum_{x_i:x_i>0} x_i\right\},$$
resume la informacion para θ_c.
3. Ambas informacion son `separadas` pero `no ajenas`. Por esto, podemos hacer inferencia sobre θ_0 y θ_c por `separado`.

Maxima verosimilitud I

Encontrar el EMV para θ_0 con base en los `datos` es relativamente simple, es dado por

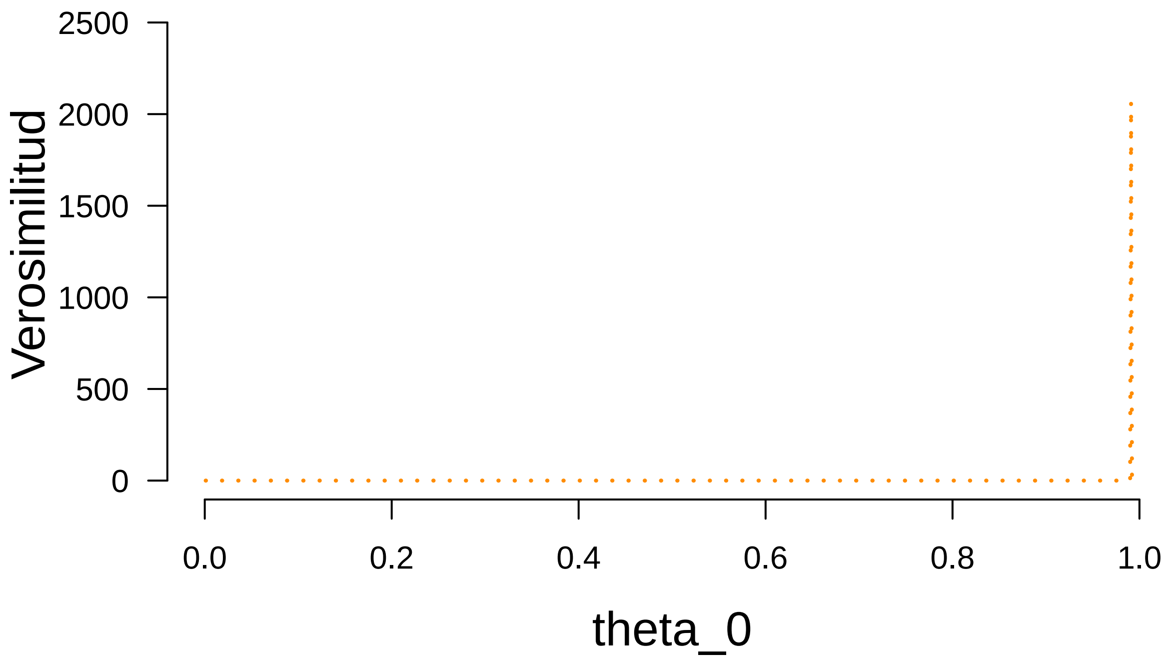
$$\begin{aligned}\theta_0^* &= \arg\max_{(0,1)} \theta_0^{n_0} \times (1-\theta_0)^{J-n_0} \\ &= \frac{n_0}{J}.\end{aligned}$$

En el caso de los datos **AllState** es

```
theta0_star <- n0/J
theta0_star
```

```
## [1] 0.9909079
```

Error epistemico I



Reflexion

- Aunque pareciera ser contundente, ¿es conveniente considerar solamente la informacion de estos "datos"?

Informacion complementaria

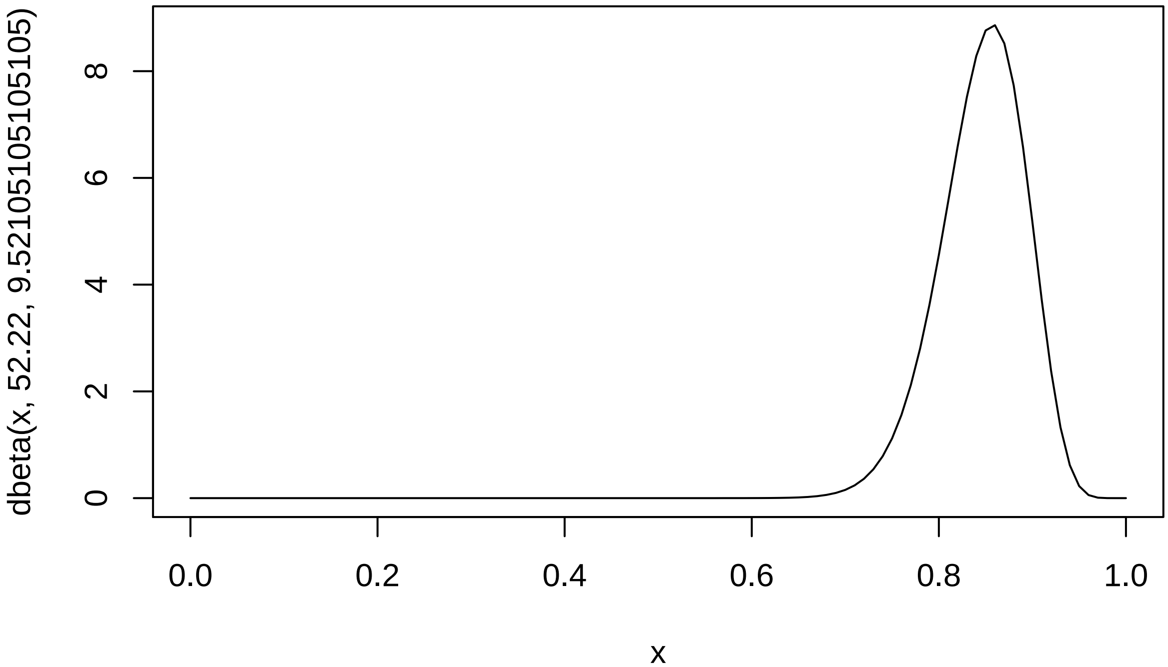
Pensemos el caso en que informacion del siguiente tipo es disponible:

- (A). Creemos que la mediana de la incidencia de siniestros es 0.85.
- (B). Creemos que el 99.999-esimo porcentil de la incidencia de siniestros es 0.95.
- (C). Creemos que el porcentil 0.001 de la incidencia de siniestros es 0.60.

Esta informacion complementaria resume la informacion de expertos / instituciones reguladoras / mercado / etc.

Visualizacion de elicitacion

```
## [1] "Elicitacion para a= 52.22 b= 9.52105105105105"
```



Consolidacion de informacion

La informacion contenida en los `datos` y `complementaria` se **consolida** en la siguiente expresion

$$q(\theta_0|\text{datos}, \text{complemento}) \propto lik(\theta_0|\text{datos}) \times q(\theta_0|\text{complemento}).$$

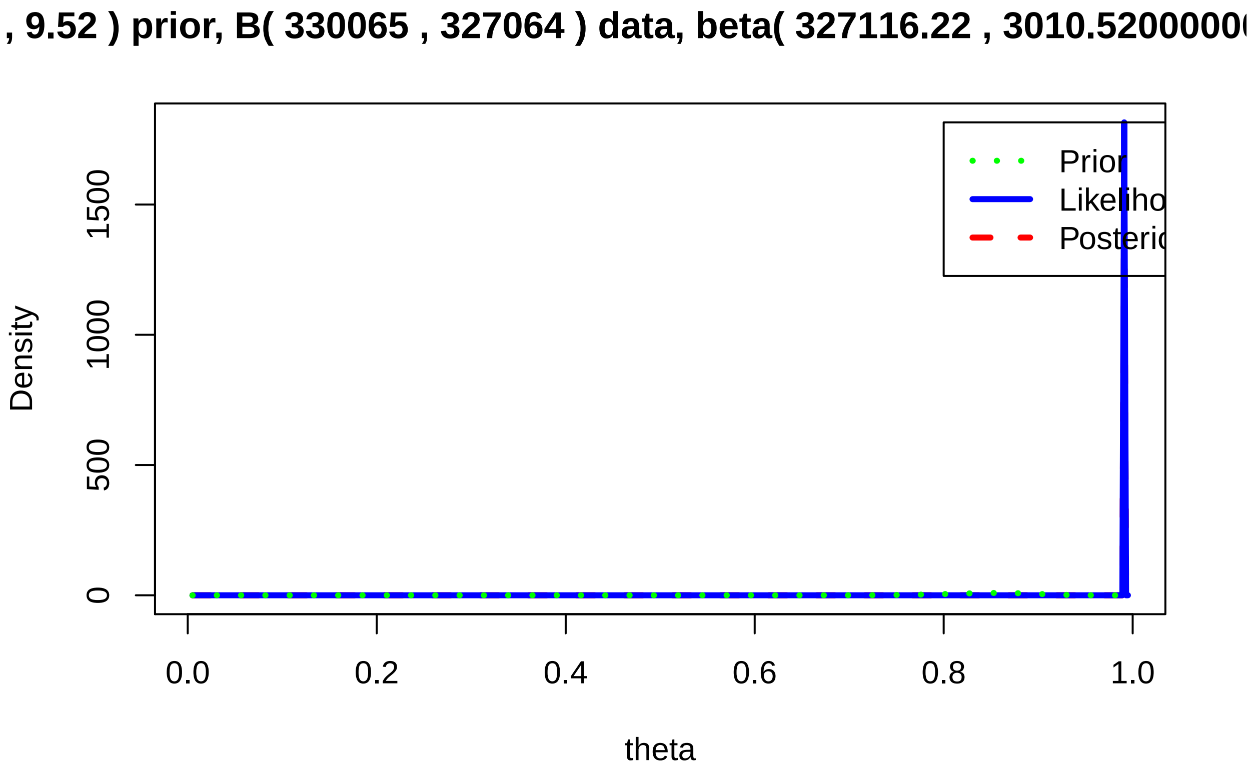
Resulta ser que si $q(\theta_0|\text{complemento})$ es una medida de probabilidad, la funcion $q(\theta_0|\text{datos}, \text{complemento})$ tambien lo es.

En este caso, tal condicion se cumple, pues $q(\theta_0|\text{complemento})$ se elicito siendo parte de la familia de distribuciones beta.

La funcion $q(\theta_0|\text{complemento})$ se refiere a la **prior** del modelo en el contexto bayesiano de inferencia.

Visualizacion de consolidacion

```r
consolidacion.binomialbeta(n0, J, 52.22, 9.52,print=TRUE,summary=FALSE)
```



Error epistemico II

```
consolidacion.binomialbeta(n0, J, 52.22, 9.52,print=FALSE,summary=TRUE)
```

```
## [1] "Modas= 0.857381988617342  |  0.990907851483799  |  0.990883688390031"
## [1] "Medias= 0.845804988662132  |  0.990904876888632  |  0.990880714479536"
## [1] "sd s= 0.0455929848904483  |  0.000165241284816415  |  0.000165443643283756"
```

Ejercicio

1. Seleccionen una submuestra aleatoria del 10% de los `datos`.
2. Replique los calculos usando los datos de la submuestra aleatoria.
3. Reflexione acerca de los **estimadores puntuales**, el **error epistemico** y relevancia de la **informacion complementaria**.