# CAPÍTULO 12. OTROS TEMAS

Se tratan algunos temas no comentados anteriormente:

- Robustez
- Inferencia no paramétrica
- Otros

# ROBUSTEZ

Ver Berger (1994), Ríos y Ruggeri (2000).

Se quiere estimar la sensibilidad de algunos resultados ante la elección de la distribución a priori $f(\boldsymbol{\theta})$. Es muy importante en situaciones donde se solicitan información de expertos para formar la distribución a priori.

También sensibilidad ante la función de pérdida y la verosimilitud son importantes. Ver Dey y Micheas (2000), Kadane et al (2000) y Shyamalkumar (2000).

Existen varios enfoques a robustez de la distribución a priori.

En una análisis informal se comparan los resultados dadas varias distribuciones a priori.

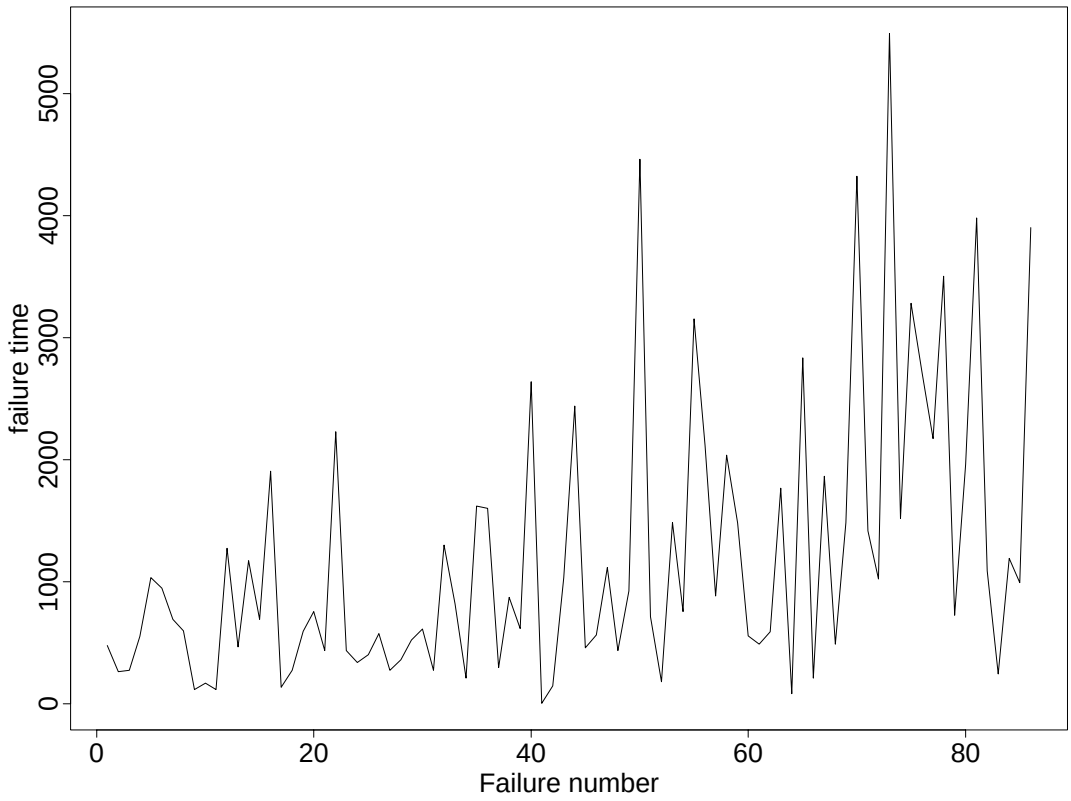
**Ejemplo 124** *Wilson y Wiper (2000)*

*Sean* $T_1, T_2, \ldots$ *los tiempos entre fallos de un programa de software. El modelo de Jelinski y Moranda (1972) supone que*

$$T_i | N, \phi \sim \mathcal{E}\left((N - i + 1)\phi\right)$$

*es decir que en principio, el programa contiene* $N$ *faltas, todos del mismo "tamaño"* $\phi$ *y después del descubrimiento de una falta, se corrige esta falta perfectamente.*

*Se observaron los primeros* $m = 86$ *tiempos entre fallos de un programa.*




*Se quería predecir el siguiente fallo y el número de faltas restantes en el programa.*

*La verosimilitud es*

$$l(N, \phi|\mathbf{t}) \propto \frac{N!}{(N-m)!}\phi^m \exp\left(-\left[(N+1)m\bar{t} - \sum_{i=1}^{m} it_i\right]\phi\right)$$

*y las distribuciones a priori típicas (semi conjugadas) son*

$$\begin{aligned} N &\sim \mathcal{P}(\lambda) \qquad (\lambda = 100) \\ \phi &\sim \mathcal{G}(\alpha, \beta) \qquad (\alpha = 1 \text{ y } \beta = ,0001) \end{aligned}$$

*Dadas estas distribuciones a priori se calcula* $E[N|data] \approx 104$ *(EMV* $= 106$*). La mediana a posteriori del tiempo al siguiente fallo es* $2440 \times 10^{-2}$ *segundos (EMV* $= 2177$*).*

*Contaminamos la distribución a priori con una distribución con colas más largas. Se define la clase de distribuciones a priori*

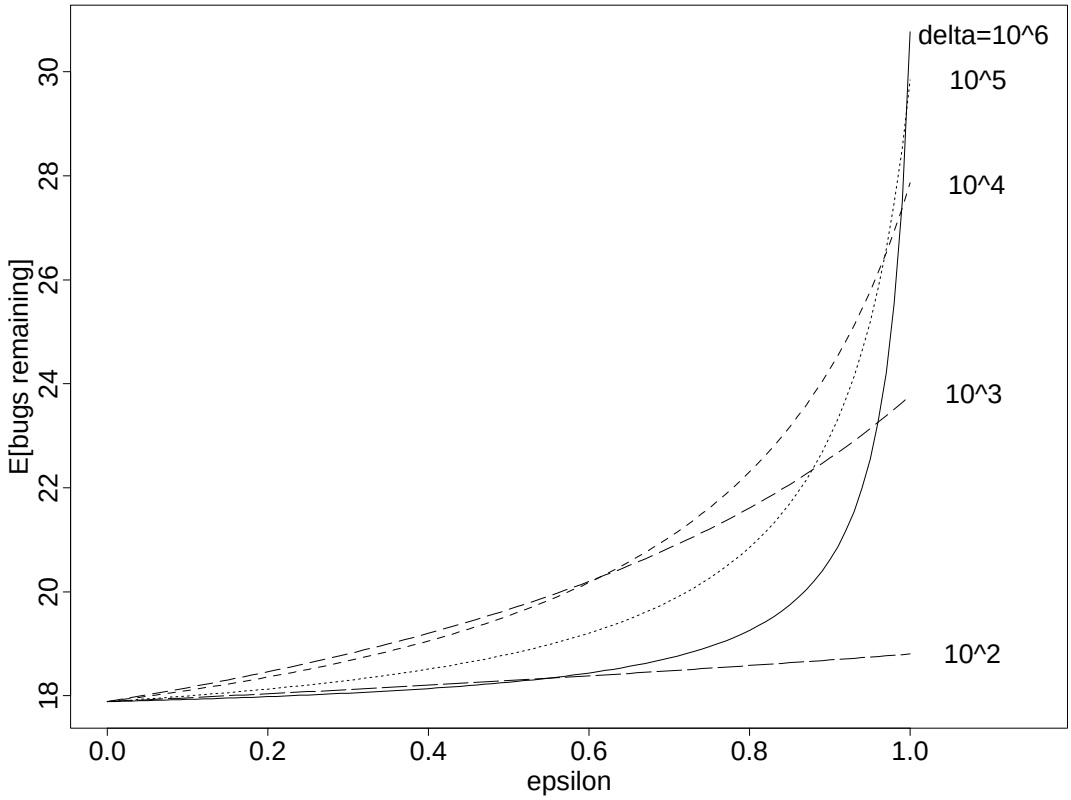
$$\Gamma = \{(1-\epsilon)f(N) + \epsilon g(N)\}$$

*donde* $f(N)$ *es la densidad Poisson y*

$$g(N) \propto \left((N - 100 + \frac{1}{2})^2 + \delta\right)^{-1}$$

$g(N)$ *es una distribución bimodal en 99 y 100 como la Poisson pero con colas más largas. Para* $\delta \approx 100$ *es parecida a la Poisson cerca de* $N = 100$.

*La figura ilustra los efectos de la contaminación en la media del número de faltas restantes.*




*Para* $\epsilon = 0{,}2$ *la mediana a posteriori del tiempo al siguiente fallo varia entre* 2410 *y* 2440 *para* $100 \leq \delta \leq 1000000$ *pero cuando* $\epsilon = 1$, *la mediana es* 1960 *en el peor caso* $\delta = 1000000$.

*Parece que los resultados son poco sensibles para contaminaciones pequeñas.*

## Robustez Global

En robustez global, se incluye la distribución a priori en una clase más amplia de distribuciones, $\Gamma$ y se examina la sensibilidad de alguna función de interés (media a posteriori ...). Si existen grandes diferencias entre el máximo y el mínimo de la función, la inferencia es sensible.

**Ejemplo 125** *(Sellke et al 1999)*

*Muchos estadísticos creen que si el p valor es* 0,05 *entonces es muy probable que la hipótesis nula sea falsa y si el p valor es* 0,01 *está casi seguro.*

*Sea* $H_0 : \theta_i = 0$ *frente a* $H_1 : \theta_i \neq 0$ *para* $i = 1, 2, \ldots$ *una sucesión de contrastes para las medias de poblaciones normales estandares.*

*Supongamos que 50 % de las hipótesis nulas son ciertas.*

*Para cada contraste, se muestran datos* $\mathbf{x}_i$ *y se calcula el p valor* $p_i$*.*

*En el subconjunto donde* $0{,}04 < p < 0{,}05$ *se puede demostrar que por lo menos 24 % de las hipótesis nulas son ciertas. En el subconjunto* $0{,}009 < p < 0{,}01$*, por lo menos 7 % son ciertas.*

*¿Cómo se lo demostró?*

*Se dejaron los parámetros* $\theta_i$ *en las hipótesis alternativas asumir todos los valores posibles o equivalentemente, se supone la clase de todas las distribuciones a priori para los* $\theta_i$*. Entonces condicionada en* $p \approx 0{,}05$ *(*$0{,}01$*), se demostró que 24 % (7 %) es el corte inferior sobre la proporción de hipótesis nulas verdaderas.*

Clases muy utilizadas son:

- clase de contaminación $\epsilon$

  $\Gamma = \{\pi : \ \pi(\theta) = (1-\epsilon)f(\theta) + \epsilon g(\theta), \ g \in G\}$

  donde $G$ es una clase de distribuciones contaminantes.

- clases de momentos (generalizados).

  Se trata de la clase de distribuciones con un conjunto de momentos o cuantiles especifacados.

- clase de bandas de densidades

  $$\Gamma = \{\pi : \ L(\theta) < \pi(\theta) < U(\theta)\}$$

  Por ejemplo, se puede fijar $L(\theta) = (1-\epsilon)f(\theta)$ y $U(\theta) = (1+\epsilon)f(\theta)$.

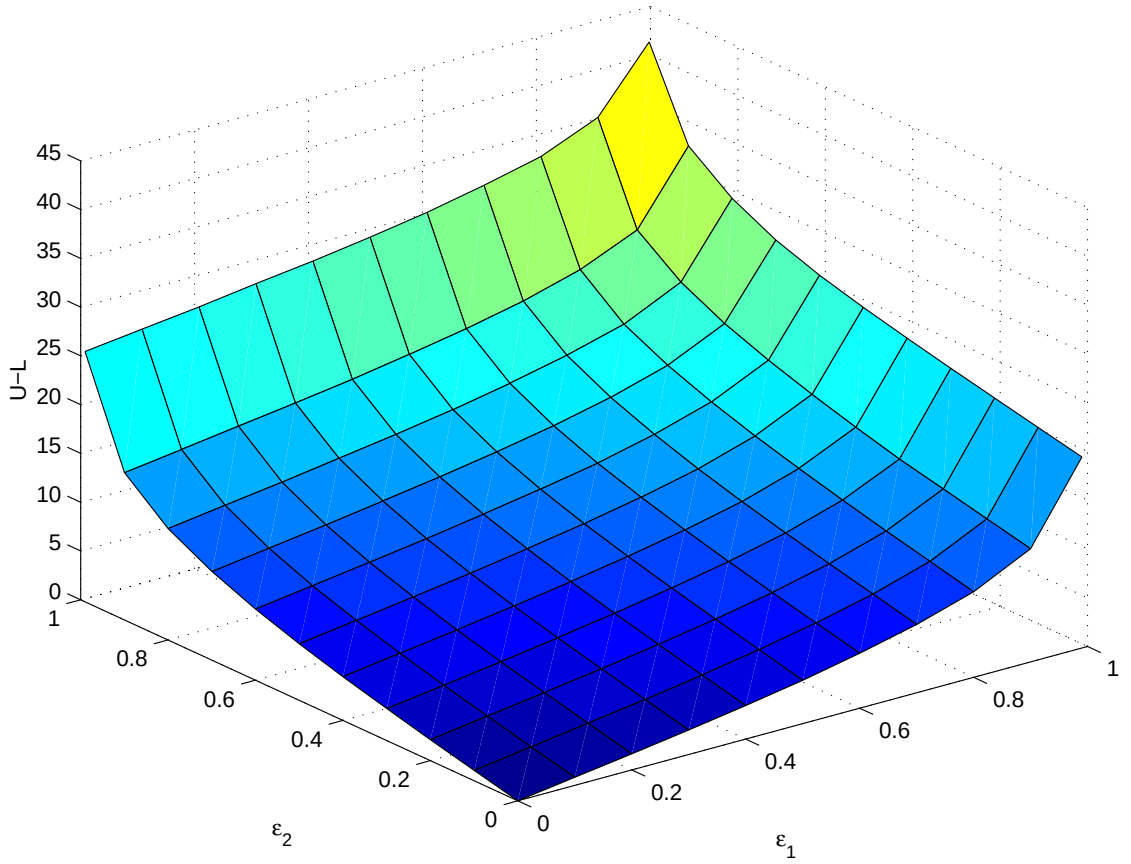
**Ejemplo 126** *Retomamos el Ejemplo 124*

*Se quiere preservar la independencia a priori entre* $N$ *y* $\phi$*. Entonces se define la clase*

$$\Gamma = \{\pi : \ \pi(N, \phi) = \pi_1(N)\pi_2(\phi)\}$$

*donde* $(1 - \epsilon_1)f(N) < \pi_1(N) < (1 + \epsilon_1)f(N)$ *y* $(1 - \epsilon_2)f(\phi) < \pi_2(\phi) < (1 + \epsilon_2)f(\phi)$ *y* $f(N)$ *y* $f(\phi)$ *son las distribuciones Poisson y gamma utilizadas anteriormente.*

*La figura ilustra las diferencias entre los cortes superiores y inferiores en el valor de la media a posteriori de* $N$ *para* $0 < \epsilon_1, \epsilon_2 < 1$*.*




*La inferencia parece más sensible a cambios en la distribución a priori de $N$ pero es bastante robusta para contaminaciones pequeñas.*

## Problemas

- ¿Cómo elegir una clase razonable?

- Las clases son demasiado grandes y contienen varias distribuciones a priori poco razonables.

- El cálculo de los mínimos y máximos es difícil.

## Robustez Local

Se quiere medir los efectos de desviaciones pequeñas en la distribución a priori $f(\theta)$ en alguna función de interés $g(\pi = f, l(\theta|\mathbf{x}))$ (por ejemplo la media a posteriori). En robustez local, se calculan derivados apropriados de $g(\pi, l)$ evaluados en $\pi = f$ (y en $l = l(\theta|\mathbf{x})$).

**Ejemplo 127** *(Gustafson 2000).*

*Supongamos contaminación $\epsilon$ de la distribución a priori, es decir consideramos la clase de distribuciones*

$$\pi = f + U \quad \textit{donde } U = \epsilon(P - f)$$

*tiene medida* 0 *y $P$ es una clase de distribuciones contaminadores.*

*Podemos medir la sensibilidad local con la función de influencia $I(\cdot)$ donde*

$$\frac{\partial}{\partial \epsilon} g(\epsilon[P - f])\Big|_{\epsilon=0} = \int I(z) d[P - f](z)$$

*y* $g(\cdot)$ *es la función de interés.*

$I(\cdot)$ *no es única (se puede añadir una constante) y entonces, a menudo se estandardiza suponiendo que* $\int I(z)df(z) = 0$.

La gran ventaja de robustez local es que a menudo, las medidas son más fáciles de calcular. El gran problema es cómo precisarlas.

# MÉTODOS NO PARAMÉTRICOS

Se tiene $X|f \sim f$ y dada una muestra $x_1, x_2, \ldots,$ se quiere hacer inferencia sobre $f$. ¿Cómo hacerla?

Se necesita definir una distribución a priori sobre un espacio (de distribuciones) de dimensión infinita. La clase más estudiado de distribuciones se llama procesos de Dirichlet (Ferguson 1973,1974).

**Definición 19** *Sea $F_0$ una función de distribución y $\nu$ un parámetro escalar.*

*Para cualquier partición finita del espacio paramétrico, $\{C_1, \ldots, C_r\}$, la distribución a priori del proceso de Dirichlet es una medida probabilística aleatoria $F$ que asigna una distribución*

$$\{F(C_1), \ldots, F(C_r)\} \sim \mathcal{D}\left(\nu F_0(C_1), \ldots, \nu F_0(C_r)\right)$$

La ventaja de esta distribución es que es fácilmente manejable. Sea $\{X\}_i$ una sucesión de variables intercambiables con $X|F \sim F$. Entonces:

- La distribución marginal de $X_i$ es $F_0$.

- La distribución condicionada de $F$ dada la muestra $x_1, \ldots, x_n$ es otro proceso Dirichlet con

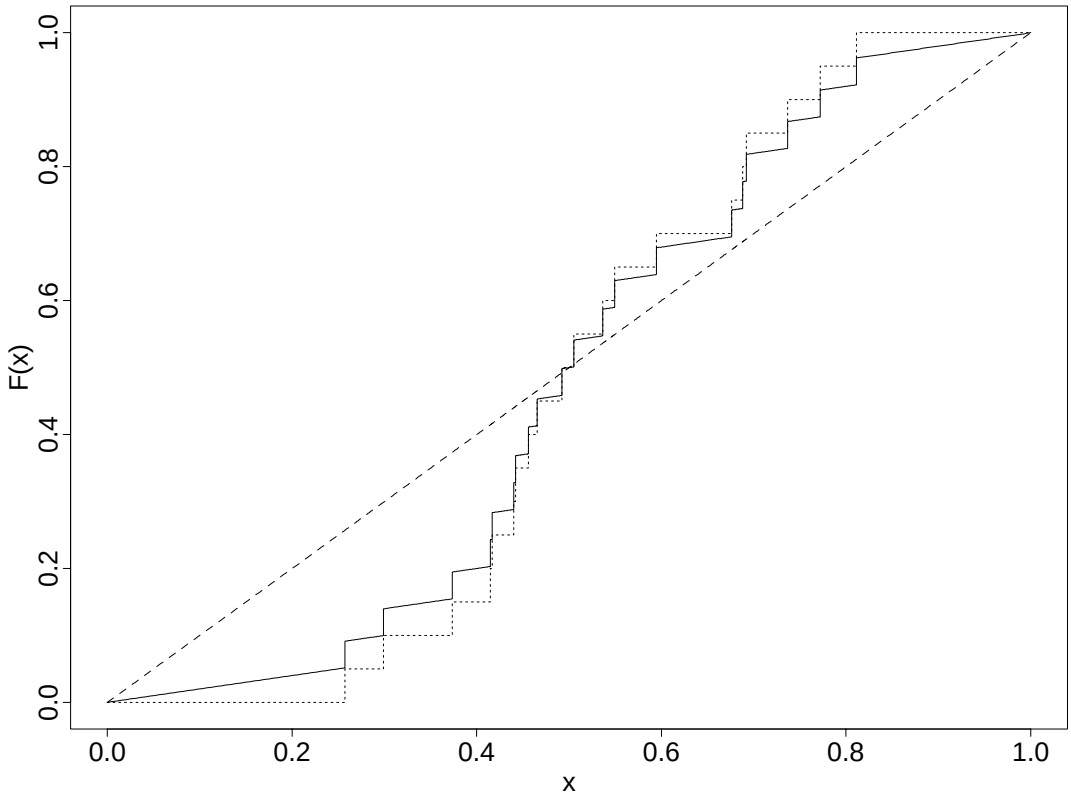
$$\{F(C_1), \ldots, F(C_r)\}|\mathbf{x} \sim \mathcal{D}\left(\nu F_0(C_1) + \sum_{i=1}^{n} I_{C_1}(x_i), \ldots, \nu F_0(C_r) + \sum_{i=1} n I_{C_r}(x_i)\right).$$

- La distribución predictiva de otra observación $X_{n+1}$ es

$$P(X_{n+1} \in C|\mathbf{x}) = \frac{\nu F_0(C) + \sum_{i=1} n I_C(x_i)}{\nu + n}.$$

Para $n$ grande, la función de distribución predictiva de $X_{n+1}$ se acerca a la función de distribución empírica.

**Ejemplo 128** *Se generaron* 20 *datos de una distribución beta* $\mathcal{B}(5,5)$*. Se uso un proceso Dirichlet con* $\nu = 5$ *y* $F_0(x) = x$ *para* $0 < x < 1$*, es decir una uniforme. La figura muestra la distribución predictiva y la distribución empírica.*



Antoniak (1974) introdujo mixturas de procesos de Dirichlet y existen varios trabajos recientes los que presentan esquemas computacionales mediante MCMC para algunos problemas.

El mayor inconveniente teórico del proceso Dirichlet es que asigna una probabilidad 1 al conjunto de medidas discretas de probabilidad. Existen alternativas como procesos gaussianos o gamma (difíciles de tratar computacionalmente) o arboles de Polya (Paddock et al 2003).

# OTROS CAMPOS

## Análisis de decisiones

Para tomar decisiones ($d$) bajo incertidumbre ($\theta$) se necesitan estimar subjetivamente tanto las probabilidades $f(\theta)$ como el valor o utilidad de una posible pérdida o ganáncia ($U(d, \theta)$).

## Ejemplo 129 Paradoja de San Petersburgo

*En un casino, se lanza una moneda insesgada hasta que salga la primera cara. Un jugador tiene 1 euro con que jugar y le ofrece la siguiente sucesión de apuestas.*

*Después de cada tirada, hay dos opciones. Puede retirarse, preservando sus ganancias o puede apostar todas sus ganancias en la siguiente tirada. Si sale una cruz, triplicará su dinero y si sale una cara, perderá todo.*

*¿Cuando debe retirarse?*

*Supongamos que el jugador quiere maximizar sus ganancias esperadas.*

*Supongamos que han salido $r$ cruces. Definimos $S_r$ = su dinero inicial más todas sus ganancias hasta ahora. Entonces el jugador tiene dos decisiones: seguir o parar.*

$$\begin{aligned} E[\textit{parar}] &= S_r \\ E[\textit{seguir}] &= \frac{1}{2}0 + \frac{1}{2}3S_r \\ &= \frac{3}{2}S_r \end{aligned}$$

*El jugador debe seguir jugando hasta que salga una cara y entonces perderá su dinero.*

El ejemplo ilustra que intentar maximizar las ganancias esperadas no siempre es lógico. También en muchos problemas se debe comparar decisiones multi-objetivas con ganancias compuestas de dinero, calidad de vida, etc.

En la teoría de utilidad se desarolla varios axiomas para definir preferencias racionales entre decisiones. Basada en esta teoría se puede definir una función de utilidad la que representa el valor esperado de cada decisión. Ver Keeney y Raiffa (1976) o French (1986).

**Modelos gráficos**

Ver Lauritzen y Spiegelhalter (1988). Se representan todas las variables de un problema como los nodos en un grafo dirigido (*influence diagram*). Si dos nodos no tienen conexión, son condicionalmente independientes. El paquete *Winbugs* se basa en el uso de grafos dirigidos.

Se pueden representar problemas de decisión con grafos parecidos.

**Métodos bayesianos lineales**

En muchos problemas no es fácil solicitar distribuciones enteros de expertos pero es bastante sencillo solicitar primeros y segundos momentos. Los métodos bayesianos lineales se basan en hacer inferencia (aproximada) utilizando estos momentos. Ver por ejemplo Goldstein (1999).