Introducción a la inferencia bayesiana

- La Inferencia Bayesiana descansa exclusivamente en el Teorema de Bayes

$$p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)$$

- θ es habitualmente un parámetro, aunque podría ser un valor puntual, una hipótesis o un modelo.
- Las p son funciones de densidad (o cuantía)
- $p(\theta)$ es la densidad a **priori**; $p(data|\theta)$, la **verosimilitud** (likelihood) de θ aportada por los datos; $p(\theta|data)$ es la densidad a **posteriori** de θ dados los datos.
- Lo que conduce al *Mantra Bayesiano*:

La densidad a posteriori es proporcional a d. a priori por la verosimilitud

$$p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)$$

- El teorema de Bayes nos dice cómo actualizar nuestras creencias sobre θ a la luz de la evidencia (datos)
- Es un método general de **inducción** o de "aprendizaje de la experiencia" a priori→ datos → a posteriori
- El teorema de Bayes (recordemos que) es un resultado incontrovertible se sigue de los axiomas de Kolmogorov y de la definición de probabilidad condicionada (La estadística bayesiana sí es discutible: la estadística clásica, p.ej. no admite que los parámetros sean aleatorios)

¿Por qué ser bayesiano?

- **Simplicidad conceptual:** se dice lo que se quiere decir y se quiere decir lo que se dice
- Un fundamento de la inferencia que no requiere pensar en experimentos repetibles que podrían dar resultados muestrales aleatorios...
- Uniformidad de aplicación: no hay que ir variando de éste a aquel análisis de datos: sólo teorema de Bayes
- La potencia de computación actual lo hace muy asequible a través de la simulación MCMC

$$p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)$$

Simplicidad conceptual

- La densidad (o cuantía) a posteriori es una completa descripción de nuestras creencias tras la observación de los datos.
- de modo que aporta todo lo necesario para hacer nuestras inferencias
- Ejemplos:

  La probabilidad a posteriori de que un coeficiente de regresión sea positivo (negativo, nulo,…)
  La probabilidad a posteriori de que un sujeto pertenezca a tal grupo
  La probabilidad a posteriori de que la hipótesis H sea cierta
  La probabilidad a posteriori de que cierto modelo estadístico sea el auténtico modelo de entre un conjunto de ellos. …

En cambio, la inferencia "frecuentista" (o clásica) :

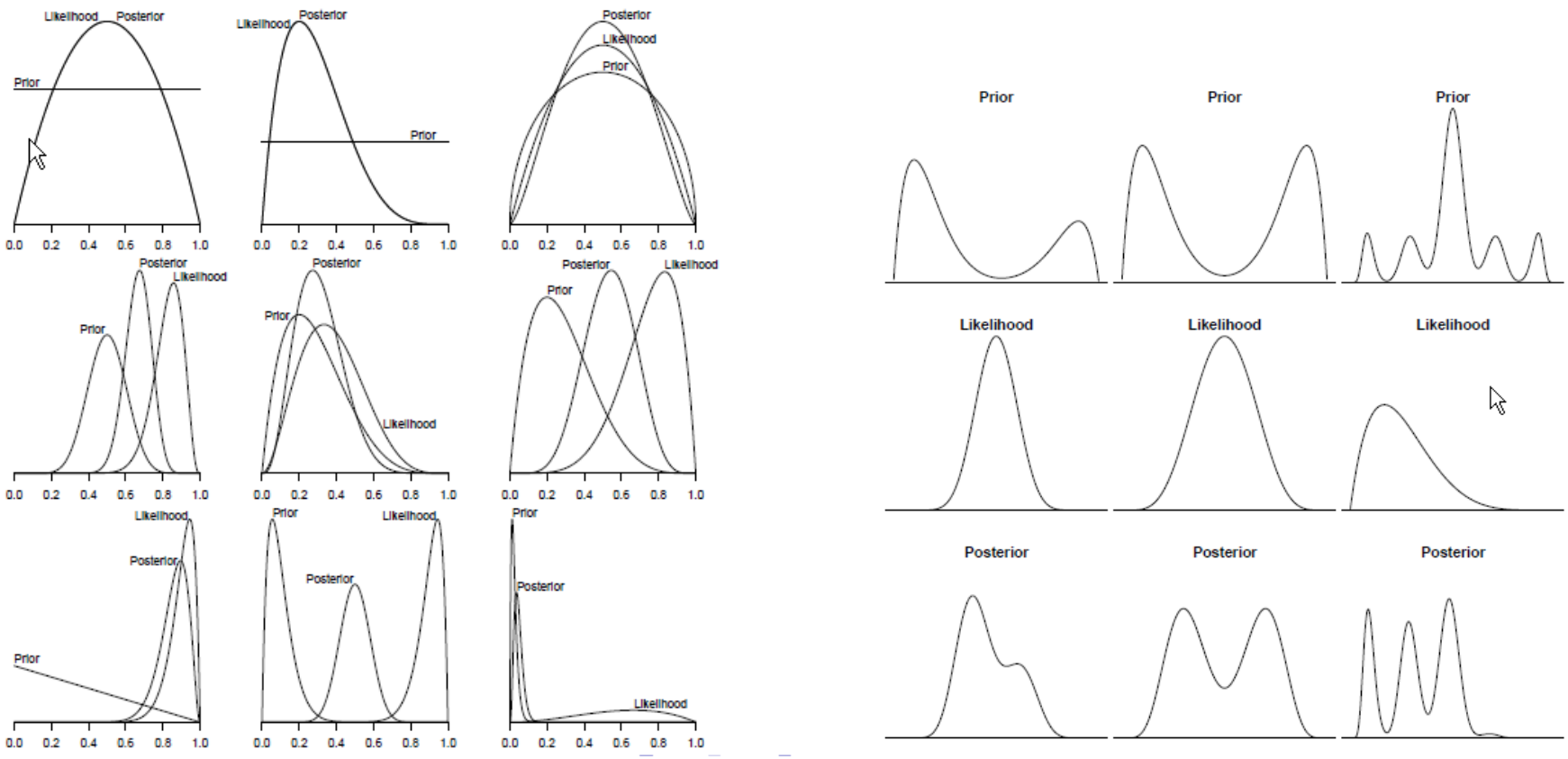
- Propone un modelo para los datos: $y \sim f(\theta)$
- La estimación de θ se basa en un estimador (una función de los datos : $\hat{\theta} \equiv \hat{\theta}(y)$
- Para el contraste de cierta hipótesis nula se trabaja con la distribución del estimador condicionada a que la hipótesis sobre el parámetro es cierta:

"asumiendo Ho en un proceso de reiterado de muestreo *cómo de frecuente* sería obtener un resultado *al menos tan extremo* como el obtenido"

Si la contestación a esta pregunta acaba siendo “muy infrecuente”; menos que el nivel de significación se rechazará la hipótesis.

- Pero actuar con esa racionalidad no es, a menudo, fácil de explicar de forma intuitiva.
- En la Inferencia Clásica trata a los estimadores ($\hat{\theta}$) y a los propios datos como “variables aleatorias” mientras que los parámetros son rasgos fijos (aunque desconocidos) de la población de la que se obtiene la muestra “aleatoriamente”.
- En la Inferencia Bayesiana, $\hat{\theta}$ es un valor **no-**aleatorio función de los datos disponibles y, es en cambio , el parámetro desconocido θ lo que es aleatorio y nuestro conocimiento (incierto) sobre θ se expresa a través de una distribución de probabilidad: Antes de los datos d. a priori y después d. a posteriori

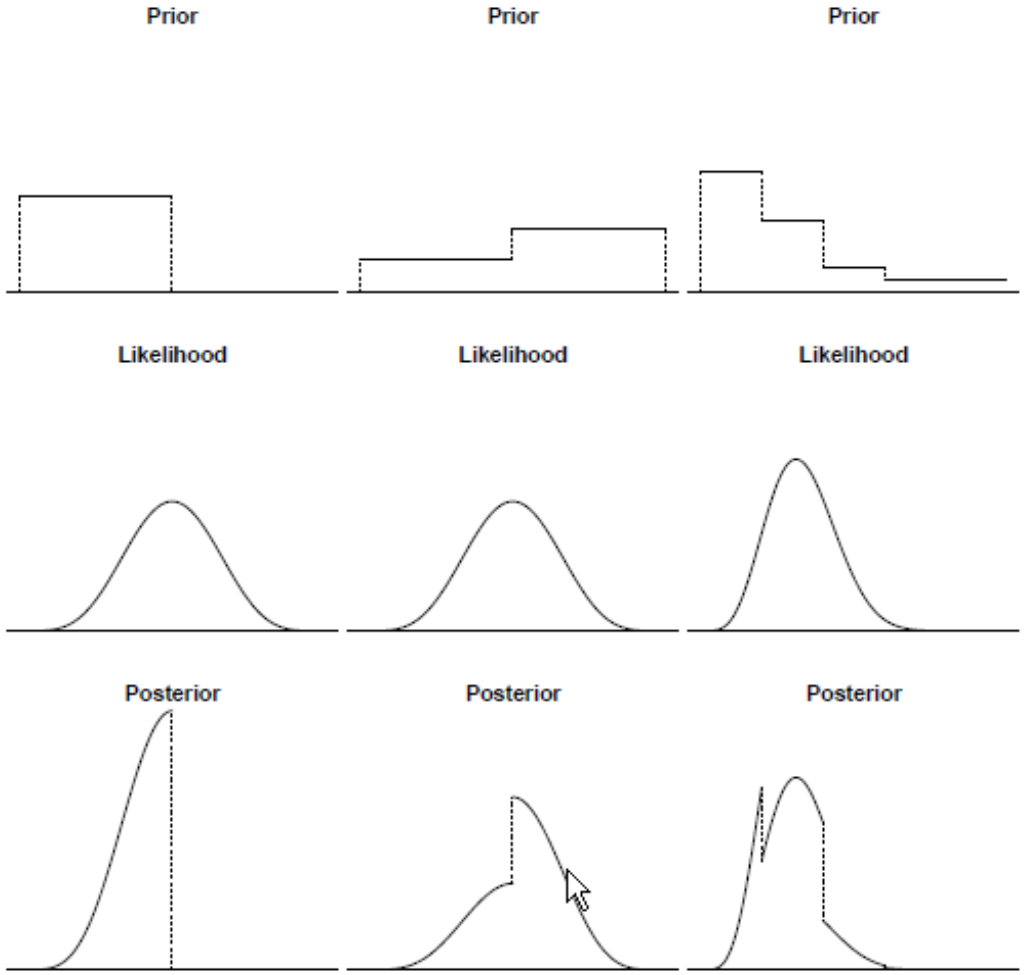
	Bayesiana	**Clásica**
θ	Aleatorio (=desconocimiento)	Constante,aunque desconocida
$\hat{\theta}$	Fijo (datos)	Aleatorio
aleatoriedad	(conocimiento parcial) Subjetivo	Viene del muestreo
Distribución importante	A posteriori	D. muestral



Fuente: Jackman S. : “Bayesian analysis for social Science”

Los peligros del dogmatismo:

- *Si a una hipótesis se le asigna una probabilidad inicial de 0 nunca podrá establecerse "a posteriori"*
- *Si a una hipótesis se le asigna una probabilidad de 1 nunca podrá rechazarse*



ESIMACIÓN PUNTUAL BAYESIANA

(en el bayesianismo no hay "estimadores")

- **La estimación puntual bayesiana** consistirá en un **número** que resuma adecuadamente la distribución a posteriori. Pero ¿Cuál? ¿Su media , su moda, su mediana algún cuantil?

- Diferentes **funciones de pérdida** conducirán a diferentes estimaciones

Sea Θ el conjunto de posibles estados de la naturaleza (o valores) de θ y sea $a \in A$ una acción (estimación) de entre las posibles que puede llevar a cabo el investigador definimos la pérdida asociada a la acción (estimación) a cuando el estado de la naturaleza (el valor del parámetro) es θ como $l(\theta,a)$

- La pérdida esperada a **posteriori** de una estimación dada una densidad a posteriori $p(\theta|y)$ y una f. de pérdida $l(\theta,a)$ será:

$$E[l(\theta,a)\,|\,y] = \int_{\Theta} l(\theta,a)\,p(\theta\,|\,y)$$

La estimación bayes será aquella que *minimice* la *pérdida* (a posteriori) *esperada*

- Pérdida cuadrática $\rightarrow$Estimación bayes ~ media de la distribución a posteriori

- Pérdida lineal-simétrica (proporcional al valor absoluto del error) $\rightarrow$ Estimación bayes ~ Mediana de la distribución a posteriori
- Pérdida "todo/nada" $\rightarrow$Estimación bayes ~ Moda de la distribución a posteriori

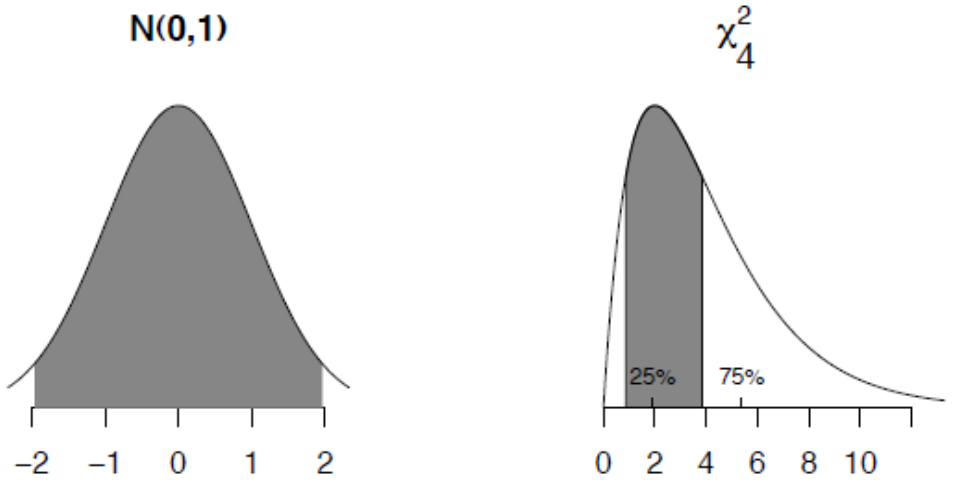
Estimación por intervalo Bayesiana y Contraste de Hipótesis

- Se resuelve por la obtención de un intervalo de la probabilidad requerida $(1-\alpha)$ en la distribución a posteriori.
- Se suele denominar intervalo de $100\times(1-\alpha)\%$ de credibilidad
- Como se suele pretender que el intervalo de estimación sea lo más preciso posible se suele considerar como intervalo de credibilidad aquél intervalo de probabilidad $(1-\alpha)$ que tenga menor amplitud y por tanto que tenga mayor densidad de probabilidad:

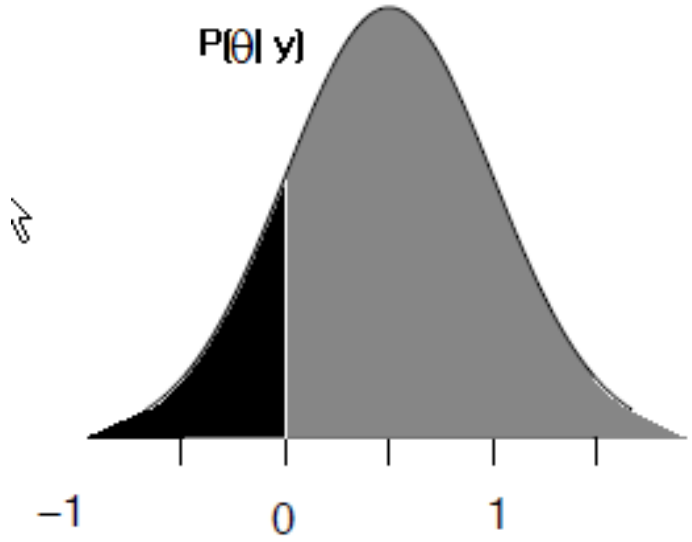
Es el llamado intervalo de mayor densidad (media) de probabilidad (o credibilidad) $(1-\alpha)$:

Intervalos de mayor densidad de probabilidad $100\times(1-\alpha)$ % ($100\times(1-\alpha)$ % HPD intervals):

- En una densidad unimodal y simétrica es único y simétrico respecto a la moda
- En una distribución asimétrica el Intervalo de máxima densidad de probabilidad diferirá bastante del formado por las cuantiles $\alpha/2$ y $1-\alpha/2$



La contrastación de hipótesis en el bayesianismo se convierte , igualmente en una discusión sobre que hipótesis goza de mayor probabilidad en la distribución a posteriori.
En este ejemplo nos decantaríamos claramente por la hipótesis de que el parámetro θ es positivo.



Introducción empírica a la inferencia bayesiana sobre proporciones o probabilidades (desconocidas)

Ejemplo: Supongamos que la proporción de accidentes de cierto tipo en un proceso productivo, p, es desconocida. Pero que basándonos en los datos de otras procesos similares podemos suponer que se encuentra entre 0,005 y 0,01 con las siguientes probabilidades asociadas:

p_i	$P(p_i)$	pi.P(pi)	
0,005	0,3	0,0015	esta distribución (a priori) a falta de otra información posterior, nos resume
0,007	0,1	0,0007	todo nuestro conocimiento sobre p Si suponemos una función de pérdida
0,008	0,2	0,0016	cuadrática sugeriríamos como estimación (con menor perdida esperada)
0,009	0,3	0,0027	
0,01	0,1	0,001	la media de la distribución a priori: E(p)=0.0075
		0,0075	

Si realizamos una experiencia: para un bayesiano, nuestra información (sobre p) mejorará y, para un clásico, supuestas algunas hipótesis previas, simplemente, podremos usar esa (única) información para estimar p.

Observamos 200 casos (independientes y aleatoriamente elegidos) del tipo analizado y se producen 1 accidentes del tipo que nos interesa. Supone una información muestral que:

- Para un estadístico clásico quedaría recogida en el estadístico proporción muestral de accidentes (un estimador **clásico** del desconocido valor de p, que en este caso arrojaría una estimación puntual (clásica) de):

$$\hat{p} = \frac{1}{200} = 0.005$$ Y, si se quiere un I. C , p.ej del 95% con p=q=0.5 de: 0.005±0.06929

- Para un Bayesiano tendríamos un dato : se ha producido 1 accidentes en 200 casos. Si llamamos X= nº de accidentes en 200 casos y si consideramos p= probabilidad de un accidente podemos determinar la **verosimilitud**

$P(X=1 \mid p) = P(X=1 \mid \sim B(200,p)) = \binom{200}{1} p\ (1-p)^{199}$ y podríamos obtener la distribución a posteriori de p aplicando el T.Bayes:

p_i	$P(p_i)$	$P(X\|p_i)$	$P(X,p_i)$	$P(p_i\|x)$
0,005	0,3	8,96729E-30	2,69019E-30	8,57952E-22
0,007	0,1	1,5678E-08	1,5678E-09	0,5
0,008	0,2	2,07476E-18	4,14952E-19	1,32336E-10
0,009	0,3	8,96729E-30	2,69019E-30	8,57952E-22
0,01	0,1	1,5678E-08	1,5678E-09	0,5
Sum →	1		3,13559E-09	1

$E(p|x)=0,0085$

Conjugación: Distribuciones conjugadas

- Sin embargo, la mayoría de las situaciones no son tan sencillas como la anterior en la que la distribución a priori era discreta y "muy" informativa. La "conjugación" suele ser un apoyo fundamental en este sentido:
- Resolver el "mantra bayesiano": posteriori es proporcional a verosimilitud por a priori; es fácil cuando se usan densidades a priori **conjugadas** con las verosimilitudes.
- Definición: Supongamos que una distribución a priori pertenece a cierta clase o familia de distribuciones paramétricas D; entonces diremos que esa distribución es conjugada respecto a la verosimilitud si la distribución a posteriori también pertenece a la clase D.
- Hasta la revolución de las MCMC (cadenas de markov de MonteCarlo) en los 90 la inferencia bayesiana se reducía a el uso de distribuciones conjugadas a las verosimilitudes de:

Ratios y proporciones (Bernouilli/binomial); conteos (Poisson), medias ,varianzas y regresiones de variables cuantitativas (Normal y derivadas)

- La estimación bayesiana (pérdida cuadrática) como $E(\theta|x)$ se reducía al cálculo (manual) del promedio de la estimaciones a priori y de la basada en los datos, ponderado por las precisiones (la precisión es una medida inversa a la varianza)

Principales conjugadas: priori-verosimilitud-posteriori

priori	**verosimilitud**	**posteriori**
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Normal	Normal($\tau=1/\sigma^2$conocida)	Normal
Normal, Gamma-inversa	Normal	Normal, Gamma-inversa
Normal, Gamma-inversa	Regresión	Normal, Gamma-inversa

Inferencia sobre proporciones/tasas/probabilidades (verosimilitud Binomial)

En un proceso experimental de Bernouilli, el número de éxitos X en n pruebas independientes de resultado dicotómico A y no-A con P(a)= θ sigue una distribución binomial: X~B(n, θ)

Por tanto dada una experiencia que arroja x éxitos la verosimilitud será:

$$l(\theta) = P(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \propto \theta^x (1-\theta)^{n-x} \quad ,$$

ya que $\binom{n}{x}$ no depende ni del parámetro θ ni de los datos x, podemos incorporarlos en la constante de normalización .

Nos planteamos cómo incorporar la información inicial sobre p para por realizar inferencias → Cuál es la conjugada de esa verosimilitud y si es capaz de dar cuenta de un adecuado rango de informaciones a priori diversas → → Distribución Beta

Beta = conjugada de la binomial:

1.-Como el parámetro θ es una tasa, una proporción o una probabilidad tendremos que $\theta \in [0,1]$ y que se deberá cumplir que la densidad a priori $p(\theta)$:

$\int_0^1 p(\theta) d\theta = 1$. Obviamente una densidad beta cumple esto.

2.- Debe ser conjugada a la d. binomial: si $p(\theta) \sim Beta(\alpha_0, \beta_0)$ la densidad a posteriori también debe seguir una Beta:

$$p(\theta \mid x) \propto p(\theta).P(x \mid \theta) \rightarrow Beta(\alpha_1, \beta_1)$$

Como la densidad a priori $p(\theta) \sim Beta(\alpha_0, \beta_0)$: $p(\theta) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0).\Gamma(\beta_0)} \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}$ con $\alpha_0, \beta_0 > 0$

La verosimilitud binomial era: $P(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ con lo que la densidad a posteriori será:

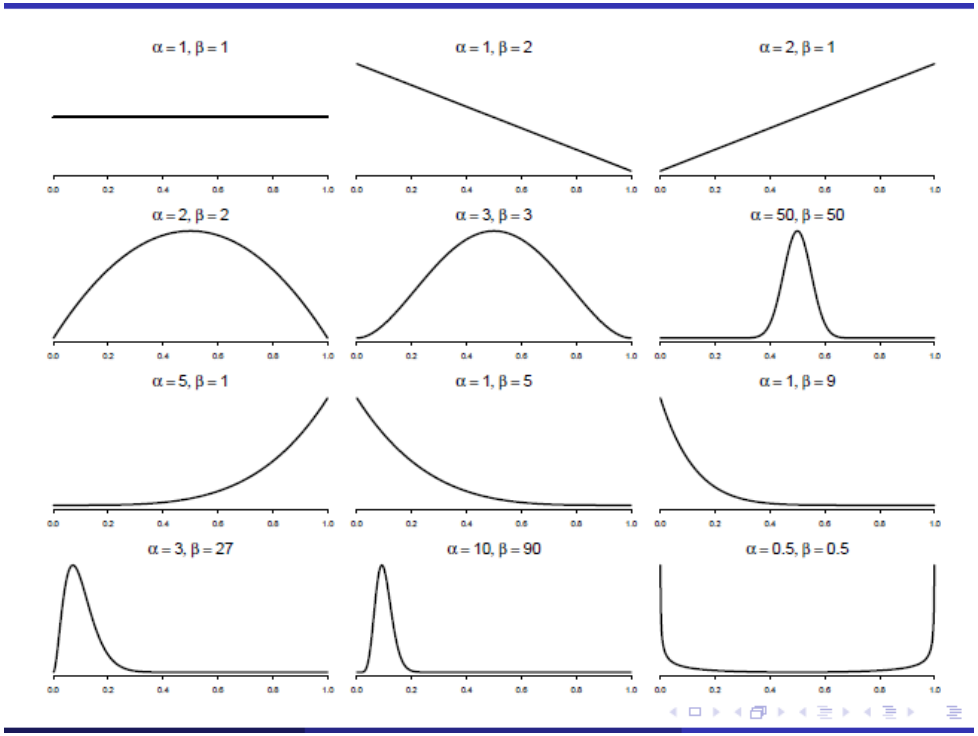
$$p(\theta \mid x) = \frac{p(\theta)\, p(x \mid \theta)}{\int_0^1 p(\theta)\, p(x \mid \theta) d\theta} \propto \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}.\theta^x (1-\theta)^{n-x}$$

$$p(\theta \mid x) \propto \theta^{x+\alpha_0 - 1}(1-\theta)^{n-x+\beta_0 - 1}$$

Que es la parte no constante de una densidad Beta con parámetros:

$$\alpha_1 = x+\alpha_0 \quad y \quad \beta_1 = \beta_0 + n - x$$

$\theta \sim \text{Beta}(\alpha_0, \beta_0)$, $x \sim \text{Binomial}(\theta, n) \Rightarrow \theta | x, n \sim \text{Beta}(\alpha_0 + x, \beta_0 + n - x)$.



Interpretación de la conjugación Beta/binomial/Beta en términos de equivalencia a "datos"

$$\theta \sim \text{Beta}(\alpha_o, \beta_o),\ x \sim \text{Binomial}(\theta, n) \Rightarrow \theta|x, n \sim \text{Beta}(\alpha_o + x, \beta_o + n - x).$$

- Es como si la información inicial representara la información de una muestra de

$(\alpha_0+\beta_0 -2)$ pruebas en la que se obtuvieran $(\alpha_0 -1)$ "éxitos" y $(\beta_0 -1)$ fracasos

- Si la distribución a priori es uniforme ≡ Beta(1,1) es como si se tuviera la información de no haber hecho ninguna prueba: $\alpha_0+\beta_0 -2= 0$

Script en R :

```
library(LearnBayes)
library(RcmdrMisc)
#Estimación proporciones
#obtener la prior a partir de dos cuantiles
#(50pcenti,90pcentil en #el ejemplo.Se puede
#cambiar) la función beta.select da como #resultado
#los valores de a y b de la d.beta ( d.a priori)
quantile1=list(p=.5,x=0.25)
quantile2=list(p=.9,x=0.45)
param_beta<-beta.select(quantile1,quantile2)
a=param_beta[1]
b=param_beta[2]
parámetros.d.beta<-list(param_beta)
parámetros.d.beta
#dibujo de la distrubución a priori
.x <- seq(0.013, 0.987, length.out=1000)
plotDistr(.x, dbeta(.x, shape1=a, shape2=b),
cdf=FALSE, xlab="p", ylab="Densidad",
main=paste("D. a priori beta", pbeta ))
remove(.x)
#muestra
exitos<-19
fallos<-31
#plot_verosimilitud
.x <- seq(0.013, 0.987, length.out=1000)
plotDistr(.x,dbinom(exitos,size=exitos+fallos,prob=.x
),xlab="valor de
p",ylab="verosimilitud",main=paste("verosimilitud"))
remove(.x)
#Posterior
aa=a+exitos
bb=b+fallos
ppost<-c(aa,bb)
ppost
posterior<-list("beta",ppost)
posterior
#estimacion supuesta pérdida cuadrática
estima<-aa/(aa+bb)
estimacion<-print(c("estimacion supuesta pérdida
cuadrática",estima))
#dibujo la d.posterior
.x <- seq(0.013, 0.987, length.out=1000)
plotDistr(.x, dbeta(.x, shape1=aa, shape2=bb),
cdf=FALSE, xlab="p",
      ylab="Density", main=paste(c("D. a posteriori",
posterior) ))
remove(.x)

#triplot:  Plot de prior, veros. y posterior
prior=c(a,b) # proporcion tiene una prior beta(a, b)
data=c(exitos,fallos) #  se observan exitos y  fallos
triplot(prior,data)
```

Inferencia sobre promedio de ocurrencias/ verosimilitud de Poisson.

Estamos interesados en la estimación del número medio, λ , de ocurrencias de ciertos hechos durante un intervalo unitario de tiempo a en el ámbito de un espacio unitario en aquellos casos en los que el tiempo o el espacio de experimentación podemos considerarlo homogéneo respecto a la factibilidad de los hechos estudiados. Si observamos independientemente n periodos-de-tiempo/ámbitos-espaciales produciéndose $(x_1,x_2,\ldots,x_n)$ hechos, la verosimilitud asociada a λ será:

$$l(\lambda) = P(x_1, x_2, ..., x_n \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = \frac{e^{-n\lambda} \lambda^{n\bar{x}}}{\prod_{i=1}^{n} x_i!} \propto e^{-n\lambda} \lambda^{n\bar{x}}$$

Nos planteamos cómo incorporar la información inicial sobre λ para por realizar inferencias $\rightarrow$ Cuál es la conjugada de esa verosimilitud y si es capaz de dar cuenta de un adecuado rango de informaciones a priori diversas $\rightarrow \rightarrow$ Distribución Gamma

Gamma = conjugada de la veros. De Poisson:

1.-Como el parámetro λ es una promedio de hechos, tendremos que $\lambda \in [0,+\infty[$ y que se deberá cumplir que la densidad a priori $p(\lambda)$:

$\int_0^{\infty} p(\lambda) d\lambda = 1$. Obviamente una densidad gamma cumple esto.

2.- Debe ser conjugada a la d. Poisson: si $p(\theta) \sim Gamma(\alpha_0,\beta_0)$ la densidad a posteriori también debe seguir una Gamma:

$$p(\lambda \mid x) \propto p(\lambda).P(x \mid \lambda) \rightarrow Gamma(\alpha_1, \beta_1)$$

Como la densidad a priori $p(\lambda) \sim Gamma(\alpha_0,\beta_0)$:

$$p(\lambda) = \frac{(\beta_0)^{\alpha_0}}{\Gamma(\alpha_0)} e^{-\beta_0 . \lambda} \lambda^{\alpha_0 - 1}$$

con $\alpha_0, \beta_0 > 0$ parámetros de forma y escala respectivamente

La verosimilitud (d. Poisson) era:

$P(x_1, x_2, ..., x_n \mid \lambda) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_{i=1}^{n} x_i!} \propto e^{-n\lambda}\lambda^{n\bar{x}}$ con lo que la densidad a posteriori será:

$$p(\lambda \mid (x_1, x_2, ..., x_n) = \frac{p(\lambda).P((x_1, x_2, ..., x_n) \mid \lambda)}{\int_0^{\infty} p(\lambda).P((x_1, x_2, ..., x_n) \mid \lambda) d\lambda} \propto p(\lambda).P((x_1, x_2, ..., x_n) \mid \lambda)$$

$$p(\lambda \mid (x_1, x_2, ..., x_n) \propto e^{-\beta_0.\lambda}\lambda^{\alpha_0 - 1}.e^{-n\lambda}\lambda^{n\bar{x}}$$

$$p(\lambda \mid (x_1, x_2, ..., x_n) \propto e^{-(\beta_0 + n).\lambda}\lambda^{(\alpha_0 + n\bar{x}) - 1}$$

Expresión que se corresponde con la parte no constante de una densidad $Gamma(\alpha_1 = \alpha_0 + n\bar{x}; \beta_1 = \beta_0 + n)$

Por lo tanto:

$$\theta \sim \text{Gamma}\ (\alpha_0, \beta_0),\ x \sim \text{Poisson}(\lambda) \Rightarrow \lambda | x_1, x_2, ..., x_n \sim \text{Gamma}\ (\alpha_0 + n\bar{x},\ \beta_0 + n)$$

Recordemos que el primer parámetro de la distribución Gamma es el parámetro de forma y que el segundo es el de escala y que se cumple que su media es α/β y su varianza es α/β^2

- La estimación Bayes asociada a una "pérdida cuadrática" será la media de la distribución a posteriori:

$$E(\lambda \mid x_1, x_2, ..., x_n) = \frac{\alpha_0 + n\bar{x}}{\beta_0 + n} = \frac{1}{\beta_0 + n}\left(\beta_0.\left(\frac{\alpha_0}{\beta_0}\right) + n\bar{x}\right)$$

- Por tanto sería una combinación lineal (promedio ponderado) de la esperanza a priori y la media muestral.
- Es como si la información inicial representara la información de una muestra de β_0 intervalos/ámbitos de experimentación en los que se produjeran α_0 "hechos".
- La ausencia de información equivaldría a $\alpha_0 = \beta_0 = 0$

Al trabajar con R hay que tener en cuenta que el segundo parámetro de la distribución gamma , β , tal como lo estamos considerando aquí ,se corresponde con "rate" .

R-script

```
#prior
#alfa
a=40
#beta
b=20
print(paste("valor esperado a priori para lambda",a/b,"con varianza",a/b**2))
q<-qgamma(c(.999),shape=a,rate=b)
.lambda <- seq(0.001, q, length.out=1000)
plot(.lambda, dgamma(.lambda, shape=a, rate=b), type="l", xlab="lambda",
        ylab="Density",
        main=paste("D. apriori:gamma con forma=",a,"escala(rate)=",b))
remove (.lambda)
#muestra:num. de periodos o tandas de osservacion y ocurrencias
n=7
ocurrencias=16
media=ocurrencias/n
```

```
#verosimilitud
.x <- seq(0.001, 4*media, length.out=1000)
plot(.x, dpois(ocurrencias, lambda=.x*n), type="l",xlab="lambda", ylab="verosimilitud",
        main=paste("verosimilitud con ocurrencias=",ocurrencias,"muestra=",n) )
#posterior
aa=a+ocurrencias
bb=b+n
q<-qgamma(c(.999),shape=aa,rate=bb)
.lambda <- seq(0.001, q, length.out=1000)
plot(.lambda, dgamma(.lambda, shape=aa, rate=bb), type="l", xlab="lambda",
        ylab="Density",
        main=paste("D. a posteriori:gamma con forma=", aa," escala(rate)=",bb))
remove (.lambda)
print(paste("valor esperado a posteriori para lambda",aa/bb))
print(paste("con varianza",aa/bb**2))
```

Análisis de la conjugación con datos normales (σ conocida)

Consideramos n datos independientemente $y_i \sim N(\mu,\sigma)$ siendo μ desconocido pero conociendo la varianza

La verosimilitud será:

$$P(y_1, y_2, ..., y_n \mid \mu, \sigma) = \prod_{i=1}^{n} \left(\frac{e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \right) = \left(2\pi\sigma^2\right)^{-(n/2)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right)$$

Buscamos una clase de distribución que tomada como d. a priori p(μ)nos permita obtener (conjugadamente) una d. a posteriori de la misma clase:

Normal = conjugada de la veros. Normal (con σ conocida):

1.-Como el parámetro μ es la media de una Normal, tendremos que $\mu \in [-\infty,+\infty[$ y que se deberá cumplir que la densidad a priori $p(\mu)$: $\int_{-\infty}^{\infty} p(\mu) d\mu = 1$

Obviamente una densidad Normal cumple esto.

2.- Debe ser conjugada a la d. Normal si $p(\mu) \sim \text{Normal}(\mu_0,\sigma_0)$ la densidad a posteriori también debe seguir una Normal: $p(\mu \mid Y, \sigma) \propto p(\mu).P(Y \mid \mu, \sigma) \rightarrow N(\mu_1, \sigma_1)$

Para ver esto y en lo sucesivo tendremos en cuenta que en Estadística Bayesiana es habitual considerar que la d. Normal se especifica en función de los parámetros media y **precisión,** τ, con $\tau = 1/\sigma^2$, y expresaremos p(μ) y P(Y|μ,τ) como:

$$p(\mu) = \sqrt{\frac{\tau_0}{2\pi}} \exp(-\frac{1}{2}\tau_0(\mu-\mu_0)^2) \qquad P(Y \mid \mu, \tau) = \sqrt{\left(\frac{\tau}{2\pi}\right)^n} \exp(-\frac{1}{2}\tau \sum_{i=1}^{n} (y_i - \mu)^2)$$

Comprobemos ahora la conjugación: La d. a posteriori vendrá dada por:

$$p(\mu \mid Y, \tau) = \frac{p(\mu).P(Y \mid \mu, \tau)}{\int_{-\infty}^{\infty} p(\mu).P(Y \mid \mu, \tau) d\mu} \propto \exp\left(-\frac{\tau_0 (\mu-\mu_0)^2}{2} \right) .\exp\left(-\frac{\tau . \sum_{i=1}^{n} (y_i - \mu)^2}{2} \right)$$

Y por lo tanto:

$$p(\mu \mid Y,\sigma) \propto \exp\left(-\frac{1}{2}\tau_0(\mu-\mu_0)^2+\tau\sum_{i=1}^{n}\left((y_i-\bar{y})-(\mu-\bar{y})\right)^2\right) \propto \exp\left(-\frac{1}{2}\left(\tau_0(\mu-\mu_0)^2+\tau\left(nS_y^2-2(\mu-\bar{y}).\sum_{i=1}^{n}(y_i-\bar{y})+n(\mu-\bar{y})^2\right)\right)\right) \propto$$

$$\propto \exp\left(-\frac{1}{2}\left(\tau_0(\mu-\mu_0)^2+\tau\left(nS_y^2+n(\mu-\bar{y})^2\right)\right)\right) \propto \exp\left(-\frac{1}{2}\left(\tau_0\mu^2-2\tau_0\mu\mu_0+\tau_0\mu_0^2+\tau nS_y^2+\tau n\mu^2-2\tau n\mu\bar{y}+\tau n\bar{y}^2\right)\right) \propto$$

$$\propto \exp\left(-\frac{1}{2}\left((\tau_0+n\tau)\mu^2-2\mu(\tau_0\mu_0+n\tau\bar{y})+\tau_0\mu_0^2+\tau nS_y^2+\tau n\bar{y}^2\right)\right) \propto \exp\left(-\frac{1}{2}(\tau_0+n\tau)\left(\mu^2-2\mu(\frac{\tau_0\mu_0+n\tau\bar{y}}{\tau_0+n\tau})+\frac{\tau_0\mu_0^2+\tau nS_y^2+\tau n\bar{y}^2}{\tau_0+n\tau}\right)\right) \propto$$

$$\propto \exp\left(-\frac{1}{2}(\tau_0+n\tau)\left(\mu^2-2\mu(\frac{\tau_0\mu_0+n\tau\bar{y}}{\tau_0+n\tau})+\frac{\tau_0\mu_0^2+\tau n\bar{y}^2}{\tau_0+n\tau}\right)\right).\exp-\frac{1}{2}(\tau nS_y^2) \propto$$

$$\propto \exp\left(-\frac{1}{2}(\tau_0+n\tau)\left(\left(\mu-\frac{\tau_0\mu_0+n\tau\bar{y}}{\tau_0+n\tau}\right)^2+\frac{\tau_0\mu_0^2+\tau n\bar{y}^2}{\tau_0+n\tau}-\frac{(\tau_0\mu_0+n\tau\bar{y})^2}{(\tau_0+n\tau)^2}\right)\right)\exp-\frac{1}{2}(\tau nS_y^2) \propto$$

$$\propto \exp\left(-\frac{1}{2}(\tau_0+n\tau)\left(\left(\mu-\frac{\tau_0\mu_0+n\tau\bar{y}}{\tau_0+n\tau}\right)^2\right)\right)$$ ya que las expresiones en azul no dependen de μ

que se corresponde con la parte no constante de una Normal:

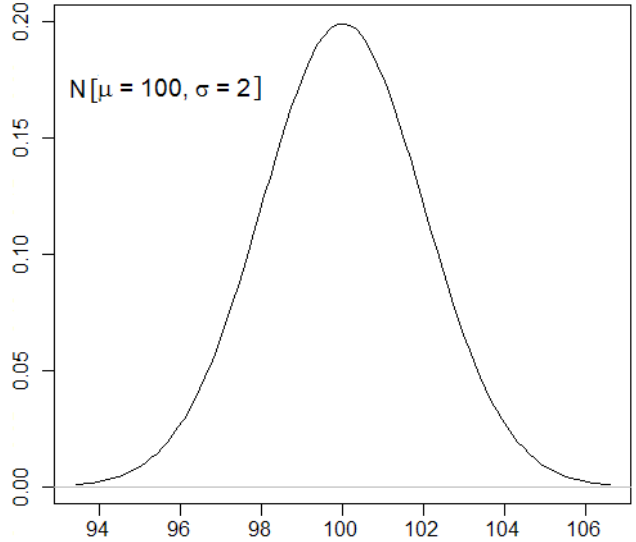
$$N\left(\mu_1=\frac{\tau_0\mu_0+n\tau\bar{y}}{\tau_0+n\tau};\tau_1=\tau_0+n\tau\right)$$

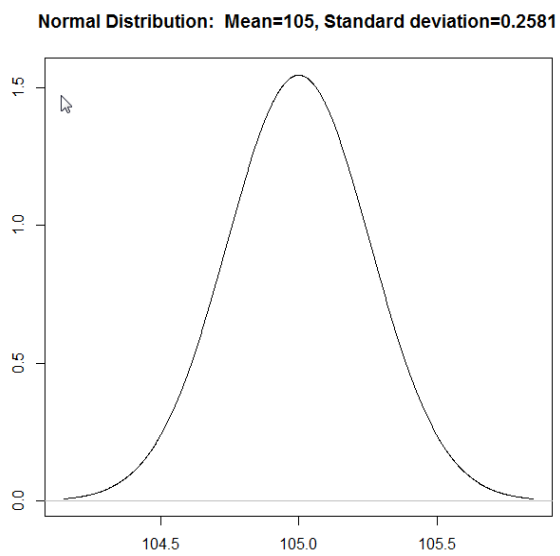
Concluimos para el caso de datos normales y σ o τ, conocidas que:

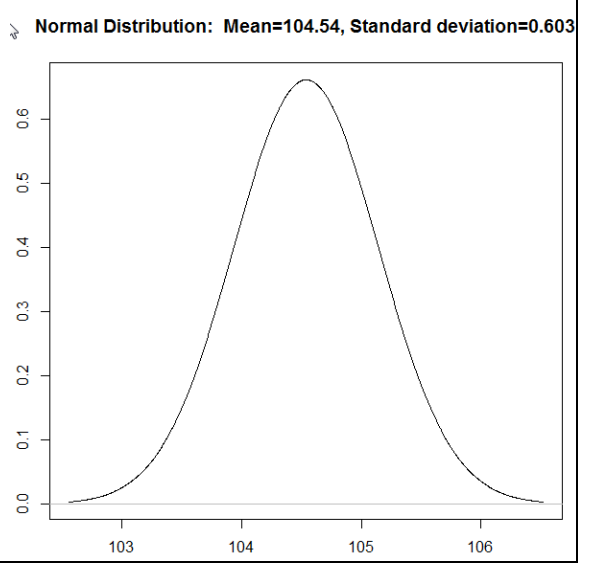
- La media de la d. a posteriori (estimación-bayes con pérdida cuadrática) es el promedio, ponderado por las precisiones, de la media a priori y la media de los datos

(supongamos que la varianza de la población es conocida $\sigma^2=8, \tau=1/8$)
(después hacemos una experiencia en la que tras analizar 20 casos observamos una media de 105 con una varianza de 6)

D. a priori	Datos: $n=20, \quad \bar{X}=105, S^2=6$	D. a posteriori
$N(100,\sigma_0=2) \equiv N(100,\tau_0=1/4)$	$verosimilitud: N(105, \sigma_v=\sqrt{\frac{\sigma^2}{nS^2}}=0.258) \equiv N(105, \tau_v=15)$	$N(\mu_1=\frac{\tau_0\mu_0+n\tau\bar{x}}{\tau_0+n\tau}; \tau_1=\tau_0+n\tau)$







Script de R

```
#
varianza de la población conocida
sigma2=8
tau=1/8

#a priori
media0=100
varianza0=4
sd0=sqrt(varianza0)
tau0=1/4

.x <- seq(media0-3.5126*sd0, media0+3.5126*sd0, length.out=1000)
plot(.x, dnorm(.x, mean=media0, sd=2), type="l", xlab="x",
        ylab="Density",
        main=paste("mu a priori-> Normal [",media0,";sigma=",sd0,  "]"))
remove(.x)



#muestra
n=20
mediamuestral=105
varianzamuestral=6
# equivalencia de la verosimilitud
varequival=sigma2/(n*varianzamuestral)
sdequival=sqrt(varequival)


.x <- seq(mediamuestral-3.5126*sdequival, mediamuestral+3.5126*sdequival,
length.out=1000)
plot(.x, dnorm(.x, mean=mediamuestral, sd=sdequival), type="l", xlab="x",
        ylab="Density",
        main=paste(" verosimilitud para mu equivale a Normal
[",mediamuestral,";sigma=",sdequival,  "]"))
remove(.x)


## a posteriori


media1=(tau0*media0+n*tau*mediamuestral)/(tau0+n*tau)
tau1=tau0+n*tau
sd1=sqrt(1/tau1)

.x <- seq(media1-3.5126*sd1, media1+3.5126*sd1, length.out=1000)
plot(.x, dnorm(.x, mean=media1, sd=sd1), type="l", xlab="x",
        ylab="Density",
        main=paste(" mu a posterriori -> Normal [",media1,";sigma=",sd1,  "]"))
remove(.x)
```

Análisis de la conjugación con datos normales (σ desconocida)

Distribución Gamma-inversa: Si cierta variable Y sigue una distribución Gamma(α,β) la variable inversa: X=1/Y seguirá una distribución Gamma-inversa(α,β) siendo su función de densidad:

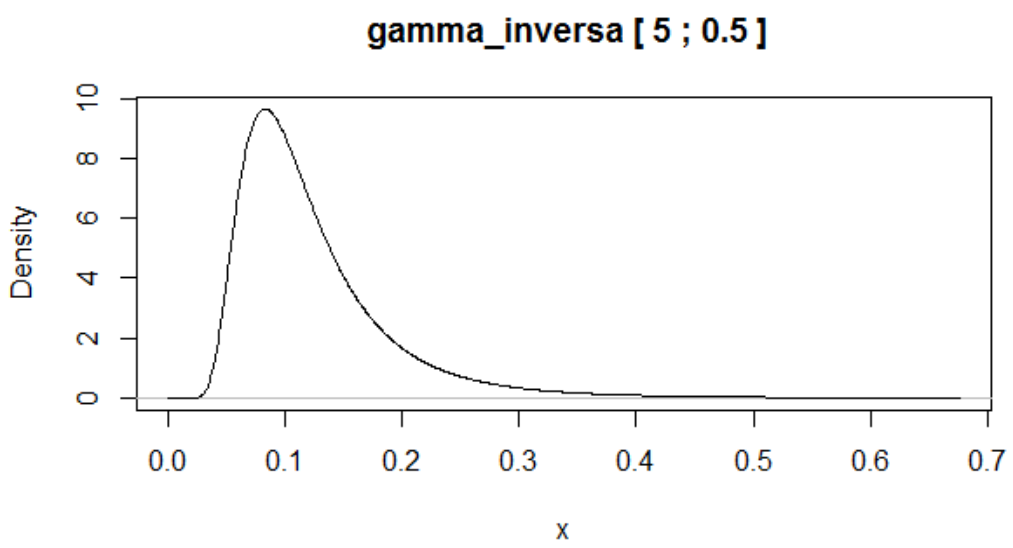
$$f(x)=\frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{-\alpha-1}e^{-\frac{\beta}{x}}$$

$$E(X)=\frac{\beta}{\alpha-1}\quad \text{si } \alpha>1 \qquad Moda=\frac{\beta}{\alpha+1}$$

$$Var(X)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\quad \text{si } \alpha>2$$

```
#igamma {pscl}
#densigamma(x,alpha,beta)
#pigamma(q,alpha,beta)
#qigamma(p,alpha,beta)
#rigamma(n,alpha,beta)
#igammaHDR(alpha,beta,content=.95,debug=FALSE)
library(pscl)

alpha=5
beta=0.5
.x <- seq(0.00123, qigamma(0.999,alpha,beta), length.out=1000)
plotDistr(.x, densigamma(.x, alpha, beta), cdf=FALSE, xlab="x",
        ylab="Density",
        main=paste(" gamma_inversa [",alpha,";",beta,  "]"))
remove(.x)
```



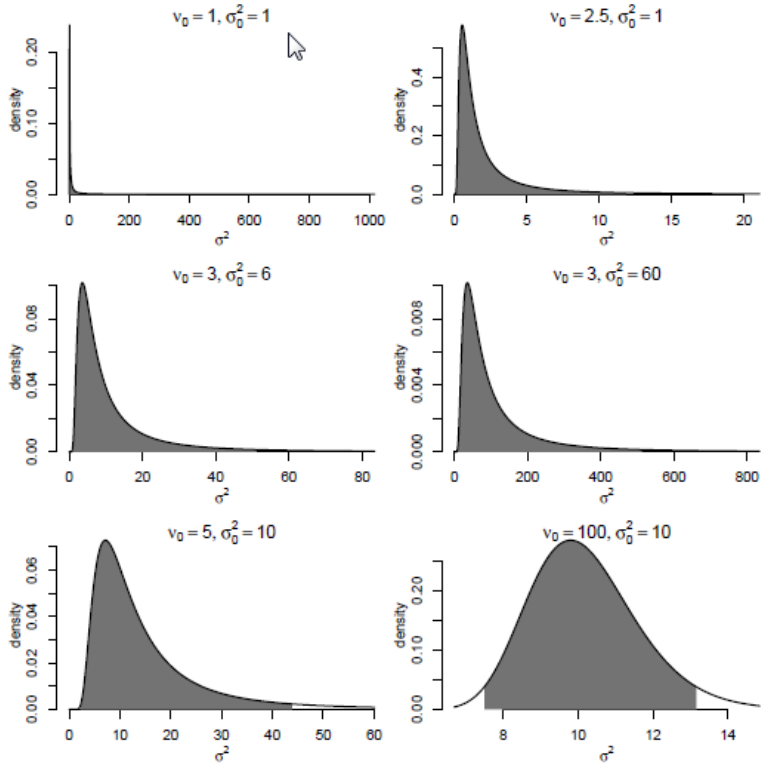
Análisis de la conjugación con datos normales (σ desconocida)

El uso bayesiano típico de la distribución Gamma-inversa es considerarla como la distribución (a priori) de la varianza de una Normal según:

$$\sigma^2 \rightarrow Gamma - inversa(\frac{\nu_0}{2}, \frac{\nu_0 \sigma^2_0}{2})$$

Lo que supone que la media será : $E(\sigma^2) = \frac{\nu_0 \sigma^2_0}{\nu_0 - 2}$ siempre que $\nu_0 > 2$ y que tendrá la moda en $\frac{\nu_0 \sigma^2_0}{\nu_0 + 2}$

Cuando $\nu_0 \rightarrow \infty$ media y moda tenderán a coincidir y la distribución irá ganando simetría hasta converger a una normal.



Análisis de la conjugación con datos normales (σ desconocida)

De hecho la densidad a priori (y también a posteriori, por conjugación) que suele usarse para el vector de parámetros (μ, σ^2)'es la de una Normal-Gamma-inversa.

Caracterizada por tener como distribución de μ condicionada a σ^2 una distribución normal y como distribución marginal de σ^2 na Gamma-inversa: esto es:

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \rightarrow NGI(\mu_0, n_0, \nu_0, \sigma_0^2) \Rightarrow \begin{cases} \mu \mid \sigma^2 \rightarrow N(\mu_0, \text{var} = \dfrac{\sigma^2}{n_0}) \\ \sigma^2 \rightarrow Gamma - Inversa(\dfrac{\nu_0}{2}, \dfrac{\nu_0 \sigma_0^2}{2}) \end{cases}$$

Y la f. densidad conjunta a priori vendrá dada por :

$$f((\mu, \sigma^2)') = f(\mu \mid \sigma^2).f(\sigma^2) = \frac{\exp[-\frac{n_0}{2\sigma^2}(\mu - \mu_0)]}{\sigma\sqrt{2\pi}\ \Gamma(\frac{\nu_0}{2})} \left(\frac{\nu_0 \sigma_0^2}{2}\right)^{\frac{\nu_0}{2}} \left(\sigma^2\right)^{-\frac{\nu_0}{2}-1} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma^2}}$$

Análisis de la conjugación con datos normales (σ desconocida)

Puede probarse que si partimos de una distribución conjunta a priori para los parámetros media y varianza :

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \rightarrow NGI(\mu_0, n_0, \nu_0, \sigma_0^2)$$

Y tomamos n datos $Y_i \rightarrow N(\mu,\sigma)$ para i=1,2,…, n

La distribución conjunta a posteriori de los dos parámetros vendrá dada también por una NormalGammaInversa:

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} | Y \rightarrow NGI(\mu_1, n_1, \nu_1, \sigma_1^2)$$ con:

$$\begin{aligned} \mu_1 &= \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \\ n_1 &= n_0 + n, \quad \nu_1 = \nu_0 + n \\ \nu_1\sigma_1^2 &= \nu_0\,\sigma_0^2 + S + \frac{n_0\, n}{n_0 + n}(\mu_0 - \bar{y})^2 \end{aligned}$$

$$\begin{aligned} \mu|\,\sigma^2, \mathbf{y} &\sim N(\mu_1, \sigma^2/n_1) \\ \sigma^2|\,\mathbf{y} &\sim \textit{inverse-Gamma}\left(\frac{\nu_1}{2}, \frac{\nu_1\,\sigma_1^2}{2}\right) \end{aligned}$$

Donde : $$S = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Análisis de la conjugación con datos normales (σ desconocida).

Distribución Marginal de μ a posteriori

- La distribución a posteriori de **μ** condicionada a **σ²** es : $\mu \mid \sigma^2, Y \rightarrow N(\mu_1; \text{var} = \frac{\sigma^2}{n_1})$, como ya hemos visto.
- Pero si estamos interesados en inferencias sobre **μ** necesitamos conocer su distribución marginal (no condicionada a la varianza)
- La densidad marginal de la media (a posteriori) será la integral con respecto a la varianza de la densidad conjunta a posteriori , **(μ,σ²)|Y,** esto es $p(\mu \mid Y) = \int_0^\infty p(\mu, \sigma^2 \mid Y)\, d(\sigma^2) = \int_0^\infty p(\mu \mid \sigma^2, Y)\, p(\sigma^2 \mid Y)\, d(\sigma^2)$
- Esta densidad marginal a posteriori acaba siendo la de una **t de Student :**En concreto **μ|Y seguirá una t de Student** (descentrada) de parámetro de localización μ_1 , y parámetro de escala $\sqrt{\frac{\sigma_1^2}{n_1}}$, con ν_1 grados de libertad

$$\mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_1} \qquad n_1 = n_0 + n \qquad \nu_1 = \nu_0 + n$$

$$\sigma_1^2 = S_1/\nu_1,\ S_1 = \nu_0\,\sigma_0^2 + (n-1)s^2 + \frac{n_0\, n}{n_1}(\bar{y} - \mu_0)^2 \qquad s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Equivalentemente: $$\frac{\mu - \mu_1}{\sqrt{\frac{\sigma_1^2}{n_1}}} \rightarrow t_{\nu_1}$$

R-script:

```
library(pscl)
##distribucion a priori
##sobre la media a priori
mu0=100 ##valor central
n0=1 ## precision inicial para N[mu0; sd=sigma/raiz(n0)]
##sobre la varianza a priori
vesp=100 ##valor eesperado a priori para la varianza
vesp=(v0*sigma2)/(v0-2)
dispers=0.5 ##dispersión a priori cvpearson=1/sqrt((vo/2)-2)
v0=((1/dispers)**2+2)*2
sigmados0=vesp*(v0-2)/v0
##varianza apriori
alpha=v0/2
beta=v0*sigmados0/2
.x <- seq(qigamma(0.001,alpha,beta), qigamma(0.99,alpha,beta),
length.out=1000)
plot(.x, densigamma(.x, alpha, beta),  xlab="varianza",
    ylab="Density", type="l",
     main=paste("varianza a priori~ gamma_inversa
[",alpha,";",beta,  "]"))
remove(.x)
##media a priori (marginal)
print(paste("a priori la variable (mu -",mu0,"/",sqrt(1/n0),")sigue
una t de Student con",v0,"grados de libertad"))
.x <-seq(qt(0.01,v0)*sqrt(n0)+mu0, qt(0.99,v0)*sqrt(n0)+mu0,
length.out=1000)
plot(.x, dt((.x-mu0)*sqrt(n0),v0),  xlab="media",
    ylab="Density", type="l",
    main=paste("(media a priori - ",mu0,")/",n0,"~ t de student
[",v0,"g.l]"))
remove(.x)
##verosimilitud
##muestra
n=20 #tamaño muestral
media=109.2 #media muestral
varianza=136#varianza muestral
```

```
S=n*varianza
###verosimilitud de la media
zz=4*sqrt(1.5*varianza/n)
sigm <- c(varianza*0.8, varianza*0.9,
varianza*1.1,varianza*1.2)
x <- seq(-zz+media, zz+media, length=100)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c(paste("var=",sigm[1]), paste("var=",sigm[2]),
paste("var=",sigm[3]),
paste("var=",sigm[4]),paste("var=",varianza))
plot(x, dnorm(media,x,sqrt(varianza/n)), type="l", lty=2,
xlab="media",
    ylab="verosimilitud", main="verosimilitudes de MU para
distintas varianzas")
for (i in 1:4){
  lines(x, dnorm(media,x,sqrt(sigm[i]/n)), lwd=2, col=colors[i])
}
legend("topright", inset=.05,
      labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
##verosimilitud de la varianza
zz=.9*varianza
med <- c(media-0.5*sqrt(varianza/n), media-
0.3*sqrt(varianza/n),  media+0.4*sqrt(varianza/n),
media+0.6*sqrt(varianza/n))
x <- seq(-zz+varianza, zz+varianza, length=100)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c(paste("mu=",med[1]), paste("mu=",med[2]),
paste("mu=",med[3]), paste("mu=",med[4]),paste("mu",media))
plot(x, dnorm(media,media,sqrt(x/n)), type="l", lty=2,
xlab="varianza",
    ylab="verosimilitud", main="verosimilitudes de sigma2 para
distintas mus")
for (i in 1:4){
  lines(x, dnorm(media,med[i],sqrt(x/n)), lwd=2, col=colors[i])
}
legend("topright", inset=.05,
```

```
      labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
###a posteriori
mu1=(n0*mu0+n*media)/(n0+n)
n1=n0+n
v1=v0+n
v1sigma1=v0*sigmados0+S+(n0*n/(n0+n)*(mu0-media)**2)
sigma1=v1sigma1/v1
##varianza a posteriori
alpha1=v1/2
beta1=v1sigma1/2
#min=qigamma(0.001,alpha1,beta1)
#max=qigamma(0.999,alpha1,beta1)
.x <- seq(qigamma(0.001,alpha1,beta1),
qigamma(0.99,alpha1,beta1),length.out=1000)
plot(.x,densigamma(.x,alpha1,beta1), xlab="varianza",
    ylab="Densidad",type="l",
    main=paste("varianza a posteriori~ gamma_inversa
[",alpha1,";",beta1,"]"))
remove(.x)
print("estimacion para la varianza asumiendo pérdida
cuadrática")
varianza.estimada=v1sigma1/(v1-2)
print(varianza.estimada)
##media a posteriori
escala=sqrt(sigma1/n1)
print(paste("a posteriori la variable (mu -",mu1,"/",escala,")sigue
una t de Student con",v1,"grados de libertad"))
.x <-seq(qt(0.01,v1)*escala+mu1,
qt(0.99,v1)*escala+mu1,length.out=1000)
plot(.x, dt((.x-mu1)/escala,v0),  xlab="media",
    ylab="Densidad", type="l",
    main=paste("(media a priori - ",mu1,")/",escala,"~ t de
student [",v1,"g.l]"))
remove(.x)
```

Regression

Theorem (Conjugate Prior Normal Regression Model)

$$
\begin{aligned}
y_i|\mathbf{x}_i &\overset{iid}{\sim} N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2) \\
\boldsymbol{\beta}|\sigma^2 &\sim N(\mathbf{b}_0, \sigma^2\mathbf{B}_0) \\
\sigma^2 &\sim \textit{inverse-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}
$$

then

$$
\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} &\sim N(\mathbf{b}_1, \sigma^2\mathbf{B}_1), \\
\sigma^2|\mathbf{y}, \mathbf{X} &\sim \textit{inverse-Gamma}(\nu_1/2,\ \nu_1\,\sigma_1^2/2) \\
\mathbf{b}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) \\
\mathbf{B}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1} \\
\nu_1 &= \nu_0 + n \qquad \textit{and} \\
\nu_1\sigma_1^2 &= \nu_0\sigma_0^2 + S + r.
\end{aligned}
$$

Bibliografía :

- Box,G. y Tiao,G : “Bayesian Inference in Statistical Analysis” Addison-Wesley,1973
- Raiffa,H. y Schlaifer,R .: “Applied Statistical Decision Theory”. M.I.T.Press,1971
- Jackman, S.: “Bayesian Analysis for Social Science” Willey, 2009.