Muestreador de Gibbs para una mixtura de distribuciones normales

```
library(openxlsx)
data<- read.xlsx("fish.xlsx")
y<-data$lenght
```

1a)Tenemos una muestra aleatoria x_i para $i = 1, \ldots, 523$, donde cada x_i representa la longitud del i-ésimo pez en la muestra. Así las cosas tenemos que la función de verosimilitud para cada x en la muestra esta dada por:

$$p(x) = \alpha * \phi(x|\theta_1, 2.8^2) + (1 - \alpha) * \phi(x|\theta_2, 8.5^2)$$

Donde $\phi(x|\theta, \sigma^2)$ es la fdp usual de una variable normal con media θ y varianza σ^2. Por lo tanto, tenemos entonces una probabilidad α de tomar una observación de $\phi(x|\theta_1, 2.8^2)$ y $(1-\alpha)$ de tomar una observación de $\phi(x|\theta_2, 8.5^2)$. Dado que es complicado calcular la distribución posterior para α,θ_1 y θ_2 con prioris conjugadas usando un modelo de mixturas normales, usaremos entonces dos variables indicadoras $z = \{z_1, z_2\}$ definidas como:

$$z_{i1} = \begin{cases} 1, & \text{si } x_i < 40 \\ 0, & e.o.c \end{cases}$$

Y z_{i2} es dada por:

$$z_{i2} = \begin{cases} 1, & \text{si } x_i > 40 \\ 0, & e.o.c \end{cases}$$

Entonces la distribución conjunta para la mixtura y la variable indicadora puede ser factorizada como:

$$p(x, z; \theta) = p(x|z, \theta) * p(z, \theta)$$

Donde

$$p(x|z, \theta) = (\phi(x|\theta_1, 2.8^2))^{z_1} * (\phi(x|\theta_2, 8.5^2))^{z_2}$$

$p(z, \theta)$ es entonces una distribución multinomial con densidad dada por:

$$p(z, \theta) \propto \alpha^{z_1} * (1 - \alpha)^{z_2}$$

$$p(z, \theta) = \prod_{j=1}^{2} \alpha_j^{z_j}$$

Donde $\alpha_1 = \alpha$ y $\alpha_2 = 1 - \alpha$. Entonces tenemos que la fdp conjunta para la variable indicadora y la mixtura es:

$$p(x,z;\theta) = \prod_{i=1}^{N}[\alpha * \phi(x|\theta_1, 2.8^2))]^{z_1} * [(1-\alpha) * \phi(x|\theta_2, 8.5^2)]^{z_2}$$

Para este problema tenemos que nuestras distibuciones priori sobre α, θ_1 y θ_2 son:

$$p(\alpha) \sim Beta(1,10)$$
$$p(\theta_1) \sim N(30,10)$$
$$p(\theta_2) \sim N(50,20)$$

Conectando nuestros hiperparámetros directamente a nuestras densidades obtenemos las siguientes distribuciones prioris:

$$p(\alpha) \propto \alpha^{1-1} * (1-\alpha)^{10-1}$$
$$p(\theta_1) \propto exp[-\frac{(\theta_1-30)^2}{2*10}]$$
$$p(\theta_2) \propto exp[-\frac{(\theta_2-50)^2}{2*20}]$$

Por lo tanto la distribución posterior para θ es:

$$p(\theta|x,z) \propto p(x,z|\theta)p(\alpha)\prod_{j=1}^{2}[p(\theta_j)]$$
$$\propto \prod_{i=1}^{N}\alpha^{z_1+1-1} * \phi(x_i|\theta_1, 2.8^2)^{z_1}\prod_{i=1}^{N}(1-\alpha)^{z_2+10-1}\phi(x_i|\theta_2, 8.5^2)^{z_2}\prod_{j=1}^{2}exp[-\frac{(\theta_j-\theta_{0j})^2}{2\sigma_{0j}^2}]$$
$$\propto \alpha^{\sum_{i=1}^{N} z_1 + 1 - 1}(1-\alpha)^{\sum_{i=1}^{N} z_2 + 10 - 1}\prod_{i=1}^{N}\prod_{j=1}^{2}\phi(x_i|\theta_j,\sigma_j^2)^{z_j}\prod_{j=1}^{2}exp[-\frac{(\theta_j-\theta_{0j})^2}{2\sigma_{0j}^2}]$$

Obtengamos entonces apartir de esta expresión anterior la distribución marginal condicional para α

$$p(\alpha|x,z) = \alpha^{\sum_{i=1}^{N} z_1 + 1 - 1}(1-\alpha)^{\sum_{i=1}^{N} z_2 + 10 - 1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{i=1}^{N}\prod_{j=1}^{2}\phi(x_i|\theta_j,\sigma_j^2)^{z_j}\prod_{j=1}^{2}exp[-\frac{(\theta_j-\theta_{0j})^2}{2\sigma_{0j}^2}]d_{\theta_1}d_{\theta_2}$$
$$\propto \alpha^{\sum_{i=1}^{N} z_1 + 1 - 1}(1-\alpha)^{\sum_{i=1}^{N} z_2 + 10 - 1}$$
$$\rightarrow p(\alpha|x,z) \sim Beta(\sum_{i=1}^{N} z_1 + 1, \sum_{i=1}^{N} z_2 + 10)$$

Bajo el mismo enfoque, obtengamos la distribución marginal condicional para θ_1, y dado que θ_2 sigue la misma distribución priori, entonces obteniendo la distribución condicional para θ_1 seria lo mismo para θ_2.

$$p(\theta_1|x,z) = \prod_{i=1}^{N}\phi(x_i|\theta_1, 2.8^2)^{z_1}exp[-\frac{(\theta_1-30)^2}{2*10}]\int_{0}^{\infty}\int_{-\infty}^{\infty}\alpha^{\sum_{i=1}^{N} z_1 + 1 - 1}(1-\alpha)^{\sum_{i=1}^{N} z_2 + 10 - 1}\prod_{i=1}^{N}\phi(x_i|\theta_2, 8.5^2)^{z_2}exp[-\frac{(\theta_2-50)^2}{2*20}]d_\pi d_{\theta_2}$$

$$\propto \prod_{i=1}^{N} \phi(x_i|\theta_1, 2.8^2)^{z_1} exp[-\frac{(\theta_1 - 30)^2}{2*10}]$$
$$\propto exp\Big[\frac{-\sum_{i=1}^{N} z_{i1}(x_i - \theta_1)^2}{2*2.8^2} - \frac{(\theta_1 - 30)^2}{2*10}\Big]$$
$$\propto exp\Big[-\frac{\sum_{i=1}^{N} z_{i1}x_i^2 - 2\theta_1 z_{i1}x_i + z_{i1}\theta_1^2}{2*2.8^2} - \frac{\theta_1^2 - 60\theta_1 + 30^2}{2*10}\Big]$$
$$\propto exp\Big[-\frac{\sum_{i=1}^{N} 10z_{i1}x_i^2 - 20\theta_1 z_{i1}x_i + 10\theta_1^2 z_{i1} + 7.84\theta_1^2 - 470.4\theta_1 + 30^2 * 2.8^2}{2*10*2.8^2}\Big]$$

Tomemos $\sum_{i=1}^{N} z_{i1}x_i = \tilde{x}$ y $\sum_{i=1}^{N} z_{i1} = n_1$, por métodos vistos en clase sabemos que aquellos términos que no dependen de θ_1 son absorbidos mediante un término constante. Así las cosas tenemos que juntar términos semejantes, y luego completar cuadrados, para obtener la distribución posterior para θ_1.

$$\propto exp\Big[-\frac{\theta_1(20\tilde{x_1} + 470.4) + \theta_1^2(10n_1 + 7.84)}{2*10*2.8^2}\Big]$$
$$\propto exp\Big[-(10n_1 + 7.84)\frac{\theta_1^2 + 2\big(\frac{20\tilde{x_1}+470.4}{10n_1+7.84}\big)\theta_1 - \big(\frac{20\tilde{x_1}+470.4}{10n_1+7.84}\big)^2 + \big(\frac{20\tilde{x_1}+470.4}{10n_1+7.84}\big)^2}{2*10*2.8^2}\Big]$$
$$\propto exp\Big[-(10n_1 + 7.84)\frac{\big(\theta_1 - \frac{20\tilde{x_1}+470.4}{10n_1+7.84}\big)^2}{2*10*2.8^2}\Big]$$
$$\rightarrow p(\theta_1|x, z) \sim N\Big(\frac{20\tilde{x_1} + 470.4}{10n_1 + 7.84}, \frac{78.4}{10n_1 + 7.84}\Big)$$
$$\rightarrow p(\theta_2|x, z) \sim N\Big(\frac{40\tilde{x_2} + 7225}{20n_1 + 72.25}, \frac{1445}{20n_2 + 72.25}\Big)$$

Obtengamos también la distribución marginal condicional para las variables indicadoras, Useamos probabilidad condicional para lograr esto:

$$p(z|\theta, x) = \frac{p(\theta, x, z)}{p(\theta, x)}$$
$$= \frac{p(x|z, \theta) * p(z|\theta) * p(\theta)}{p(x|\theta)p(\theta)}$$
$$= \frac{p(x|z, \theta) * p(z|\theta)}{p(x|\theta)}$$
$$= \frac{\alpha_j \phi(x_i|\theta_j, \sigma_j^2)}{\sum_{j=1}^{2} \alpha_j \phi(x_i|\theta_j, \sigma_j^2)}, i = 1, \ldots, 523$$

Algoritmo

1) Comenzar con valores iniciales para α, θ_1 y θ_2

2) repetir para $t = 1, 2, \ldots$

a) calcular la probabilidad de que salga una obsservación de la población 1:

$$z_t \sim Ber\Big(1 - \frac{(1-\alpha_t)\phi(x_i|\theta_2, 8.5^2)}{\alpha_t\phi(x_i|\theta_1, 2.8^2) + (1-\alpha_t)\phi(x_i|\theta_2, 8.5^2)}\Big)$$

$$\alpha_t \sim Beta\Big(\sum_{i=1}^{N} z_1 + 1, \sum_{i=1}^{N} z_2 + 10\Big)$$

b) si $z_t = 1$, entonces actualizar θ_1, de lo contrario actualizar θ_2.

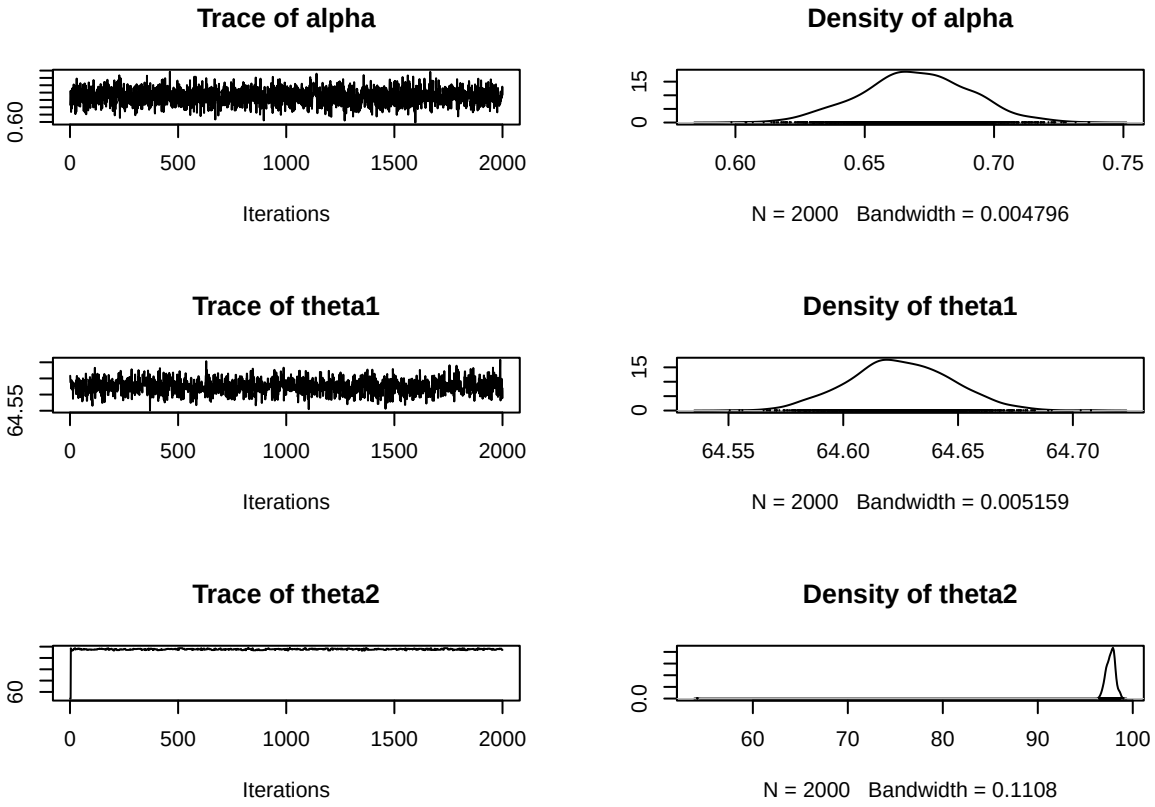
3) Repetir 2 hasta alcanzar una distribución estacionaria para los parámetros.

Muestreador de Gibbs en R

```
#punto a
z1_cond<-function(y,alpha,theta_1,theta_2){
  a=(alpha)*sum(dnorm(y,theta1_ini,2.8))
  b=(1-alpha)*sum(dnorm(y,theta2_ini,8.5))
  pi_i=1-(b/(a+b)) #probabilidad de escogencia de la población 1
  z1=rbinom(1,1,prob =pi_i)
  return(z1)
}
theta1_cond<-function(data){
  theta1<-rnorm(1,mean=(20*sum(data)+470.4)/(10*length(data)+7.84),sd=(78.4)/(10*length(data)+7.84))
  return(theta1) #Distribucion normal que se obtuvo analiticamente
}
theta2_cond<-function(data){
  theta2<-rnorm(1,mean=(40*sum(data)+7225)/(20*length(data)+72.25),sd=(1445)/(20*length(data)+72.25))
  return(theta2) #Distribucion normal que se obtuvo analiticamente

}
alpha_cond<-function(N,n){
  #N es el tamaño de la poblacion total
  #n es el tamaño de la poblacion seleccionada
  return(rbeta(1,n+1,(N-n)+10))
}
gibbs<-function(data,alpha,theta_1,theta_2,n_iter){
  alpha_salida<-c()
  theta1_salida<-c()
  theta2_salida<-c()
  alpha_actual<-alpha
  theta1_actual<-theta_1
  theta2_actual<-theta_2
  datap<-0
  for(i in 1:n_iter){
    z1<-z1_cond(y=data,alpha=alpha_actual,theta_1=theta1_actual,theta_2=theta2_actual)
    if(z1==1){datap<-data[data<40]
    theta1_actual<-theta1_cond(datap)}
    else if(z1==0){datap<-data[data>40]
    theta2_actual<-theta2_cond(datap)}
    alpha_actual<-alpha_cond(length(data),length(data[data<40]))
    alpha_salida[i]<-alpha_actual
    theta1_salida[i]<-theta1_actual
    theta2_salida[i]<-theta2_actual
```

```
  }
  return(data.frame(alpha=alpha_salida, theta1=theta1_salida,theta2=theta2_salida))
}
alpha_ini<-rbeta(1,1,10)
theta1_ini<-rnorm(1,30,sqrt(10))
theta2_ini<-rnorm(1,50,sqrt(20))
set.seed(42)
posterior<-gibbs(data=y,alpha=alpha_ini,theta_1 =theta1_ini,theta_2 =theta2_ini,n_iter =2000)
plot(as.mcmc(posterior))
```



Notemos que para los tres parámetros obtuvimos distribuciones estacionarias de manera inmediata, lo que quiere decir que el muestreador de Gibbs funciono bien, además las formas de las distribuciones posteriores concuerdan con la teoría.

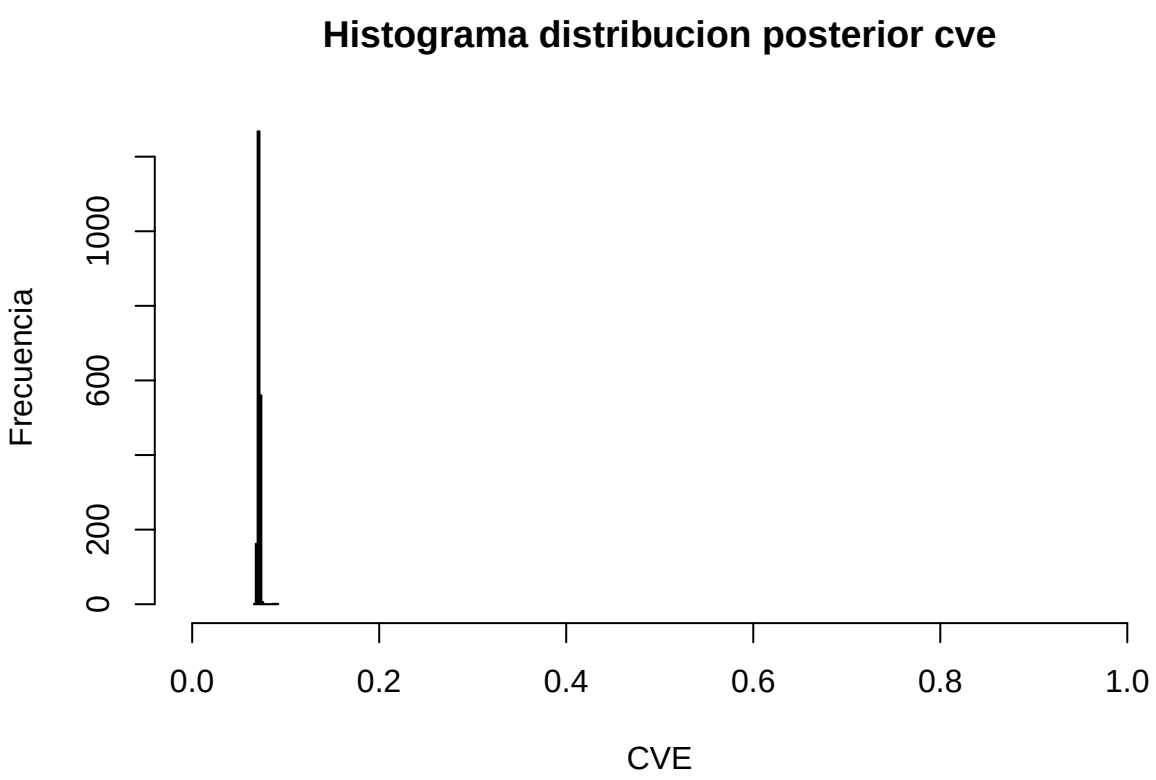
1b)Obtengamos entonces el valor esperado y la desviación estándar de esta mixtura.

$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} xp(x)dx \\ &= \int_{-\infty}^{\infty} x(\alpha * \phi(x|\theta_1, 2.8^2) + (1-\alpha) * \phi(x|\theta_2, 8.5^2))dx \\ &= \alpha \int_{-\infty}^{\infty} x\phi(x|\theta_1, 2.8^2)dx + (1-\alpha) \int_{-\infty}^{\infty} x\phi(x|\theta_2, 8.5^2)dx \\ &= \alpha\theta_1 + (1-\alpha)\theta_2 \\ &= \alpha(\theta_1 - \theta_2) + \theta_2 \end{aligned}$$

Utilizando la definición de varianza, tenemos que:

$$
\begin{aligned}
Var(x) &= \int_{-\infty}^{\infty} (x - E(x))^2 p(x) dx \\
&= \int_{-\infty}^{\infty} (x - E(x))^2 (\alpha * \phi(x|\theta_1, 2.8^2) + (1-\alpha) * \phi(x|\theta_2, 8.5^2)) dx \\
&= \alpha \int_{-\infty}^{\infty} (x - E(x))^2 \phi(x|\theta_1, 2.8^2) dx + (1-\alpha) \int_{-\infty}^{\infty} (x - E(x))^2 \phi(x|\theta_2, 8.5^2) dx \\
&= 2.8^2\alpha + (1-\alpha)8.5^2 \\
&= 72.25 - 64.41\alpha \\
\rightarrow \frac{SD(x)}{E(x)} &= \frac{\sqrt{72.25 - 64.41\alpha}}{\alpha(\theta_1 - \theta_2) + \theta_2}
\end{aligned}
$$

```
#punto b
set.seed(43)
posterior$cve<-sqrt(72.25-64.41*posterior$alpha)/(posterior$alpha*(posterior$theta1-posterior$theta2)+p
hist(posterior$cve,xlim=c(0,1),main = "Histograma distribucion posterior cve",xlab = 'CVE',ylab='Frecue
```



Notemos que para la distribución posterior del CVE, sus valores simulados nos dan menores a 0.05 aproximadamente, por lo que tenemos entonces un conjunto de datos homogéneo el cual su valor esperado es significativo, por lo tanto podemos decir que hay una dependencia alta sobre los parámetros θ_1 y θ_2 que son las medias de las dos poblaciones en el problema y la probabilidad de su escogencia α.