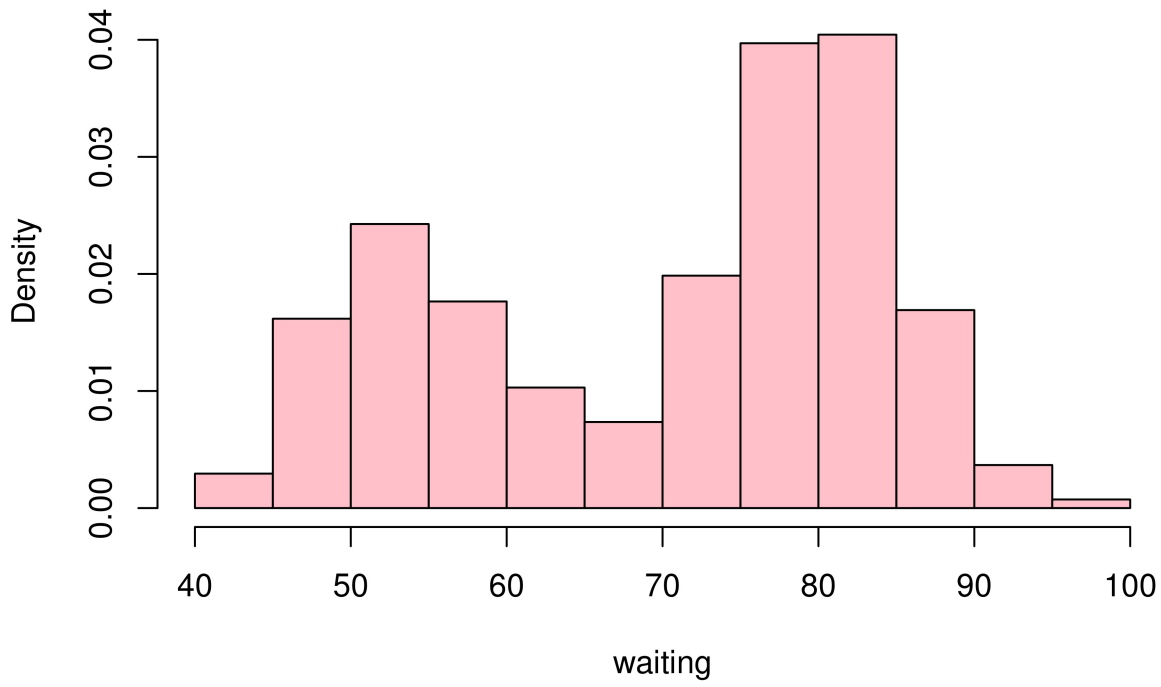
PBSR Assignment 2 Q.3

2022-11-14

Problem 3: Analysis of `faithful` datasets.

Consider the `faithful` datasets:

```
attach(faithful)
hist(faithful$waiting,xlab = 'waiting',probability = T,col='pink',main='')
```



Fit following three models using MLE method and calculate **Akaike information criterion** (aka., AIC) for each fitted model. Based on AIC decides which model is the best model? Based on the best model calculate the following probability

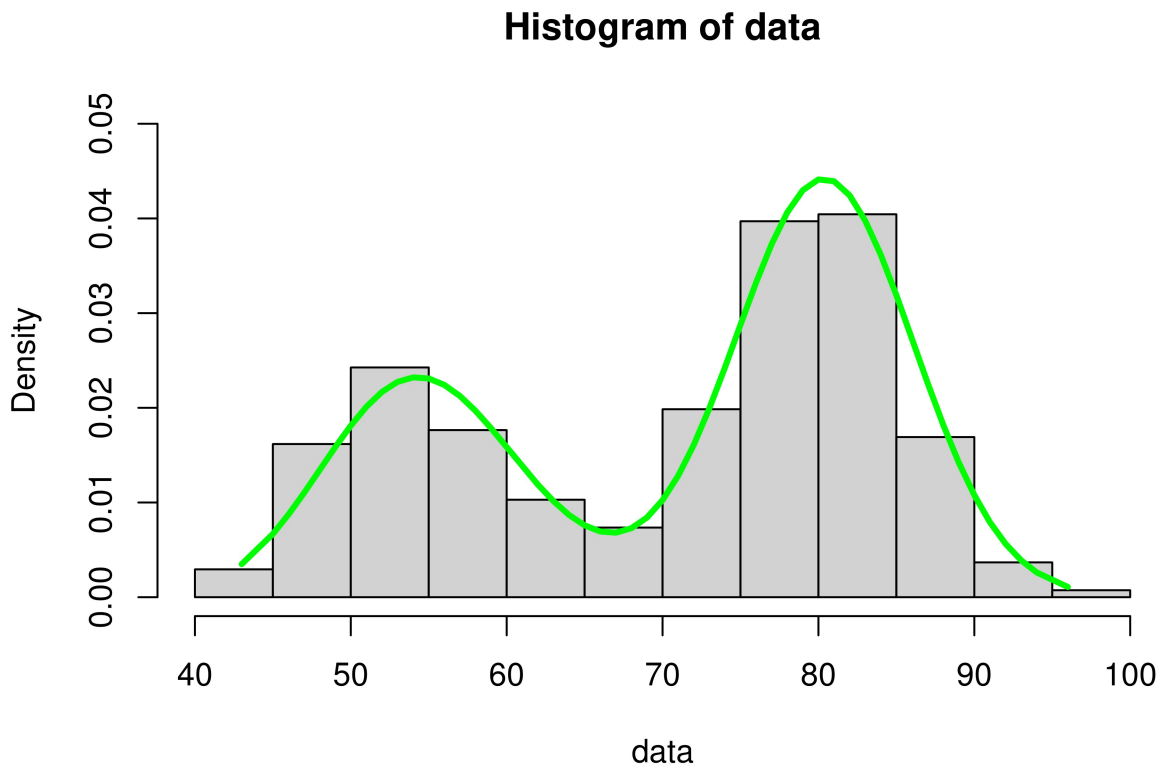
$$\mathbb{P}(60 < \texttt{waiting} < 70)$$

```
data = sort(faithful$waiting)
```

(i) Model 1:

$$f(x) = p * Gamma(x|\alpha, \sigma_1) + (1-p)N(x|\mu, \sigma_2^2), \quad 0 < p < 1$$

```
loglike1 = function(theta,data){
   alpha = exp(theta[1])
   beta = exp(theta[2])
   mu = theta[3]
   sigma = exp(theta[4])
   p = exp(theta[5])/(1+exp(theta[5]))
   n = length(data)
   l=0
   for(i in 1:n){
      l = l + log(p*dgamma(data[i],shape = alpha, rate = beta)
                  +(1-p)*dnorm(data[i], mean = mu, sd = sigma))
   }
   return(-l)
}

theta_initial=c(4.4,0.47,75,8,0.35)

fit = optim(theta_initial, loglike1, data = data, control = list(maxit=2000))

theta_hat = fit$par
alpha_hat = exp(theta_hat[1])
beta_hat = exp(theta_hat[2])
mu_hat = theta_hat[3]
sigma_hat = exp(theta_hat[4])
p_hat = exp(theta_hat[5])/(1+exp(theta_hat[5]))

d_mle = p_hat*dgamma(data, shape = alpha_hat, rate = beta_hat)+
   (1-p_hat)*dnorm(data, mean = mu_hat,sd = sigma_hat)

hist(data, probability = T, ylim = c(0, 0.05))
lines(data, d_mle,lwd=3,col='green')
```




```
AIC = 2*length(fit$par) - 2*(-fit$value)
## AIC Value for model 1
AIC
```

```
## [1] 2076.506
```

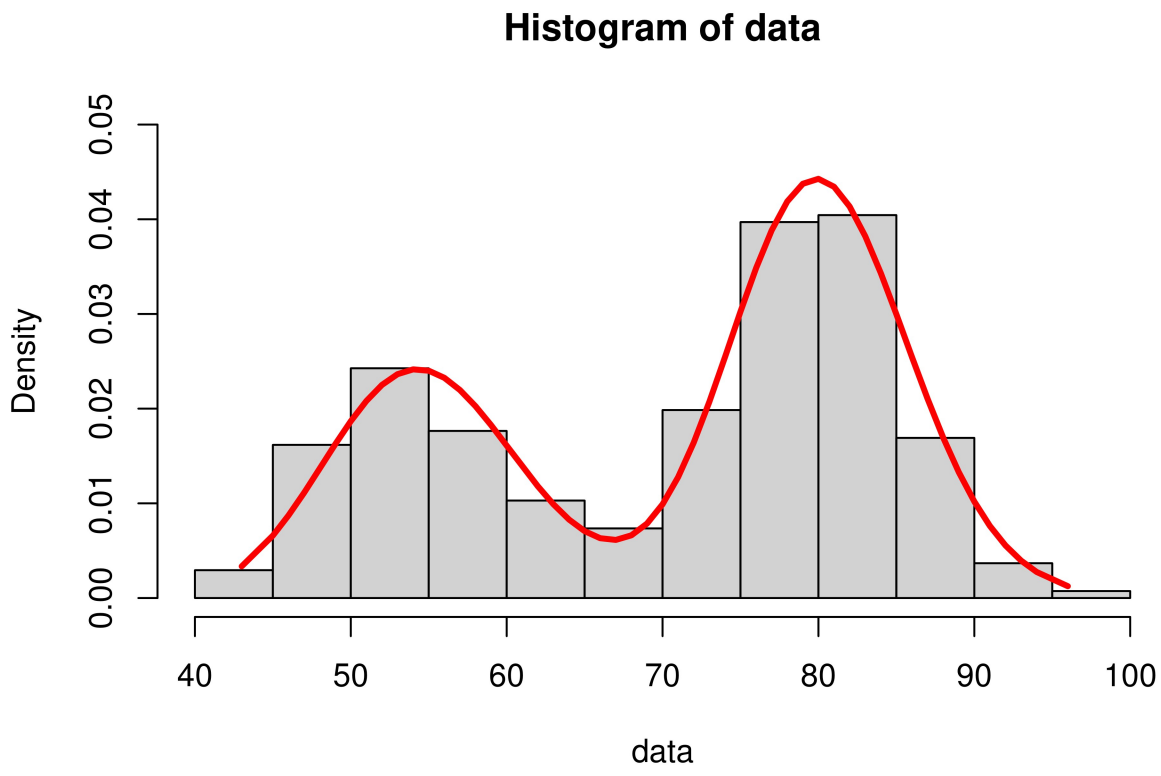
AIC value for model 1 is 2076.506

(ii) Model 2:

$$f(x) = p * Gamma(x|\alpha_1, \sigma_1) + (1-p)Gamma(x|\alpha_2, \sigma_2), \quad 0 < p < 1$$

```
loglike2 = function(theta,data){
   alpha1 = exp(theta[1])
   beta1 = exp(theta[2])
   alpha2 = exp(theta[3])
   beta2 = exp(theta[4])
   p = exp(theta[5])/(1+exp(theta[5]))
   n = length(data)
   l=0
   for(i in 1:n){
      l = l + log(p*dgamma(data[i],shape = alpha1, rate = beta1)
                  +(1-p)*dgamma(data[i], shape = alpha2, rate = beta2))
```

```
   }
   return(-l)
}

theta_initial = c(4,0,4.4,0,0.35)

fit = optim(theta_initial, loglike2, data = data, control = list(maxit=2000))

theta_hat = fit$par
alpha1_hat = exp(theta_hat[1])
beta1_hat = exp(theta_hat[2])
alpha2_hat = exp(theta_hat[3])
beta2_hat = exp(theta_hat[4])
p_hat = exp(theta_hat[5])/(1+exp(theta_hat[5]))

d_mle = p_hat*dgamma(data, shape = alpha1_hat, rate = beta1_hat)+
   (1-p_hat)*dgamma(data, shape = alpha2_hat, rate = beta2_hat)

hist(data, probability = T, ylim = c(0, 0.05))
lines(data, d_mle,lwd=3,col='red')
```




```
AIC = 2*length(fit$par) - 2*(-fit$value)
## AIC Value for model 1
AIC
```

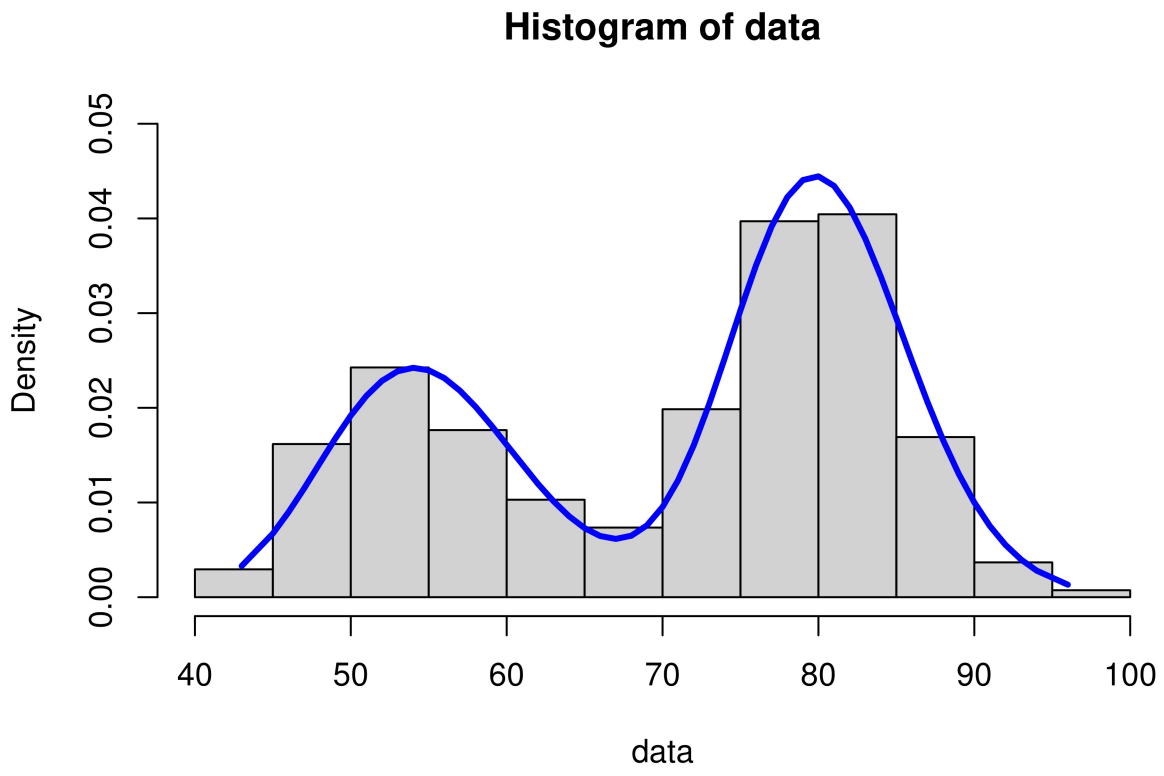
```
## [1] 2076.117
```

AIC value for model 2 is 2076.117

(iii) Model 3:

$$f(x) = p * logNormal(x|\mu_1, \sigma_1^2) + (1-p)logNormal(x|\mu_1, \sigma_1^2), \quad 0 < p < 1$$

```
loglike3 = function(theta,data){
   mu1 = theta[1]
   sigma1 = exp(theta[2])
   mu2 = theta[3]
   sigma2 = exp(theta[4])
   p = exp(theta[5])/(1+exp(theta[5]))
   n = length(data)
   l=0
   for(i in 1:n){
      l = l + log(p*dlnorm(data[i],meanlog = mu1,sdlog = sigma1)
             +(1-p)*dlnorm(data[i],meanlog = mu2,sdlog = sigma2))
   }
   return(-l)
}

theta_initial = c(2.76,-2.25,4.4,-2.6,0.35)

fit = optim(theta_initial, loglike3, data = data, control = list(maxit=2000))

theta_hat = fit$par
mu1_hat = theta_hat[1]
sigma1_hat = exp(theta_hat[2])
mu2_hat = theta_hat[3]
sigma2_hat = exp(theta_hat[4])
p_hat = exp(theta_hat[5])/(1+exp(theta_hat[5]))

d_mle = p_hat*dlnorm(data,meanlog = mu1_hat,sdlog = sigma1_hat)+
      (1-p_hat)*dlnorm(data,meanlog = mu2_hat,sdlog = sigma2_hat)

hist(data, probability = T, ylim = c(0, 0.05))
lines(data, d_mle,lwd=3,col='blue')
```




```
AIC = 2*length(fit$par) - 2*(-fit$value)
## AIC Value for model 1
AIC
```

```
## [1] 2075.433
```

AIC value for model 3 is 2075.433

Akaike information criterion(AIC)

Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is the following.

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

Conclusion:

Comparing AIC values for the three given models, we can observe that AIC value of model 3 is minimum among them making it the best model for the given data.

Required probability using best model.

```
dMix = function(x,theta){
  mu1 = theta[1]
  sigma1 = theta[2]
  mu2 = theta[3]
  sigma2 = theta[4]
  p = theta[5]
  f = p*dlnorm(x,meanlog = mu1,sdlog = sigma1)+(1-p)*dlnorm(x, meanlog = mu2, sdlog = sigma2)
  return(f)
}

integrate(dMix,60,70,c(mu1_hat,sigma1_hat,mu2_hat,sigma2_hat,p_hat))
```

```
## 0.09112692 with absolute error < 1e-15
```

$$\mathbb{P}(60 < \texttt{waiting} < 70) = 0.09112692$$