# Assignment 3

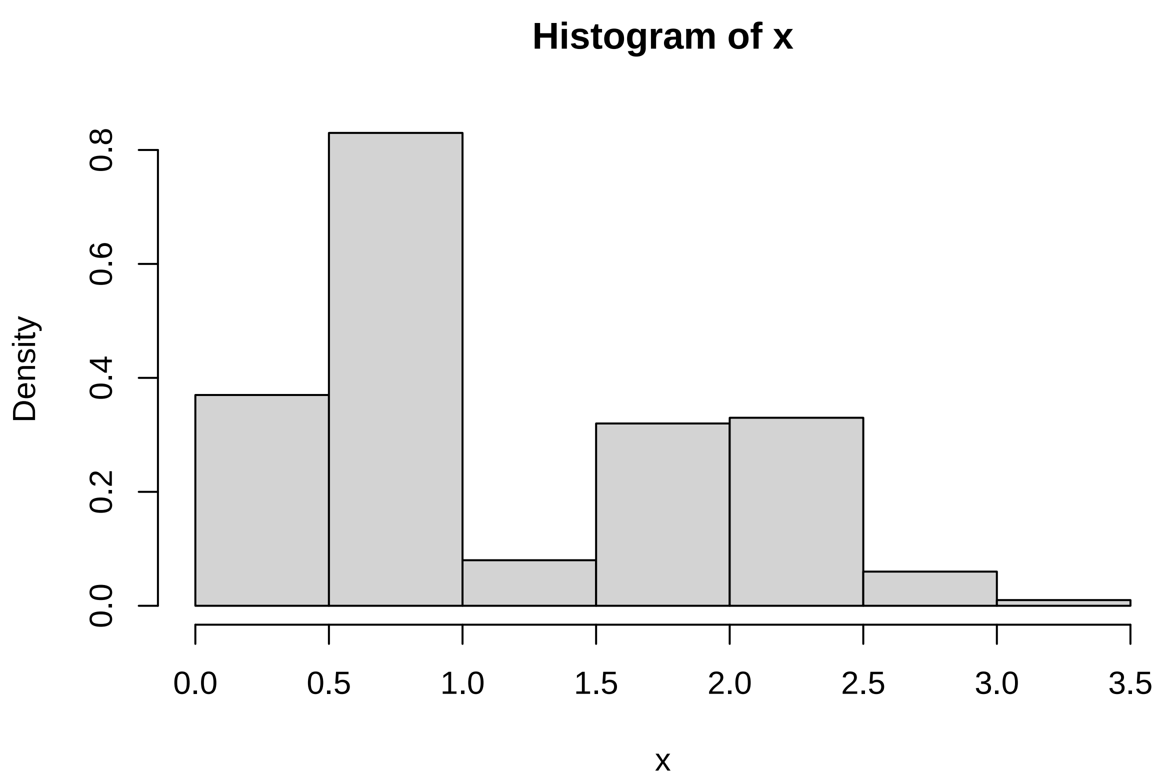
251159206 Andy Yuan

2025-04-06

## Question 1

### a)

```r
set.seed(1) # reproducibility
num <- 200 # number of observations total
p1 <- 0.4 # mixture probability
x1 <- rgamma(num*p1, shape = 20, rate = 10) # first gamma distribution
x2 <- rgamma(num*(1-p1), shape = 15, rate = 25) # second gamma distribution
x <- c(x1, x2) # combined together
hist(x, prob = T)
```



## b)

In a Gamma distribution, $\beta$ represents the rate parameters. Those are the target parameters we want in the EM algorithm in this case. We have the known shape parameters $\alpha$

Starting with the log of the Gamma distribution:

$$\begin{aligned}
\gamma(y) &= \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \\
log\gamma(y) &= \sum_{i=1}^{n} \alpha log(\beta) + (\alpha-1)logy - \beta y - log(\Gamma(\alpha)) \\
&= \alpha log(\beta y) - \beta y - logy - log(\Gamma(\alpha))
\end{aligned}$$

**E step**

If we assume that the responses given the $z_i$'s are independent and the $z_i$'s are also independent:

$$\begin{aligned}
logp(y|\theta) &= log \sum_{z} p(y|z,\gamma)p(z|\alpha) \\
&= \sum_{i=1}^{n} log\left[\sum_{j=1}^{M} p(y_i|\beta_i,\alpha_i) \times p_j\right] \quad w.r.t\theta \\
\text{Complete data log-likelihood: } \rightarrow log(p,z|\theta) &= logp(y|z,\gamma) + logp(z|\alpha) \\
&= \sum_{i=1}^{n} logp(y_i|z_i,\alpha_{z_i},\beta_{z_i}) + \sum_{i=1}^{n} logp_{z_i} \\
&= \mathcal{L}_1(\gamma) + \mathcal{L}_2(\alpha)
\end{aligned}$$

Then,

$$E_{\theta^{(c)}}(\mathcal{L}_1(\gamma)|y) = \sum_{i=1}^{n}\sum_{j=1}^{M} p(z_i = j|y,\theta^{(c)})logp(y_i|z_i = j,\alpha_j,\beta_j)$$

and, since we assume independence,

$$\begin{aligned}
p(z_i = j|y,\theta^{(c)}) = p(z_i = j|y_i,\theta^{(c)}) &= p_{ij}^{(c)} \\
&= \frac{p(y_i|z_i = j,\alpha_j^{(c)},\beta_j^{(c)}) \times p_j^{(c)}}{\sum_{k=1}^{M} p(y_i|z_i = k,\alpha_k^{(c)},\beta_j^{(c)}) \times p_k^{(c)}}
\end{aligned}$$

Thus,

$$E_{\theta^{(c)}}(\mathcal{L}_1(\gamma)|y) = \sum_{i=1}^{n}\sum_{j=1}^{M} p_{ij}^{(c)}[\alpha_j log(\beta_j y_i) - \beta_j y_i - logy_i - log(\Gamma(\alpha_j))]$$

and then since,

$$\mathcal{L}_2(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{M} l(z_i = j) \times logp_j$$

$$E_{\theta^{(c)}}(\mathcal{L}_2(\alpha)|y) = \sum_{i=1}^{n}\sum_{j=1}^{M} p(z_i = j, y_i, \theta^{(c)}) \times logp_j$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{M} p_{ij}^{(c)} \times logp_j$$

Therefore,

$$Q(\theta, \theta^{(c)}) = \sum_{i=1}^{n}\sum_{j=1}^{M} p_{ij}^{(c)}[\alpha_j log(\beta_j) + \alpha_j log(y_i) - \beta_j y_i - logy_i - log(\Gamma(\alpha_j)) + logp_j]$$

**M step**

Now, to maximize $Q$ with respect to the rate parameter $\beta$, we must take the derivative and set it to 0 so we can extract an estimate for $\beta$

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n}\sum_{j=1}^{M}[\alpha_j/\beta_j + 0 - y_i - 0 - 0 + 0]p_{ij}^{(c)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\alpha_j}{\beta_j} p_{ij}^{(c)} - \sum_{i=1}^{n}\sum_{j=1}^{M} y_i p_{ij}^{(c)}$$

$$\rightarrow \text{Set to 0: } 0 = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\hat{\alpha}_j}{\hat{\beta}_j} p_{ij}^{(c)} - \sum_{i=1}^{n}\sum_{j=1}^{M} y_i p_{ij}^{(c)}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{M} y_i p_{ij}^{(c)} = \sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\hat{\alpha}_j}{\hat{\beta}_j} p_{ij}^{(c)}$$

$$\sum_{i=1}^{n} y_i p_{ij}^{(c)} = \frac{\hat{\alpha}_j}{\hat{\beta}_j} \sum_{i=1}^{n} p_{ij}^{(c)}$$

$$\hat{\beta}_j = \hat{\alpha}_j \frac{\sum_{i=1}^{n} p_{ij}^{(c)}}{\sum_{i=1}^{n} y_i p_{ij}^{(c)}}$$

Using lagrange multipliers and the restriction that $\sum_{j=1}^{M} p_j = 1$,

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(c)}$$

## c)

1. Create a function with necessary arguments for:
    - $p =$ the mixture probability for the first distribution, the probability of the second distribution is $1 - p$
    - $shape1 =$ the shape parameter of the first distribution

- $shape2$ = the shape parameter of the second distribution
- $rate1$ = the rate parameter of the first distribution
- $rate2$ = the rate parameter of the second distribution
- $x$ = the combination of our two distributions
- $maxiter$ = the maximum times we want to run our algorithm
- $tol$ = the difference between updated parameters when we stop our algorithm

2. Set diff = 1 and iter = 0. iter is the number of iterations we will use in our 'while' loop, where it will become 1 after the first loop, 2 after the second loop, etc, until either the updated parameters do not change drastically or we reach our maximum set iterations. diff is the difference between our old and new parameters after updating them.
3. WHILE our diff is greater than our set 'tol' AND iter is greater than our set 'maxiter',
   - E step:
     - Produce two distributions using the `dgamma` function, using the arguments for `x, shape,` and `rate`.
     - Produce a new variable `pz` that is the probability of our data coming from the first distribution.
   - M step:
     - Set our new `p` to the mean of `pz`. We aim to update our parameters from their original value to another value when we get new information
     - Set our new rate parameters using the estimator obtained from the EM algorithm. In this case, we must use the shape parameter of the respective distributions, that we know and are fixed, in conjunction with the new estimates for the mixture probability and our previous combination of our two distributions.
   - Calculate the square root sum of squared differences for all of our parameters, which is essentially the euclidean distance between our first set and second set of parameters.
   - Set the name of the old parameters to the value of the new ones, so we can use the new values in this iteration for the next iteration.
   - Add 1 to `iter` so we can repeat the cycle.
   - END WHILE
4. Return a list of the final parameters.

## d)

Creating function

```
EM_TwoMixtureGamma = function(p, shape1, shape2, rate1,
                              rate2, x, maxiter=1000, tol=1e-5){
  diff=1
  iter=0
  while (diff>tol & iter<maxiter) {
    d1=dgamma(x, shape=shape1, rate=rate1)
    d2=dgamma(x, shape=shape2, rate=rate2)
    pz=d1*p/(d1*p+d2*(1-p)) ## P(Z_i=1|x_i and current parameter estimates)
    p.new=mean(pz)
    rate1.new=shape1*sum(pz) / sum(x* pz)
    rate2.new=shape2*sum(1-pz) / sum(x*(1 - pz))
    diff=sqrt(sum((rate1.new-rate1)**2+(rate2.new-rate2)**2
                  +(p.new-p)**2))
    p=p.new
    rate1=rate1.new
    rate2=rate2.new
    iter=iter+1
```

```
  }
  out <- list(rate1=rate1, rate2=rate2, p=p, iter=iter)
  return(out)
}
```

Executing function

```
p_0 <- 0.5
rate1_0 <- quantile(x,0.1)
rate2_0 <- quantile(x, 0.9)
shape1 <- 20
shape2 <- 15

EM_results <- EM_TwoMixtureGamma(p=p_0, shape1 = shape1,
                                  shape2 = shape2, rate1 = rate1_0,
                                  rate2 = rate2_0, x=x)

print(t(EM_results))
```

```
##      rate1    rate2    p         iter
## [1,] 9.971406 25.31634 0.4013324 17
```

My final estimates are:

- $\beta_1 = 9.971406$
- $\beta_2 = 25.31634$
- $p = 0.4013324$

This is remarkably close to the values that we input in part a)

## e)

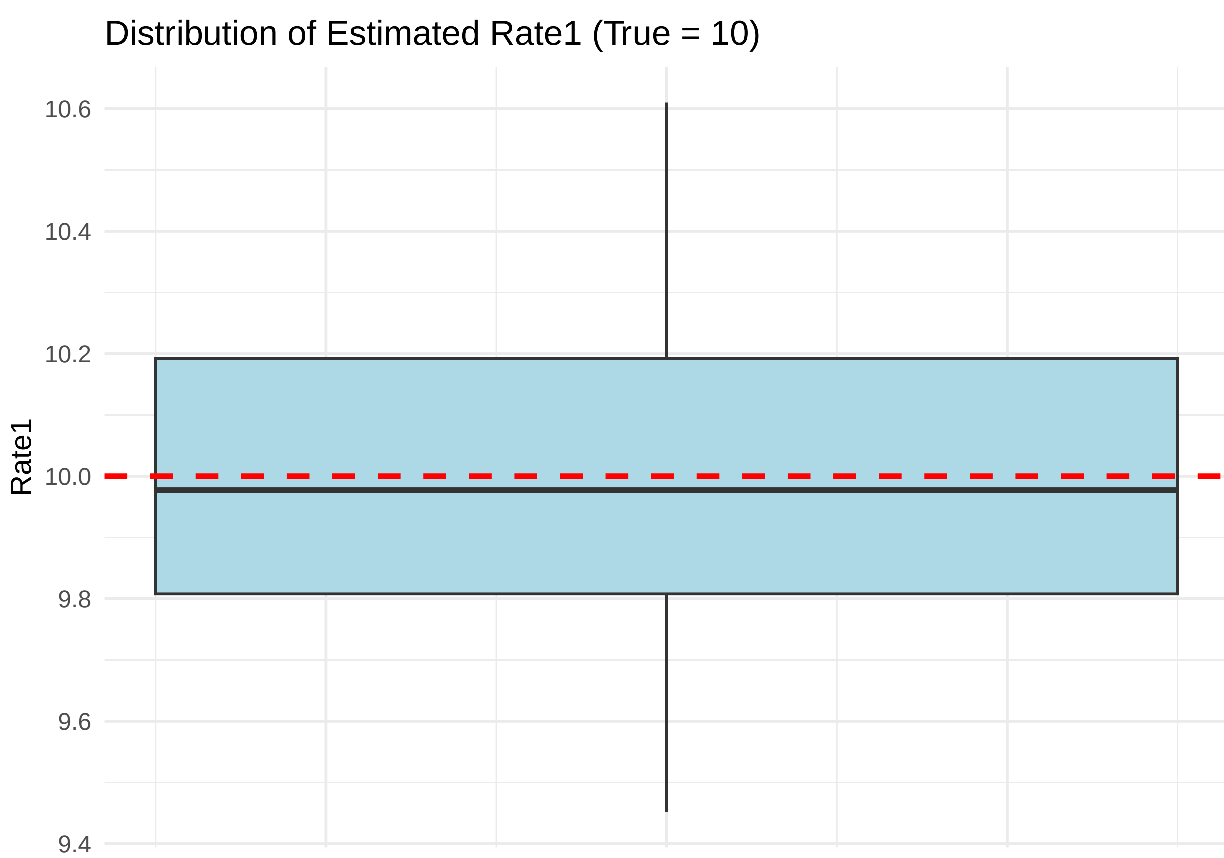
```
suppressMessages(library(ggplot2))
suppressMessages(library(tidyr))
suppressWarnings(library(patchwork))

set.seed(1)  # For reproducibility

# True parameters
true_p1 <- 0.4
true_shape1 <- 20
true_rate1 <- 10
true_shape2 <- 15
true_rate2 <- 25
num <- 200  # Observations per dataset
S <- 100    # Number of simulated datasets

results <- data.frame(
  rate1 = numeric(S),
  rate2 = numeric(S),
```

```r
  p = numeric(S),
  iter = numeric(S)
)

for (s in 1:S) {
  # Generate data
  x1 <- rgamma(num * true_p1, shape = true_shape1, rate = true_rate1)
  x2 <- rgamma(num * (1 - true_p1), shape = true_shape2, rate = true_rate2)
  x <- c(x1, x2)

  # Initial guesses
  p_0 <- 0.5
  rate1_0 <- quantile(x, 0.1)  # Low percentile for rate1
  rate2_0 <- quantile(x, 0.9)  # High percentile for rate2

  # Run EM
  fit <- EM_TwoMixtureGamma(
    p = p_0,
    shape1 = true_shape1,
    shape2 = true_shape2,
    rate1 = rate1_0,
    rate2 = rate2_0,
    x = x
  )

  # Store results
  results$rate1[s] <- fit$rate1
  results$rate2[s] <- fit$rate2
  results$p[s] <- fit$p
  results$iter[s] <- fit$iter
}

# Boxplot for rate1
plot_rate1 <- ggplot(results, aes(y = rate1)) +
  geom_boxplot(fill = "lightblue", width = 0.3) +
  geom_hline(yintercept = true_rate1, color = "red",
             linetype = "dashed", linewidth = 1) +
  labs(
    title = "Distribution of Estimated Rate1 (True = 10)",
    y = "Rate1"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank())

# Boxplot for rate2
plot_rate2 <- ggplot(results, aes(y = rate2)) +
  geom_boxplot(fill = "lightgreen", width = 0.3) +
  geom_hline(yintercept = true_rate2, color = "red",
             linetype = "dashed", linewidth = 1) +
  labs(
    title = "Distribution of Estimated Rate2 (True = 25)",
    y = "Rate2"
  ) +
```

```
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank())

# Boxplot for p (mixture probability)
plot_p <- ggplot(results, aes(y = p)) +
  geom_boxplot(fill = "lightpink", width = 0.3) +
  geom_hline(yintercept = true_p1, color = "red",
             linetype = "dashed", linewidth = 1) +
  labs(
    title = "Distribution of Estimated Mixture Probability (True = 0.4)",
    y = "Mixture Probability (p)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank())
```

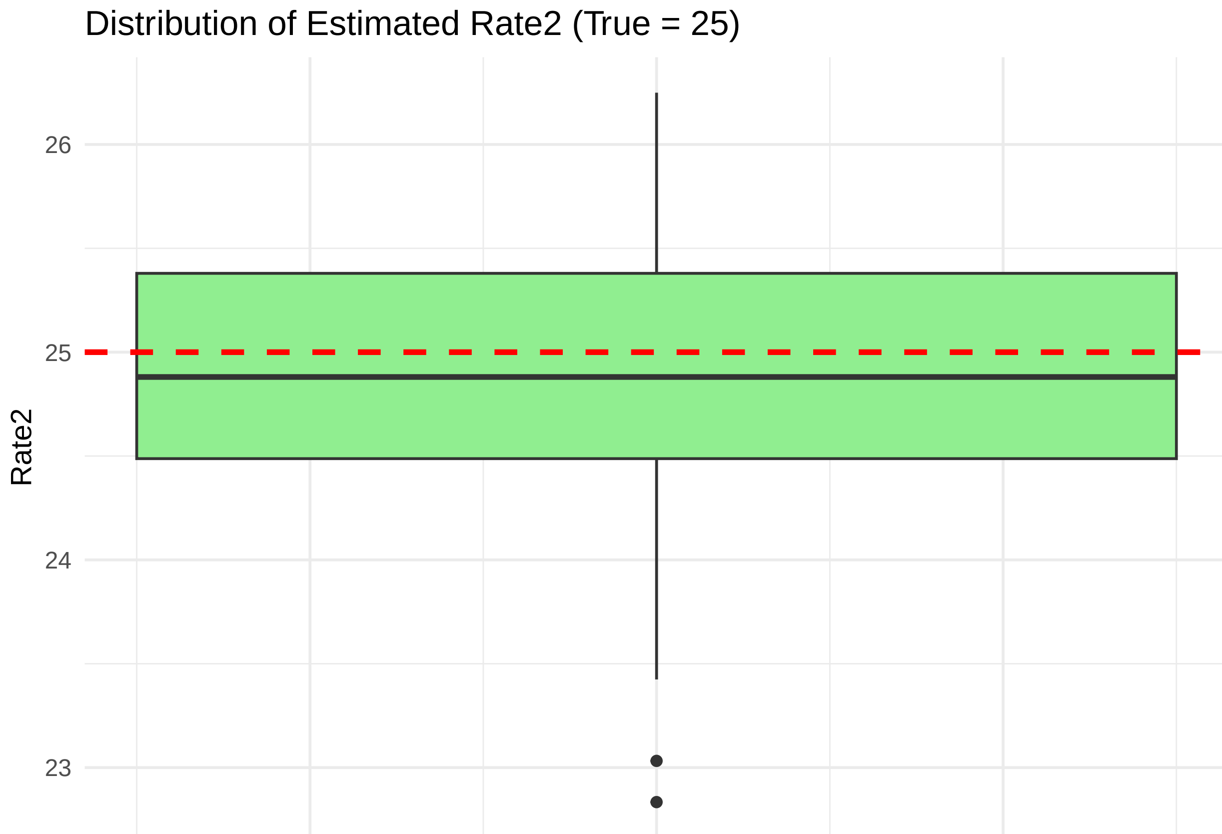
First plot of rate 1: This plot is almost centered around the true value of 10.

```
plot_rate1
```



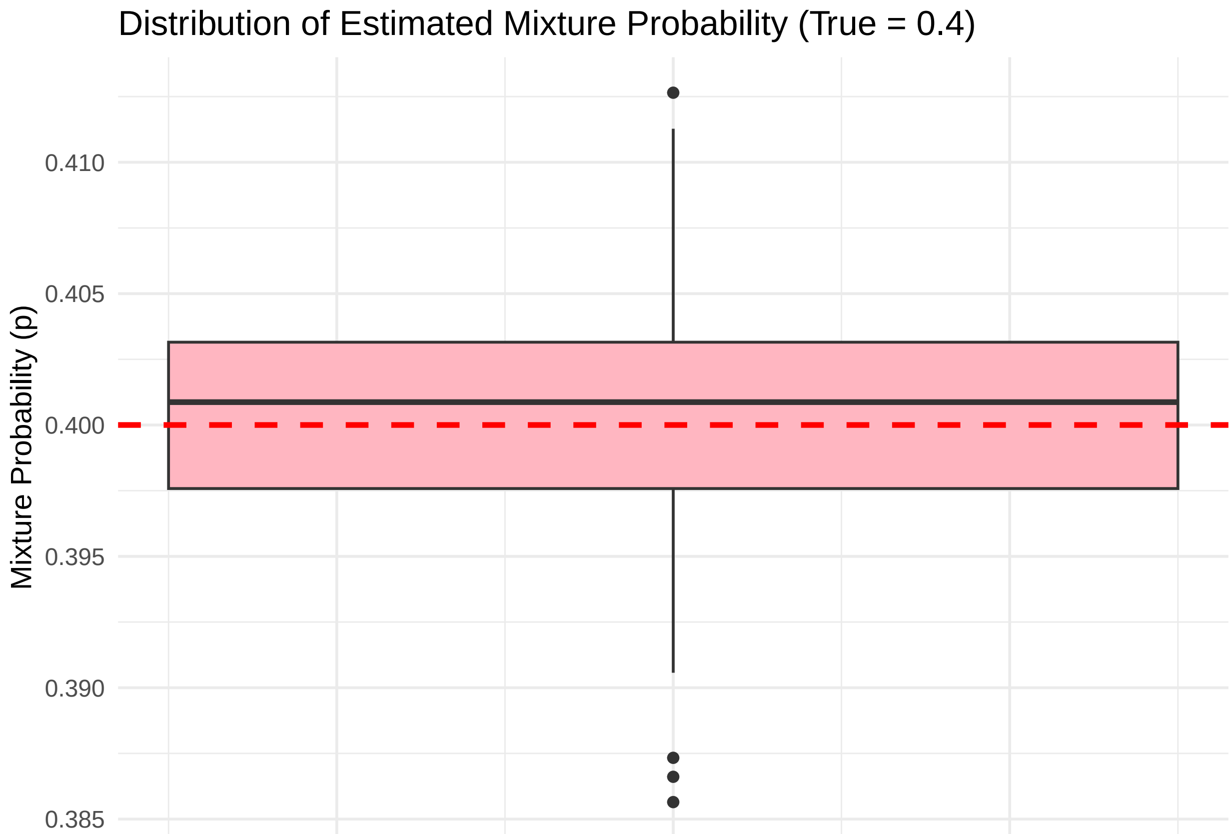

Second plot of rate2: This plot is almost centered to the true value of 25.

```
plot_rate2
```




Third plot of p: This plot is almost centered around the true value of 0.4.

```
plot_p
```




## Question 2

From our assumptions, we can express the joint posterior distribution $p(\sigma^2, \theta|Y)$ as proportionate to $p(Y|\sigma^2, \theta) \cdot p(\sigma^2, \theta)$ below:

$$\begin{aligned} p(\sigma^2, \theta|Y) &\propto p(Y|\sigma^2, \theta) \cdot p(\sigma^2, \theta) \\ &\propto (2\pi\sigma^2)^{-n/2} \cdot (2\pi\sigma^2/k_0)^{-1/2} \cdot (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \cdot exp(-\frac{\nu_0\sigma_0^2}{2\sigma^2}) \cdot exp(-\frac{\sum_{i=1}^n (y_i-\theta)^2 + k_0(\theta-\mu_0)^2}{2\sigma^2}) \\ &\propto (\sigma^2)^{-\frac{n+1+\nu_0}{2}-1}(2\pi)^{-n/2}(2\pi/k_0)^{-1/2} \cdot exp(\frac{-1}{2\sigma^2}[\sum_{i=1}^n (y_i-\theta)^2 + k_0(\theta-\mu_0)^2 + \nu_0\sigma_0^2]) \\ &\propto (\sigma^2)^{-\frac{n+1+\nu_0}{2}-1} \cdot exp(-\frac{A}{2\sigma^2}) \end{aligned}$$

$$p(\theta|Y) = \int p(\sigma^2, \theta|Y) d\sigma^2 \propto \int (\sigma^2)^{-\frac{n+1+\nu_0}{2}-1} \cdot exp(-\frac{A}{2\sigma^2}) d\sigma^2$$

This is similar to inverse-Gamma function:

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

Where:

- $x = \sigma^2$
- $\alpha = \frac{(n+\nu_0+1)}{2}$
- $\beta = \frac{A}{2}$

The general integral of the inverse-gamma wrt $\theta$ is:

$$\int x^{-\alpha-1} exp(-\frac{\beta}{x})dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

And because we only want terms with $\theta$, we can express as $\beta^{-\alpha}$

$$\therefore \int (\sigma^2)^{-\frac{n+1+\nu_0}{2}-1} \cdot exp(-\frac{A}{2\sigma^2})d\sigma^2 \propto (\frac{A}{2})^{-\frac{n+1+\nu_0}{2}}$$

$$\begin{aligned}
(\frac{A}{2})^{-\frac{n+1+\nu_0}{2}} &\propto [\sum_{i=1}^{n}(y_i-\theta)^2 + k_0(\theta-\mu_0)^2 + \nu_0\sigma_0^2]^{-\frac{n+1+\nu_0}{2}} \\
&= [n(\bar{y}-\theta)^2 + (n-1)s^2 + k_0(\theta-\mu_0)^2 + \nu_0\sigma_0^2]^{-\frac{n+1+\nu_0}{2}} \\
&= [n(\bar{y}^2 - 2\bar{y}\theta + \theta^2) + (n-1)s^2 + k_0(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \nu_0\sigma_0^2]^{-\frac{n+1+\nu_0}{2}} \\
&= \left[(n+k_0)\left[\theta^2 - 2\left(\frac{n\bar{y}+k_0\mu_0}{n+k_0}\right)\theta\right] + n\bar{y}^2 + (n-1)s^2 + k_0\mu_0^2 + \nu_0\sigma_0^2\right]^{-\frac{n+1+\nu_0}{2}} \\
&= \left[(n+k_0)\left[\left(\theta - \frac{n\bar{y}+k_0\mu_0}{n+k_0}\right)^2 - \left(\frac{n\bar{y}+k_0\mu_0}{n+k_0}\right)^2\right] + n\bar{y}^2 + (n-1)s^2 + k_0\mu_0^2 + \nu_0\sigma_0^2\right]^{-\frac{n+1+\nu_0}{2}}
\end{aligned}$$

We notice now that the mean of $\theta|\sigma^2, Y$ is present in

$$mean = \frac{n\bar{y}+k_0\mu_0}{n+k_0}$$

, we can also define $G$ as the terms after our term with $\theta$:

$$G = n\bar{y}^2 + (n-1)s^2 + k_0\mu_0^2 + \nu_0\sigma_0^2 - (n+k_0)\left[\frac{n\bar{y}+k_0\mu_0}{n+k_0}\right]^2$$

Therefore,

$$\begin{aligned}
(A)^{-\frac{n+1+\nu_0}{2}} &= [(n+k_0)(\theta - mean)^2 + G]^{-\frac{n+1+\nu_0}{2}} \\
p(\theta|y) &\propto [(n+k_0)(\theta - mean)^2 + G]^{-\frac{n+1+\nu_0}{2}} \\
&\propto \left[G\left(\frac{(\theta - mean)^2}{G/(n+k_0)} + 1\right)\right]^{-\frac{n+1+\nu_0}{2}} \\
&\propto G^{-\frac{n+1+\nu_0}{2}}\left[\left\{\frac{(\theta - mean)}{\sqrt{G/(n+k_0)}}\right\}^2 + 1\right]^{-\frac{n+1+\nu_0}{2}}
\end{aligned}$$

This ending expression is very similar to the t-distribution, where:

- $G^{-\frac{n+1+\nu_0}{2}}$ represents the first portion with the Gamma functions
- $\frac{n+1+\nu_0}{2} = \frac{\nu+1}{2}$
- $\sigma = \sqrt{G/(n+k_0)}$
- $\mu = mean = \frac{n\bar{y}+k_0\mu_0}{n+k_0}$