Lab 1 report

Christian Kammerer(chrka821), Jakob Lindner (jakli758)

1. Daniel Bernoulli

Let $y_1, ..., y_n|\theta \sim Bern(\theta)$, and assume that you have obtained a sample with $f = 35$ failures in $n = 78$ trials. Assume a $Beta(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 7$.

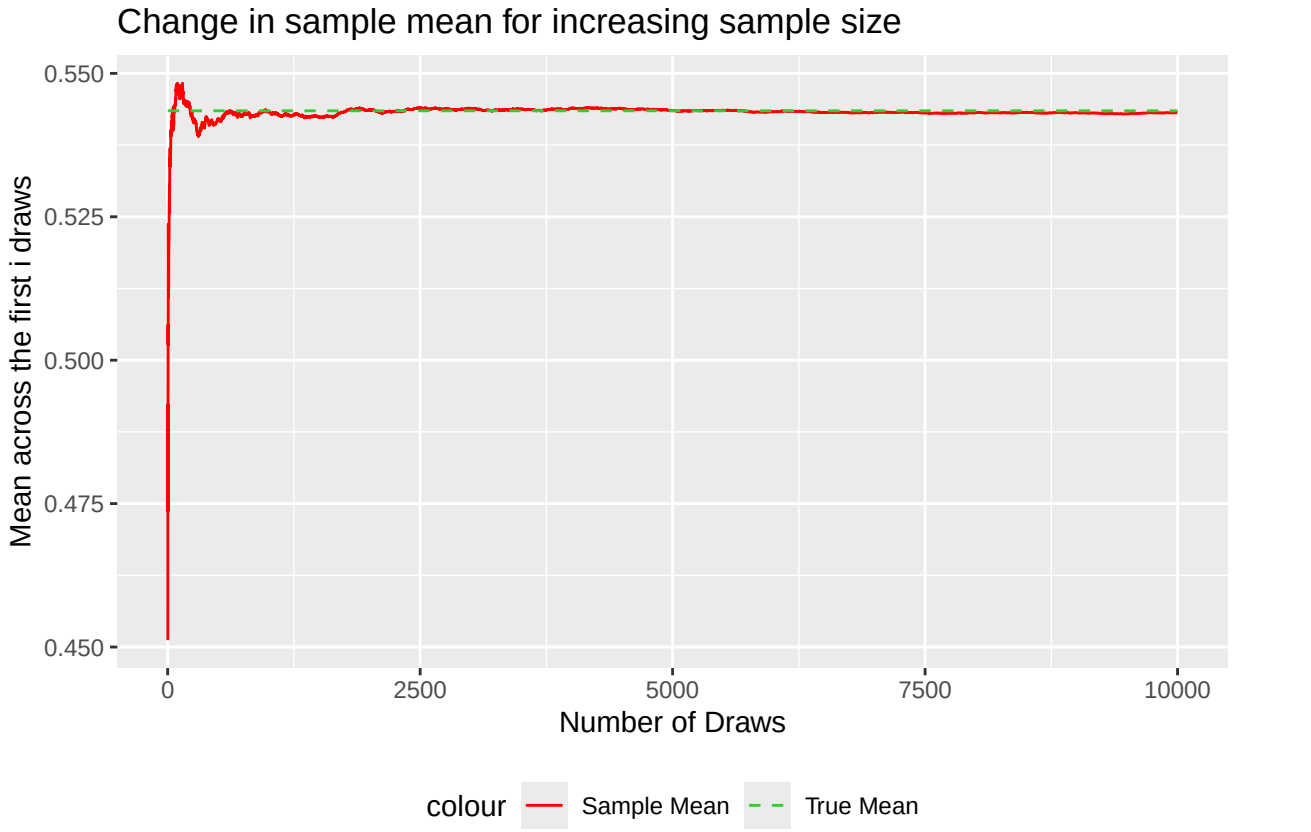
1.1

Draw 10000 random values (nDraws = 10000) from the posterior $\theta|y \sim Beta(\alpha_0 + s, \beta_0 + f)$, where $y = (y_1, ..., y_n)$, and verify graphically that the posterior mean $E[\theta|y]$ and standard deviation $SD[\theta|y]$ converges to the true values as the number of random draws grows large. [Hint: use rbeta() to draw random values and make graphs of the sample means and standard deviations of θ as a function of the accumulating number of drawn values].

To verify that the posterior mean and standard deviation converge to the true values as the number of random draws grows large, we draw from the posterior and plot the mean and standard deviation of the accumulating number of values and add lines for their true values. The resulting graphs both show that the mean converges to the true mean from around 6500 drawn values on, the standard deviation converges earlier at around 2600 drawn values.

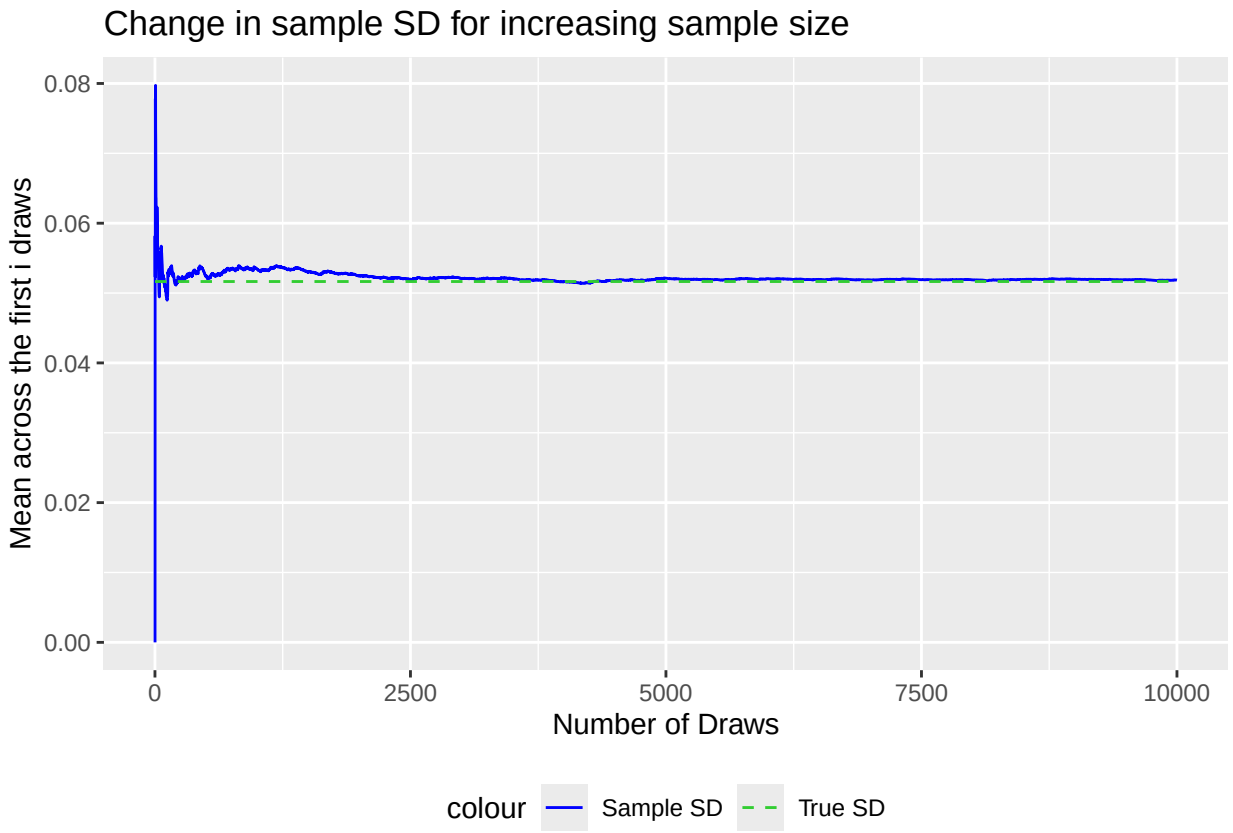
```
library(ggplot2)
n <- 78
f <- 35
s <- n - f
alpha <- 7
beta <- 7
drawsize <- 10000

posterior_sample <- function(n, alpha, beta, f, s){
  return(rbeta(n, alpha + s, beta + f))
}

n_draws <- posterior_sample(drawsize, alpha, beta, f, s)
n_means <- numeric(drawsize)
n_sds <- numeric(drawsize)

for(i in 1:drawsize){
  n_means[[i]] <- mean(n_draws[1:i])
  n_sds[[i]] <- sd(n_draws[1:i])
}

n_sds[[1]] <- 0

a <- alpha + s
b <- beta + f
```

```r
# Calculate mean and standard deviation according to beta distribution
true_mean <- a / (a + b)
true_sd <- sqrt((a * b) / ((a + b)^2 * (a + b + 1)))

# Create a data frame for horizontal line to allow for color mapping
hline_data <- data.frame(x = c(1, 10000), y = true_mean)

p <- ggplot() +
  aes(x = (1:10000), y = n_means, color = "Sample Mean") +
  geom_line() +
  xlab("Number of Draws") +
  ylab("Mean across the first i draws") +
  ggtitle("Change in sample mean for increasing sample size") +
  geom_line(data = hline_data, aes(x = x, y = y, color = "True Mean"), linetype = "dashed") +
  scale_color_manual(values = c("Sample Mean" = "red", "True Mean" = "limegreen")) +
  theme(legend.position = "bottom")

print(p)
```




```r
hline_data <- data.frame(x = c(1, 10000), y = true_sd)
p <- ggplot() +
  aes(x = (1:10000), y = n_sds, color = "Sample SD") +
  geom_line() +
  xlab("Number of Draws") +
  ylab("Mean across the first i draws") +
  ggtitle("Change in sample SD for increasing sample size") +
  geom_line(data = hline_data, aes(x = x, y = y, color = "True SD"), linetype = "dashed") +
```

```
  scale_color_manual(values = c("Sample SD" = "blue", "True SD" = "limegreen")) +
  theme(legend.position = "bottom")

print(p)
```



1.2

Draw 10000 random values from the posterior to compute the posterior probability $Pr(\theta > 0.5|y)$ and compare with the exact value from the Beta posterior. [Hint: use pbeta() to calculate the exact value].

```
prob <- mean(n_draws > 0.5)
actual_prob <- 1 - pbeta(q = 0.5, shape1 = alpha + s, shape2 = beta + f)
cat("Posterior probability: ", prob, ", exact value: ", actual_prob)
```

```
## Posterior probability:  0.7976 , exact value:  0.7990936
```

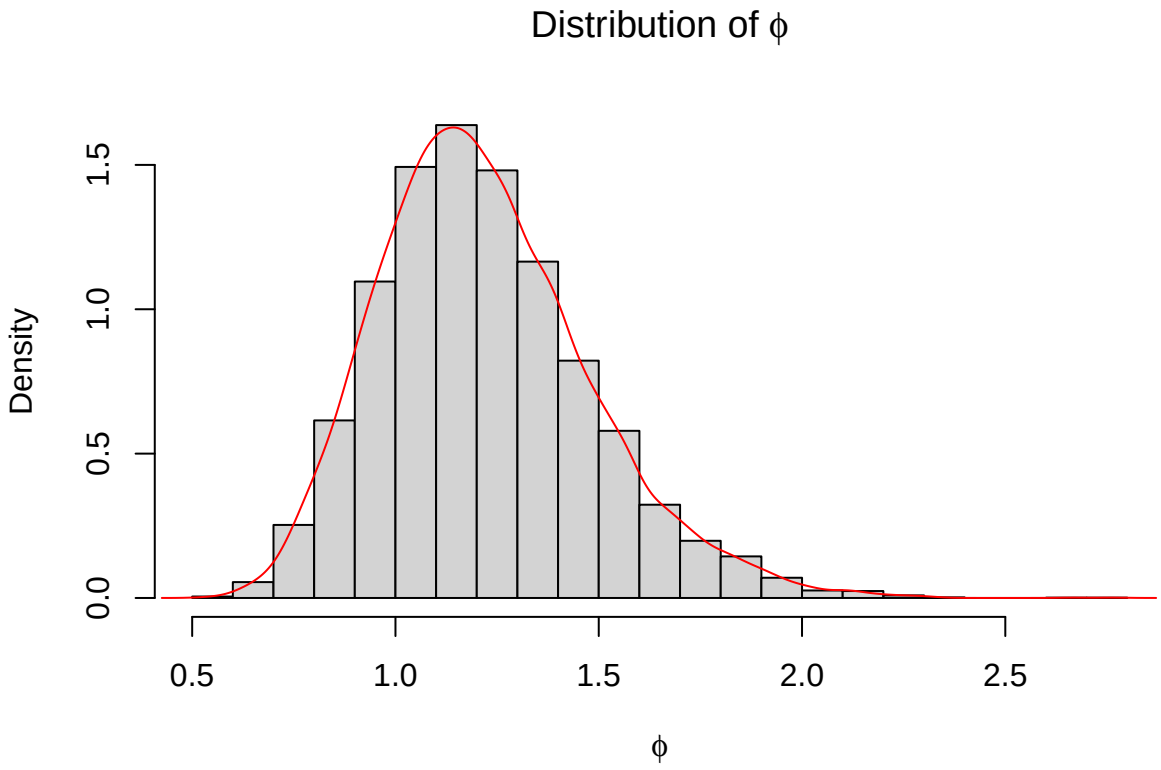
$Pr(\theta > 0.5|y)$ is the same as $1 - Pr(\theta \leq 0.5|y)$. Using the pbeta() function we get the value for the probability. The exact value we get is the same as the posterior probability in two decimals.

1.3

Draw 10000 random values from the posterior of the odds $\phi = \theta/(1-\theta)$ by using the previous random draws from the Beta posterior for θ and plot the posterior distribution of ϕ. [Hint: hist() and density() can be utilized]

```
phis <- n_draws / (1 - n_draws)
density_phi <- density(phis)
```

```r
# Plot histogram
hist(phis, main = expression(paste("Distribution of ", phi)),
     xlab = expression(phi), freq = FALSE, breaks=20)

# Add density line
lines(density_phi, col = "red")
```



The resulting histogram has a similar shape to the actual density of the odds.

2. Log-normal distribution and the Gini coefficient.

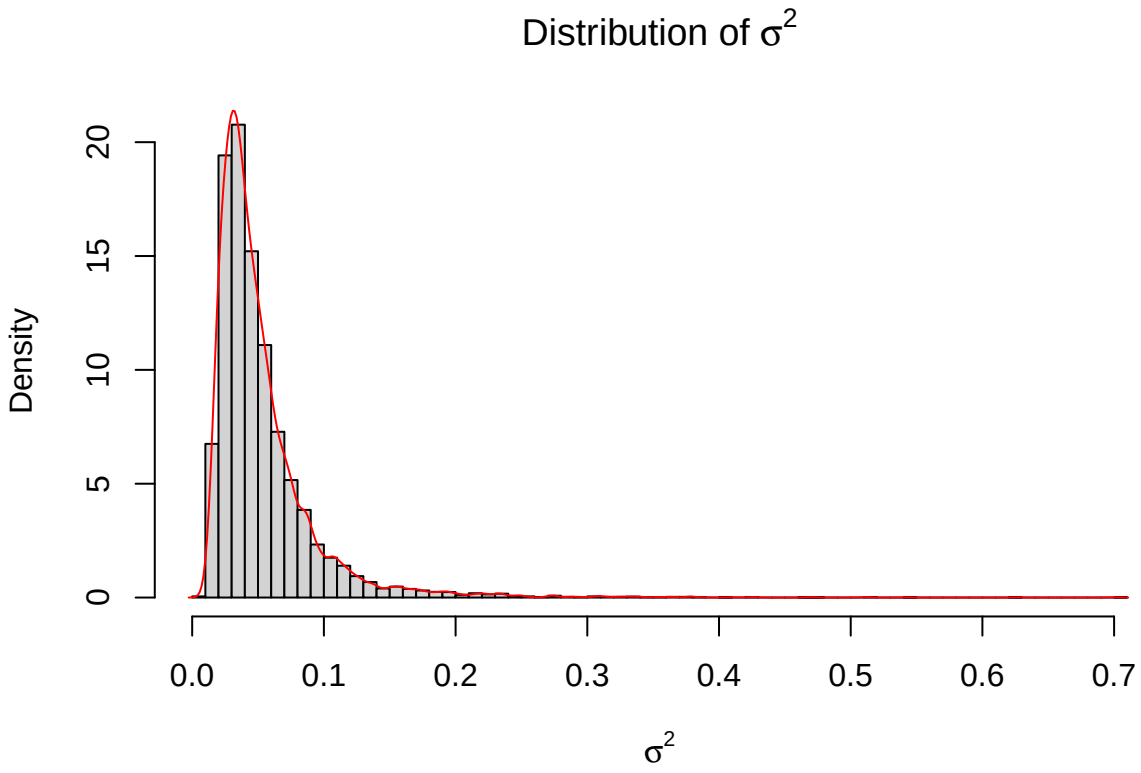
Assume that you have asked 8 randomly selected persons about their monthly income (in thousands Swedish Krona) and obtained the following eight observations: 22, 33, 31, 49, 65, 78, 17, 24. A common model for non-negative continuous variables is the log-normal distribution. The log-normal distribution $logN(\mu, \sigma 2)$ has density function $p(y|\mu,\sigma^2) = \frac{1}{y\cdot\sqrt{2\pi\sigma^2}} exp[-\frac{1}{2\sigma^2}(\log y - \mu^2)]$, where $y > 0$, $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The log-normal distribution is related to the normal distribution as follows: if $y \sim \log N(\mu, \sigma^2)$ then $\log y \sim N(\mu, \sigma^2)$. Let $y1, ..., yn|\mu, \sigma^2 iid \sim logN(\mu, \sigma^2)$, where $\mu = 3.65$ is assumed to be known but σ^2 is unknown with non-informative prior $p(\sigma^2) \propto 1/\sigma^2$. The posterior for σ^2 is the scaled inverse chi-squared distribution, $Scale - inv - \chi^2(n, \tau^2)$, where $\tau^2 = \frac{\sum_{i=1}^{n}(\log y_i - \mu)^2}{n}$.

2.1

Draw 10000 random values from the posterior of σ^2 by assuming $\mu = 3.65$ and plot the posterior distribution.

```r
y_vec <- c(22, 33, 31, 49, 65, 78, 17, 24)
mu <- 3.65

calc_tau_sq <- function(y_vec, mu){
```

```r
  return(sum((log(y_vec) - mu)^2)/length(y_vec))
}

tau_sq <- calc_tau_sq(y_vec, mu)
sample_var <- rinvchisq(n = 10000, df = length(y_vec) - 1) * tau_sq

density_sample_var <- density(sample_var)

# Plot histogram
hist(sample_var, main = expression(paste("Distribution of ", sigma^2)),
     xlab = expression(sigma^2), breaks = 50, freq = FALSE)

# Add density line
lines(density_sample_var, col = "red")
```

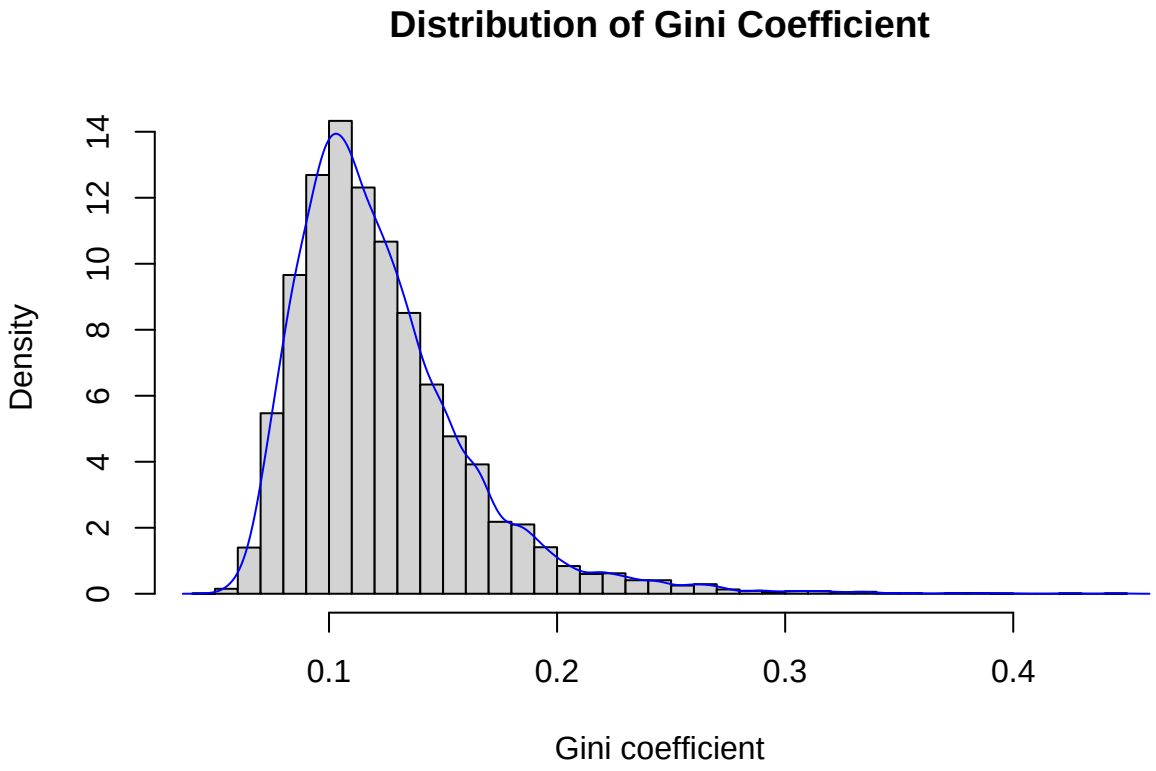


2.2

The most common measure of income inequality is the Gini coefficient, G, where $0 \leq G \leq 1$. $G = 0$ means a completely equal income distribution, whereas $G = 1$ means complete income inequality (see e.g. Wikipedia for more information about the Gini coefficient). It can be shown that $G = 2\phi(\sigma/\sqrt{2}) - 1$ when incomes follow a $\log N(\mu, \sigma^2)$ distribution. $\phi(z)$ is the cumulative distribution function (CDF) for the standard normal distribution with mean zero and unit variance. Use the posterior draws in a) to compute the posterior distribution of the Gini coefficient G for the current data set.

```r
ginis <- sapply(sample_var, function(sigma_sq) 2 * pnorm(q = sqrt(sigma_sq)/sqrt(2)) - 1)

density_ginis <- density(ginis)
```

```r
# Plot histogram
hist(ginis, main = "Distribution of Gini Coefficient",
     xlab = "Gini coefficient", breaks = 50, freq = FALSE)

lines(density_ginis, col = "blue")
```



Using the posterior draws from 2.1 we get the histogram above for G of the current data set.

2.3

Use the posterior draws from b) to compute a 95% equal tail credible interval for G. A 95% equal tail credible interval (a, b) cuts off 2.5% percent of the posterior probability mass to the left of a, and 2.5% to the right of b.

To obtain the 95%-CI we use the quantile()-method in R.

```r
ci <- quantile(ginis, probs = c(0.025, 0.975))
```

2.4

Use the posterior draws from b) to compute a 95% Highest Posterior Density Interval (HPDI) for G. Compare the two intervals in (c) and (d). [Hint: Use the hdi() function from the bayestestR package.]

```r
hpdi <- hdi(ginis, ci = 0.95)
cat("CI: ", ci)
```

```
## CI:  0.07248247 0.2211406
```

```r
print(hpdi)
```

```
## 95% HDI: [0.07, 0.20]
```

Both intervals are similar, the CI is slightly shifted to larger values though.

3.Bayesian inference for the rate parameter in the Poisson distribution.

This exercise aims to show you that a grid approximation can be used to obtain information of the posterior distribution when the distributions for prior and posterior are not conjugate. In some cases, it is sufficient to evaluate the posterior pdf $p(\lambda|y)$ at a finite set of λ values to understand its structure and obtain important quantities. The data points below represent the number of goals scored in each match of the opening week of the 2024 Swedish women's football league (Damallsvenskan):

$$y = (0, 2, 5, 5, 7, 1, 4).$$

We assume that these data points are independent observations from the Poisson distribution with rate parameter $\lambda > 0$ which tells us about the average goals rate in a match. Let the prior distribution of λ be the half-normal distribution with prior pdf

$$p(\lambda|\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} exp(-\frac{\lambda^2}{2\sigma^2}), \lambda \geq 0,$$

and the scale parameter that is set to be $\sigma = 5$.

3.1

Derive the expression for what the posterior pdf $p(\lambda|y, \sigma)$ is proportional to. Then, plot the posterior distribution of the average goals rate parameter λ over a fine grid of λ values. [Hint: you need to normalize the posterior pdf $p(\lambda|y, \sigma)$ so that it integrates to one.]

At first we need the likelihood to compute the posterior. It is the product of the pdf. As the denominator is not dependent on λ, we only use the nominator in the following.

$$p(y|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \frac{e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^{n} k_i}}{\prod_{i=1}^{n} k_i!} \propto e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^{n} k_i}$$

The given data gives us the following likelihood:

$$p(y|\lambda) = e^{-7\lambda} \cdot \lambda^{24}$$

The posterior is the product of likelihood and prior:

$$p(\lambda|y, \sigma) \propto p(y|\lambda) \cdot p(\lambda|\sigma)$$

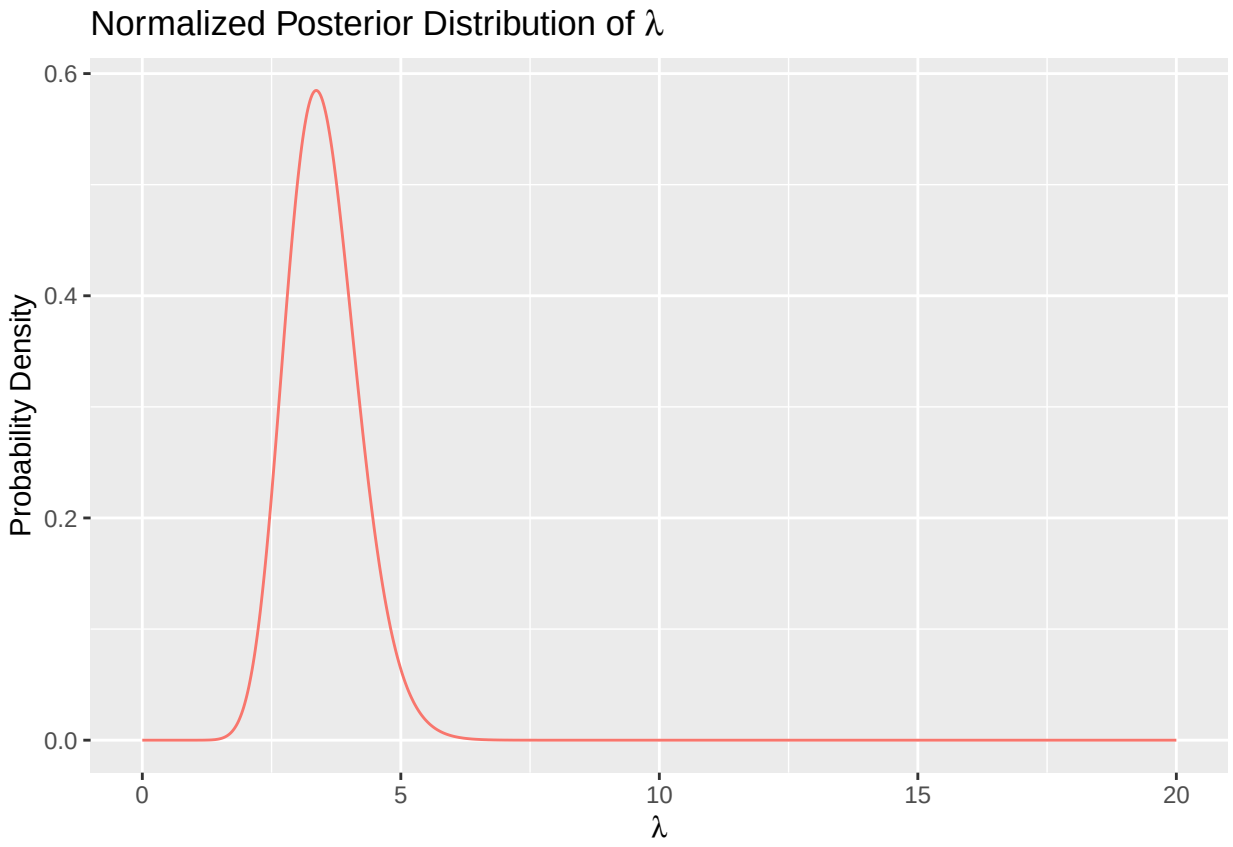
In the prior we can leave out the first part, as it does not depend on λ. As we know that $\sigma = 5$, we get the following prior:

$$p(\lambda|\sigma) \propto e^{\frac{-\lambda^2}{50}} \cdot \lambda^{24}$$

Plugging in the likelihood and prior, we finally get the posterior:

$$p(\lambda|y, \sigma) \propto \lambda^{24} e^{-7\lambda} \cdot e^{-\frac{\lambda^2}{50}}$$

```r
lambda_grid <- seq(0, 20, length.out = 1000)
unnorm_posterior <- lambda_grid^24 * exp(-7 * lambda_grid - lambda_grid^2 / 50)
posterior <- unnorm_posterior / sum(unnorm_posterior * diff(lambda_grid)[1])

p <- ggplot() +
  aes(x=lambda_grid, y=posterior, col = "red") +
  geom_line() +
  ggtitle(expression(paste("Normalized Posterior Distribution of ", lambda))) +
  xlab(expression(lambda)) +
  ylab("Probability Density") +
  theme(legend.position="none")
print(p)
```



3.2

Find the (approximate) posterior mode of λ from the information in a).

```r
mode <- lambda_grid[which.max(posterior)]
cat("mode: ", mode)
```

```
## mode:  3.363363
```

The mode is approximately 3.36. As we are only searching on the grid, it is only an approximation.