ST4234: Bayesian Statistics

Tutorial 3 Solution, AY 19/20

Solutions

1. (a) Let $x_{.1} = \sum_{i=1}^n x_{i1}$, $x_{.2} = \sum_{i=1}^n x_{i2}$, and $x_{.3} = \sum_{i=1}^n x_{i3}$. They are counts of the three categories in the sample $\boldsymbol{x}$. The posterior of $(\theta_1, \theta_2, \theta_3)$ is

$$\begin{aligned}
p(\theta_1, \theta_2, \theta_3 | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) &\propto p(\theta_1, \theta_2, \theta_3) \prod_{i=1}^n p(\boldsymbol{x}_i | \theta_1, \theta_2, \theta_3) \\
&\propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1} \prod_{i=1}^n [\theta_1^{x_{i1}} \theta_2^{x_{i2}} \theta_3^{x_{i3}}] \\
&\propto \theta_1^{\alpha_1 + x_{.1} - 1} \theta_2^{\alpha_2 + x_{.2} - 1} \theta_3^{\alpha_3 + x_{.3} - 1}
\end{aligned}$$

Therefore $p(\theta_1, \theta_2, \theta_3 | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is a Dirichlet distribution with parameters $(\alpha_1 + x_{.1}, \alpha_2 + x_{.2}, \alpha_3 + x_{.3})$. Since both the posterior belongs to the same class of distributions (Dirichlet) as the prior, the prior $p(\theta_1, \theta_2, \theta_3)$ is a conjugate prior for $(\theta_1, \theta_2, \theta_3)$ with respect to $p(\boldsymbol{x} | \theta_1, \theta_2, \theta_3)$.

(b) For abbreviation, let $\beta_j = \alpha_j + x_{.j}$ for $j = 1, 2, 3$. Let $\Theta = \{(\theta_1, \theta_2, \theta_3) : \ \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 \geq 0, \theta_1 + \theta_2 + \theta_3 = 1\}$. Then the posterior predictive distribution is given by

$$\begin{aligned}
& p(\boldsymbol{x}_{n+1} | \boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n) \\
= & \int_\Theta p(\boldsymbol{x}_{n+1} | \theta_1, \theta_2, \theta_3) p(\theta_1, \theta_2, \theta_3 | \boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\theta_3 \\
\propto & \int_\Theta \frac{\theta_1^{x_{n+1,1}} \theta_2^{x_{n+1,2}} \theta_3^{x_{n+1,3}}}{x_{n+1,1}! x_{n+1,2}! x_{n+1,3}!} \theta_1^{\beta_1 - 1} \theta_2^{\beta_2 - 1} \theta_3^{\beta_3 - 1} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\theta_3 \\
\propto & \int_\Theta \theta_1^{\beta + x_{n+1,1} - 1} \theta_2^{\beta_2 + x_{n+1,2} - 1} \theta_3^{\beta_3 + x_{n+1,3} - 1} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \mathrm{d}\theta_3 \\
\propto & \frac{\Gamma(\beta_1 + x_{n+1,1}) \Gamma(\beta_2 + x_{n+1,2}) \Gamma(\beta_3 + x_{n+1,3})}{\Gamma(\beta + x_{n+1,1} + \beta_2 + x_{n+1,2} + \beta_3 + x_{n+1,3})} \\
\propto & \frac{\Gamma(\beta_1 + x_{n+1,1}) \Gamma(\beta_2 + x_{n+1,2}) \Gamma(\beta_3 + x_{n+1,3})}{\Gamma(\beta + \beta_2 + \beta_3 + 1)} \\
\propto & \Gamma(\beta_1 + x_{n+1,1}) \Gamma(\beta_2 + x_{n+1,2}) \Gamma(\beta_3 + x_{n+1,3}) \\
\overset{(i)}{\propto} & (\beta_1)^{x_{n+1,1}} (\beta_2)^{x_{n+1,2}} (\beta_3)^{x_{n+1,3}}.
\end{aligned}$$

The step (i) above is because of the following: only one of $x_{n+1,1}, x_{n+1,2}, x_{n+1,3}$ is 1 and the other two are zeros. If $x_{n+1,1} = 1$, then $x_{n+1,2} = x_{n+1,3} = 0$, and

$$\Gamma(\beta_1 + x_{n+1,1})\Gamma(\beta_2 + x_{n+1,2})\Gamma(\beta_3 + x_{n+1,3}) = \beta_1\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3) \propto \beta_1,$$

and similarly for the case when $x_{n+1,2} = 1$ and the case when $x_{n+1,3} = 1$. The three cases can be summarized in the formula $\propto (\beta_1)^{x_{n+1,1}}(\beta_2)^{x_{n+1,2}}(\beta_3)^{x_{n+1,3}}$.

Note that $\boldsymbol{x}_{n+1} = (1,0,0)$ or $(0,1,0)$ or $(0,0,1)$. Therefore the predictive distribution of $\boldsymbol{x}_{n+1}$ is a multinomial$(1; \frac{\beta_1}{\beta}, \frac{\beta_2}{\beta}, \frac{\beta_3}{\beta})$ distribution, where $\beta = \beta_1 + \beta_2 + \beta_3 = \alpha_1 + \alpha_2 + \alpha_3 + x_{.1} + x_{.2} + x_{.3} = \alpha_1 + \alpha_2 + \alpha_3 + n$.

2. (a) The posterior distribution $p(\theta_1, \theta_2|\boldsymbol{x}, \boldsymbol{y})$ is

$$\begin{aligned}
& p(\theta_1, \theta_2|\boldsymbol{x}, \boldsymbol{y}) \\
\propto\ & p(\boldsymbol{x}, \boldsymbol{y}|\theta_1, \theta_2)p(\theta_1, \theta_2) \\
=\ & p(\boldsymbol{x}|\theta_1)p(\boldsymbol{y}|\theta_2)p(\theta_1)p(\theta_2) \\
=\ & \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \theta_1)^2}{2\sigma^2}\right\} \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_j - \theta_2)^2}{2\sigma^2}\right\} \\
& \quad \times \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left\{-\frac{\theta_1^2}{2\tau_1^2}\right\} \times \frac{1}{\sqrt{2\pi\tau_2^2}} \exp\left\{-\frac{\theta_2^2}{2\tau_2^2}\right\} \\
\propto\ & \exp\left\{-\frac{\sum_{i=1}^{m}(x_i - \theta_1)^2}{2\sigma^2} - \frac{\theta_1^2}{2\tau_1^2}\right\} \cdot \exp\left\{-\frac{\sum_{j=1}^{n}(y_j - \theta_2)^2}{2\sigma^2} - \frac{\theta_2^2}{2\tau_2^2}\right\} \\
\propto\ & \exp\left\{-\frac{m(\theta_1 - \bar{x})^2}{2\sigma^2} - \frac{\theta_1^2}{2\tau_1^2}\right\} \cdot \exp\left\{-\frac{n(\theta_2 - \bar{y})^2}{2\sigma^2} - \frac{\theta_2^2}{2\tau_2^2}\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\left(\frac{m}{\sigma^2} + \frac{1}{\tau_1^2}\right)\theta_1^2 + \left(\frac{m\bar{x}}{\sigma^2}\right)\theta_1\right\} \\
& \quad \cdot \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau_2^2}\right)\theta_2^2 + \left(\frac{n\bar{y}}{\sigma^2}\right)\theta_2\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\left(\frac{m}{\sigma^2} + \frac{1}{\tau_1^2}\right)\left(\theta_1 - \frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2} + \frac{1}{\tau_1^2}}\right)^2\right\} \\
& \quad \cdot \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau_2^2}\right)\left(\theta_2 - \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_2^2}}\right)^2\right\}.
\end{aligned}$$

This shows that (θ_1, θ_2) are jointly independent in the posterior, since the joint posterior density can be factorized into the product of two density functions for θ_1 and θ_2, respectively, i.e.

$$p(\theta_1|\boldsymbol{x},\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}\right)\left(\theta_1 - \frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}}\right)^2\right\},$$

$$p(\theta_2|\boldsymbol{x},\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}\right)\left(\theta_2 - \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}\right)^2\right\}.$$

Furthermore, we can recognize that

$$\theta_1|\boldsymbol{x},\boldsymbol{y} \sim \mathrm{N}\left(\frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}}, \frac{1}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}}\right), \quad \theta_2|\boldsymbol{x},\boldsymbol{y} \sim \mathrm{N}\left(\frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}\right).$$

Jointly, $p(\theta_1, \theta_2|\boldsymbol{x},\boldsymbol{y})$ is the bivariate normal distribution with mean and covariance matrix given by

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}\Bigg|\boldsymbol{x},\boldsymbol{y} \sim \mathrm{N}\left(\begin{bmatrix}\frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}}\\ \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}\end{bmatrix}, \begin{bmatrix}\frac{1}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}} & 0\\ 0 & \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}\end{bmatrix}\right).$$

(b) Using the independence and the normality of θ_1 and θ_2 given $\boldsymbol{x},\boldsymbol{y}$, we can immediately obtain that $\delta = \theta_1 - \theta_2$ given $\boldsymbol{x},\boldsymbol{y}$ follows a normal distribution. The mean and variance of this normal distribution are

$$\mathrm{E}(\delta|\boldsymbol{x},\boldsymbol{y}) = \mathrm{E}(\theta_1|\boldsymbol{x},\boldsymbol{y}) - \mathrm{E}(\theta_2|\boldsymbol{x},\boldsymbol{y}) = \frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}} - \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}},$$

$$\mathrm{Var}(\delta|\boldsymbol{x},\boldsymbol{y}) = \mathrm{Var}(\theta_1|\boldsymbol{x},\boldsymbol{y}) + \mathrm{Var}(\theta_2|\boldsymbol{x},\boldsymbol{y}) = \frac{1}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}} + \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}.$$

Hence,

$$\delta|\boldsymbol{x},\boldsymbol{y} \sim \mathrm{N}\left(\frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}} - \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}, \ \frac{1}{\frac{m}{\sigma^2}+\frac{1}{\tau_1^2}} + \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_2^2}}\right). \tag{1}$$

(c) In the frequentist statistics (when θ_1 and θ_2 are treated as fixed parameters), we have that

$$\bar{x} \sim \mathrm{N}\left(\theta_1, \frac{\sigma^2}{m}\right), \quad \bar{y} \sim \mathrm{N}\left(\theta_2, \frac{\sigma^2}{n}\right),$$

and that $\bar{x}$ and $\bar{y}$ are independent. Therefore, $\bar{x} - \bar{y}$ is normally distributed, with mean and variance given by

$$\begin{aligned}
\mathrm{E}(\bar{x} - \bar{y}|\theta_1, \theta_2) &= \theta_1 - \theta_2, \\
\mathrm{Var}(\bar{x} - \bar{y}|\theta_1, \theta_2) &= \frac{\sigma^2}{m} + \frac{\sigma^2}{n} = \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right).
\end{aligned}$$

Hence,

$$\bar{x} - \bar{y} \mid \theta_1, \theta_2 \sim \mathrm{N}\left(\theta_1 - \theta_2, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right). \tag{2}$$

The statistic

$$T = \frac{(\bar{x} - \bar{y}) - (\theta_1 - \theta_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

is the standardized version of $\bar{x} - \bar{y}$, so $T \sim \mathrm{N}(0, 1)$.

We can relate the distribution of T (or the distribution of $\bar{x} - \bar{y}$) to the posterior $p(\delta|\boldsymbol{x}, \boldsymbol{y})$ in (b) in the following way: When the sample sizes m and n are both sufficiently large, the distribution in Equation (1) approximately become

$$\begin{aligned}
\delta|\boldsymbol{x}, \boldsymbol{y} &\overset{\text{approx.}}{\sim} \mathrm{N}\left(\frac{\frac{m\bar{x}}{\sigma^2}}{\frac{m}{\sigma^2}} - \frac{\frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2}}, \frac{1}{\frac{m}{\sigma^2}} + \frac{1}{\frac{n}{\sigma^2}}\right) \\
&= \mathrm{N}\left(\bar{x} - \bar{y}, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right).
\end{aligned}$$

Compared to the distribution in (2), they have the same variance, but the roles of $\delta = \theta_1 - \theta_2$ and $\bar{x} - \bar{y}$ are swapped. In the frequentist inference, it is the data/statistic $(\bar{x} - \bar{y})$ centered at the parameter $(\delta = \theta_1 - \theta_2)$. In the Bayesian inference, it is the parameter $(\delta = \theta_1 - \theta_2)$ centered at the data/statistic $(\bar{x} - \bar{y})$.

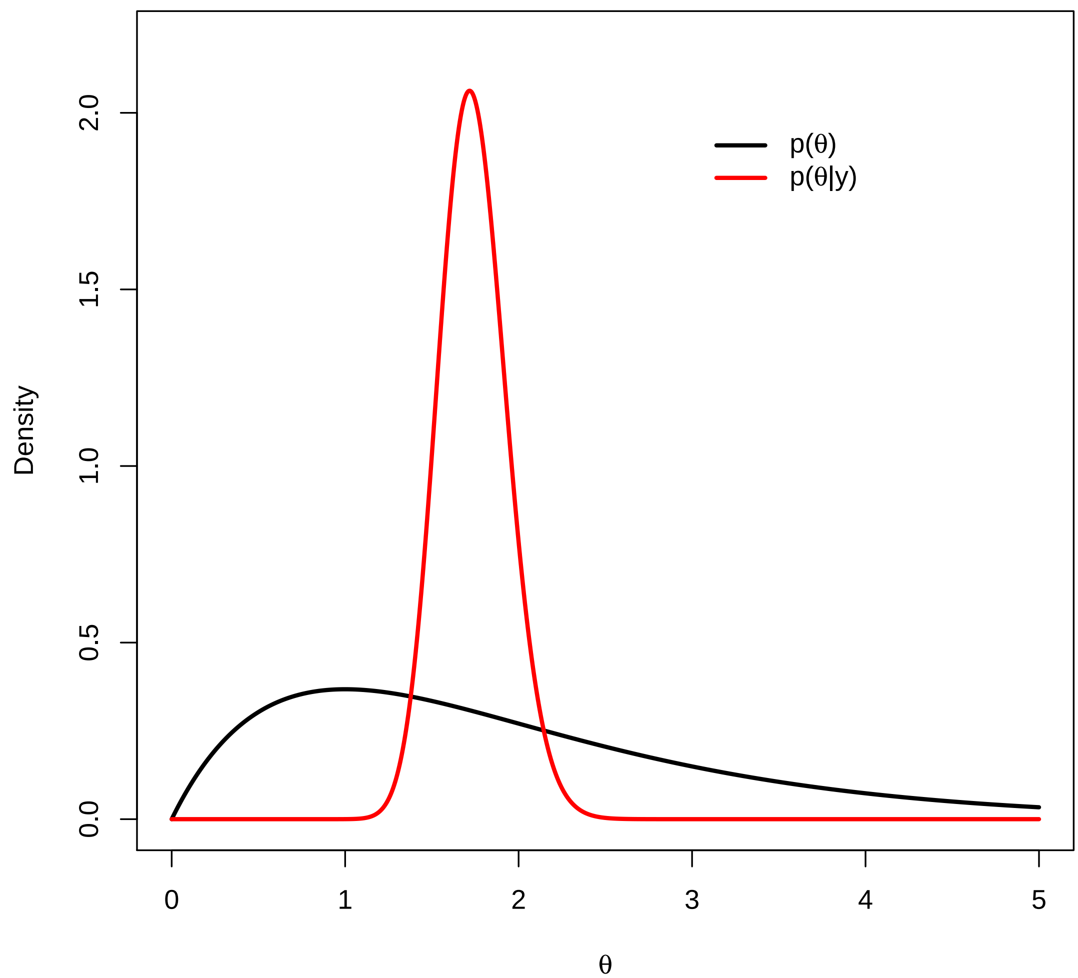
3. (a) The Poisson model has the likelihood $p(\boldsymbol{y}|\theta) = \prod_{i=1}^n \theta^{y_i} e^{-\theta}/y_i! \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}$. The Gamma(2,1) prior of θ is $p(\theta) \propto \theta^{2-1} e^{-\theta}$. The posterior is

$$
\begin{aligned}
p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\
&\propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \cdot \theta e^{-\theta} \\
&\propto \theta^{\sum_{i=1}^n y_i + 2 - 1} e^{-(n+1)\theta}.
\end{aligned}
$$

This is the density of Gamma($\sum_{i=1}^n y_i + 2, n+1$). The plots of the prior and the posterior is given below.

```r
y <- c(0,0,0,0,0,0,1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,5)
n <- length(y)
a0 <- 2          # prior parameters
b0 <- 1
a <- a0 + sum(y)         # posterior parameters
b <- b0 + n

theta <- seq(from=0, to=5, by=0.001)
plot(theta, dgamma(theta,shape=a0,rate=b0), type="l", lwd=2.5,
     xlab=expression(theta), ylab="Density", ylim=c(0,2.2))
points(theta, dgamma(theta,shape=a,rate=b), type="l", lwd=2.5, col="red")
legend(3,2,legend = c(expression(paste("p(", theta, ")")),
       expression(paste("p(", theta, "|y)"))), lty=1,
       lwd=2.5, cex=1, col=c("black","red"), bty="n")
```




```r
# posterior mean
a/b
```

```
## [1] 1.73913
```

```r
# posterior standard deviation
sqrt(a/b^2)
```

```
## [1] 0.1944407
```

```r
# posterior mode
(a-1)/b

## [1] 1.717391
```

Let $a = \sum_{i=1}^n +2$ and $b = n + 1$. The posterior mean is $a/b = 80/46 = 1.7391$. The posterior standard deviation is $\sqrt{a/b^2} = \sqrt{80/46^2} = 0.1944$. The posterior mode is $(a-1)/b = 79/46 = 1.7174$. The sample mean is 1.7333. This prior is sensible: the posterior mean and mode are close to the sample mean. The posterior mean is between the prior mean (2) and the sample mean (1.7333), but is much closer to the sample mean. The posterior variation is also much smaller than the prior variation (observed from the plots).

(b) If we use the noninformative prior $p(\theta) \propto \theta^{-1/2}$, then the posterior becomes

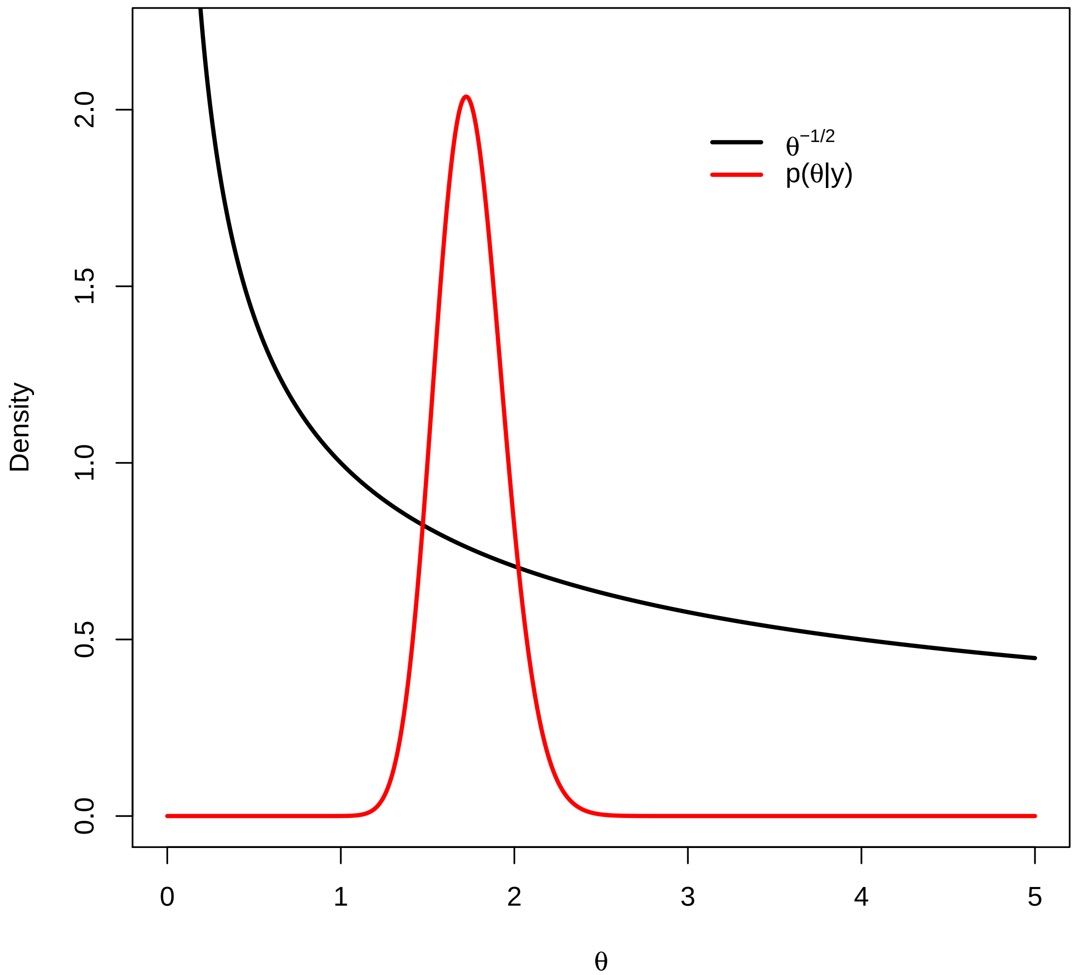
$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \cdot \theta^{-1/2} \\ &\propto \theta^{\sum_{i=1}^n y_i + 1/2 - 1} e^{-n\theta}. \end{aligned}$$

This is the density of Gamma$(\sum_{i=1}^n y_i + 1/2, n)$. The posterior can be plotted, but the prior is not a proper density, so we cannot plot it (we can only plot the function $\theta^{-1/2}$).

```r
y <- c(0,0,0,0,0,0,1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,5)
n <- length(y)
a0 <- 1/2               # prior parameters
b0 <- 0
a <- a0 + sum(y)        # posterior parameters
b <- b0 + n

theta <- seq(from=0, to=5, by=0.001)
plot(theta, theta^(-1/2), type="l", lwd=2.5,
     xlab=expression(theta), ylab="Density", ylim=c(0,2.2))
```

```r
points(theta, dgamma(theta,shape=a,rate=b), type="l", lwd=2.5, col="red")
legend(3,2,legend = c(expression(paste(theta^"-1/2")),
       expression(paste("p(", theta, "|y)"))), lty=1,
       lwd=2.5, cex=1, col=c("black","red"), bty="n")
```




```r
# posterior mean
a/b
```

```
## [1] 1.744444

# posterior standard deviation
sqrt(a/b^2)

## [1] 0.1968894

# posterior mode
(a-1)/b

## [1] 1.722222
```

Let $a = \sum_{i=1}^{n} +1/2$ and $b = n$. The posterior mean is $a/b = 78.5/45 = 1.7444$. The posterior standard deviation is $\sqrt{a/b^2} = 0.1969$. The posterior mode is $(a-1)/b = 1.7222$. The sample mean is 1.7333. This prior, although not a proper density, is still sensible: the posterior distribution has a proper density; the posterior mean and mode are close to the sample mean. (Note that the prior does not have a mean in this case.)