Compulsory HWA1, Bayesian Statistics

1. (2p) A survey study is conducted to investigate customer satisfaction. The parameter $\theta \in \{0, 1, 2\}$ is the level of satisfaction. The probability $\mathrm{P}(X = x \mid \theta)$ of the responses is given in the following table

	x			
θ	1: satisfied	2: OK	3: not satisfied	4: no answer
2	0.6	0.1	0	0.3
1	0.1	0.2	0.1	0.6
0	0	0.2	0.79	0.01

   (a) We set the prior as $\mathrm{P}(\theta = 2) = 0.2$, $\mathrm{P}(\theta = 1) = 0.5$, and $\mathrm{P}(\theta = 0) = 0.3$. Find the corresponding posterior distribution.
   Solution: The posterior distribution is computed from $\mathrm{P}(\theta \mid x) \propto \mathrm{P}(x \mid \theta)\,\mathrm{P}(\theta)$. Then, we can tabulate $\mathrm{P}(x \mid \theta)\,\mathrm{P}(\theta)$ as

	x			
θ	1: satisfied	2: OK	3: not satisfied	4: no answer
2	0.12	0.02	0	0.06
1	0.05	0.10	0.05	0.3
0	0	0.06	0.237	0.003

   Normalizing this table columnwise yields the posterior probability

	x			
θ	1: satisfied	2: OK	3: not satisfied	4: no answer
2	12/17	1/9	0	20/121
1	5/17	5/9	50/287	100/121
0	0	3/9	237/287	1/121

   (b) Find the prior that maximizes the Shannon entropy.
   Solution: Denote $p_i = \mathrm{P}(\theta = i)$. The Shannon entropy is computed as

$$\begin{aligned} S(p) &= -\mathrm{E}[\log p] = -\sum_{i=0}^{2} p_i \log p_i \\ &= -\left[p_0 \log p_0 + p_1 \log p_1 + (1 - p_0 - p_1) \log (1 - p_0 - p_1)\right]. \end{aligned}$$

   Note that

$$\begin{aligned} \frac{\partial S(p)}{\partial p_0} &= -\left[\log p_0 + 1 - \log p_2 - 1\right] \\ \frac{\partial S(p)}{\partial p_1} &= -\left[\log p_1 + 1 - \log p_2 - 1\right] \end{aligned}$$

   which means that $p_0 = p_1 = p_2$. Hence, the discrete uniform distribution maximizes the Shannon entropy. In fact, we already stated the result that the discrete uniform distribution maximizes the Shannon entropy in the slide.

(c) Suppose that our prior information is that $\mathrm{P}(\theta = 0) = \mathrm{P}(\theta = 1)$ and $\mathrm{E}[\theta] = 1$. Find the corresponding prior.
Solution: Denote $p_i = \mathrm{P}(\theta = i)$. Then, we have

$$\begin{aligned} \mathrm{E}[\theta] = 1 &= 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2, \\ 1 &= p_0 + p_1 + p_2, \\ 0 &= p_0 - p_1. \end{aligned}$$

Solving the linear system yields $p_0 = p_1 = p_2 = 1/3$.

2. (2p) Suppose that we have observed an iid sample X_1, ..., X_n, $n > 2$, from a multinomial distribution with 3 categories and 1 trial. The probability mass function is

$$\mathrm{P}(X_i = x_i) = \left[p^2\right]^{I(x_i=1)} \left[2p(1-p)\right]^{I(x_i=2)} \left[(1-p)^2\right]^{I(x_i=3)},$$

where $i = 1, \cdots, n$ and

$$I(x_i = j) = \begin{cases} 1, & \text{if } x_i = j, \\ 0, & \text{if } x_i \neq j. \end{cases}$$

(a) Find the conjugate prior and the corresponding posterior.
Solution: Let $n_1 = \sum_{i=1}^n I(x_i = 1)$, $n_2 = \sum_{i=1}^n I(x_i = 2)$, and $n_3 = \sum_{i=1}^n I(x_i = 3)$. The joint mass function can be expressed as

$$f(x|p) = 2^{n_2} p^{2n_1+n_2} (1-p)^{n_2+2n_3}.$$

Hence, the Beta distribution $\mathrm{Beta}(a_0, b_0)$ is a conjugate prior. The posterior is proportional to

$$\begin{aligned} \pi(p|x) &\propto f(x \mid p)\, \pi(p) \\ &= 2^{n_2} p^{2n_1+n_2} (1-p)^{n_2+2n_3} \cdot \frac{1}{B(a_0, b_0)} p^{a_0-1} (1-p)^{b_0-1} \\ &\propto p^{2n_1+n_2+a_0-1} (1-p)^{n_2+2n_3+b_0-1}. \end{aligned}$$

Hence, the posterior is $\mathrm{Beta}(2n_1 + n_2 + a_0, n_2 + 2n_3 + b_0)$.

(b) Find the MAP of p.
Solution: The log of posterior is

$$\log \pi(p|x) = \text{constant} + (2n_1 + n_2 + a_0 - 1) \log p + (n_2 + 2n_3 + b_0 - 1) \log(1-p).$$

Note that

$$\frac{d \log \pi(p|x)}{dp} = \frac{2n_1 + n_2 + a_0 - 1}{p} - \frac{n_2 + 2n_3 + b_0 - 1}{1-p}.$$

The stationary point is

$$\hat{p} = \frac{2n_1 + n_2 + a_0 - 1}{a_0 + b_0 + 2n_1 + 2n_2 + 2n_3 - 2}.$$

The second derivative is

$$\begin{aligned} \frac{d^2 \log \pi(\hat{p}|x)}{dp^2} &= -\frac{2n_1 + n_2 + a_0 - 1}{\hat{p}^2} - \frac{n_2 + 2n_3 + b_0 - 1}{(1-\hat{p})^2} \\ &= -\frac{(a_0 + b_0 + 2n_1 + 2n_2 + 2n_3 - 2)^3}{(2n_1 + n_2 + a_0 - 1)(n_2 + 2n_3 + b_0 - 1)} < 0 \end{aligned}$$

if $a_0 > 1$, $b_0 > 1$, and $n > 2$. Hence, the MAP is $\hat{p}$.

(c) Derive the Jeffreys prior.
Solution: Note that

$$\log f(x|p) = n_2 \log 2 + (2n_1 + n_2) \log p + (n_2 + 2n_3) \log (1-p).$$

Thus,

$$\begin{aligned} \frac{d \log f(x|p)}{dp} &= \frac{2n_1 + n_2}{p} - \frac{n_2 + 2n_3}{1-p}, \\ \frac{d^2 \log f(x|p)}{dp^2} &= -\frac{2n_1 + n_2}{p^2} - \frac{n_2 + 2n_3}{(1-p)^2}. \end{aligned}$$

Hence, the Fisher information is

$$\mathcal{I}(p) = -\mathrm{E}\left[\frac{d^2 \log f(x|p)}{dp^2}\right] = \frac{2}{p(1-p)}.$$

It means that the Jeffreys prior is $\pi(p) \propto \sqrt{\frac{2}{p(1-p)}}$.

3. (1p) Consider one observation $X \mid \theta \sim N(\theta, 1)$. Find the reference prior.
Solution: We start with the Jeffreys prior for θ, that is $\pi^*(\theta) \propto 1$. Suppose that we have k iid copies $x = (x_i, ..., x_k)$. The posterior is $\theta \mid x \sim N\left(\bar{x},\ k^{-1}\right)$, which is proper. Then,

$$\begin{aligned} \int f(x \mid \theta) \log \pi^*(\theta \mid x)\, dx &= \int \left[\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \theta)^2}{2}\right\}\right] \left[-\frac{1}{2}\log\left(2\pi k^{-1}\right) - \frac{(\theta - \bar{x})^2}{2k^{-1}}\right] dx \\ &= -\frac{1}{2}\log\left(\frac{2\pi}{k}\right). \end{aligned}$$

Thus, for an θ_0,

$$\frac{p_k(\theta)}{p_k(\theta_0)} = \frac{\exp\left\{-\frac{1}{2}\log\left(\frac{2\pi}{k}\right)\right\}}{\exp\left\{-\frac{1}{2}\log\left(\frac{2\pi}{k}\right)\right\}} = 1,$$

which means that the uniform prior is the reference prior.
We also know that under some conditions the reference prior can be approximated by the Jeffreys prior. So you can also use the Jeffreys prior to approximate the reference prior.

4. (3p) Consider the normal linear model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$, where $\theta = (\beta, \sigma)$ is unknown, and β is a $p \times 1$ vector. We also assume $n > p$.

(a) Find the independent Jeffreys prior.
Solution: The likelihood is

$$f(y \mid \beta, \sigma) = \exp\left\{-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\right\}$$

The Fisher information is derived from

$$\begin{aligned} \frac{\partial \log f(y \mid \beta, \sigma)}{\partial \begin{bmatrix}\beta \\ \sigma\end{bmatrix}} &= \begin{bmatrix} -\frac{1}{\sigma^2}\left(X^T X\beta - X^T y\right) \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}(y - X\beta)^T (y - X\beta) \end{bmatrix}, \\ \frac{\partial^2 \log f\left(y \mid \beta, \sigma^2\right)}{\partial \begin{bmatrix}\beta \\ \sigma\end{bmatrix} \partial \begin{bmatrix}\beta \\ \sigma\end{bmatrix}^T} &= \begin{bmatrix} -\frac{1}{\sigma^2} X^T X & \frac{2}{\sigma^3}\left(X^T X\beta - X^T y\right) \\ \frac{2}{\sigma^3}\left(X^T X\beta - X^T y\right)^T & \frac{n}{\sigma^2} - \frac{3}{\sigma^4}(y - X\beta)^T (y - X\beta) \end{bmatrix} \\ &\Downarrow \\ \mathcal{I}(\beta, \sigma) &= \begin{bmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}. \end{aligned}$$

Hence, the independent Jeffreys prior is

$$\pi(\beta,\sigma)=\pi(\beta)\,\pi(\sigma) \quad \propto \quad \sqrt{\det\left(\mathcal{I}(\beta)\right)}\sqrt{\det\left(\mathcal{I}(\sigma)\right)}=\sigma^{-1}.$$

(b) Find the posterior of $\theta=(\beta,\sigma)$.
Solution: The posterior satisfies

$$\begin{aligned}\pi(\beta,\sigma\mid y) &\propto \exp\left\{-n\log(\sigma)-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\}\cdot\frac{1}{\sigma}\\ &\propto \frac{1}{(\sigma^2)^{(n+1)/2}}\exp\left\{-\frac{\beta^TX^TX\beta-2y^TX\beta+y^Ty}{2\sigma^2}\right\}.\end{aligned}$$

We observe a quadratic expression in β as $\beta^TX^TX\beta-2y^TX\beta$. Hence, we can match it to some normal density. If you are not confident with creating a normal density from a quadratic form, one general useful result is that, for a symmetric matrix $A>0$, we have

$$\beta^TA\beta-2b^T\beta \quad = \quad \left(\beta-A^{-1}b\right)^TA\left(\beta-A^{-1}b\right)-b^TA^{-1}b.$$

Using this result with $A=X^TX$ and $b=X^Ty$, we have

$$\begin{aligned}\beta^TX^TX\beta-2y^TX\beta &= \left(\beta-\left(X^TX\right)^{-1}X^Ty\right)^TX^TX\left(\beta-\left(X^TX\right)^{-1}X^Ty\right)-y^TX\left(X^TX\right)^{-1}X^Ty,\\ &= (\beta-\mu_n)^TX^TX(\beta-\mu_n)-\mu_n^TX^TX\mu_ny,\end{aligned}$$

where $\mu_n=\left(X^TX\right)^{-1}X^Ty$. Hence,

$$\pi(\beta,\sigma\mid y) \quad \propto \quad \frac{1}{(\sigma^2)^{p/2}}\exp\left\{-\frac{(\beta-\mu_n)^TX^TX(\beta-\mu_n)}{2\sigma^2}\right\}\times\frac{1}{(\sigma^2)^{(n-p+1)/2}}\exp\left\{-\frac{y^Ty}{2\sigma^2}+\frac{\mu_n^TX^TX\mu_n}{2\sigma^2}\right\},$$

which means that

$$\begin{aligned}\beta\mid\sigma,y &\sim N_p\left(\mu_n,\ \sigma^2\left(X^TX\right)^{-1}\right)\\ \pi(\sigma\mid y) &\propto \frac{1}{(\sigma^2)^{(n-p+1)/2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y^Ty-\mu_n^TX^TX\mu_n\right)\right\}=\frac{1}{(\sigma^2)^{(n-p+1)/2}}\exp\left\{-\frac{y^T(I_n-H)y}{2\sigma^2}\right\},\end{aligned}$$

where $H=X\left(X^TX\right)^{-1}X^T$ and

$$y^Ty-\mu_n^TX^TX\mu_n \quad = \quad y^T\left[I_n-X\left(X^TX\right)^{-1}X^T\right]y=y^T(I_n-H)y.$$

(c) Suppose that we observe an independent new observation x_0 and want to predict the corresponding y_0. Find the predictive distribution.
Solution: The predictive distribution is

$$\begin{aligned}f(y_0\mid y) &= \int f(y_0\mid y,\theta)\,\pi(\theta\mid y)\,d\theta\\ &= \int\int f(y_0\mid y,\beta,\sigma)\cdot\pi(\beta\mid y,\sigma)\,\pi(\sigma\mid y)\,d\beta d\sigma,\end{aligned}$$

where $y_0\mid y,\beta,\sigma\sim N\left(x_0^T\beta,\sigma^2\right)$ and $\beta\mid y,\sigma\sim N_p\left(\mu_n,\ \sigma^2\left(X^TX\right)^{-1}\right)$. Thus,

$$\begin{bmatrix}y_0\\ \beta\end{bmatrix}\mid y,\sigma \quad \sim \quad N_{1+p}\left(\begin{bmatrix}x_0^T\mu_n\\ \mu_n\end{bmatrix},\ \begin{bmatrix}\sigma^2\left[1+x_0^T\left(X^TX\right)^{-1}x_0\right] & \sigma^2x_0^T\left(X^TX\right)^{-1}\\ \sigma^2\left(X^TX\right)^{-1}x_0 & \sigma^2\left(X^TX\right)^{-1}\end{bmatrix}\right).$$

Hence, $y_0 \mid y, \sigma \sim N\left(x_0^T \mu_n, \sigma^2\left[1 + x_0^T \left(X^T X\right)^{-1} x_0\right]\right)$. Or equivalently,

$$y_0 \mid y, \sigma^2 \quad \sim \quad N\left(x_0^T \mu_n, \sigma^2\left[1 + x_0^T \left(X^T X\right)^{-1} x_0\right]\right).$$

By change of variables, we can see that the posterior distribution of σ^2 is inverse-Gamma as

$$\begin{aligned}
\pi\left(\sigma^2 \mid y\right) &\propto \frac{1}{\left(\sigma^2\right)^{(n-p+1)/2}} \exp\left\{-\frac{y^T\left(I_n - H\right) y}{2\sigma^2}\right\} \cdot \left|\frac{d\sigma}{d\sigma^2}\right| \\
&\propto \frac{1}{\left(\sigma^2\right)^{(n-p)/2+1}} \exp\left\{-\frac{y^T\left(I_n - H\right) y}{2\sigma^2}\right\} \\
&\sim \text{InvGamma}\left(\frac{n-p}{2},\ \frac{1}{2} y^T\left(I_n - H\right) y\right).
\end{aligned}$$

This means that

$$\left(y_0, \sigma^2\right) \mid y \quad \sim \quad \text{NIV}\left(\frac{n-p}{2},\ \frac{1}{2} y^T\left(I_n - H\right) y,\ x_0^T \mu_n, 1 + x_0^T\left(X^T X\right)^{-1} x_0\right).$$

Thus, using the property of normal-inverse-gamma distribution, we obtain

$$y_0 \mid y \quad \sim \quad t_{n-p}\left(x_0^T \mu_n,\ \frac{y^T\left(I_n - H\right) y}{n-p}\left[1 + x_0^T\left(X^T X\right)^{-1} x_0\right]\right).$$

5. (2p) Suppose that we want to model the delay time (in minutes) of the SL commute train service between Stockholm and Uppsala. We assume that the delay time x_1, ..., x_n are iid and follow an exponential distribution $\text{Exp}(\theta)$ with mean θ^{-1}.

   (a) Show that the gamma prior $\text{Gamma}(a_0, b_0)$ with density

   $$\pi(\theta) \quad = \quad \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} \exp\left(-b_0 \theta\right)$$

   is a conjugate prior.
   Solution: The posterior satisfies

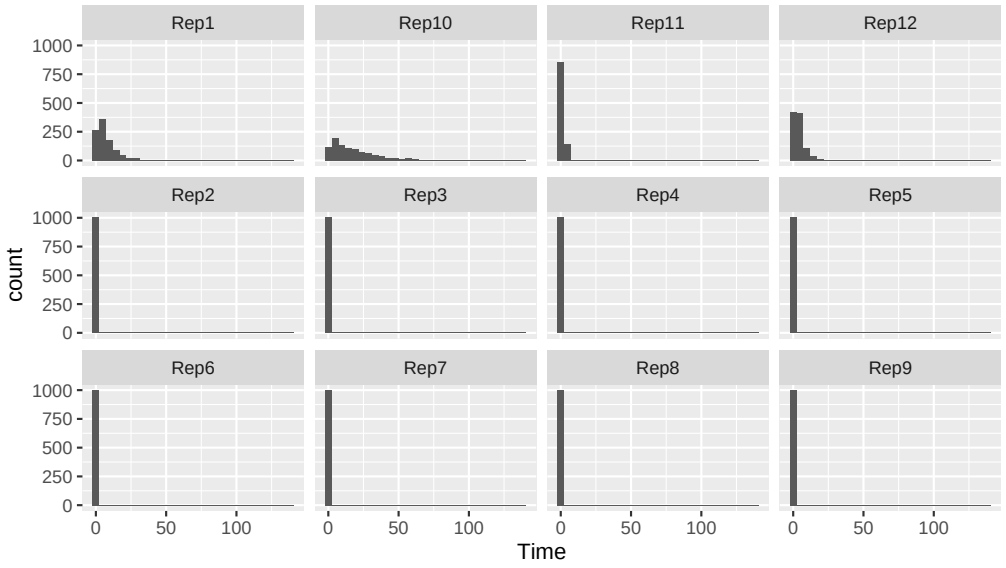
   $$\begin{aligned}
   \pi(\theta \mid x) &\propto \left[\prod_{i=1}^{n} \theta \exp\left(-\theta x_i\right)\right] \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} \exp\left(-b_0 \theta\right) \\
   &\propto \theta^{a_0 + n - 1} \exp\left\{-\left(b_0 + \sum_{i=1}^{n} x_i\right)\theta\right\} \sim \text{Gamma}\left(a_0 + n, b_0 + \sum_{i=1}^{n} x_i\right).
   \end{aligned}$$

   (b) Two statisticians propose their respective priors: $\text{Gamma}(0.25, 0.05)$ and $\text{Gamma}(25, 5)$. Both priors have mean 5 but different variances. Which prior is more reasonable by prior predictive check?
   Solution:

```
set.seed(12345)
Exp1 <- Exp2 <- NULL
for(i in 1 : 12){
   ## Simulate theta
   Lambda1 <- rgamma(n = 1, shape = 0.25, rate = 0.05)
   Lambda2 <- rgamma(n = 1, shape = 25, rate = 5)
   ## Simulate Exponential random numbers
   Exp1 <- c(Exp1, rexp(n = 1000, rate = 1 / Lambda1))
   Exp2 <- c(Exp2, rexp(n = 1000, rate = 1 / Lambda2))
}
library(ggplot2)
```

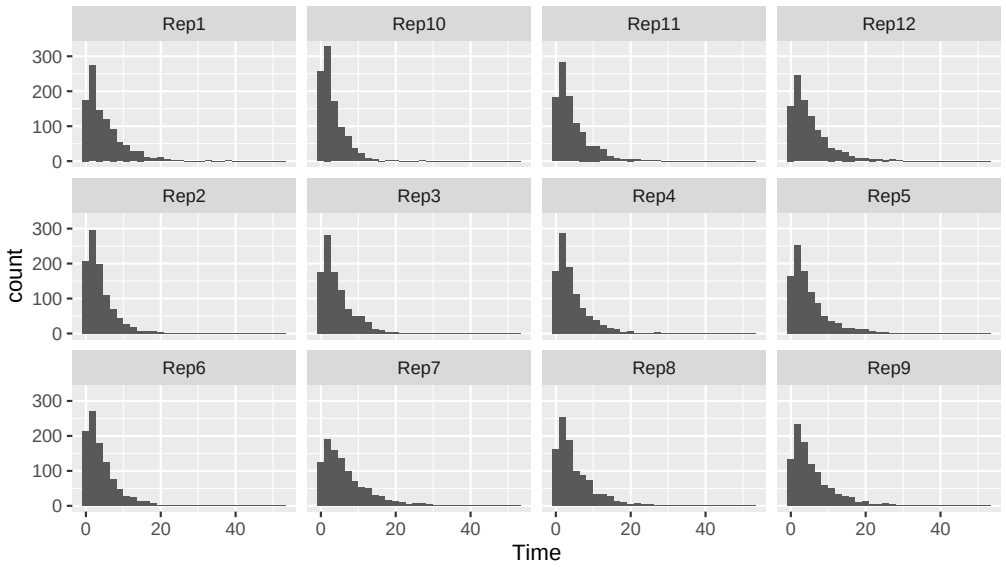
The first prior yields

```
DF1 <- data.frame(Time = Exp1, rep = rep(paste0("Rep", c(1 : 12)), each = 1000))
ggplot(DF1, aes(x = Time)) + geom_histogram() + facet_wrap(~ rep, nrow = 3)
```



It sometimes yields too small delay times. And the second prior yields

```
DF2 <- data.frame(Time = Exp2, rep = rep(paste0("Rep", c(1 : 12)), each = 1000))
ggplot(DF2, aes(x = Time)) + geom_histogram() + facet_wrap(~ rep, nrow = 3)
```



The second prior is more reasonable than the first prior.

(c) Suppose that we have observed the data in Delay.csv, accessible on Studium. Find the predictive distribution using the prior you chose in (b), and perform posterior predictive check. If you don't know how to generate random numbers from the predictive distribution, describe how the posterior predictive check can be done if you knew how to generate random numbers from such distribution.

Solution: The predictive distribution is

$$\begin{aligned}
f\left(x_{\text{new}} \mid x\right) &= \int f\left(x_{\text{new}} \mid x, \theta\right) \pi\left(\theta \mid x\right) d\theta \\
&= \int \theta \exp\left(-\theta x_{\text{new}}\right) \cdot \frac{\left(b_0 + \sum_{i=1}^{n} x_i\right)^{a_0+n}}{\Gamma\left(a_0+n\right)} \theta^{a_0+n-1} \exp\left\{-\left(b_0 + \sum_{i=1}^{n} x_i\right)\theta\right\} d\theta \\
&= \frac{\left(b_0 + \sum_{i=1}^{n} x_i\right)^{a_0+n}}{\Gamma\left(a_0+n\right)} \int \theta^{a_0+n} \exp\left\{-\left(b_0 + \sum_{i=1}^{n} x_i + x_{\text{new}}\right)\theta\right\} d\theta \\
&= \frac{\left(b_0 + \sum_{i=1}^{n} x_i\right)^{a_0+n}}{\Gamma\left(a_0+n\right)} \cdot \frac{\Gamma\left(a_0+n+1\right)}{\left(b_0 + \sum_{i=1}^{n} x_i + x_{\text{new}}\right)^{a_0+n+1}} \\
&= \frac{\left(a_0+n\right)\left(b_0 + \sum_{i=1}^{n} x_i\right)^{a_0+n}}{\left(b_0 + \sum_{i=1}^{n} x_i + x_{\text{new}}\right)^{a_0+n+1}}.
\end{aligned}$$

The cumulative distribution function is

$$F\left(x_{\text{new}} \mid x\right) = 1 - \left(\frac{b_0 + \sum_{i=1}^{n} x_i}{b_0 + \sum_{i=1}^{n} x_i + x_{\text{new}}}\right)^{a_0+n}.$$

We can easily generate x_{new} from uniform random variable U by solving

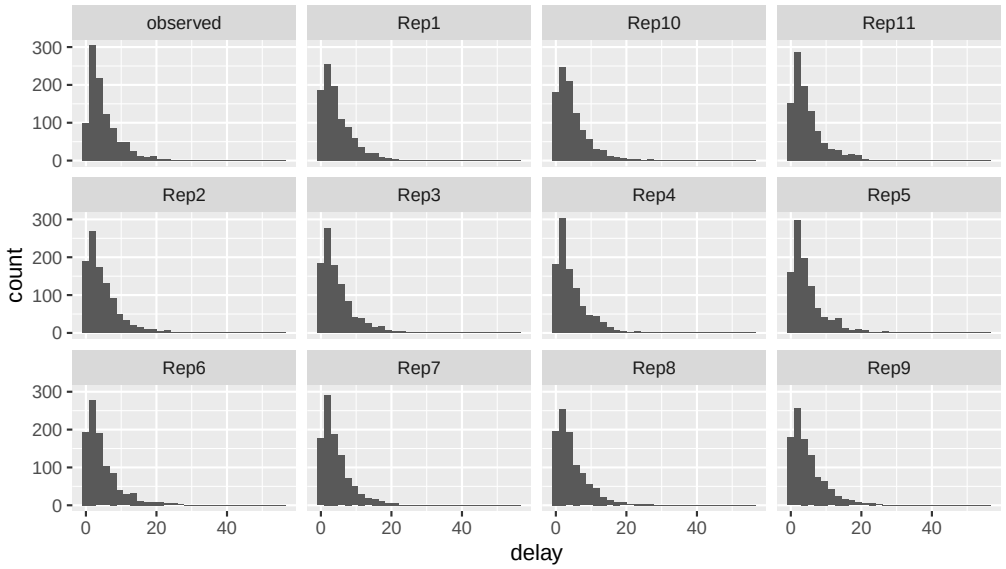
$$u = 1 - \left(\frac{b_0 + \sum_{i=1}^{n} x_i}{b_0 + \sum_{i=1}^{n} x_i + x_{\text{new}}}\right)^{a_0+n},$$

which yields

$$x_{\text{new}} = \frac{b_0 + \sum_{i=1}^{n} x_i}{(1-u)^{1/(a_0+n)}} - b_0 - \sum_{i=1}^{n} x_i,$$

where $U \sim \text{Uniform}\left[0, 1\right]$.

```r
## Import data
Delay <- read.csv("Delay.csv", sep="")
Total <- sum(Delay$x)
a0 <- 25; b0 <- 5; n <- nrow(Delay)
## Simulate from predictive distribution
ysim <- matrix(NA, n, 11)
for(i in 1 : 11){
   u <- runif(n, 0, 1)
   ysim[, i] <- (b0 + Total) / ((1 - u) ^ (1 / (a0 + n))) - b0 - Total
}
## Plot
library(ggplot2)
DF <- data.frame(delay = c(Delay$x, c(ysim)),
      rep = rep(c("observed", paste0("Rep", c(1 : 11))), each = n))
ggplot(DF, aes(x = delay)) + geom_histogram() + facet_wrap(~ rep, nrow = 3)
```



It is seen that the posterior simulations look similar to the observed data.
In fact, you don't need the closed form expression to simulate from the posterior predictive distribution. You can simulate θ from the posterior distribution, and simulate data from the likelihood given the simulated θ.