# 1 Summary

Let us assume $\theta$ is the parameter of interest and $\mathbf{y} = (y_1, \ldots, y_n)$ is the set of data points.

1. The posterior density of $\theta$ is $P(\theta \mid \mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})}$, where $P(\mathbf{y}) = \int P(\mathbf{y}|\theta)P(\theta)d\theta$.

2. Since $P(\mathbf{y})$ does not depend on $\theta$ (effect of $\theta$ is integrated out), we tend to write $P(\theta|\mathbf{y}) \propto P(\mathbf{y}|\theta)P(\theta)$. Note that $P(\mathbf{y}|\theta)P(\theta) = P(\mathbf{y}, \theta)$, the joint distribution of the data ($\mathbf{y}$) and parameter ($\theta$). Technically it is NOT the posterior probability.

3. Another term for $P(\mathbf{y}|\theta)$ is likelihood which is denoted as $L(\theta|\mathbf{y})$. Thus $L(\theta|\mathbf{y}) = P(\mathbf{y}|\theta) = P(y_1, \ldots, y_n|\theta)$. If the data points are independent, then $P(y_1, \ldots, y_n \mid \theta) = P(y_1 \mid \theta)P(y_2 \mid \theta) \ldots P(y_n \mid \theta) = \prod_{i=1}^{n} P(y_i \mid \theta)$. Hence, for independent data points $L(\theta|\mathbf{y}) = \prod_{i=1}^{n} P(y_i \mid \theta)$. The formula for $P(y_i \mid \theta)$ depends on distributional assumptions on the data whether they are normal or Poisson or binomial etc.

4. Here $P(\mathbf{y}) = \int P(\mathbf{y}|\theta)P(\theta)d\theta$ is called *prior predictive distribution* and $\int P(\tilde{\mathbf{y}}|\theta)P(\theta \mid \mathbf{y})d\theta$ is called posterior predictive distribution (PPD) of $\tilde{\mathbf{y}}$. In PPD, the prior probability $P(\theta)$ of *prior predictive distribution* is replaced by posterior probability (parameter given observed data $\mathbf{y}$) $P(\theta \mid \mathbf{y})$.

5. Bayesian inference is about drawing inferences of $\theta$ from $P(\theta \mid \mathbf{y})$. Now if the conditional distribution $P(\theta \mid \mathbf{y})$ belongs to any known class of probability distributions, you can directly use the summary statistics such as mean, median, mode, variance, etc. of that distribution. Some popular cases are: a) if $L(\theta \mid \mathbf{y}) = P(\mathbf{y}|\theta)$ is normal and $P(\theta)$ is also normal, we showed $P(\theta \mid \mathbf{y})$ is also normal, b) if $L(\theta \mid \mathbf{y}) = P(\mathbf{y}|\theta)$ is binomial and $P(\theta)$ is beta, we showed $P(\theta \mid \mathbf{y})$ is a beta distribution, c) if $L(\theta \mid \mathbf{y}) = P(\mathbf{y}|\theta)$ is Poisson and $P(\theta)$ is gamma, we showed $P(\theta \mid \mathbf{y})$ is a gamma distribution, etc.

6. Now, if $P(\theta \mid \mathbf{y})$ does not follow any known parametric distribution family. We then approximately learn the distribution $P(\theta \mid \mathbf{y})$ using samples $\theta_1, \ldots, \theta_K$ such that $\theta_i \sim P(\theta \mid \mathbf{y})$. To see that this is indeed valid.

```
theta <- rbeta(10000, 19, 133)

#Density of log(\theta) for page 40 of the book is given. Use that to compute density values in followin

thetagrid <- (1:1000)/1000#seq(range(theta)[1], range(theta)[2], length.out = 1000)
densijacobian <- (1/beta(19,133)) * exp(log(thetagrid))^19 * (1-exp(log(thetagrid)))^(133-1)

plot(density(log(theta)), col=1, type = 'l') #ploting density using Monte Carlo method,
                                             #just transformed the generated data in the
                                             #first line and computed numeric desity
points(log(thetagrid), densijacobian, col=2, type = 'l') #Jacobian computed densities

#Some standard distributions:

plot(density(theta), col=1, type = 'l') #ploting density using Monte Carlo method
```

```
                                        #(computed numeric density)
points(thetagrid, dbeta(thetagrid, 19, 133), col=2, type = 'l') #r function computed density


x <- rnorm(1000, 0, 1)
xgrid <- seq(-3,3,length.out = 1000)

plot(density(x), col=1, type = 'l') #ploting density using Monte Carlo method
                                    #(computed numeric density)
points(xgrid, dnorm(xgrid, 0, 1), col=2, type = 'l') #r function computed density
```

Above examples show numerically computed densities from samples are matching with exact densities.

7. Then we need to learn sampling techniques for "non-standard" probability distributions. By non-standard, I mean the ones for which there does not exist any R program to sample $\theta$ directly. For example if $L(\theta \mid \mathbf{y}) = P(\mathbf{y}|\theta)$ is Poisson and $P(\theta)$ is gamma, we showed $P(\theta \mid \mathbf{y})$ is a gamma distribution. So in this case, we can sample $\theta$ directly using `rgamma` in R. However, the prior $P(\theta)$ is changed to log-normal, we cannot use any standard functions.

8. To sample from non-standard densities, there are methods such as accept-reject which may be implemented using R package `AR`

```
#Example from HW 2
library("AR")
n <- 10 #number of trials
y <- 8  #number of success
thetagivenY <- function(theta){
  likelihood <- dbinom(y, n, theta)
  prior      <- dnorm(theta, 0.5, sd = 0.05)

  out <- likelihood * prior
  return(out)
}
#thetagivenY(theta) is upper bounded by theta^{y}(1-theta)^{n-y} ignoring
#constant. A good choice for instrument is Beta(y+1, n-y+1)

samples <- AR.Sim(n=300, thetagivenY, Y.dist = "beta", Y.dist.par = c(y+1,n-y+1))
plot(density(samples))


######Importance sample#####
J <- 100000
K <- 300 #Need to be K << J, K is the number of posterior samples you finally want
#Sample from the instrument distribution
betasamples <- rbeta(J, y+1, n+y-1)
#Importance weights
weights      <- thetagivenY(betasamples)/dbeta(betasamples, y+1,n+y-1)
```

```
#Generate K posterior samples using Importance sampling
postsamplesIS <- betasamples[sample(1:J, 300, prob = weights)]

points(density(postsamplesIS), col=2, type = 'l')
```

Samples from Accept/Reject and Importance sampling show the same numeric density. (I will provide one example on ARS using R package `Runuran`. There were some issues. ARS is very difficult to handle in general)

9. If there are more than one parameter, you need to draw joint samples. Say $\theta_1$ and $\theta_2$ are two parameters. Examples include normal distribution with mean and sigma both unknown, generalized linear models with at least 2 predictors. There are two ways. First is method of composition write $P(\theta_1, \theta_2 \mid \mathbf{y}) = P(\theta_1 \mid \theta_2, \mathbf{y})P(\theta_2 \mid \mathbf{y})$ or $P(\theta_1, \theta_2 \mid \mathbf{y}) = P(\theta_2 \mid \theta_1, \mathbf{y})P(\theta_1 \mid \mathbf{y})$. If you choose the first decomposition,

   1) Draw a sample of $\theta_2$ from $P(\theta_2 \mid \mathbf{y})$ (in this case, this conditional distribution cannot involve $\theta_1$ and need to be a valid distribution). If $P(\theta_2 \mid \mathbf{y})$ is "standard", use standard R packages to samples. Otherwise, use Accept/Reject, Importance sampling etc. Say the drawn sample be $\theta_2'$
   2) Draw a sample of $\theta_1$ form $P(\theta_1 \mid, \theta_2', \mathbf{y})$

   Repeat the above two steps for say 1000 times to draw 1000 "independent" samples of $(\theta_1, \theta_2)$.

10. Method of composition requires you compute the "marginal" posterior for one of the two parameters. When there are more than 1 parameter, then $P(\theta_2 \mid \mathbf{y})$, $P(\theta_1 \mid \mathbf{y})$ are called marginal posteriors. It may or may not be easy. It may or may not take a "standard" form. The marginal posterior may be "non-standard" which force you to use Accept/reject sampler etc. which are in general less efficient.

11. Jeffrey's prior: the Jeffrey's prior for $\theta$ is $P(\theta) \propto |J(\theta)|^{1/2}$, where $J(\theta) = -\mathbb{E}\left(\frac{d^2 \log P(\mathbf{y}|\theta)}{d\theta^2}\right)$.

12. Flat prior element corresponding to Jeffrey: Due to transformation of variable the prior for $\psi$, where $\psi = h(\theta)$ is $P(\psi) = P(\theta)\frac{d\theta}{d\psi}$. We will find the $\psi$ for which $P(\psi)$ is flat. To ensure that we need $P(\psi) = P(\theta)\frac{d\theta}{d\psi} = C$, we thus need $d\psi \propto P(\theta)d\theta \implies \psi \propto \int P(\theta)d\theta$. Thus, for this transformation $\psi \propto \int P(\theta)d\theta$, the prior $P(\psi)$ is flat.

13. In a Gibbs sample, we sample from full conditionals. Let $P(\theta_1, \theta_2 \mid \mathbf{y})$ be the joint posterior. Here samples of $\theta_1$ and $\theta_2$ are generated from corresponding full conditional distributions $P(\theta_1 \mid \theta_2, \mathbf{y})$ and $P(\theta_2 \mid \theta_1, \mathbf{y})$, respectively.

14. Note that $P(\theta_1, \theta_2 \mid \mathbf{y}) \neq P(\theta_1 \mid \theta_2, \mathbf{y})P(\theta_2 \mid \theta_1, \mathbf{y})$. Thus, the pairs of draws $(\theta_1, \theta_2)$ from each iteration are not independent draws.

15. Bayes factor for null to alternative is defined as $\text{BF}_{01} = \frac{P(\mathbf{y}|H_0)}{P(\mathbf{y}|H_1)}$. Let us assume null is indeed true. Then your Bayes factor computation method is consistent if $\text{BF}_{01}$ with increasing sample size i.e. the data vector $\mathbf{y}$ has more number of observations.

16. We can write $P(\mathbf{y}|H_0) = \int P(\mathbf{y}|\theta, H_0)P(\theta|H_0)d\theta$. Here $P(\mathbf{y}|\theta, H_0)$ is the likelihood under null hypothesis and $P(\theta|H_0)$ is the prior probability. Given $H_0$ part essentially put constraints on $\theta$. For example, if $H_0 : \theta \leq 0$, then $P(\theta|H_0)$ ensures that the prior is only supported in the range $(-\infty, 0]$. Similarly, $P(\mathbf{y}|H_1) = \int P(\mathbf{y}|\theta, H_1)P(\theta|H_1)d\theta$.

17. Bayes factor requires us to compute $P(\mathbf{y}|H_0)$ and $P(\mathbf{y}|H_1)$ which are difficult for complex Bayesian model. Approximate methods are available.

18. First one is the Monte Carlo method where you draw sample $\{\theta_{1,0}, \ldots, \theta_{K,0}\}$ from the prior $P(\theta|H_0)$ and draw $\{\theta_{1,1}, \ldots, \theta_{K,1}\}$ from $P(\theta|H_1)$. Then compute the likelihood for each sampled $\theta$ and take the average $\frac{1}{K}\sum_{i=1}^{K} P(\mathbf{y}|\theta_{i,j})$ as your estimate of $P(\mathbf{y}|H_j)$ for $j = 0, 1$. This estimate is not very good.

19. Another option is using the Harmonic Mean Identity which is more efficient than above Monte Carlo method. The identity essentially comes from importance sampling estimates of $P(\mathbf{y}|H_0)$ and $P(\mathbf{y}|H_1)$ based on the posterior samples of the underlying parameters. In this method $P(\mathbf{y}|H_j) = \left(\sum_{i=1}^{K} \frac{P(\mathbf{y}|\theta_{i,j})^{-1}}{K}\right)^{-1}$ where $\theta_{1,j}, \ldots, \theta_{K,j}$ are sampled from the posterior distribution $P(\theta|\mathbf{y}, H_j)$. The only difference lies in how samples of $\theta$ are generated. When sampled from posterior, it is more efficient as it used information from the data. [1]

# References


[1] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.