Bayesian Inference

Tutorial 4

1. (From May 2007 exam - number in brackets correspond to number of marks - the total was 50 marks.) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables such that $X_i \sim N(\mu, \omega^{-1})$ for $i = 1, \ldots, n$, where ω^{-1} is known.

   (a) Define Jeffreys' prior and suggest why we may wish to specify such a prior, in general. [2]

   (b) Show that Jeffreys' prior for μ is of the form, $p(\mu) \propto 1$. [3]

   (c) Hence show that the posterior distribution for μ is also Normal, where the posterior mean and variance should be specified. [3]

   (d) Suppose that we observe data $x_1, \ldots, x_{10}$ such that $\bar{x} = 10.1$. Assuming that $\omega = 1$, show that a 95% highest posterior density interval for μ is (9.480, 10.720). [4]

   (e) Now, suppose that a further observation, y, is independently generated from the same distribution as each X_i, so that,

   $$Y|\mu \sim N(\mu, \omega^{-1}).$$

   Calculate the posterior predictive distribution for Y. [5]

2. (a) Suppose an urn contains red and blue balls and we are interested in the proportion of red balls, θ. Our prior belief is that $\theta \sim Beta(9, 14)$. We extract 10 samples and 8 are red. We might ask whether the observed data is 'compatible' with the expressed prior distribution. One method is to calculate the predictive probability of observing such an extreme number of successes under this prior: this is a standard p-values but where the null hypothesis is a distribution. Using a `for` loop in `R` (and the `beta` function), find the probability of getting at least 8 red balls in 10 extractions. Are the data incompatible with the prior?

   (b) Suppose that 9 out of 10 experts suggest the $Beta(9, 14)$ prior but one expert disagrees and suggests instead that the proportion of red balls should have a prior distribution with mean 0.8 and standard deviation 0.1. What $Beta(a, b)$ distribution might represent the belief of this expert? Use the prior information from all the experts to set up a mixture prior and repeat the prior/data compatibility test. Use `R` to check whether the data are more compatible with this new prior.

   (*Note*: a mixture distribution has the form $f(x) = \sum_{i=1}^{k} w_k p_k(x)$ where $p_k(x)$ are pdfs or pmfs and w_k are the relative weights which add up to 1).

   (c) Use `R` to compute the posterior probability that θ is greater than 0.7.

3. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma)^2$, given μ and σ, where both μ and σ^2 are unknown. The following priors were specified:

   $$\mu \sim N(0, s^2); \qquad \text{and } \sigma \sim U[0, T],$$

   where T is "large".

(a) Using the transformation of variables formula, calculate the corresponding prior on σ^2.

(b) Calculate the posterior conditional distributions $f(\mu|\boldsymbol{x}, \sigma)$ and $f(\sigma^2|\boldsymbol{x}, \mu)$, after observing data $\boldsymbol{x} = (x_1, \ldots, x_n)$.

4. Consider parameters $\theta = \{\theta_1, \ldots, \theta_p\}$ (for $p \geq 2$), with posterior density function $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. We wish to obtain posterior summary statistics of interest using Monte Carlo integration. We have n sampled values of θ denoted $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^n$, such that $\boldsymbol{\theta}^i = \left\{\theta_1^i, \ldots, \theta_p^i\right\}$. Discuss how we would obtain estimates of the following summary statistics:

   (a) $\mathbb{E}_\pi(\theta_1)$;

   (b) $\text{Var}_\pi(\theta_1)$

   (c) 95% highest posterior density interval of θ_1 (you can assume that the distribution is unimodal);

   (d) $\mathbb{P}_\pi(\theta_1 > 0)$;

   (e) Posterior correlation between θ_1 and θ_2.

5. Use Monte Carlo integration in R to compute the following integral:

$$\int_{-1}^{1} 2\sqrt{1 - x^2} dx = \pi$$

   Run the code multiple times to see how the estimates improve with increasing sample size N (you could use $N = 10^i$ and `i=seq(2,6,by=0.05))` and the error $= |\pi - \hat{\pi}|$ decreases. Produce a log-log plot of the error as a function of N and show that the data can be fit to a straight line of slope -1/2.

Bayesian Inference

Tutorial 4: Solutions

1. (a) Jeffreys' prior for θ is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

where,

$$I(\theta|\boldsymbol{x}) = \mathbb{E}\left[\frac{d\log f(\boldsymbol{x}|\theta)}{d\theta}\right]^2 = -\mathbb{E}\left[\frac{d^2\log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

We may use Jeffrey's prior as as uninformative prior on the parameter. It is invariant to bijective transformations of the parameter.

(b) We have that,

$$\begin{aligned} f(x|\mu) &= \sqrt{\frac{\omega}{2\pi}}\exp\left(-\frac{\omega}{2}(x-\mu)^2\right) \\ \Rightarrow \log f(x|\mu) &= -\frac{1}{2}\log(2\pi/\omega) - \frac{\omega}{2}(x-\mu)^2 \\ \Rightarrow \frac{d\log f(x|\mu)}{d\mu} &= \omega(x-\mu) \\ \Rightarrow \frac{d^2\log f(x|\mu)}{d\mu^2} &= -\omega. \end{aligned}$$

Thus,

$$I(\mu|x) = -\mathbb{E}\left(-\omega\right) = \omega.$$

Then,

$$p(\mu) \propto \sqrt{\omega} \propto 1,$$

since ω is a known constant.

(c) The posterior distribution for μ is given by,

$$\begin{aligned} \pi(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu)p(\mu) \\ &= \prod_{i=1}^{n}\sqrt{\frac{\omega}{2\pi}}\exp\left(-\frac{\omega}{2}(x_i-\mu)^2)\right) \\ &\propto \exp\left(-\frac{\omega}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \exp\left(-\frac{\omega}{2}\left(\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}x_i\mu + n\mu^2\right)\right) \\ &\propto \exp\left(-\frac{\omega}{2}(-2n\mu\bar{x} + n\mu^2)\right) \\ &= \exp\left(-\frac{\omega n}{2}(-2\mu\bar{x} + \mu^2)\right) \\ &\propto \exp\left(-\frac{\omega n(\mu-\bar{x})^2}{2}\right) \\ \Rightarrow \mu|\boldsymbol{x} &\sim N\left(\bar{x}, \frac{1}{\omega n}\right). \end{aligned}$$

(d) Since the Normal distribution is symmetrical, the 95% highest posterior density interval will be equal to the 95% symmetrical credible interval. Now, consider,

$$Z = (\mu - \bar{x})\sqrt{\omega n} \sim N(0,1).$$

This implies that,

$$\mathbb{P}\left(-1.96 \leq (\mu - \bar{x})\sqrt{\omega n} \leq 1.96\right) = 0.95.$$

Rearranging this standard formula we obtain the 95% HPDI for μ:

$$\bar{x} \pm \frac{1.96}{\sqrt{\omega n}}.$$

Thus, for $\bar{x} = 10.1$, $\omega = 1$ and $n = 10$, we obtain the 95% HPDI of (9.480, 10.720).

(e) The posterior predictive distribution for y is given by,

$$\begin{aligned}
f(y|x) &= \int f(y|\mu)\pi(\mu|\boldsymbol{x})d\mu \\
&= \int \sqrt{\frac{\omega}{2\pi}} \exp\left(-\frac{\omega}{2}(y-\mu)^2\right) \times \sqrt{\frac{\omega n}{2\pi}} \exp\left(-\frac{\omega n}{2}(\mu - \bar{x})^2\right) d\mu \\
&\propto \int \exp\left(-\frac{\omega}{2}(y-\mu)^2 - \frac{\omega n}{2}(\mu-\bar{x})^2\right) d\mu \\
&\propto \exp(-\frac{\omega y^2}{2}) \int \exp\left(-\frac{\omega}{2}\mu^2 + \omega y \mu - \frac{\omega n}{2}\mu^2 + \omega n \mu \bar{x}\right) d\mu \\
&= \exp(-\frac{\omega y^2}{2}) \int \exp\left(-\frac{\omega}{2}\mu^2(n+1) + \mu(\omega y + \omega n \bar{x}\right) d\mu \\
&\propto \exp\left(-\frac{\omega y^2}{2}\right) \int \exp\left\{\left(-\frac{\omega(n+1)}{2}\right)\left(\mu - \frac{\omega y + \omega n \bar{x}}{\omega(n+1)}\right)^2 + \frac{(\omega y + \omega n \bar{x})^2}{2\omega(n+1)}\right\} d\mu \\
&\propto \exp\left(-\frac{\omega y^2}{2}\right) \exp\left(\frac{\omega(y+n\bar{x})^2}{2(n+1)}\right) \\
&\propto \exp\left(-\frac{\omega}{2}\frac{n}{n+1}(y^2 - 2y\bar{x})\right) \\
&\propto \exp\left(-\frac{\omega(y-\bar{x})^2}{2(1+1/n)}\right) \\
\Rightarrow Y &\sim N\left(\bar{x}, \frac{1}{\omega}\left(1 + \frac{1}{n}\right)\right).
\end{aligned}$$

Alternatively, we could note that the joint pdf of Y and μ, given $\boldsymbol{x}$, is,

$$f(y|\mu)\pi(\mu|\boldsymbol{x}) = \sqrt{\frac{\omega}{2\pi}} \exp\left(-\frac{\omega}{2}(y-\mu)^2\right) \times \sqrt{\frac{\omega n}{2\pi}} \exp\left(-\frac{\omega n}{2}(\mu-\bar{x})^2\right)$$

Then, let $Z = Y - \mu$. The corresponding joint pdf of Z and μ, given $\boldsymbol{x}$, is,

$$f(z|\mu)\pi(\mu|\boldsymbol{x}) = \sqrt{\frac{\omega}{2\pi}} \exp\left(-\frac{\omega z^2}{2}\right) \times \sqrt{\frac{\omega n}{2\pi}} \exp\left(-\frac{\omega n}{2}(\mu-\bar{x})^2\right)$$

So that Z and μ are independent, where $Z \sim N(0, 1/\omega)$ and $\mu \sim N(\bar{x}, 1/(\omega n))$. Hence,

$$Y = Z + \mu \sim N\left(\bar{x}, \frac{1}{\omega}\left(1 + \frac{1}{n}\right)\right),$$

using the standard result of the sum of two independent Normal random variables.

2. (a) We start by calculating the prior predictive distribution

$$
\begin{aligned}
f(y) &= \int_{\Theta} f(y \mid \theta)p(\theta)\mathrm{d}\theta \\
&= \int_0^1 \left(\begin{array}{c} n \\ y \end{array}\right) \theta^y(1-\theta)^{n-y}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\ \mathrm{d}\theta \\
&= \left(\begin{array}{c} n \\ y \end{array}\right)\frac{1}{B(a,b)}\int_0^1 \theta^{a+y-1}(1-\theta)^{b+n-y-1}\ \mathrm{d}\theta \\
&= \left(\begin{array}{c} n \\ y \end{array}\right)\frac{B(a+y,b+n-y)}{B(a,b)}
\end{aligned}
$$

The prior predictive probability of observing at least 8 positive responses can then be computed from the last expression using, e.g., a simple loop in R.

```
result0=0
for(i in 8:10){
result=result0+choose(10,i)*beta(9+i,14+10-i)/beta(9,14)
result0=result
}
print(result)
```

and it is 0.0269. This suggests some evidence that the data and the prior are incompatible.

(b) Solving for a and b gives a $Beta(12,3)$ prior. We compute again the prior predictive distribution using the mixture prior:

$$
\begin{aligned}
&f(y) = \int_{\Theta} f(y \mid \theta)p(\theta)\mathrm{d}\theta \\
&= \int_0^1 \left(\begin{array}{c} n \\ y \end{array}\right)\theta^y(1-\theta)^{n-y}\left\{\pi\frac{1}{B(a_1,b_1)}\theta^{a_1-1}(1-\theta)^{b_1-1} + (1-\pi)\frac{1}{B(a_2,b_2)}\theta^{a_2-1}(1-\theta)^{b_2-1}\right\}\mathrm{d}\theta \\
&= \pi\left(\begin{array}{c} n \\ y \end{array}\right)\frac{1}{B(a_1,b_1)}\int_0^1 \theta^{a_1+y-1}(1-\theta)^{b_1+n-y-1}\ \mathrm{d}\theta + \\
&(1-\pi)\left(\begin{array}{c} n \\ y \end{array}\right)\frac{1}{B(a_2,b_2)}\int_0^1 \theta^{a_2+y-1}(1-\theta)^{b_2+n-y-1}\ \mathrm{d}\theta \\
&= \pi\left(\begin{array}{c} n \\ y \end{array}\right)\frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)} + (1-\pi)\left(\begin{array}{c} n \\ y \end{array}\right)\frac{B(a_2+y,b_2+n-y)}{B(a_2,b_2)}
\end{aligned}
$$

The prior predictive probability of observing at least 8 red balls is now 0.091 (by updating the R function in (a)), which does not provide strong evidence of incompatibility.

(c) We will start by finding the posterior distribution of θ.

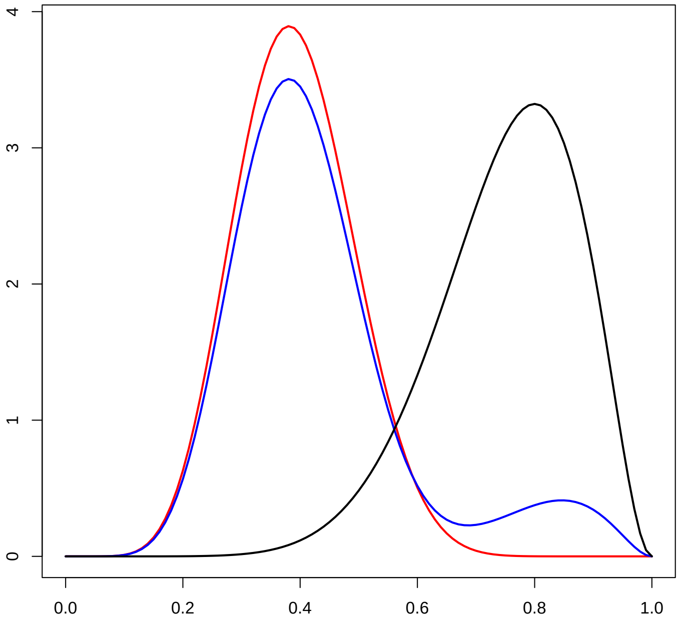

Figure 1: Beta prior (red), mixture prior (blue) and (scaled) likelihood (black)

$$
\begin{aligned}
p(\theta \mid y) \propto & \binom{n}{y} \theta^{y}(1-\theta)^{n-y}\left\{\pi \frac{1}{B\left(a_{1}, b_{1}\right)} \theta^{a_{1}-1}(1-\theta)^{b_{1}-1}+(1-\pi) \frac{1}{B\left(a_{2}, b_{2}\right)} \theta^{a_{2}-1}(1-\theta)^{b_{2}-1}\right\} \\
& \propto \pi \frac{1}{B\left(a_{1}, b_{1}\right)} \theta^{a_{1}+y-1}(1-\theta)^{b_{1}+n-y-1}+(1-\pi) \frac{1}{B\left(a_{2}, b_{2}\right)} \theta^{a_{2}+y-1}(1-\theta)^{b_{2}+n-y-1} \\
& =\pi \frac{B\left(a_{1}+y, b_{1}+n-y\right)}{B\left(a_{1}, b_{1}\right)} \frac{1}{B\left(a_{1}+y, b_{1}+n-y\right)} \theta^{a_{1}+y-1}(1-\theta)^{b_{1}+n-y-1} \\
& +(1-\pi) \frac{B\left(a_{2}+y, b_{2}+n-y\right)}{B\left(a_{2}, b_{2}\right)} \frac{1}{B\left(a_{2}+y, b_{2}+n-y\right)} \theta^{a_{2}+y-1}(1-\theta)^{b_{2}+n-y-1} \\
= & \pi \frac{B\left(a_{1}+y, b_{1}+n-y\right)}{B\left(a_{1}, b_{1}\right)} \operatorname{Beta}\left(\theta \mid a_{1}+y, b_{1}+n-y\right) \\
& +(1-\pi) \frac{B\left(a_{2}+y, b_{2}+n-y\right)}{B\left(a_{2}, b_{2}\right)} \operatorname{Beta}\left(\theta \mid a_{2}+y, b_{2}+n-y\right) \\
& =\omega_{1} \operatorname{Beta}\left(\theta \mid a_{1}+y, b_{1}+n-y\right)+\left(1-\omega_{1}\right) \operatorname{Beta}\left(\theta \mid a_{2}+y, b_{2}+n-y\right)
\end{aligned}
$$

with $\omega_1 = \pi \frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)} \left(\pi \frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)} + (1-\pi)\frac{B(a_2+y,b_2+n-y)}{B(a_2,b_2)}\right)^{-1}$
We are now ready to compute the required probability (see R script), which turned out to be 0.4959

```
w1=(0.9*beta(9+8,14+10-8)/beta(9,14))*((0.9*beta(9+8,14+10-8)/beta(9,14))+
(0.1*beta(12+8,3+10-8)/beta(12,3)))^(-1)
#
w2=(1-w1)
#
cat("Pr(theta> 0.7)=",w1*(1-pbeta(0.7,9+8,14+10-8))+
w2*(1-pbeta(0.7,12+8,3+10-8)),"\n")
```

3. (a) According to the transformation of variables formula, for $Y = g(X)$, and for probability functions f_X and f_Y,

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|.$$

Here, $g(\sigma) = \sigma^2$, i.e. $\sigma = g^{-1}(\sigma^2) = \sqrt{\sigma^2}$. Then, we have that (for $\sigma^2 \leq T^2$),

$$\begin{aligned} p_{\sigma^2}(\sigma^2) &= p_\sigma(\sqrt{\sigma^2}) \left| \frac{d\sigma}{d\sigma^2} \right| \\ &= p_\sigma(\sqrt{\sigma^2}) \left| \frac{d(\sigma^2)^{1/2}}{d\sigma^2} \right| \\ &= \frac{1}{T} \left| \frac{1}{2}(\sigma^2)^{-1/2} \right| \\ &= \frac{1}{2T}\frac{1}{\sigma}. \end{aligned}$$

Thus, the prior on σ^2 is of the form,

$$p(\sigma^2) = \begin{cases} \frac{1}{2T}\frac{1}{\sigma} & 0 < \sigma^2 \leq T^2 \\ 0 & \text{otherwise} \end{cases}$$

Note that,

$$\int_0^{T^2} \frac{1}{2T}\frac{1}{\sigma} d\sigma^2 = \int_0^{T^2} \frac{1}{2T}\frac{1}{\sigma} 2\sigma d\sigma = \int_0^T \frac{1}{T} d\sigma = 1.$$

Note also that, in practice, T is set very high to ensure that $\sigma > T$ essentially has probability equal to 0.

(b) To calculate the posterior conditional distributions, we typically calculate the joint posterior distribution and then the corresponding conditional distributions of interest by considering the single parameter of interest, and all other parameters as "fixed". For $-\infty < \mu < \infty$ and $\sigma^2 \in (0, T^2]$, the joint posterior distribution is,

$$\pi(\mu, \sigma^2|\boldsymbol{x}) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{\mu^2}{2s^2}\right) \frac{1}{2T}\frac{1}{\sigma}.$$

Consider each parameter in turn. We begin with μ. Since we condition on σ^2, we consider only the terms that contain μ and derive,

$$\mu|\sigma^2, \boldsymbol{x} \sim N\left(\frac{s^2 n\bar{x}}{ns^2+\sigma^2}, \frac{\sigma^2 s^2}{s^2 n + \sigma^2}\right)$$

(setting $\phi = 0$, $\tau^2 = s^2$ in the relevant result from the lecture notes). The posterior conditional distribution of σ^2 is given by,

$$\begin{aligned} \pi(\sigma^2|\mu, \boldsymbol{x}) &\propto (\sigma^2)^{-(n/2+1/2)} \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right) I(\sigma^2 < T^2) \\ \Rightarrow \quad \sigma^2|\mu, \boldsymbol{x} &\sim \Gamma^{-1}\left(\frac{n}{2} - \frac{1}{2}, \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right) I(\sigma^2 < T^2). \end{aligned}$$

In other words $\sigma^2|\mu, \boldsymbol{x}$ could be regarded as a truncated inverse Gamma distribution (truncated so that $\sigma^2 \leq T^2$).

4. In order to obtain Monte Carlo estimates of the marginal distribution of θ_1 we simply look at the sampled values of θ_1 (ignoring the associated values of the other parameters).

   (a) Use $\mathbb{E}_\pi(\theta_1) \approx \frac{1}{n}\sum_{i=1}^{n}\theta_1^i$ (i.e. sample mean).

   (b) Use

$$\mathrm{Var}_\pi(\theta_i) \approx \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_i - \mathbb{E}_\pi(\theta_1)\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\theta_1^i\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\theta_1^i\right)^2\right)$$

   (i.e. sample variance). Note that there are ways of writing this e.g. $\mathrm{Var}_\pi(\theta_i) = \frac{n}{n-1}\left(\mathbb{E}_\pi(\theta_1)^2 - \mathbb{E}_\pi\left(\theta_1^2\right)\right)$ where $\mathbb{E}_\pi(\theta_1)$ is given above and $\mathbb{E}_\pi\left(\theta_1^2\right) \approx \frac{1}{n}\sum_{i=1}^{n}\left(\theta_1^i\right)^2$.

   (c) Calculate the width of each possible 95% credible interval of θ_i and take the interval with the shortest width. In other words take the difference between the lower $j\%$ quantile and upper $95 + j\%$ quantile of θ_i for $j \in [0, 5]$, and take the shortest interval to be the 95% HPDI. This can be done by writing a function in R and using the quantile function, for say $j = 0.01, 0.02, \dots, 4.99$.

   (d) Use $\mathbb{P}_\pi(\theta_1 > 0) \approx \frac{1}{n}\sum_{i=1}^{n} I(\theta_1 > 0)$, where $I(\cdot)$ denotes the indicator function (i.e. take the proportion of simulated values of θ_1 that are greater than 0 .

   (e) Use

$$\mathrm{corr}_\pi(\theta_1, \theta_2) \approx \frac{\sum_{i=1}^{n}\left(\theta_1^i\theta_2^i\right) - \frac{1}{n}\left(\sum_{i=1}^{n}\theta_1^i\theta_2^i\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(\theta_1^i\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\theta_1^i\right)^2\right)}\sqrt{\left(\sum_{i=1}^{n}\left(\theta_2^i\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\theta_2^i\right)^2\right)}}$$

   (i.e. sample correlation).

5. See code and plots below. You can see from the right panel that the rate of convergence of Monte Carlo integration is in fact $\sqrt{n}$. This means that, to double the precision of a certain estimate, you will need 4 times the original sample! However, MC integration can be easily extended to multiple dimensions and the convergence rate for Monte Carlo is independent of the number of dimensions in the integral (which is not the case for many other standard integration techniques)

```
ex5= function(N){
samp=runif(10^N,-1,1)
f_x=sqrt(1-samp^2)
estim=(2*2/10^N)*sum(f_x)
err=abs(pi-estimate)
estim<<-estim
err<<-err
}
i=0
estimate=samples=error=NULL
#
for(N in seq(2,6,by=0.05)){
i=i+1
set.seed(i+2)
ex5(N)
estimate[i]=estim
samples[i]=10^N
error[i]=err
}
#
plot(samples, sqrt(samples),type="l",col="red",lwd=2,log="xy",
ylim=c(10^-5,10^0),xlab="Samples",ylab="Error")
lines(samples,error,col="blue",lwd=2)
legend("topright", legend=c("log_y=-1/2*log_x"),
col=c("red"), lty=1, lwd=2, cex=1.2, box.lty=0)
```

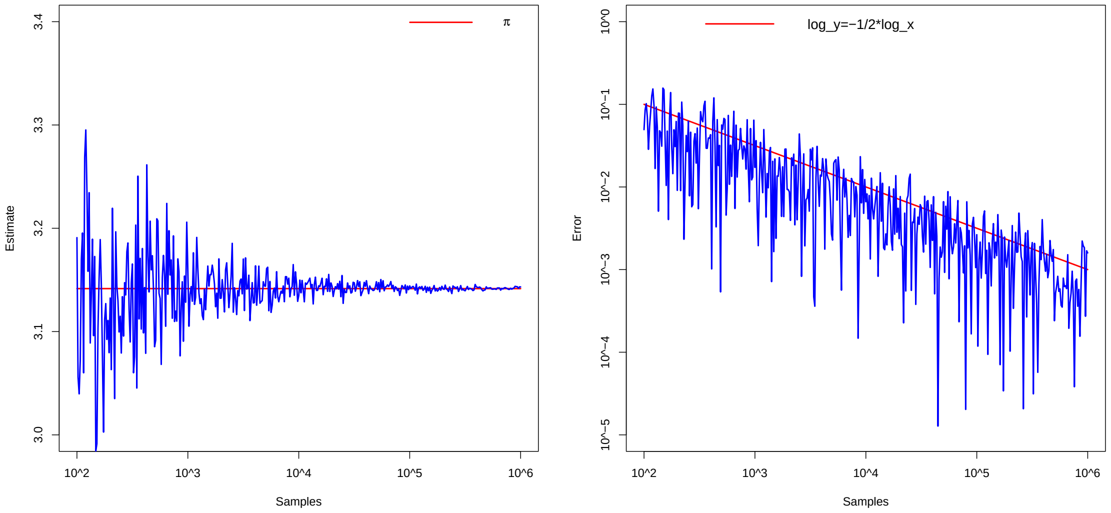

Figure 2: MC convergence to true value π (left) and log-log plot of the error (right)