Bayesian

Bill Last Updated:

24 January, 2020

Contents

Preface: Motivation

All the notes I have done here are about bayesian. While I have tried my best, probably there are still some typos and errors. Please feel free to let me know in case you find one. Thank you!

Chapter 1

Bayesian - 1

The following is the part of the class note that I took from the online course of "Bayesian Statistics: From Concept to Data Analysis." (https://www.coursera.org/learn/bayesian-statistics/home/welcome)

Important note: All the notes here are just for my own study purpose. I do not clain any copyright. You can use it for study purpose as well, but not for any business purposes.

1.1 Frequentist perspective

$$\theta = \{fair, loaded\}$$

$$x \sim Bin(5, \theta)$$

$$f(x|\theta) = \begin{cases} \binom{5}{x}(\frac{1}{2})^5 & if\ \theta = fair \\ \binom{5}{x}(0.7)^x(0.3)^{5-x} & if\ \theta = loaded \end{cases}$$

$$= \binom{5}{x}(\frac{1}{2})^5 I_{\{\theta=fair\}} + \binom{5}{x}(0.7)^x(0.3)^{5-x} I_{\{\theta=loaded\}}$$

When $x = 2$

$$f(\theta|x=2) = \begin{cases} \binom{5}{x}(\frac{1}{2})^5 = 0.3125 & if\ \theta = fair \\ \binom{5}{x}(0.7)^x(0.3)^{5-x} = 0.1323 & if\ \theta = loaded \end{cases}$$

Thus, based on MLE, it suggests that it should be "fair", since it has a greater probablity if we observe 2 head out of 5 trials.

However, we can not know the following probability: given that we observe $x = 2$, what is the probability that θ is fair?

$$P(\theta = fair|X = 2)$$

From the frequentist's perspective, the coin is the fixed coin. And thus, the probablity of $P(\theta = fair|x = 2)$ is equal to $P(\theta = fair)$.

$$P(\theta = fair|x = 2) = P(\theta = fair)$$

As,

$$P(\theta = fair) \in C(0, 1)(i.e., either\ 0\ or\ 1)$$

1.2 Bayesian perspective

Prior $P(loaded) = 0.6$

$$\begin{aligned} f(\theta|X) &= \frac{f(x|\theta)f(\theta)}{\sum_\theta f(x|\theta)f(\theta)} \\ &= \frac{\binom{5}{x}[(\frac{1}{2})^5 \times 0.4 \times I_{\{\theta=fair\}} + (0.7)^x(0.3)^{5-x} \times 0.6 \times I_{\{\theta=loaded\}}]}{\binom{5}{x}[(\frac{1}{2})^5 \times 0.4 + (0.7)^x(0.3)^{5-x} \times 0.6]} \end{aligned}$$

$$\begin{aligned} f(\theta|X = 2) &= \frac{0.0125I_{\{\theta=fair\}} + 0.0079I_{\{\theta=loaded\}}}{0.0125 + 0.0079} \\ &= 0.612I_{\{\theta=fair\}} + 0.388I_{\{\theta=loaded\}} \end{aligned}$$

Thus, we can say that:

$$P(\theta = loaded|X = 2) = 0.388$$

We can change the prior, and get different posterior probabilities:

$$P(\theta = loaded) = \frac{1}{2} \rightarrow P(\theta = loaded|X = 2) = 0.297$$

$$P(\theta = loaded) = \frac{9}{10} \rightarrow P(\theta = loaded|X = 2) = 0.792$$

1.3 Continous parameters

In the examples above, θ is discrete. In contrast, the examples below use continous θ.

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta} = \frac{likelihood \times prior}{normalizing - constant} \propto likelihood \times prior$$

Note that, the posterior is a PDF of θ, which is not in the function of $f(y)$. Thus, removing the denominator (i.e., the normalizing constant) does not change the form of the posterior.

1.3.1 Uniform

Suppose that θ is the probablity of a coin getting head. We could assign a uniform distribution.

$$\theta \sim U[0, 1]$$

$$f(\theta) = I_{\{0 \leqq \theta \leqslant 1\}}$$

(It is interesting to see how to write the pdf for uniform distribution.)

$$f(\theta|Y = 1) = \frac{\theta^1(1-\theta)^0 I_{\{0 \leqq \theta \leqslant 1\}}}{\int_{-\infty}^{+\infty} \theta^1(1-\theta)^0 I_{\{0 \leqq \theta \leqslant 1\}} d\theta} = \frac{\theta I_{\{0 \leqq \theta \leqslant 1\}}}{\int_0^1 \theta d\theta} = 2\theta I_{\{0 \leqq \theta \leqslant 1\}}$$

If we ignore the normalizing constant, we will get

$$f(\theta|Y = 1) \propto \theta^1(1-\theta)^0 I_{\{0 \leqq \theta \leqslant 1\}} = \theta I_{\{0 \leqq \theta \leqslant 1\}}$$

Thus, we can see that with vs. without the noramlizing constant is the "2".

1.3.2 Uniform: prior versus posterior

When θ follows uniform distribution:

Prior

$$P(0.025 < \theta < 0.975) = 0.95$$
$$P(0.05 < \theta) = 0.95$$

Posterior

$$P(0.025 < \theta < 0.975) = \int_{0.025}^{0.975} 2\theta d\theta = 0.95$$

$$P(0.05 < \theta) = 1 - P(\theta < 0.05) = \int_{0}^{0.05} 2\theta d\theta = 1 - 0.05^2 = 0.9975$$

Thus, we can see that, while $P(0.025 < \theta < 0.975)$ is the same for prior and posterior, $P(0.05 < \theta)$ is not the same.

1.3.3 Uniform: equal tailed versus HPD

Equal tailed

$$P(\theta < q|Y = 1) = \int_{0}^{q} 2\theta d\theta = q^2$$

$$P(\sqrt{0.025} < \theta < \sqrt{0.975}|Y = 1) = P(0.158 < \theta < 0.987) = 0.95$$

We can say that: there's a 95% probability that θ is in between 0.158 and 0.987.

Highest Posterior Density

$$P(\theta > \sqrt{0.05}|Y = 1) = P(\theta > 0.224|Y = 1) = 0.95$$

1.4 Bernoulli/binomial likelihood with uniform prior

$$\begin{aligned} f(\theta|Y = 1) &= \frac{\theta^{\sum y_i}(1-\theta)^{\sum n-y_i} I_{\{0 \leqq \theta \leqslant 1\}}}{\int_{-\infty}^{+\infty} \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} I_{\{0 \leqq \theta \leqslant 1\}} d\theta} \\ &= \frac{\theta^{\sum y_i}(1-\theta)^{\sum n-y_i} I_{\{0 \leqq \theta \leqslant 1\}}}{\frac{\Gamma(\sum y_i+1)\Gamma(n-\sum y_i+1)}{\Gamma(n+2)} \int_{-\infty}^{+\infty} \frac{\Gamma(n+2)}{\Gamma(\sum y_i+1)\Gamma(n-\sum y_i+1)} \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} I_{\{0 \leqq \theta \leqslant 1\}} d\theta} \\ &= \frac{\Gamma(n+2)}{\Gamma(\sum y_i+1)\Gamma(n-\sum y_i+1)} \theta^{\sum y_i}(1-\theta)^{\sum n-y_i} I_{\{0 \leqq \theta \leqslant 1\}} \end{aligned}$$

Thus,

$$\theta|y \sim Beta(\sum y_i + 1, n - \sum y_i + 1)$$

Side note: $Beta(1,1) = Uniform(0,1)$:

$Beta(\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} I_{\{0 \leqq x \leqslant 1\}}$

Thus, we can get the following since the support for beta distribution is $[0,1]$:

$Beta(1,1) = \frac{x^{0}(1-x)^{0}}{B(\alpha,\beta)} = 1 \times I_{\{0 \leqq x \leqslant 1\}}$

1.5 Conjugate priors

As noted above, beta prior (or, Uniform) leads to beta posterior. In a more general sense, Beta prior always leads to beta posterior.

For instance,

$$\begin{aligned} f(\theta|y) \propto f(y|\theta)f(\theta) &= \theta^{\sum y_i}(1-\theta)^{\sum n-y_i}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} I_{\{0 \leqq \theta \leqslant 1\}} \\ &= \frac{1}{B(\alpha,\beta)}\theta^{\sum y_i+\alpha-1}(1-\theta)^{\sum n-y_i+\beta-1} I_{\{0 \leqq \theta \leqslant 1\}} \\ &\propto \theta^{\sum y_i+\alpha-1}(1-\theta)^{\sum n-y_i+\beta-1} I_{\{0 \leqq \theta \leqslant 1\}} \end{aligned}$$

Thus,

$$f(\theta|y) \sim Beta(\alpha + \sum y_i, \beta + \sum n - y_i)$$

Conjugate prior: prior and posterior share the same distribution. As we can see, both the data and the prior contribute to the posterior.

For the prior of $Beta(\alpha,\beta)$, the mean is

$$Mean_{prior} = \frac{\alpha}{\alpha+\beta}$$

Posterior mean is,

$$\begin{aligned} Mean_{posterior} &= \frac{\alpha + \sum y_i}{\alpha + \sum y_i + \beta + n - \sum y_i} \\ &= \frac{\alpha + \sum y_i}{\alpha + \beta + n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{\sum y_i}{n} \\ &= Weight_{prior} \times Mean_{prior} + Weight_{data} \times Mean_{data} \end{aligned}$$

Side Note

(1) Binomial proportion confidence interval:

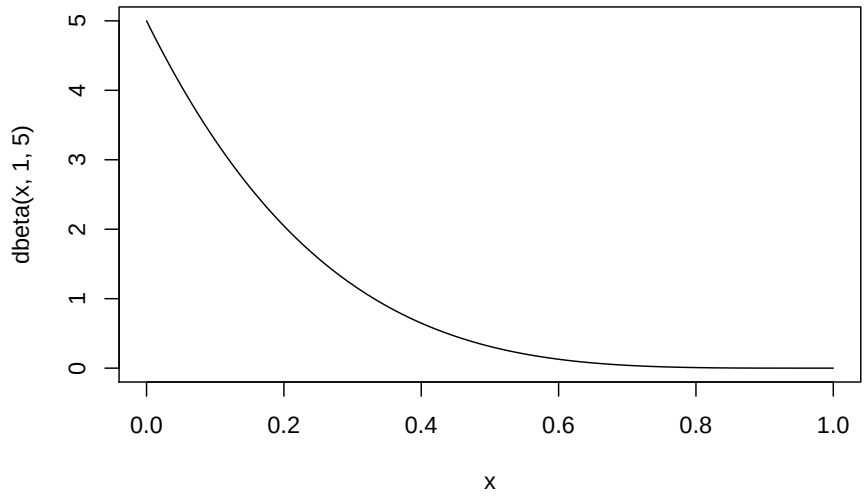
$$\hat{p} + / - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(2) Mean of Beta distribution:

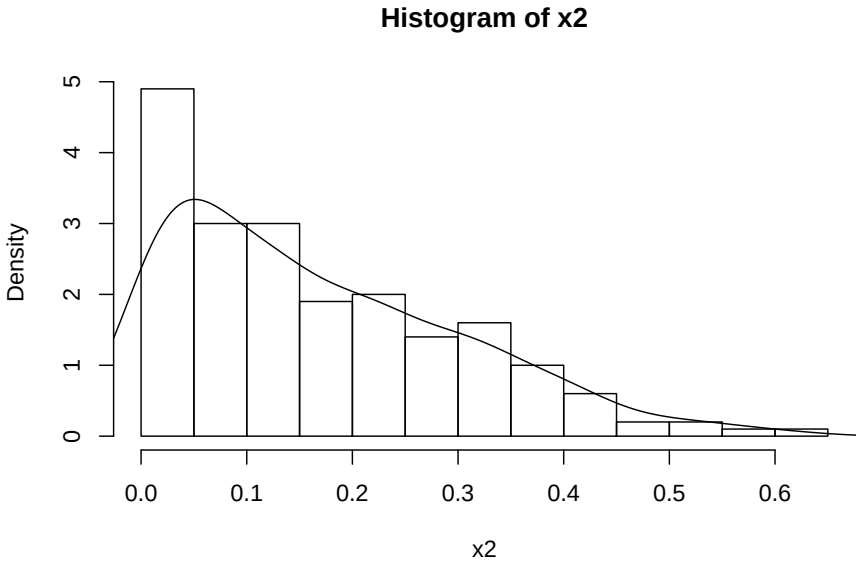
$$\frac{\alpha}{\alpha+\beta}$$

(3) Plot of Beta distribution

```
# Method 1
x<-seq(0,1,length=200)
plot(x,dbeta(x,1,5),type = "l")
```




```
# Method 2
x2<-rbeta(200,1,5)
hist(x2,prob = TRUE)
lines(density(x2))
```



1.6 Poisson distribution

Pmf of Poisson distribution:

$$Pois(\lambda) \sim \frac{\lambda^k e^{-\lambda}}{k!}$$

We can replace k with the notation of y, and assume that we observe n y_i:

$$y_i \sim \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

$$f(y|\lambda) = \frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}$$

We assume that λ follows Gamma distribution (i.e.,Gamma prior):

$$\lambda \sim \Gamma(\alpha, \beta)$$

The pdf for Gamma distribution is:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Thus, the posterior is as follows:

$$\begin{aligned} f(\lambda|y) &\propto f(y|\lambda)f(\lambda) \\ &= \frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &\propto \lambda^{\sum y_i} e^{-n\lambda} \times \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^{(\alpha+\sum y_i)-1} e^{-(\beta+n)\lambda} \end{aligned}$$

Thus, the posterior is:

$$\Gamma(\alpha + \sum y_i, \beta + n)$$

As we know that, the mean of prior for Gamma is $\frac{\alpha}{\beta}$. Thus, we can get the mean for the posterior for Gamma is:

$$\begin{aligned} &= \frac{\alpha + \sum y_i}{\beta + n} \\ &= \frac{\beta}{\beta + n} \frac{\alpha}{\beta} + \frac{n}{\beta + n} \frac{\sum y_i}{n} \end{aligned}$$

To determine the prior of α and β:

(1) Prior mean $\frac{\alpha}{\beta}$

(a) Prior std. dev. $\frac{\sqrt{\alpha}}{\beta}$
(b) Effective sample size β

(2) Vague prior Small $\varepsilon > 0$: $\Gamma(\varepsilon, \varepsilon)$. Thus, the posterior mean is primarily driven by the data:

$$\frac{\varepsilon + \sum y_i}{\varepsilon + n} \approx \frac{\sum y_i}{n}$$

(1) As we know, beta prior lead the Bernoulli trial to a beta posterior. That is, we know $f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$. What is the prior predictive distribution of $f(y)$?

(2) If Beta is Beta (3,3), what is the prior predictive probablity that we wil observe $y = 0$ in the next trial?

1.7 Exponential data

For instance, suppose that on average you need to wait for 10 minutes for a fast food delivery, and thus we can assume that $y \sim Exp(\lambda)$. Furthermore, we assume that the prior λ follows Gamma distribution $Gamma(\alpha, \beta)$, thus it is with a mean of $\frac{\alpha}{\beta} = \frac{1}{10}$.

$$if \;\; \Gamma(100, 1000)$$

(Note that, it has a mean of $\frac{100}{1000} = \frac{1}{10}$).

Thus, we can get:

$$\begin{aligned} f(\lambda|y) &\propto f(y|\lambda)f(\lambda) \\ &\propto \lambda e^{-\lambda y}\lambda^{\alpha-1}e^{-\beta\lambda} \\ &\propto \lambda^{(\alpha+1)-1}e^{-(\beta+y)\lambda} \end{aligned}$$

Thus, we get

$$\lambda|y \sim \Gamma(\alpha+1, \beta+y)$$

Thus, if we observe a data point that we need to wait for 12 minutes for a fast food delivery, we can update the posterior:

$$\lambda|y \sim \Gamma(101, 1012)$$

Thus, the mean for the posterior is

$$\frac{101}{1012} = \frac{1}{10.02}$$

Note:

(1) Typically, we know that the pdf of Gamma is $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$. We replace x with λ since now the random variable of x is to represent the parameter λ in the exponential distribution.

(2) In the above, we drop the constant part ($\frac{\beta^\alpha}{\Gamma(\alpha)}$) in the Gamma distribution as long as it does not include x (i.e., λ).

(3) Suppose that you have 4 observations in total, then

$$\begin{aligned} f(\lambda|y) &\propto f(y_1|\lambda)f(y_2|\lambda)f(y_3|\lambda)f(y_4|\lambda)f(\lambda) \\ &\propto \lambda e^{-\lambda y_1}\lambda e^{-\lambda y_2}\lambda e^{-\lambda y_3}\lambda e^{-\lambda y_4}\lambda^{\alpha-1}e^{-\beta\lambda} \\ &\propto \lambda^{(\alpha+4)-1}e^{-(\beta+\sum_{i=1}^4 y_i)\lambda} \end{aligned}$$

Thus, the generalized form is as follows:

$$\lambda^{(\alpha+n)-1} e^{-(\beta+\sum_{i=1}^{n} y_i)\lambda}$$

1.8 Normal likelihood

1.8.1 When variance is known

$$x_i \sim N(\mu, \sigma_0^2)$$

(σ_0 is assumed to be known. Thus, the only unknown parameter is μ.)

The conjugate prior for normal distribution is normal distribution itself.

$$f(\mu|x) \sim f(x|\mu)f(\mu)$$

Assume that

$$\mu \sim N(m_0, s_0^2)$$

$$\mu|x \sim N(\frac{\frac{n\bar{x}}{\sigma_o^2} + \frac{m_o}{S_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{s_0^2}}, \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{s_0^2}})$$

Thus, where

$$\begin{aligned} \frac{\frac{n\bar{x}}{\sigma_o^2} + \frac{m_o}{S_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{s_0^2}} &= \frac{\frac{n\bar{x}}{\sigma_o^2}}{\frac{n}{\sigma_0^2} + \frac{1}{s_0^2}} + \frac{\frac{m_o}{S_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{s_0^2}} \\ &= \frac{n}{n + \frac{\sigma_0^2}{s_0^2}}\bar{x} + \frac{\frac{\sigma_0^2}{S_0^2}}{n + \frac{\sigma_0^2}{s_0^2}} m_o \end{aligned}$$

Note:

(1) As we can see, the posterior mean is a weighted mean – a combination of prior mean and sample mean.

(2) When n is larger, the sample mean $\bar{x}$ gets more weight.

(3) For the prior mean m_0, the smaller the prior variance s_0^2 is, the prior mean gets more weight. If the prior variance s_0^2 is big, the prior mean will get less weight in the final posterior mean.

1.8.2 When variance is unknown

$$x_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$$

$$\mu|\sigma^2 \sim N(m,\frac{\sigma^2}{w})$$

Side note

$w = \frac{\sigma^2}{\sigma_\mu^2}$ effective sample size

$$\sigma^2 \sim \Gamma^{-1}(\alpha,\beta)$$

Thus, we can get that,

$$\sigma^2|x \sim \Gamma^{-1}(\alpha+\frac{n}{2},\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{nw}{2(n+w)}(\bar{x}-m)^2)$$

$$\mu|\sigma^2,x \sim N(\frac{n\bar{x}+wm}{n+w},\frac{\sigma^2}{n+w})$$

Where,

$\frac{n\bar{x}+wm}{n+w} = \frac{w}{n+w}m + \frac{n}{n+w}\bar{x}$

$$\mu|x \sim t-distribution$$

1.9 Non-informative priors

1.9.1 Bernoulli

$Y_i \sim B(\theta)$

$$\theta \sim U[0,1] = Beta(1,1)$$

(Effective sample size is 1+1=2)

If we get $Beta(\frac{1}{2},\frac{1}{2})$ and $Beta(0.001,0.001)$ have less impact on the posterior.

Improper prior, for instance, THe prior

$$Beta(0,0)$$

$$f(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

In this case,

$$f(\theta|y) \propto \theta^{y-1}(1-\theta)^{n-y-1} \sim Beta(y, n-y)$$

Posterior mean: $\frac{y}{n} = \hat{\theta}$

1.9.2 Gaussian

$$Y_i \sim N(\mu, \sigma^2)$$

Vague prior:

$$\mu \sim N(0, 1000000^2)$$

or,

$$f(\mu) \sim 1$$

$$\begin{aligned} f(\mu|y) &\propto f(y|\mu)f(\mu) \\ &\propto exp(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2) \times 1 \\ &\propto exp(-\frac{1}{2\frac{\sigma^2}{n}}\sum(y_i-\bar{y})^2) \end{aligned}$$

Thus,

$$\mu|y \sim N(\bar{y}, \frac{\sigma^2}{n})$$

This is exactly the same as the esitimate from MLE estimate.

NOTE

In case that the variance is unknown,

$$f(\sigma^2) \propto \frac{1}{\sigma^2}$$

This is equivalent to the following:

$$\Gamma^{-1}(0, 1)$$

Thus the posteiro for σ^2

$$\sigma^2|y \sim \Gamma^{-1}(\frac{n-1}{2}, \frac{1}{2}\sum(y_i-\bar{y})^2)$$

1.10 Jeffreys Priors

Jeffreys Prior

$$f(\theta) \propto \sqrt{I(\theta)}$$

For instance,

1.10.1 Gaussian

$$Y_i \sim (\mu, \sigma^2) \rightarrow f(\mu) \propto 1, f(\sigma^2) \propto \frac{1}{\sigma^2}$$

1.10.2 Bernoulli

$$Y_i \sim B(\theta) \rightarrow f(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \sim Beta(\frac{1}{2}, \frac{1}{2})$$

1.10.3 Side Note: Fisher Information

The Fisher information (for one paramter):

$$?(\theta) = E[(\frac{d}{d\theta} log(f(X|\theta)))^2]$$

The Expectation is with respect to X with a PDF $f(x|\theta)$.

For instance, suppose the $X|\theta \sim N(\theta, 1)$

$$\begin{aligned} f(x|\theta) &= \frac{1}{\sqrt{2\pi}} exp[-\frac{1}{2}(x-\theta)^2] \\ log(f(x|\theta)) &= log(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}(x-\theta)^2 \\ \frac{d}{d\theta} log(f(x|\theta)) &= x - \theta \\ (\frac{d}{d\theta} log(f(x|\theta)))^2 &= (x-\theta)^2 \\ E[(\frac{d}{d\theta} log(f(x|\theta)))^2] &= (x-\theta)^2 = Var(X) = 1 \end{aligned}$$

1.10.4 Prior predictive distribution

Prior distribution vs. prior predictive distribution

https://stats.stackexchange.com/questions/394648/differences-between-prior-distribution-and-prior-predictive-distribution

Stochastic Processes

https://www.youtube.com/watch?v=TuTmC8aOQJE

1.11 Linear regression

1.11.1 Review

For single covariate of x:

$$E[y] = \beta_0 + \beta_1 x$$

Thus,

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

For multiple covariates,

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Side Note It is interesting to think y as the expected value of of the combination of βx.

1.11.2 When σ^2 is known

If we use a Jeffreys prior and assume σ^2 is known, we will get β that has the same mean as the standard OLS. Assuming we only has one covariate:

$$\beta_1 | y \sim N(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

For mutiple covariates, we can use matrix notation,

$$\beta | y \sim N((X^t X)^{-1} X^t y, (X^t X)^{-1} \sigma^2)$$

1.11.3 When σ^2 is unknown

When both β and σ^2 are unknown, the standard prior is the non-informative Jeffreys prior:

$$f(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

The posterior mean for β is the same as standard OLS estimates. The posterior for β conditional on σ^2 is the same normal distribution when σ^2 is known. However, the marginal posteior distribution for β, with σ^2 integrated out, is a t distribution.

The t distribution has the mean of $(X^tX)^{-1}X^ty$ and variance matrix, $s^2(X^tX)^{-1}$, where $s^2 = \sum_{i=1}^{n}(y_i - \hat{y_i})^2/(n-k-1)$

The variance of σ^2 is an inverse gamma distribution

$$\sigma^2|y \sim \Gamma^{-1}(\frac{n-k-1}{2}, \frac{n-k-1}{2}s^2)$$

Chapter 2

Bayesian - 2

2.1 Components of Bayesian models

$$y_i = \mu + \epsilon_i$$

Where,

$$\epsilon_i \sim N(0, \sigma^2)$$

$$y_i \sim N(\mu, \sigma^2)$$

(Thus, y_i is = to a fixed μ plus with a ϵ_i, whereas $y_i \sim N(\mu, \sigma^2)$. These two expressions are not exactly the same, but they are connected.)

Likelihood: $P(y|\theta)$ $(P(y, \theta) = P(\theta)P(y|\theta))$

Prior: $P(\theta)$

Posterior:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)} = \frac{P(\theta, y)}{\int P(\theta, y)d\theta} = \frac{P(\theta)P(y|\theta)}{\int P(\theta, y)d\theta} = \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta}$$

Markts

(1) The only random variables in frequentist models are the data. In contrast, Bayesian paradigm also uses probability to describe one's uncertainty about unknown model parameters.

(2) Consider the following model for binary outcome y:

$y_i|\theta_i \sim Bern(\theta_i), i = 1, 2, 3...6$

$\theta_i|\alpha \sim Beta(a, b_0), i = 1, 2, 3...6$

$\alpha \sim Exp(r_0)$

Thus, the joint distribution of all variable:

$$\prod_{i=1}^{6}[\theta_i^{y_i}(1-\theta_i)^{1-y_i}\frac{\Gamma(a+b_0)}{\Gamma(a)\Gamma(b_0)}\theta_i^{a-1}(1-\theta_i)^{b_0-1}]r_0e^{-r_0\alpha}$$

(Question: Why not write it as a_i?)

2.2 Monte Carlo Estimation

2.2.1 Mean and Variance: Application of Central Limit Theorem

If we simulate 100 samples from a Gamma(2,1), what is the approximate distribution of the sample average $\bar{x^*} = \frac{1}{100}\sum_{i=1}^{100} x_i^*$?

As we know, based on the central limit theorem, the approximating distribution is normal with the mean equal to the sample mean, and the variance equal to the variance of the orignal variable divided by the sample size. We know that the mean for Gamma(2,1) is $\frac{2}{1} = 2$ and variance is $\frac{2}{1^2} = 2$. Thus, we know that the mean for the distribution for $\bar{x^*}$ is 2, whereas the variance is $\frac{2}{100}$. Thus, it is $N(2, 0.02)$.

Thus, we can get the generalized formula as follows:

$$\bar{\theta^*} \sim N(E(\theta), \frac{Var(\theta)}{m})$$

Note that, it is the variance for $\bar{\theta^*}$, not θ. The approximate for the variance for θ is $\frac{1}{m}\sum(\theta_i^* - \bar{\theta^*})^2$.

The following R code is for the θ:

```
sample_data_gamma<-rgamma(10000,4,2)
mean(sample_data_gamma)
```

```
## [1] 2.007437
```

```
var(sample_data_gamma)
```

```
## [1] 1.020376
```

The following R code is for the $\bar{\theta^*}$. As we can see, the variance is 0.0016, which is close to the true value of $\frac{1}{1000} = 0.001$.

```
set.seed(123)
mean_c<-c()
for(i in 1:10)
{mean_c[i]<-mean(rgamma(1000,4,2))}
var(mean_c)
```

```
## [1] 0.001634891
```

Side Note

Note that the definition of variance is:

$$V(x) = E[(x-\mu)^2]$$

Thus, we can calculate the variance using integral:

$$V(x) = \int (x-\mu)^2 f(x) dx$$

When we using samples from the simulation, we will get the following:

$$V(x) = \int (x-\mu)^2 f(x) dx = \frac{1}{m} \sum (x_i^* - \bar{x^*})^2$$

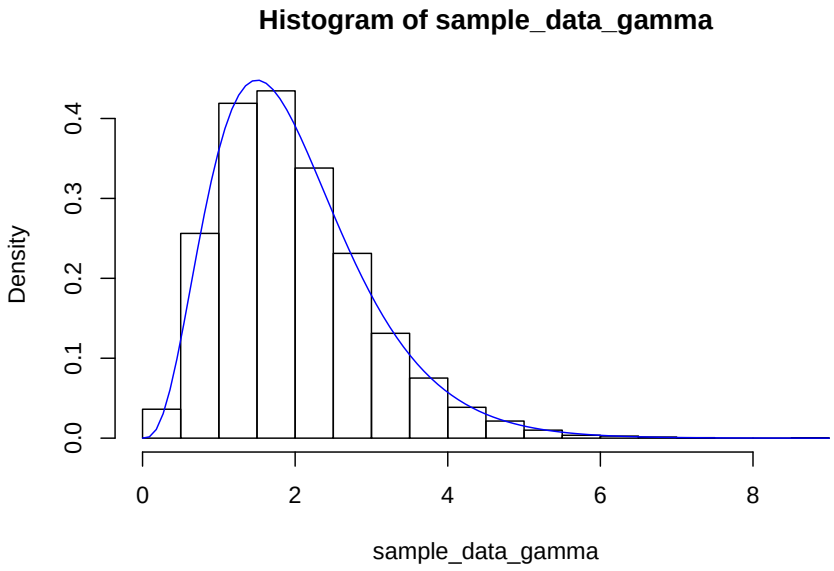
Thus, we can use Var function in R with simulated sample to calculate the variance.

2.2.2 Monte Carlo error (Standard Error)

For the $\bar{x^*}$, we can use the CLT to approximate how accurate the Monte Carlo estimates are.

$$\frac{SD(sample)}{\sqrt{m}}$$

```
set.seed(123)
sample_data_gamma<-rgamma(10000,4,2)

hist(sample_data_gamma, freq=FALSE)
curve(dgamma(x=x, shape=4, rate=2), col="blue", add=TRUE)
```




```
var(sample_data_gamma)
```

```
## [1] 0.9682145
```

```
sqrt(var(sample_data_gamma))
```

```
## [1] 0.9839789
```

```
sd(sample_data_gamma)
```

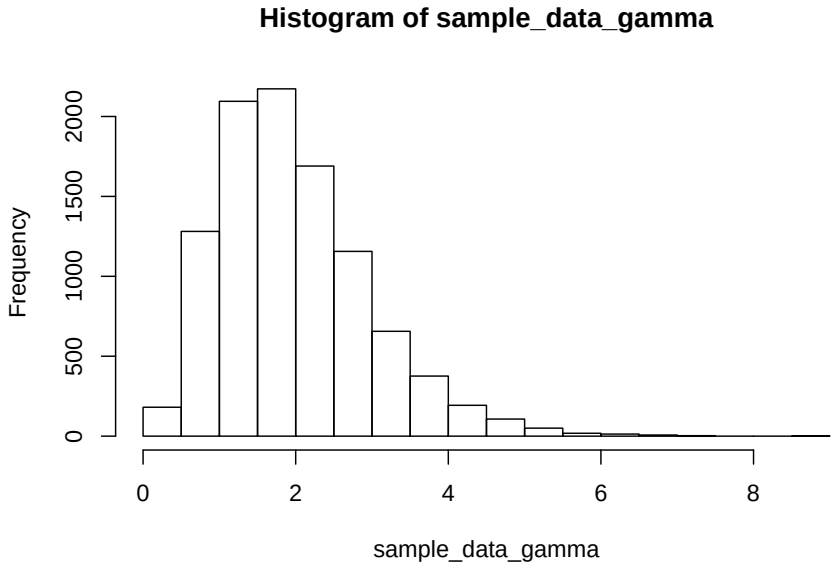
```
## [1] 0.9839789
```

```
sd(sample_data_gamma) / sqrt(10000)
```

```
## [1] 0.009839789
```

We can also calculate Standard Error for the probablity

```
hist(sample_data_gamma)
```




```
se = sd(sample_data_gamma<5) / sqrt(10000)
se
```

```
## [1] 0.0009547917
```

2.2.3 Expected value and probability

If you know that $\theta \sim Beta(5,3)$, what is the approximate for the $E(\frac{\theta}{1-\theta})$? You can use the following R code to calculate it.

```
sample_data<-rbeta(1000,5,3)
mean(sample_data/(1-sample_data))
```

```
## [1] 2.789685
```

If you want to calculate the approximate for the probability that $\frac{\theta}{1-\theta}$ is greater 1.

```
mean((sample_data/(1-sample_data))>1)
```

```
## [1] 0.789
```

2.2.4 Quantile

Use Monte Carlo to approximate the 0.3 quantile of N(0,1). Note that, the idea of quantile is to quantify the probability. Thus, the number we will get for 0.3 quantile if the value for the random variable's cdf $\int_{-\infty}^{quantile-number} dx$. As we can see below, $quantile(sample_data2, 0.3)$ gets the result of -0.52, which is the same from the $qnorm(0.3)$. Note that, the cdf of $pnorm(-0.52)$ will get the probablity of 0.3.

```
sample_data2<-rnorm(10000,0,1)
quantile(sample_data2,0.3)
```

```
##        30%
## -0.5260847
```

```
qnorm(0.3)
```

```
## [1] -0.5244005
```

```
pnorm(-0.52)
```

```
## [1] 0.3015318
```

2.2.5 Prior predictive distributions (Marginalization)

$$y|\theta \sim Bin(10, \theta)$$
$$\theta \sim Beta(2, 2)$$

Simulate:

(1) θ^* from beta

(2) Given θ, draw $y_i^* \sim Bin(10, \theta_i^*)$

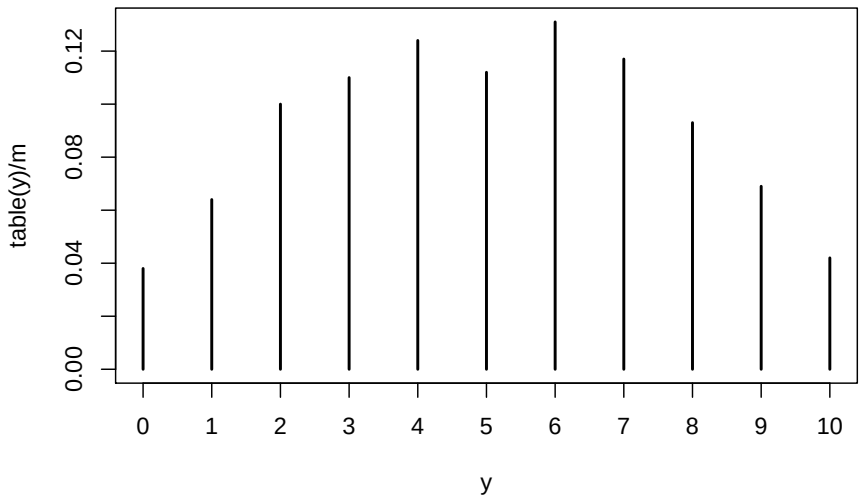
(3) Get pairs (y_i^*,θ_i^*)

```
m=1000

y=numeric(m)
phi=numeric(m)

for (i in 1:m)
```

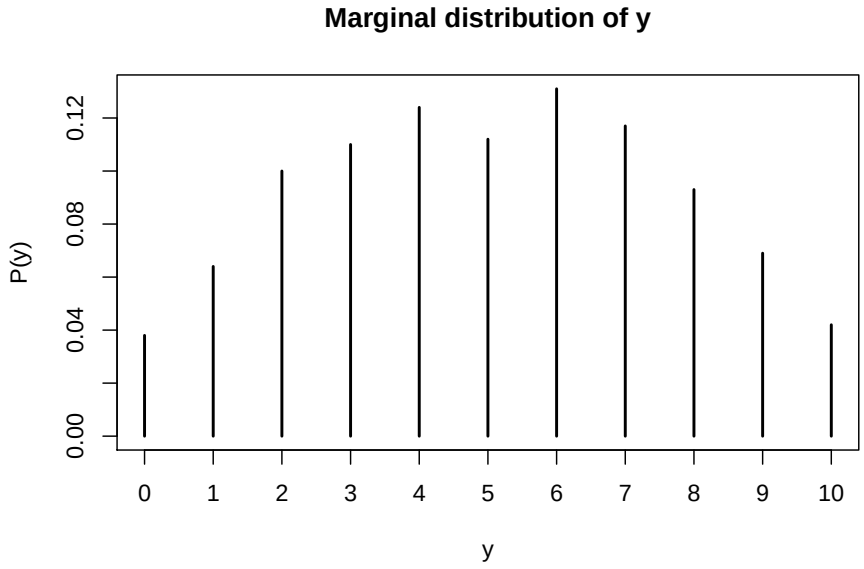
```
{
  phi[i]=rbeta(1,shape1=2,shape2 = 2)
  y[i]=rbinom(1,size=10,prob = phi[i])
}

# we can use vector method

phi=rbeta(m,shape1 = 2,shape2 = 2)
y=rbinom(m,size=10,prob = phi)
table(y)/m
```

```
## y
##     0     1     2     3     4     5     6     7     8     9    10
## 0.038 0.064 0.100 0.110 0.124 0.112 0.131 0.117 0.093 0.069 0.042
```

```
# The marginal distribution of y
plot(table(y)/m)
```




```
# Another way to plot
plot(prop.table(table(y)), ylab="P(y)", main="Marginal distribution of y")
```




```
# the marginal expected value of y
mean(y)
```

```
## [1] 5.04
```

2.3 Markov chains

2.3.1 Definition

A sequence of random variable $X_1, X_2, ..., X_n$, with $1, 2, ..., 3$ indicating the successitve points in time. Thus, based on the chain rule, we can write the following:

$$p(X_1, X_2, ..., X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)...p(X_n|X_{n-1}, X_{n-2}, ..., X_2, X_1)$$

For Markov chains, it puts an assumption, called Markov assumption: The random variable at the next time step only depends on the current variable. Mathematically,

$$p(X_{t+1}|X_t, X_{t-1}, ..., X_2, X_1) = p(X_{t+1}|X_t)$$

where,

$$t = 2, ...n$$

Thus, we can write the expression above as follows.

$$p(X_1, X_2, ..., X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2)P(X_4|X_3)...p(X_n|X_{n-1})$$

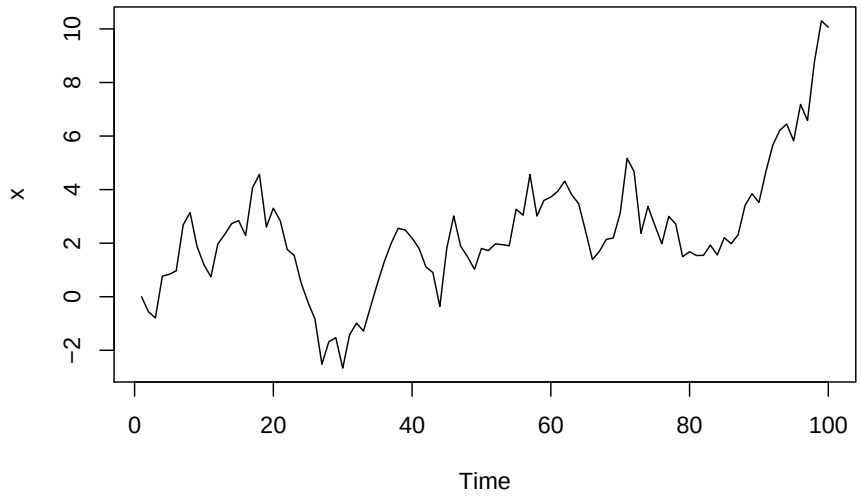
2.3.2 Discrete example

Suppose that you flip a coin. You have a set of number {1,2,3,4,5}. If it is head, you increase 1 in the next number (for instance, if you are 2 now, you will be get 3 in the next). In contrast, if it is tail, you will decrease the number (e.g., 2 is now and 1 is next). If is is 5, increase 1 will lead to 1. Logically, 1 and then it is reduced by 1, leading to 5.

2.3.3 Continuous example

$$p(X_{t+1}|X_t = x_t) = N(x_t, 1)$$

```
set.seed(123)
n=100
x=numeric(n)

for(i in 2:n)
{x[i]=rnorm(1,mean=x[i-1],1)}

plot.ts(x)
```



2.3.4 Discrete example and transition matrix

For the discrete example above, we can write the following transition matrix. We know that $p(X_{t+1} = 5|X_t = 4)$ can be found in the 4th row, 5th column, namely 0.5.

$$Q = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 & 0 \end{bmatrix}$$

The transition matrix is especially useful when there are mutiple steps in the chain. such as $p(X_{t+1} = 5|X_t = 4)$. It can be calculated as $\sum_{k=1}^{5} p(X_{t+2} = 3|X_{t+1} = k) \cdot p(X_{t+1} = k|X_t = 1)$. We can use R code the impletement this.

```
Q = matrix(c(0.0, 0.5, 0.0, 0.0, 0.5,
             0.5, 0.0, 0.5, 0.0, 0.0,
             0.0, 0.5, 0.0, 0.5, 0.0,
             0.0, 0.0, 0.5, 0.0, 0.5,
             0.5, 0.0, 0.0, 0.5, 0.0),
           nrow=5, byrow=TRUE)

(jj<-Q %*% Q)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] 0.50 0.00 0.25 0.25 0.00
## [2,] 0.00 0.50 0.00 0.25 0.25
## [3,] 0.25 0.00 0.50 0.00 0.25
## [4,] 0.25 0.25 0.00 0.50 0.00
## [5,] 0.00 0.25 0.25 0.00 0.50
```

```
jj[1,3]
```

```
## [1] 0.25
```

2.3.5 Stationary distribution

Discrete Cases_

```
Q5 = Q %*% Q %*% Q %*% Q %*% Q # h=5 steps in the future

Q5
```

```
##         [,1]    [,2]    [,3]    [,4]    [,5]
## [1,] 0.06250 0.31250 0.15625 0.15625 0.31250
## [2,] 0.31250 0.06250 0.31250 0.15625 0.15625
## [3,] 0.15625 0.31250 0.06250 0.31250 0.15625
## [4,] 0.15625 0.15625 0.31250 0.06250 0.31250
## [5,] 0.31250 0.15625 0.15625 0.31250 0.06250
```

```
round(Q5, 3)
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]
## [1,] 0.062 0.312 0.156 0.156 0.312
## [2,] 0.312 0.062 0.312 0.156 0.156
## [3,] 0.156 0.312 0.062 0.312 0.156
## [4,] 0.156 0.156 0.312 0.062 0.312
## [5,] 0.312 0.156 0.156 0.312 0.062
```

```
Q30 = Q
for (i in 2:30) {
  Q30 = Q30 %*% Q
}
round(Q30, 3) # h=30 steps in the future
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]
## [1,] 0.201 0.199 0.200 0.200 0.199
## [2,] 0.199 0.201 0.199 0.200 0.200
## [3,] 0.200 0.199 0.201 0.199 0.200
## [4,] 0.200 0.200 0.199 0.201 0.199
## [5,] 0.199 0.200 0.200 0.199 0.201
```

Thus, as we can see, as the time gap gets bigger, the transition distributions appear to converge.

```
Q = matrix(c(0.0, 0.5, 0.0, 0.0, 0.5,
             0.5, 0.0, 0.5, 0.0, 0.0,
             0.0, 0.5, 0.0, 0.5, 0.0,
             0.0, 0.0, 0.5, 0.0, 0.5,
             0.5, 0.0, 0.0, 0.5, 0.0),
           nrow=5, byrow=TRUE)
Q
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.0  0.5  0.0  0.0  0.5
## [2,]  0.5  0.0  0.5  0.0  0.0
## [3,]  0.0  0.5  0.0  0.5  0.0
## [4,]  0.0  0.0  0.5  0.0  0.5
## [5,]  0.5  0.0  0.0  0.5  0.0
```

```
c(0.2, 0.2, 0.2, 0.2, 0.2) %*% Q
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.2  0.2  0.2  0.2  0.2
```

The definition of stationary distribution of a chain:

The initial state distribution for which performing a transition will not change the probability of ending up in any given state.

For a Markov chain after many iterations, the samples can be used as a Monte Carlo sample from the stationary distribution. Thus, we can use Markov chains for Bayesian inference. In order to simulate from a complicated posterior distribution, we will set up and run a Markov chain whose stationary distribution is the posterior distribution. (Note that: stationary distribution doesn't always exist for any given Markov chain)

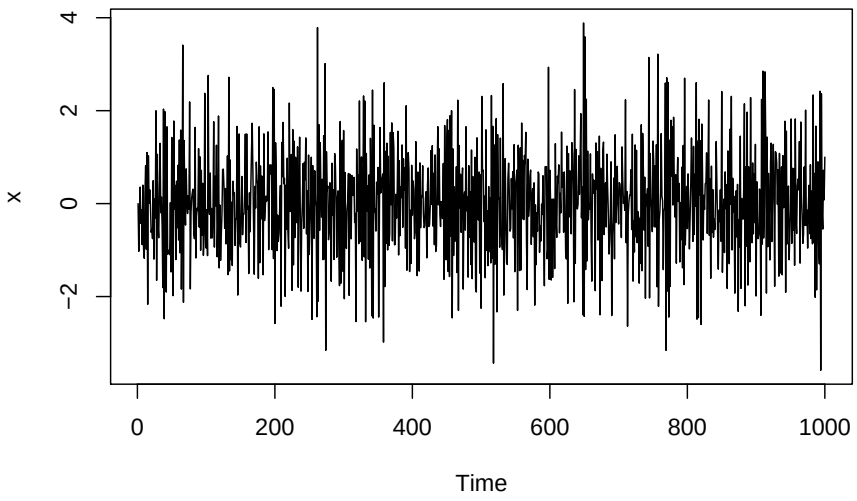
Continuous Cases_

As we can see, the example shown earlier does not reach a stationary distribution. But, we can modify it to make it happen.

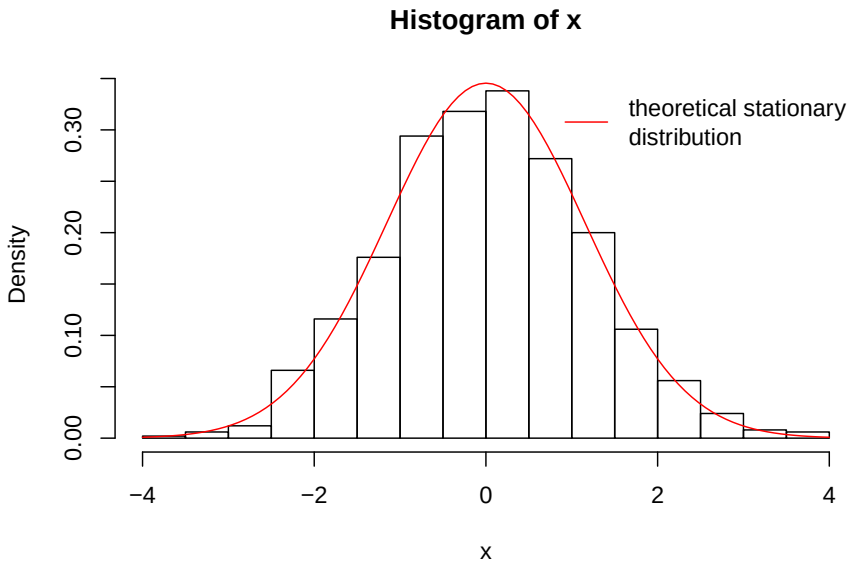
$$p(X_{t+1}|X_t = x_t) = N(\phi x_t, 1), -1 < \phi < 1$$

It will reach a stationary distribution as long as $-1 < \phi < 1$.

```
n = 1000
x = numeric(n)
phi = -0.5

for (i in 2:n) {
  x[i] = rnorm(1, mean=phi*x[i-1], sd=1.0)
}

plot.ts(x)
```




```
hist(x, freq=FALSE)
curve(dnorm(x, mean=0.0, sd=sqrt(1.0/(1.0-phi^2))), col="red", add=TRUE)
legend("topright", legend="theoretical stationary\ndistribution", col="red", lty=1, bty="n")
```



It will reach the stationary distribution of $N(0, \frac{1}{1-\phi^2})$

2.4 Metropolis-Hastings

2.4.1 Procedure

(1) Select initial value θ_0

(2) For $i = 1, ..., m$, repeat:

(a) Draw candidate $\theta^* \sim q(\theta^*|\theta_{i-1})$

(b) $\alpha = \frac{g(\theta^*)/q(\theta^*|\theta_{i-1})}{g(\theta_{i-1})/q(\theta_{i-1}|\theta^*)} = \frac{g(\theta^*)q(\theta_{i-1}|\theta^*)}{g(\theta_{i-1})q(\theta^*|\theta_{i-1})}$

(c) $\alpha \geq 1$ accept θ^* and set $\theta_i \leftarrow \theta^*$

$0 < \alpha < 1$ accept θ^* and set $\theta_i \leftarrow \theta^*$ with probablity α

2.4.2 Demonstration

Prior $P(loaded) = 0.6$

$$f(\theta|X) = \frac{f(x|\theta)f(\theta)}{\sum_\theta f(x|\theta)f(\theta)}$$
$$= \frac{\binom{5}{x}[(\frac{1}{2})^5 \times 0.4 \times I_{\{\theta=fair\}} + (0.7)^x(0.3)^{5-x} \times 0.6 \times I_{\{\theta=loaded\}}]}{\binom{5}{x}[(\frac{1}{2})^5 \times 0.4 + (0.7)^x(0.3)^{5-x} \times 0.6]}$$

$$f(\theta|X=2) = \frac{0.0125I_{\{\theta=fair\}} + 0.0079I_{\{\theta=loaded\}}}{0.0125 + 0.0079}$$
$$= 0.612I_{\{\theta=fair\}} + 0.388I_{\{\theta=loaded\}}$$

Thus, we can say that:

$$P(\theta = loaded|X = 2) = 0.388$$

(1) Start at either $\theta_0 = fair$ or $\theta_0 = loaded$

(2) For $i = 1, ...m$

(a) Propose candidate θ^* to be the other state such as θ_{i-1}

(b)

$$\alpha = \frac{g(\theta^*)/q(\theta^*|\theta_{i-1})}{g(\theta_{i-1})/q(\theta_{i-1}|\theta^*)} = \frac{f(x=2|\theta^*)/1}{f(x=2|\theta_{i-1})/1}$$

If $\theta^* = loaded, \alpha = \frac{0.0794}{0.0125} = 0.635$ If $\theta^* = fair, \alpha = \frac{0.0125}{0.0794} = 1.574$

(c) if $\theta^* = fair, \alpha > 1$, accept θ^* set $\theta_i = fair$

if $\theta^* = loaded, \alpha > 0.635$, accept θ^* set $\theta_i = loaded$ w.p. 0.635.

Thus, we can get the following transition matrix:

$$Q = \begin{bmatrix} 0.365 & 0.635 \\ 1 & 0 \end{bmatrix}$$

This means that:

(1) When θ_{i-1} is fair, θ_i is also fair, the $P = 0.365$.

(2) When θ_{i-1} is fair, θ_i is loaded, the $P = 0.635$.

(3) When θ_{i-1} is loaded, θ_i is fair, the $P = 1$.

(4) When θ_{i-1} is loaded, θ_i is loaded, the $P = 0$.

Thus,

$$\pi = [0.612, 0.388]$$

$$[0.612, 0.388] \begin{bmatrix} 0.365 & 0.635 \\ 1 & 0 \end{bmatrix} = [0.612, 0.388]$$

2.4.3 R code

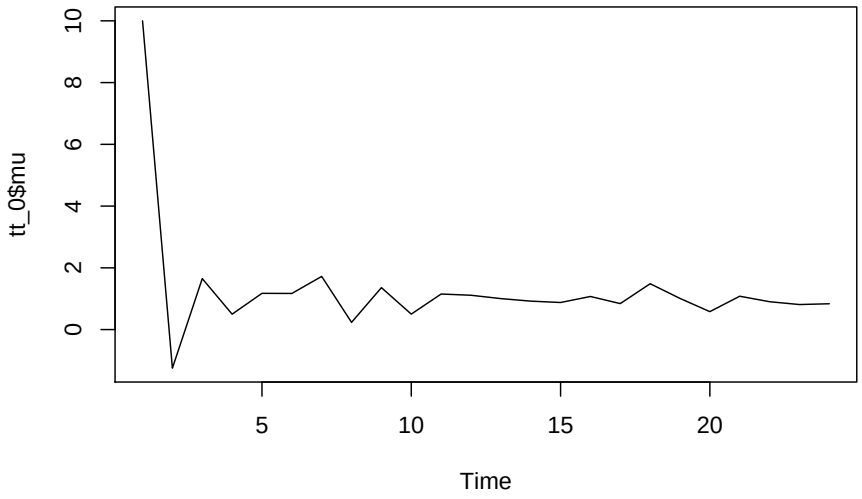
The data are the percent change in total personnel for n=10 companies (1.2,1.4,-0.5,0.3,0.9,2.3,1.0,0.1,1.3,1.9). We assume such changes follow a normal distribution with known variance (but mean is unknown). For the mean (i.e., μ), we assume it follows a t distribution (i.e., the prior for the μ is t-distribution).Since t-distribution prior is not conjugate to a normal distribution, the posterior distribution is not in a standard form that we can easily sample. Thus, we need to set up a Markov chain whose stationary distribution is the posterior distribution.

$$p(\mu|y_1, ..., y_n) \propto \frac{exp[n(\bar{y}\mu - \mu^2/2)]}{1 + \mu^2}$$

$$log(g(\mu)) = n(\bar{y}\mu - \mu^2/2) - log(1 + \mu^2)$$

```
lg=function(mu, n, ybar)
{ mu2=mu^2
  n*(ybar*mu-mu2/2)-log(1+mu2)}

mu_now=10
cand_sd=5
y=c(1.2,1.4,-0.5,0.3,0.9,2.3,1.0,0.1,1.3,1.9)
ybar=mean(y)
m=1000
n=length(y)
mu_storage<-c()
mu_storage<-c(mu_storage,mu_now)
acceptance_count=0

mh<-function(mu_now,type_MH,cand_sd)
{

for (i in 1:m)
```

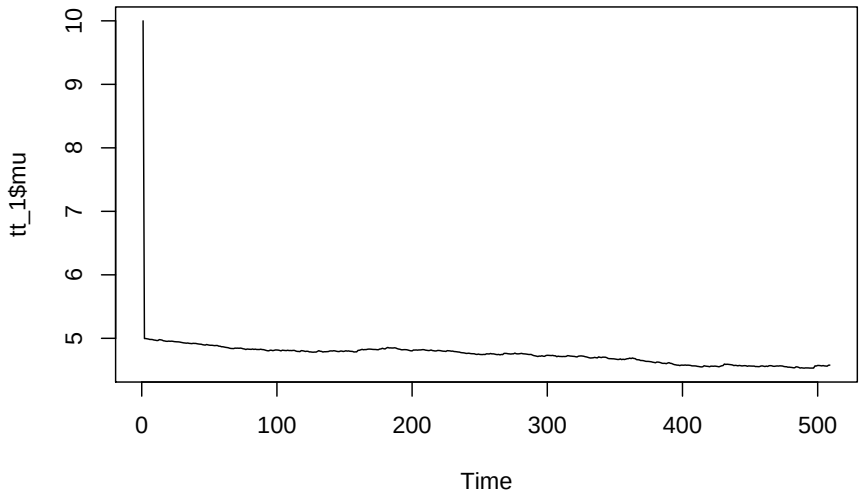
```
{
  if (type_MH=="randomwalk")  # random walk, as mean is updated with mu_now
  {
  mu_cand = rnorm(n=1, mean=mu_now, sd=cand_sd)
  }
  else
  {
    if(type_MH=="independent")  # independent, as mean is fixed
    {mu_cand = rnorm(n=1, mean=3, sd=cand_sd)}
    else
    {print("woring type of MH")
     break
    }
  }

  lg_alpha=lg(mu_cand,n=n,ybar = ybar)-lg(mu_now,n=n,ybar=ybar)
  alpha=exp(lg_alpha)
  if(alpha>1)
    {mu_now=mu_cand
    mu_storage<-c(mu_storage,mu_cand)
    acceptance_count=acceptance_count+1}
  else
    {random_p=runif(1)
    if(alpha>random_p)
      {mu_now=mu_cand
      acceptance_count=acceptance_count+1}
    else
      {mu_now=mu_now}
    }
}
#return the following
list(mu=mu_storage,accpt=acceptance_count/m)
}

tt_0<-mh(mu_now=5,type_MH="randomwalk",cand_sd=10)
plot.ts(tt_0$mu)
```




```
tt_0$accpt
```

```
## [1] 0.039
```

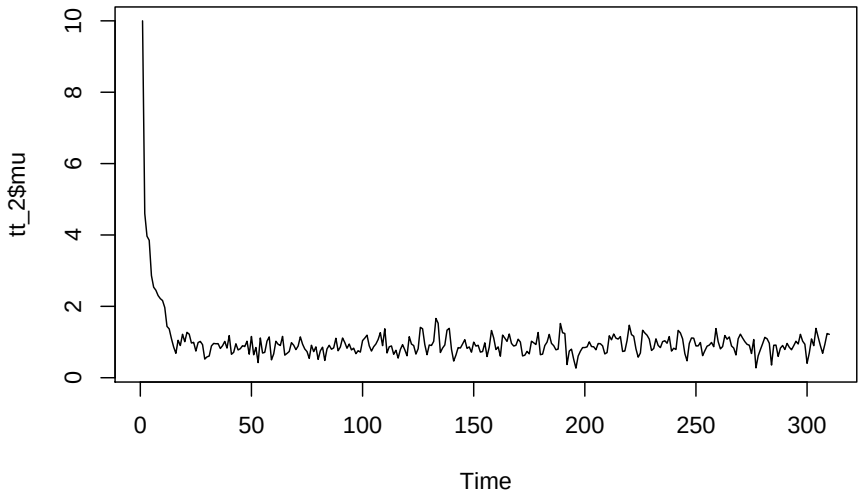
```
tt_1<-mh(mu_now=5,type_MH="randomwalk",cand_sd=0.005)
plot.ts(tt_1$mu)
```




```
tt_1$accpt
```

```
## [1] 0.931
```

```
tt_2<-mh(mu_now=5,type_MH="randomwalk",cand_sd=0.5)
plot.ts(tt_2$mu)
```

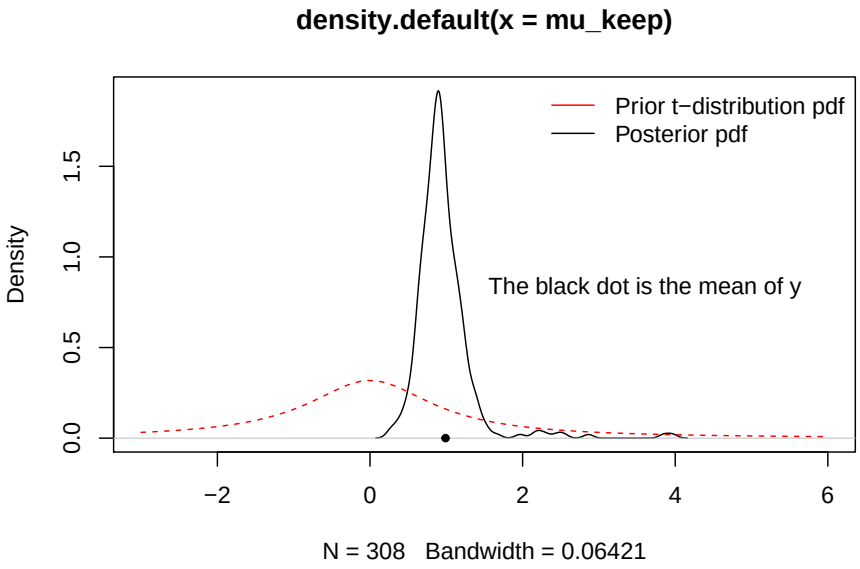



```
tt_2$accpt
```

```
## [1] 0.587
```

As we can see, when the variance is big, the acceptance rate is low. When the variance is low, in conrast, the acceptance rate is high. Note that, changing the mean does not change the acceptance rate too much. In the following, I plotted the prior, posteior, and the sample(i.e., y).

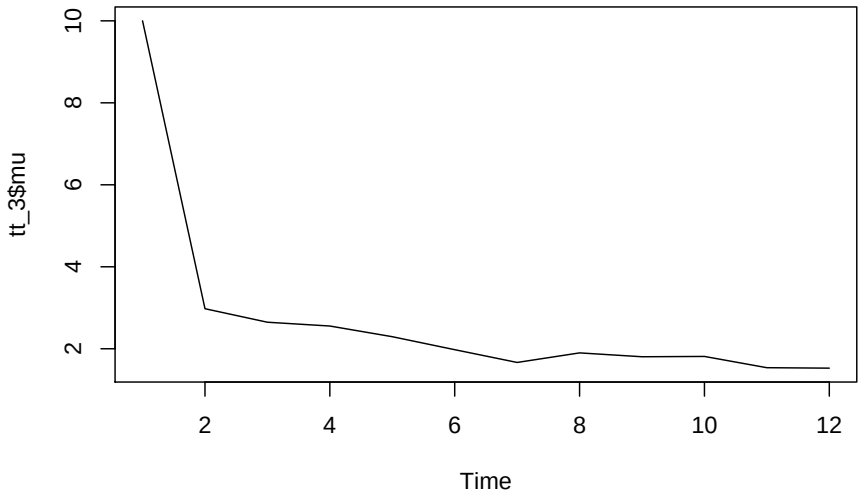
```
mu_keep<-tt_2$mu[-c(1:2)]
plot(density(mu_keep),xlim=c(-3,6))
curve(dt(x,df=1),lty=2,add = TRUE,col="red")
legend("topright",legend=c("Prior t-distribution pdf", "Posterior pdf"), col=c("red","
points(ybar,0,pch=20)
legend(1,1, legend="The black dot is the mean of y",bty="n")
```



In the following, I show the situation that the proposal distribution is not random walk, but rather independent. That is, the mean for the normal distribution is fixed. Refer to the procedure of MH, there is a setence of:

Draw candidate $\theta^* \sim q(\theta^*|\theta_{i-1})$

```
tt_3<-mh(mu_now=5,type_MH="independent",cand_sd=0.5)
plot.ts(tt_3$mu)
```




```
tt_3$accpt
```

```
## [1] 0.016
```

```
#tt_4<-mh(mu_now=5,type_MH="hello",cand_sd=0.5)
#tt_4
```