Statg012 Package Vignette

YI JIN

2017-08-24

The package **statg012** focuses on conjugate priors and hypothesis testing of the Bayesian Inference, which is suitable for students studying STATG012 course at University College London. Conjugate prior families were first introduced and discussed by Raiffa and Schlaifer (1961). They showed that the posterior distribution, resulting from the conjugate prior, is itself a member of the same family as the conjugate prior (Raiffa & Schlaifer 1961). When the prior and posterior distribution are from the same family, the formulas for updating the prior into the posterior will become relatively simple. Due to this property, Fink (1997) demonstrated that the conjugate prior families are attractive that go beyond analytical tractability. In STATG012 course, it pointed out that Bayesian hypothesis testing is built on the posterior distribution, and the decision is to not reject or reject the null hypothesis according to whichever decision minimises the expected posterior loss. The objective of this package is to help students study the course, and understand the conjugate priors and Bayesian hypothesis testing more clearly. Students can utilize the package to check answers of examples in slides and see corresponding plots. The probability distributions from the exponential family all have conjugate prior families (Robert 2007). In general, we have 8 functions in this package. Specifically, the STATG012 course mentioned 6 distributions of them which we have already explained in the section 4. Hence, there are 5 functions about conjugate priors related with 6 different models in the package. Moreover, it also has 1 function to summaries statistics of these distinctive models, and 1 function to show plots of their prior and posterior distributions. The rest 1 function is about hypothesis testing of the Bayesian Inference. Notice that the purpose of this vignette is only to show that how to use the functions in this package with some examples. For more details about the arguments required by each function, please refer to the help files or STATG012 notes.

Functions for Conjugate Priors

As we mentioned, there are 6 examples about conjugate prior families introduced in STATG012 slides: the binomial sampling distribution with a beta distribution of prior, the negative binomial sampling distribution with a beta distribution of prior, the Poisson sampling distribution with a gamma distribution of prior, the exponential sampling distribution with a gamma distribution of prior, the gamma sampling distribution with a gamma distribution of prior and the normal sampling distribution with a normal distribution of prior. We will explain these 6 models with 5 functions simply below.

The `binombeta()` defines the posterior distribution function for $\pi(\theta \mid r)$, with a beta prior distribution $\pi(\theta; \alpha, \beta)$ and a binomial sampling distribution $p(r \mid \theta)$, where r denotes the number of successes in n trials. The function will return the prior distribution, the posterior distribution, the likelihood function, the range of theta, the model name, and the parameters of beta posterior distribution, `pos.alpha`, `pos.beta` which can be compared with parameters of the beta prior distribution.

The `nbinombeta()` focuses on the posterior distribution function for $\pi(\theta \mid k)$, with a beta prior distribution $\pi(\theta; \alpha, \beta)$ and a negative binomial sampling distribution $p(k \mid \theta)$, where k is the number of failures in rth successes. It will return the same values as `binombeta()` function.

The `poisgamma()` defines the posterior distribution function for $\pi(\mu \mid x)$, with a gamma prior distribution $\pi(mu; \alpha, \lambda)$ and a poisson sampling distribution $f(x \mid \mu)$, where μ is the average rate of the sample data x from Poisson distribution. It will return the prior and posterior distributions, the likelihood function, the range of μ, the model name, and the parameters of gamma distribution for posterior `pos.shape`, `pos.rate`. Then we can compare parameters between prior distribution and posterior distribution.

The `gamgam()` focuses on the posterior distribution function for $\pi(\theta \mid t)$, with a gamma prior distribution $\pi(\theta; \alpha, \beta)$ and a gamma sampling distribution t with known shape parameter a and unknown rate parameter θ. The return of this function will be the prior and posterior distributions, the likelihood function, the unknown parameter θ, the model name, and the parameters of gamma distribution for posterior `pos.shape`, `pos.rate`. When a equals 1, it will become a special case that a random sample from the exponential distribution has a gamma prior distribution.

The `normnorm()` defines the posterior distribution function for $\pi(\mu \mid x)$ if we have a normal sampling distribution $f(x \mid \mu, \sigma^2)$ where μ is unknown and σ^2 is known, and the normal prior distribution $\pi(\mu; m, s)$ for unknown μ. It will return the the prior and posterior distributions, the likelihood function, the unknown μ, the model name, and the mean, standard deviation and precision of the posterior normal distribution. If we have a normal sampling distribution where both μ and σ^2 are unknown, and the normal-gamma prior distribution $\pi(\mu, \tau; m, s, \alpha, \beta)$ for unknown mean and precision $\tau = 1/\sigma^2$, then we can define the posterior distribution $\pi(\mu, \tau \mid x)$. Except the prior, likelihood and posterior, this return will include the prior and posterior distribution for μ, the prior and posterior distribution for τ, the parameters of posterior normal-gamma distribution.

The `summary()` is a generic function used to produce summaries of the results of above six various models. It will return a list of summary statistics such as mean, variance and quantiles of prior and posterior distribution respectively.

The `plot()` is a function to generate plotting of prior and posterior distributions of above six models.

We will introduce how to use `binombeta()` and `normnorm()` individually, accompanied by `summary()` and `plot()` in the following parts

Using the `binombeta()` with the Example

Consider the example in the binomial distribution. Suppose X is a random sample distribution from the binomial distribution $X \sim Bin(10, \theta)$ with a beta prior distribution for $\theta \sim Beta(4, 6)$. Given the number of successes is 3, we can have

```r
library(statg012)
ex <- binombeta(4, 6, 10, 3)
#> Prior Alpha :  4
#> Prior Beta  :  6
#> Posterior alpha :  7
#> Posterior beta  :  13
```

We can check that the above results are the same as the results calculated by the formula. Thus, we could utilize this function to calculate the posterior parameters α, β or check the answers.

Using `summary()` to show summarized statistics of prior and posterior respectively.

```r
summary(ex)
#> Prior Mean           :  0.4
#> Prior Variance       :  0.0218
#> Prior Std. Deviation :  0.1477
#> prob. quantiles
#> 0.05 0.169
#> 0.25 0.291
#> 0.5  0.393
#> 0.75 0.502
#> 0.95 0.655
#>
```

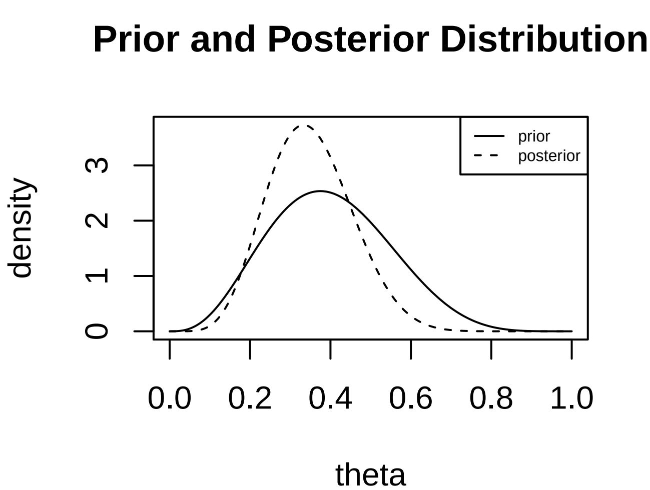
```
#> Posterior Mean           :  0.35
#> Posterior Variance       :  0.0108
#> Posterior Std. Deviation :  0.1041
#> prob. quantiles
#> 0.05 0.1875
#> 0.25 0.2752
#> 0.5  0.3449
#> 0.75 0.4195
#> 0.95 0.53
```

It can be recognized that the posterior mean and variance are the same as the results calculated by the formula as well. As Bolstad (2007) mentioned, it is necessary for us to know where the posterior distribution is located on the number line. Here, we provide two possible methods to measure the location: posterior mean and posterior median. The posterior mean, or named the expected value, is a frequent measure of location. If distributions have heavy tails such as skewed distributions, then the mean of posterior could be affected strongly, a distance away from the most probabilities. However, for beta distributions, which do not have the heavy tails, mean is a good choice to measure the location. The 50% quantile of the posterior distribution, also called posterior median, is another good measure of location. From the function results, the posterior mean and median are similar here, 0.35 and 0.3449 respectively.

The second thing we want to know is the spread of the posterior distribution. There are three possible measures of spread we can consider: posterior variance, posterior variance, posterior standard deviation and posterior quantiles. The variance of posterior is also affected by the heavy-tail distributions. With the squared units, it is difficult to explain the size with reference to the size of the mean. Due to this, we should use the posterior standard deviation, which is in terms of units (Bolstad, 2007). Moreover, we provide the 5%, 25%, 50%, 75% and 95 % quantiles of the posterior distribution to show the spread of widely probabilities lies between 0.1875 and 0.53.

Then, we use `plot()` to draw prior and posterior distribution

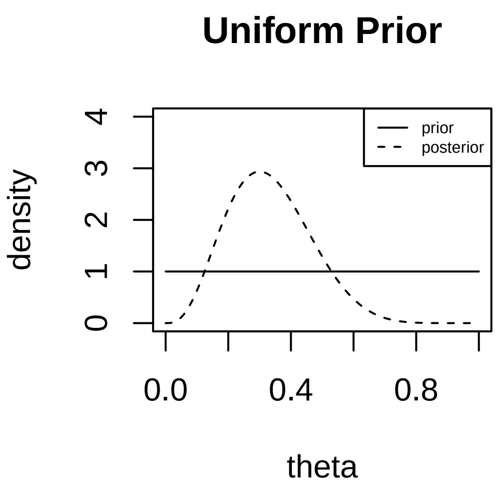
```
plot(ex, lty = 1:2, cex=0.5)
```

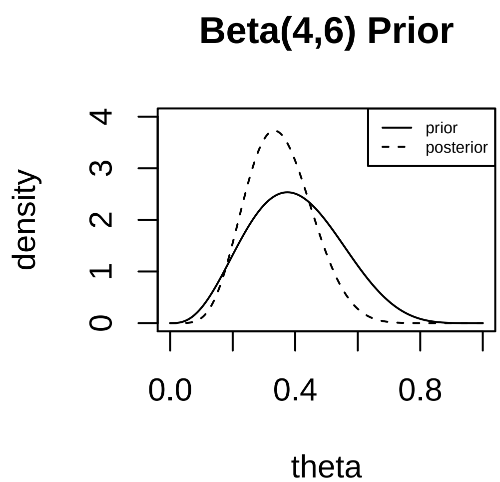


Since prior and posterior distributions are all from the Beta family, we can see that they have the similar shape. Moreover, the posterior distribution is much more concentrated than the prior, because the posterior is the degree of beliefs after observing the X sample distribution.

In the given example, we have already known the prior distribution. If we do not have any information about the θ, we may select the uniform prior which gives equal weight to all possible values of θ.

```r
uni <- binombeta(1, 1, 10, 3)
plot(uni, lty = 1:2, cex=0.5, main = "Uniform Prior", xlim=c(0,1), ylim=c(0,4))
plot(ex, lty = 1:2, cex=0.5, main = "Beta(4,6) Prior",xlim=c(0,1), ylim=c(0,4))
```

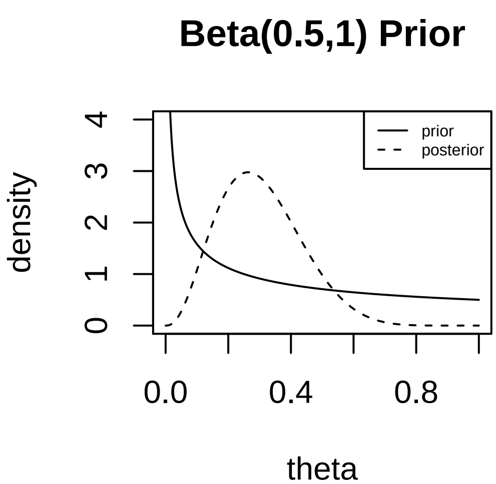


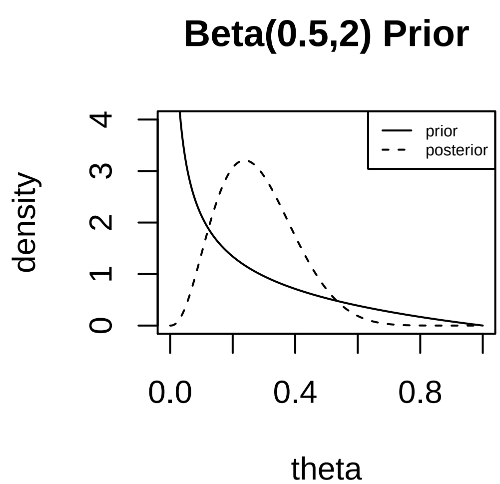


It can be seen that the two resulting posterior distributions are fairly similar, though the posterior distribution with Beta prior is more concentrated.

Moreover, if the prior information for θ is vague, such as θ is small, we may utilize the shape properties of Beta distribution to select Beta(0.5,1) or Beta(0.5,2) as the prior distribution.

```r
be1 <- binombeta(0.5, 1, 10, 3)
be2 <- binombeta(0.5, 2, 10, 3)
plot(be1, lty = 1:2, cex=0.5, main = "Beta(0.5,1) Prior", xlim=c(0,1), ylim=c(0,4))
plot(be2, lty = 1:2, cex=0.5, main = "Beta(0.5,2) Prior",xlim=c(0,1), ylim=c(0,4))
```





We can also recognize that given the data, the resulting posterior distributions are similar though beginning with different priors, and they are not as concentrated as the posterior with known Beta(4,6) prior.

Using the `normnorm()` with the Example

Consider the example in the normal distribution with unknown mean parameter at first. Suppose X_i, $i = 1, \cdots, 9$ are independent $\mathcal{N}(\theta, 4)$ and that the prior distribution of θ is $\mathcal{N}(25, 10)$. And the observed sample mean is 20. Then, we can use `normnorm()` to find the mean and variance of normal posterior distribution as follow.

```
## generate a sample of 9 from a normal distribution with sd=2
x <- rnorm(9, sd = 2)
## given that the observed sample mean is 20
xx <- x- mean(x) + 20
## find the posterior density
exmp1 <- normnorm(xx, m = 25, s = sqrt(10), sigma = 2)
#> Posterior precision      : 2.35
#> Posterior mean           : 20.2128
#> Posterior variance       : 0.4255
#> Posterior std. deviation : 0.6523
```

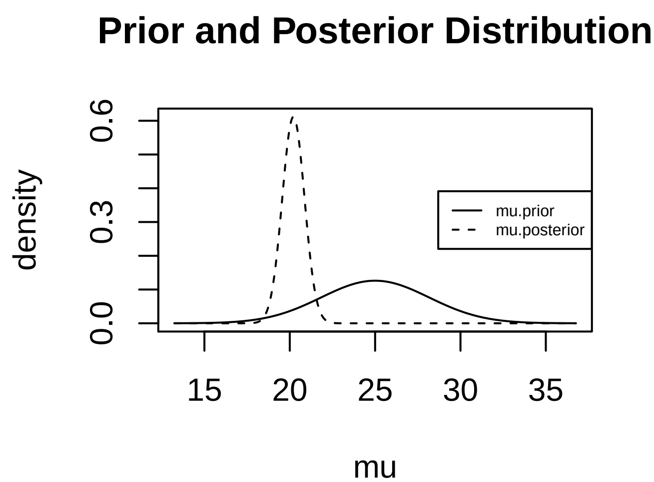
Checking the outputs, they are same as the results obtained from calculating by formula. Then we can utilize `summary()` to summarized statistics of prior and posterior distribution.

```
summary(exmp1)
#> Prior Precision      :  0.1
#> Prior Mean           :  25
#> Prior Variance       :  10
#> Prior Std. Deviation :  3.1623
#> prob. quantiles
#> 0.05 19.7985
#> 0.25 22.8671
#> 0.5  25
#> 0.75 27.1329
#> 0.95 30.2015
#> Posterior Precision      :  2.35
#> Posterior Mean           :  20.2128
#> Posterior Variance       :  0.4255
#> Posterior Std. Deviation :  0.6523
#> prob. quantiles
#> 0.05 19.1398
#> 0.25 19.7728
#> 0.5  20.2128
#> 0.75 20.6528
#> 0.95 21.2858
```

Based on the posterior mean and standard deviation, we can know the location and spread of the posterior distribution. The posterior median is the same as the posterior mean, equals 20.2128, since it is the normal distribution. From the quantiles, we know that the majority posterior probabilities lies between 19.1398 and 21.2858.

Utilizing `plot()` to show prior and posterior distribution together.

```r
plot(exmp1, leg_pos = "right", cex = 0.5)
```



We can recognize that they have similar shapes because they are both normal distributions. Given the sample data X, the distribution of posterior is much more concentrated than that of prior.

Now, consider another example in the normal distribution with unknown mean and unknown variance. Suppose Y_i, $i = 1, \cdots, 9$ are independent $\mathcal{N}(\mu, \sigma^2)$, and the prior distribution of unknown μ and precision τ $(= 1/\sigma^2)$ follows the normal-gamma distribution $\mathcal{NG}(\mu, \tau; m = 1, s = 2, \alpha = 1, \beta = 1)$. Using `normnorm()` to find the posterior distribution and `summary()` to summarized statistics.

```r
## generate a sample of 9 from a normal distribution
y <- rnorm(9)
## Given parameters, find the posterior density
exmp2 <- normnorm(y, m = 1, s = 2, a = 1, b = 1)
#> Posterior mu       : 0.1793
#> Posterior s        : 11
#> Posterior alpha    : 5.5
#> Posterior beta     : 3.985
summary(exmp2)
#> Mu Mean              :  0.1793
#> Mu Variance          :  0.0805
#> Mu Std. Deviation    :  0.2837
#>
#> Tau Mean      :  1.3802
#> Tau Variance  :  0.3463
#> Tau Deviation :  0.5885
```

The `plot()` function for this situation is different from the above two examples. It will show five plots, which are prior and posterior distribution of μ, prior and posterior distribution of *tau*, prior contour, posterior contour and two contours together.

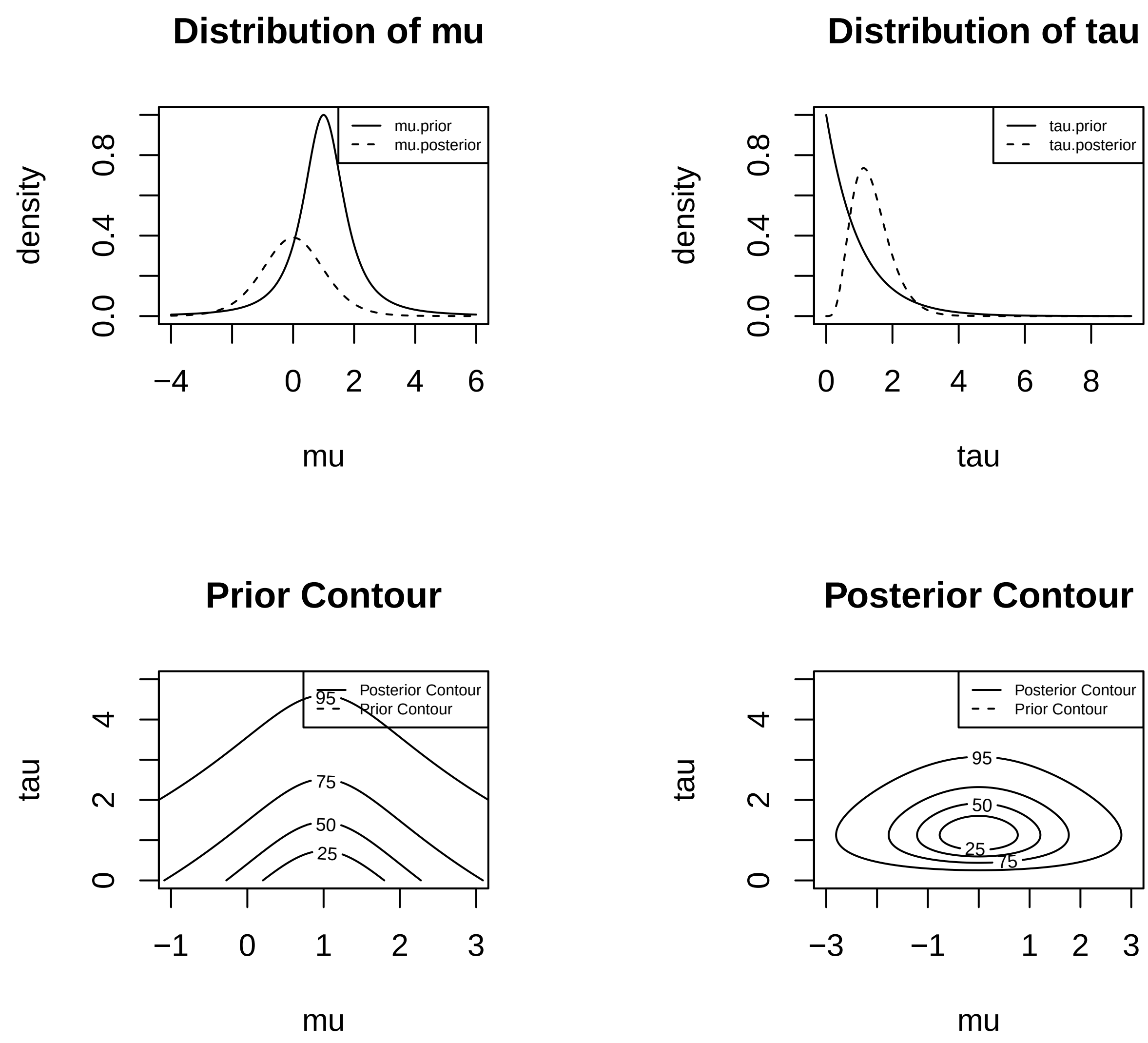
```r
## show the first plot : Prior and Posterior Distribution of mu
 plot(exmp2, which = 1, main = " Distribution of mu", cex = 0.5)
```

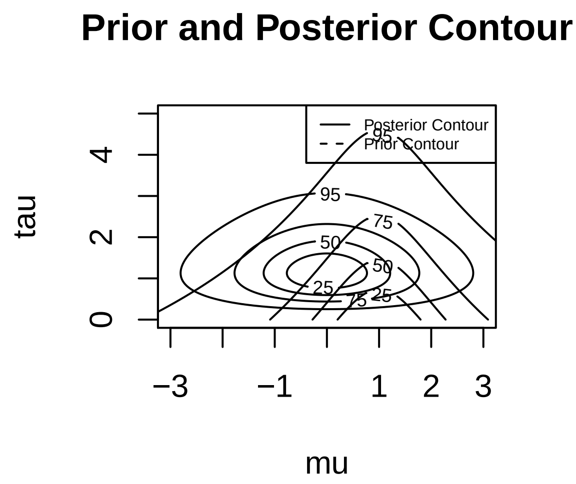
```r
## show the second plot : Prior and Posterior Distribution of tau
 plot(exmp2, which = 2, main = " Distribution of tau", cex = 0.5)

## show the third plot : Prior Contour
plot(exmp2, which = 3, main = "Prior Contour",
       xlim = c(-1,3),ylim = c(0,5), cex = 0.5)

## show the fourth plot : Posterior Contour
plot(exmp2, which = 4, main = "Posterior Contour",
           xlim = c(-3,3), ylim = c(0,5), cex = 0.5)

## show the fifth plot : Prior and Posterior Contour
 plot(exmp2, which = 5, main = "Prior and Posterior Contour",
           xlim = c(-3,3), ylim = c(0,5), cex = 0.5)
```





We can notice that the contour plot looks like a "squashed egg". Compared with the prior contour, the posterior contour, after observing the data, is much more smoothly and narrow.

Function for Bayesian Hypothesis Testing

The `normaltest()` function defines the Bayesian hypothesis testing with normal distribution. The Bayesian hypothesis testing is based on the posterior distribution $\pi(\theta \mid x)$, and the decision is to reject or accept the null hypothesis according to which decision provides the smaller losses. The loss of rejecting the null hypothesis is a times the probability of the null is true, where a is the loss due to type I error. The loss of accepting the null hypothesis is b times the probability of the null is false, where b is the loss due to type II error.

Using the Code with Example

Suppose a sample distribution $x_1, \cdots, x_{16}$ is the independent normal distribution $\mathcal{N}(\theta, 4)$, with the prior $\theta \sim \mathcal{N}(4.5, 10)$, the observed sample mean $\bar{x}$ is 5.2. Given that the losses a and b are both 1, let us utilize `normaltest()` function to get the result.

```
## generate a sample of 16 from a normal distribution with sd=4
y <- rnorm(16, mean = 5.2, sd = 4)
## the observed sample mean is 5.2
yy <- y- mean(y)+5.2
## find the posterior density
exmp1 <- normnorm(yy, m = 4.5, s = sqrt(10), sigma = 4)
#> Posterior precision      : 1.1
#> Posterior mean           : 5.1364
#> Posterior variance       : 0.9091
#> Posterior std. deviation : 0.9535

## make the hypothesis testing
normaltest(exmp1, 5, 1, 1)
#> We reject H0 and the expected posterior loss is : 0.44314
```

The testing result and expected posterior loss are the same as the outcome we calculated by the formula.

References

Bolstad, W. M. (2007). *Introduction to bayesian statistics*(2nd ed.). John Wiley & Sons.