

# STAT 4010 Bayesian Learning

TUTORIAL 1

Spring 2023

Cheuk Hin (Andy) CHENG ([Email](#) | [Homepage](#))

Di SU ([Email](#) | [Homepage](#))

## 1 Review

### 1.1 Comparison of Bayesian and Frequentist Philosophy

The differences between the two schools can be summarized in the following table.

	Bayesian	Frequentist
View on probability	Degree of belief (Subjective)	Limiting relative frequencies (Objective)
$\theta$ is a	Random variable	Unknown constant
Infer on $\theta$ by	Modeling a probability distribution	Procedures with well-defined long run frequency properties
Common methods	Bayes estimator, MAP, Bayes factor	MME, MLE, C.I.

\*This is a summary from Ch.11, All of Statistics (Wasserman 2003).

\*An insightful discussion between Frequentist and Bayesian methods is given in example 1.7 from lecture note 1.

### 1.2 Posterior Calculation

Let  $\Theta$  be the parameter space. Define the sampling density as  $f(x | \theta)$  and the prior density be  $f(\theta)$  (these are known).

- Prior predictive:  $f(x) = \int_{\Theta} f(x, \theta) d\theta = \int_{\Theta} f(x | \theta) f(\theta) d\theta$
- Posterior:  $f(\theta | x_{1:n}) \propto f(x_{1:n} | \theta) f(\theta)$
- Posterior predictive:

$$\begin{aligned} f(x_{n+1} | x_{1:n}) &= \int_{\Theta} f(\theta | x_{1:n}) f(x_{n+1} | x_{1:n}, \theta) d\theta \\ &\stackrel{*}{=} \int_{\Theta} f(\theta | x_{1:n}) f(x_{n+1} | \theta) d\theta \end{aligned}$$

\*The second equality holds if  $x_{n+1}$  and  $x_{1:n}$  are conditional independent given  $\theta$ .

**Remark 1.1.** The building blocks of Bayesian inference is the prior  $f(\theta)$  and the sampling distribution  $f(x_{1:n} | \theta)$ . They are assumed by us based on our own belief. After observing

data, inference is conducted through the following formula

$$\begin{aligned} f(\theta | x_{1:n}) &= \frac{f(x_{1:n} | \theta)f(\theta)}{f(x_{1:n})} \\ &\propto f(x_{1:n} | \theta)f(\theta) \end{aligned} \quad (1.1)$$

We call  $f(x_{1:n} | \theta)f(\theta)$  the **kernel** of the probability density function. The above formula will be used intensively through out the whole course.

**Remark 1.2.** There are two reasons why we only care about the proportional part in 1.1,

1. The denominator  $f(x_{1:n})$  is a “constant” with respect to  $\theta$ , because it doesn’t depend on  $\theta$ . On the other hand,  $f(\theta | x_{1:n})$  is the density of  $\theta$  given the data, it is not a density of  $x_{1:n}$ . Hence we can “discard”  $f(x_{1:n})$  from this density.
2. Using the following relation, once we know  $f(x_{1:n} | \theta)f(\theta)$ , we immediately know  $f(x_{1:n})$ .

$$f(x_{1:n}) = \int_{\Theta} f(x_{1:n} | \theta)f(\theta)d\theta \quad (1.2)$$

Relation 1.2 comes from the fact that  $f(\theta | x_{1:n})$  is a density function, and it satisfies that

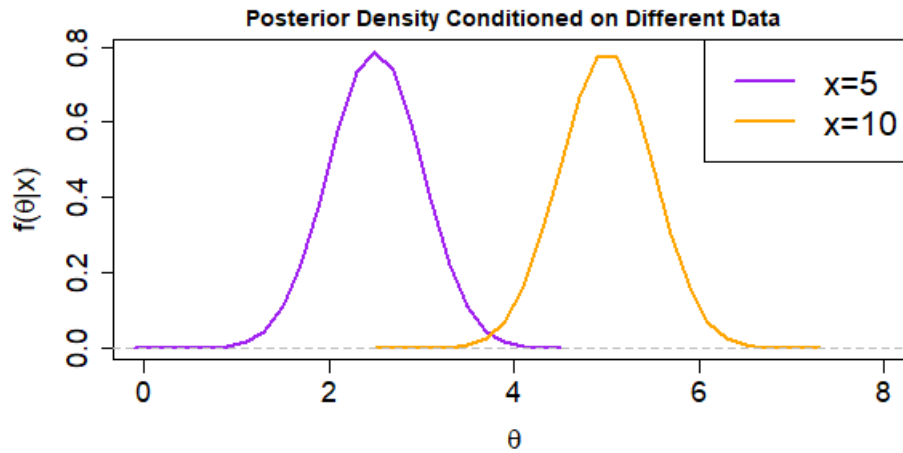
$$\int_{\Theta} f(\theta | x_{1:n})d\theta = \int_{\Theta} \frac{f(x_{1:n} | \theta)f(\theta)}{f(x_{1:n})}d\theta = 1.$$

**Remark 1.3.** The posterior  $f(\theta | x_{1:n})$  is a function of  $\theta$ , and it depends on  $x_{1:n}$  as well.

1. The joint density  $f(x_{1:n}, \theta)$  is a function of both  $x_{1:n}$  and  $\theta$ .
2. Given  $x_{1:n}$ ,  $\theta$  is a **random variable**, and  $f(\theta | x_{1:n})$  is the **density function of  $\theta$**  only. It also depends on  $x_{1:n}$  because  $f(\theta | X_{1:n} = x_{1:n}) = f(X_{1:n} = x_{1:n}, \theta)/f_X(X_{1:n} = x_{1:n})$ .

**⚙ Experiment:** Revisit Example 1.3. The model is  $[x | \theta] \stackrel{\text{i.i.d.}}{\sim} N(\theta, \sigma^2), \theta \sim N(\theta_0, \tau_0^2)$ . Assume  $\sigma^2 = 1, \theta_0 = 0, \tau_0^2 = 1$ . Then  $[\theta | x] \sim N(x/2, 1/2)$ .

- When  $x = 5$ ,  $[\theta | x = 5] \sim N(5/2, 1/2)$ . It is represented in the following figure in purple.
- When  $x = 10$ ,  $[\theta | x = 10] \sim N(5, 1/2)$ . It is represented in the following figure in orange.



🔗 **Takeaway:** The data decides which curve the posterior density is.

**Example 1.1.** Let  $x$  denote the random variable of interest. The following table summarizes the density of common distribution (more details are given in chapter 2 lecture note).

Distribution	Kernel (in red)
$N(\mu, \sigma^2)$	$(2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$
$\theta \text{Exp}(1)$	$\frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{\{x>0\}}$
$\text{Ga}(\alpha)/\beta$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\{x>0\}}$
$\beta/\text{Ga}(\alpha)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} \mathbb{1}_{\{x>0\}}$
$\chi_k^2$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \mathbb{1}_{\{x>0\}}$
$\text{Po}(\theta)$	$e^{-\theta} \frac{\theta^x}{x!} \mathbb{1}_{\{x=0,1,\dots\}}$
$\text{Bin}(m, \theta)$	$\binom{m}{x} \theta^x (1 - \theta)^{m-x} \mathbb{1}_{\{x=0,1,\dots\}}$
$\text{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1 - x)^{\beta-1} \mathbb{1}_{\{x \in (0,1)\}}$

Table 1: Density of Common Distribution

**Example 1.2.** Let  $\Theta$  be the support of  $\theta$  and the density of  $\theta$  be  $f(\theta) = cK(\theta)$ , where  $K(\theta)$  is the probability kernel and  $c$  is a constant (with respect to  $\theta$ ). Then, we have,

$$1 = \int_{\Theta} f(\theta) d\theta \Rightarrow \frac{1}{c} = \int_{\Theta} K(\theta) d\theta.$$

For example, if  $\theta \sim \text{Ga}(\alpha)/\beta$ , then  $c = \beta^\alpha/\Gamma(\alpha)$  and  $K(\theta) = \theta^{\alpha-1} e^{-\beta\theta} \mathbb{1}(\theta > 0)$ . Therefore,

$$\int_0^\infty \theta^{\alpha-1} e^{-\beta\theta} \mathbb{1}(\theta > 0) d\theta = \frac{1}{c} = \frac{\Gamma(\alpha)}{\beta^\alpha},$$

where  $\Gamma(\alpha) = (\alpha - 1)!$ .

And if  $\theta \sim \text{Beta}(\alpha, \beta)$ , then  $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ ,  $K(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$ , and

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

🔗 **Takeaway:** Some integrals can be found without direct calculation. Try to find the integral of the kernels of other common distributions.

**Example 1.3.** (Poisson-Gamma Model) Consider the following model,

$$\begin{aligned} x_{1:n} | \theta &\stackrel{\text{iid}}{\sim} \text{Po}(\theta) \\ \theta &\sim \text{Ga}(\alpha)/\beta \end{aligned}$$

1. Find the posterior.
2. Find the prior predictive density, posterior distribution given  $x_{1:n}$  and also the posterior predictive.
3. Let  $\alpha = 12, \beta = 2$ . Given  $x_{1:4} = (6, 7, 5, 4)$ .
  - (a) Produce plots of prior and posterior  $[\theta | x_{1:n}]$  PDF on the same graph for  $n = 1, \dots, 4$ . Comment.
  - (b) Produce plots of prior predictive and posterior predictive  $[x_{n+1} | x_{1:n}]$  PDF on the same graph for  $n = 1, \dots, 4$ . Comment.

**SOLUTION:**

1. For the posterior,

$$\begin{aligned} f(\theta | x_{1:n}) &\propto f(\theta)f(x_{1:n} | \theta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \mathbb{1}_{\{\theta>0\}} \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!} \mathbb{1}_{\{x_i=0,1,\dots\}} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \mathbb{1}_{\{\theta>0\}} e^{-n\theta} \theta^{\sum_{i=1}^n x_i} \prod_{i=1}^n \frac{1}{x_i!} \mathbb{1}_{\{x_i=0,1,\dots\}} \\ &\propto \theta^{\alpha+\sum_{i=1}^n x_i-1} e^{-\theta(\beta+n)} \mathbb{1}_{\{\theta>0\}}. \end{aligned}$$

Therefore,  $\theta | x_{1:n}$  follows  $\text{Ga}(\alpha_n)/\beta_n$ , where  $\alpha_n = \alpha + \sum_{i=1}^n x_i$  and  $\beta_n = \beta + n$ .

2. For prior predictive, we have

$$\begin{aligned} f(x) &= \int_0^\infty f(x | \theta)f(\theta)d\theta \\ &= \int_0^\infty \frac{e^{-\theta} \theta^x}{x!} \mathbb{1}_{\{x=0,1,\dots\}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\ &= \frac{\beta^\alpha}{x! \Gamma(\alpha)} \mathbb{1}_{\{x=0,1,\dots\}} \int_0^\infty \theta^{\alpha+x-1} e^{-\theta(\beta+1)} d\theta \\ &= \frac{\beta^\alpha}{x! \Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+x)}{(\beta+1)^{\alpha+x}} \mathbb{1}_{\{x=0,1,\dots\}} \end{aligned}$$

Finally for the posterior predictive, since  $x_{n+1}$  and  $x_{1:n}$  are conditionally independent given  $\theta$ , we have by similar calculation in prior predictive

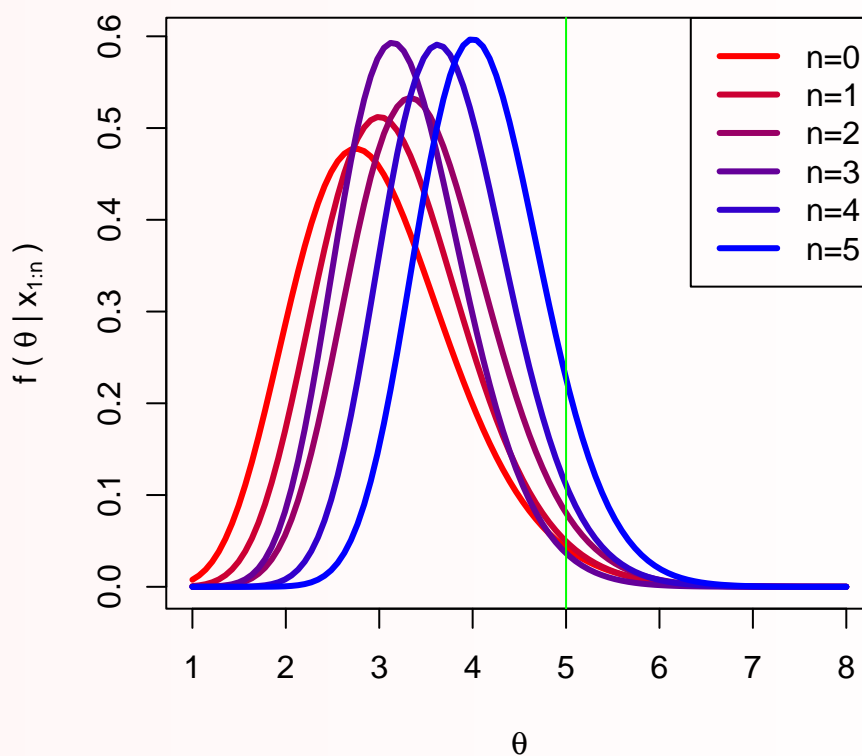
$$\begin{aligned} f(x_{n+1} | x_{1:n}) &= \int_0^\infty f(\theta | x_{1:n}) f(x_{n+1} | \theta) d\theta \\ &= \frac{\beta_n^{\alpha_n}}{x_{n+1}! \Gamma(\alpha_n)} \cdot \frac{\Gamma(\alpha_n + x_{n+1})}{(\beta_n + 1)^{\alpha_n + x_{n+1}}} \mathbb{1}_{\{x_{n+1}=0,1,\dots\}} \end{aligned}$$

Notice that we can replace the parameters  $\alpha, \beta$  in the prior predictive by  $\alpha_n, \beta_n$  to quickly get the posterior predictive.

3. As a remark, we compute the log-density first and then take exponential for numerical stability; see Experiment 1.4 for more discussion. Note that the true DGP is generated as followed,

```
1 ## true DGP, pretend you do not know it
2 set.seed(999)
3 theta = 5
4 x = rpois(5, theta)
5 x
6 #[1] 4 5 2 7 7
```

- (a) It is clear that given more data, the posterior changes gradually. The mode moves towards the true  $\theta = 5$ . Intuitively, our belief is corrected by more observations.



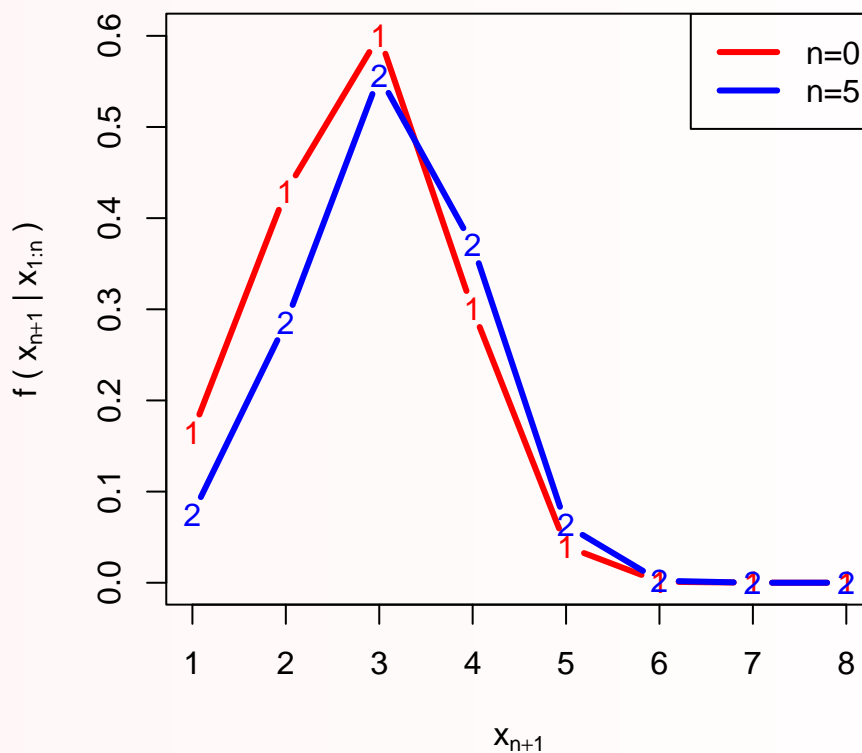
```
1 n = length(x)
```

```

2 alpha = 12
3 beta = 4
4
5
6 # Step 1: compute density for a grid of theta
7 theta=seq(1, 8,length.out=100)
8 post=array(NA, dim=c(100,n+1))
9 ##prior density
10 post[,1] = dgamma(theta,shape = alpha,rate = beta)
11 #posterior density
12 for(i in 1:length(x)){
13   alpha_n = alpha+sum(x[1:i])
14   beta_n = beta+i
15   post[,i+1] = dgamma(theta,shape = alpha_n,rate = beta_n)
16 }
17
18 # Step 2: plot
19 col= colorRampPalette(c("red", "blue"))(ncol(post))
20 windows(height=5,width = 5)
21 matplot(theta, post, col=col, type="l", lwd=3, lty=1, xaxt='n',
22         xlab=expression(theta),ylab=bquote(f~"("~theta~"|"~x[1:n]~") "
23         ))
24 axis(side = 1, at = 1:8)
25 abline(v = 5,col = 'green')
26 legend("topright", paste0("n=",0:length(x)), lty=1, lwd=3, col=col)

```

- (b) The posterior predictive is different from the prior predictive. This is reflected in the difference between the prior and posterior.



```

1  dprior_pred <- function(y,alpha=4, beta=2){
2  n = length(y)
3  lnf = array(NA,n)
4  for (i in 1:n) {
5    lnf[i] = alpha*log(beta)-sum(log(gamma(1:(y[i]))))-log(gamma(
6      alpha))+
7      log(gamma(alpha+y[i]))-(alpha+y[i])*log(beta+1)
8    }
9  return(exp(lnf))
10 }
11 dpost_pred <- function(y,d,alpha=4, beta=2){ ##d is array of data
12 alphaN = alpha + sum(d)
13 betaN = beta+length(d)
14 return(dprior_pred(y,alphaN,betaN))
15 }
16
17 # Step 2 plot
18 y=1:8
19 pred_den = array(NA,c(8,2))
20 pred_den[,1] = dprior_pred(y,alpha,beta)
21 pred_den[,2] = dpost_pred(y,x,alpha,beta)
22
23
24 col= colorRampPalette(c("red","blue"))(ncol(pred_den))
25 windows(height=5,width = 5)
26 matplot(y, pred_den, col=col, type="b", lwd=3, lty=1, xaxt = 'n',
27         xlab=expression(x[n+1]),ylab=bquote(f~"("~x[n+1]~"|~"~x[1:n]~"
28         ~)"))
29 axis(side = 1, at = y)
30 legend("topright", paste0("n=",c(0,5)), lty=1, lwd=3, col=col)

```

### 1.3 Commonly Used Models and Representation

**Example 1.4.** The following distributions will be frequently encountered throughout the course.

Distribution	Representation	
$N(\mu, \sigma^2)$	$x = \mu + \sigma z$	$z \sim N(0, 1)$
$\theta \text{Exp}(1)$	$x = \theta z$	$z \sim \text{Exp}(1)$
$\text{Ga}(\alpha)/\beta$	$x = \frac{1}{\beta} \sum_{i=1}^{\alpha} z_i$	$z_i \stackrel{\text{iid}}{\sim} \text{Ga}(1) = \text{Exp}(1)$
$\beta/\text{Ga}(\alpha)$	$x = 1/z$	$z \sim \text{Ga}(\alpha)/\beta$
$\chi_k^2$	$x = \sum_{i=1}^k z_i$	$z_i \stackrel{\text{iid}}{\sim} \chi_1^2$
$\text{Po}(\theta)$	$x = \sum_{i=1}^{\theta} z_i$	$z_i \stackrel{\text{iid}}{\sim} \text{Po}(1)$
$\text{Bin}(m, \theta)$	$x = \sum_{i=1}^m z_i$	$z_i \stackrel{\text{iid}}{\sim} \text{Bern}(\theta)$
$\text{Beta}(\alpha, \beta)$	$x = \frac{z_{\alpha}}{z_{\alpha} + z_{\beta}}$	$z_j \stackrel{\text{iid}}{\sim} \text{Ga}(j)$

Table 2: Representation of Common Distribution

The following examples show how representation can be a helpful tool.

**Example 1.5.** Let  $X, Y \sim \text{Ga}(1)$ . Find the distribution of  $Z = \frac{X}{X+Y}$ .

**SOLUTION:** By representation,

$$Z \stackrel{d}{=} \frac{\text{Ga}(1)}{\text{Ga}(1) + \text{Ga}(1)} \stackrel{d}{=} \text{Beta}(1, 1).$$

That is  $Z \sim \text{Beta}(1, 1)$ .

**Takeaway:** Representation technique helps us to determine the distributions of random variables.

**Example 1.6.** (Optional\*) Let  $[X | Y] \sim N(Y, Y^2)$  and  $Y \sim \text{Unif}(0, 1)$ . Prove that  $X/Y \perp\!\!\!\perp Y$ .

**SOLUTION:** We can jointly represent  $(X, Y)$  as

$$\begin{cases} X = Y + YZ \\ Y = U \end{cases} \Rightarrow \begin{cases} X/Y = 1 + Z \\ Y = U \end{cases}$$

where  $Z \sim N(0, 1)$  and  $U \sim \text{Unif}(0, 1)$  are independent. Note that

- $X/Y$  depends only on  $Z$  and  $Y$  depends only on  $U$ , and
- $Z$  and  $U$  are independent.

We conclude that  $X/Y$  and  $Y$  are independent.

\*This example is from STAT4003 Lecture Note 1 (Keith 2020).



## 1.4 R Tips

**Example 1.7.** There are several methods/tricks that will be helpful in this course.

1. Four density-related functions in R. Take the normal distribution for an example.
  - (a) The function `rnorm(nRep, mean, sd)` samples from  $N(\text{mean}, \text{sd})$  for `nRep` times.
  - (b) The function `qnorm(prob, mean, sd)` returns the  $100\% \times \text{prob}$ -th quantile of  $N(\text{mean}, \text{sd})$ .
  - (c) The function `pnorm(q, mean, sd)` returns the value of  $F(q)$  where  $F$  is the CDF of  $N(\text{mean}, \text{sd})$ .
  - (d) The function `dnorm(q, mean, sd)` returns the value of  $f(q)$  where  $f$  is the PDF of  $N(\text{mean}, \text{sd})$ .

The functions for other distributions are similarly used by changing `norm` to `binom`, `beta`, `exp`, `gamma`, `t`, `chisq`, etc., and specifying the corresponding parameters as arguments.


2. Usually, we want to find the expectation or variance of a complicated variables. Analytical results may be difficult to derive, instead, we can use Monte Carlo method to find them based on a large number of samples from that distribution.
3. Remember to use `set.seed(4010)` to generate reproducible samples. You can change 4010 to other numbers. It will help you to debug your codes and to justify your results.
4. To debug, comment out your codes section by section and run again. If the bug disappears, then the bugs are from the section you just commented out.
5. If you want to achieve something in R but have no clue at all, you can always find references by Googling it. The best way to succeed in R programming is to practice.

### Experiment:

```

1  #sampling standard normal
2  set.seed(4010)
3  x = rnorm(10^4, 0, 1) #x is a sample of 10^4 IID standard normal random
   variables
4  z = qnorm(0.975, 0, 1) # z = 1.959964
5  p = pnorm(1.96, 0, 1) # p = 0.9750021
6  d = dnorm(0, 0, 1) # d = 0.3989423
7
8  # using Monte Carlo to find quantities related to a certain random
   variable
9  expectation = mean(x) # expectation = -0.0100194
10 variance = var(x) # variance = 1.006437
11 quart = quantile(x, 0.25) # quart = -0.6778637
12 prob = mean(x < 1.96) # prob = 0.975



```

 **Experiment:** In previous example, we compute log-density first. Here is a motivating example. How to compute  $\frac{\exp(4010)}{10^{1585}\Gamma(100)}$  using R?

```
1 ## standard but fail
2 exp(4010) / (gamma(100) * (10^1585))
3 # [1] NaN
4
5 ## take log and then take exp
6 exp(4010 - log(gamma(100)) - 1585 * log(10))
7 # [1] 3.555239
```

## 2 Remarks on Assignment 1

Please check the [platform](#) first if you have questions.

-  Remember to write indicators in the density.
-  Make use of the fact that  $\int f(x)dx = 1$  if  $f(\cdot)$  is a probability density function.