Log transform

Xueyuan Li

2022-10-03

Introduction

It is common in the analysis of biological data to log transform data representing concentrations or data representing dose response.

In this assignment, I will write a blog post to answer a series of questions related to the transformation of data. I will use both analytic methods and simulation to answer the questions.

Method

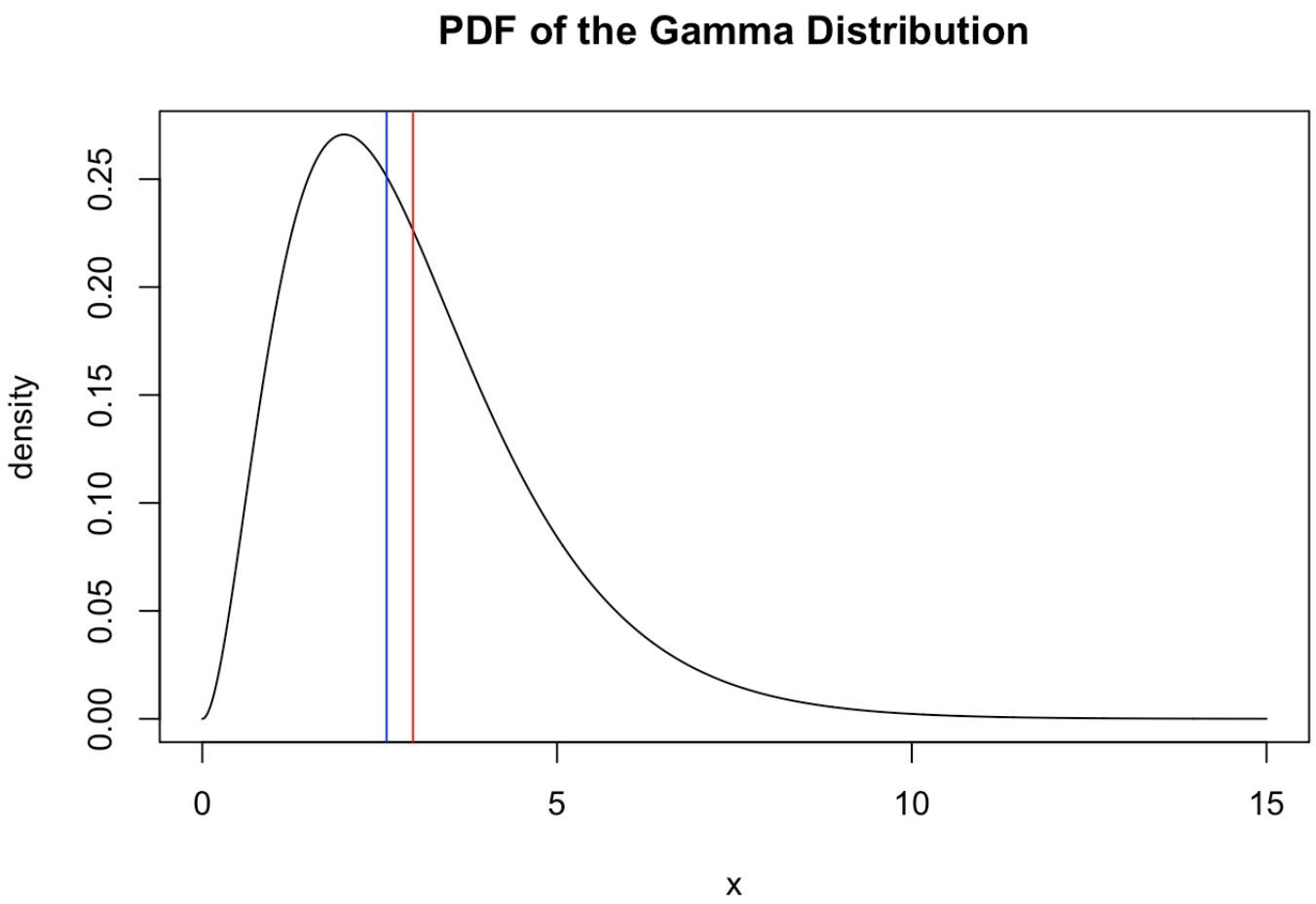
Part 1

Distribution 1

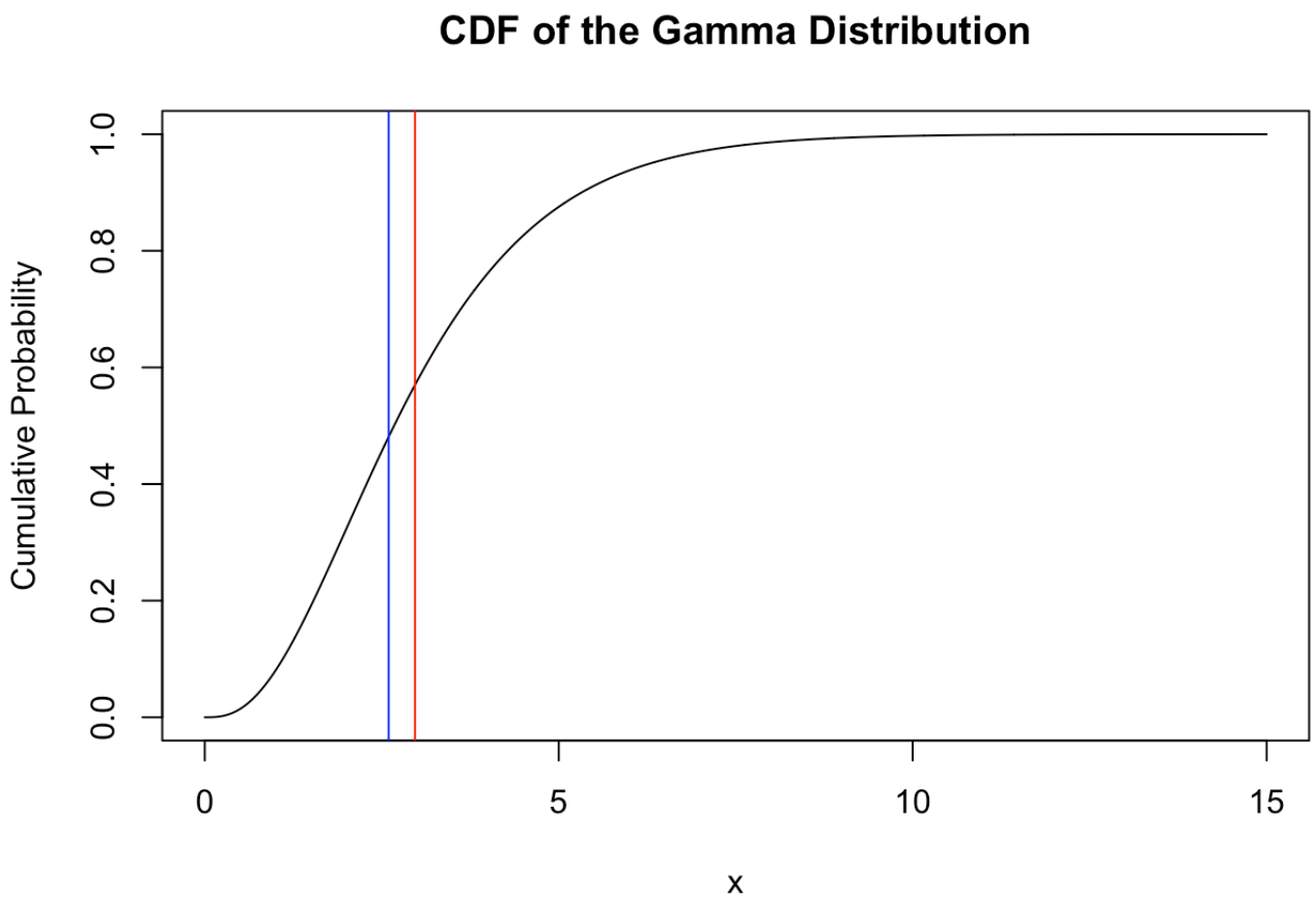
X~GAMMA(shape= 3, scale= 1)

For each distribution below, generate a figure of the PDF and CDF.Mark the mean and median in the figure.

```
#PDF
#set the range of the x
x=seq(0,15,by=0.01)
dat = dgamma(x,shape=3,scale=1)
#data.frame(x,dat)
#plot(x,dat,type="l",main="PDF of the Gamma Distribution",ylab="density")

#get mean and median
#method 2
rand_gamma = rgamma(1000,shape=3, scale=1)
mean_gamma<-mean(rand_gamma)
median_gamma<-median(rand_gamma)
plot(x,dat,type="l",main="PDF of the Gamma Distribution",ylab="density")
abline(v=mean_gamma,col="red")
abline(v=median_gamma,col="blue")
```

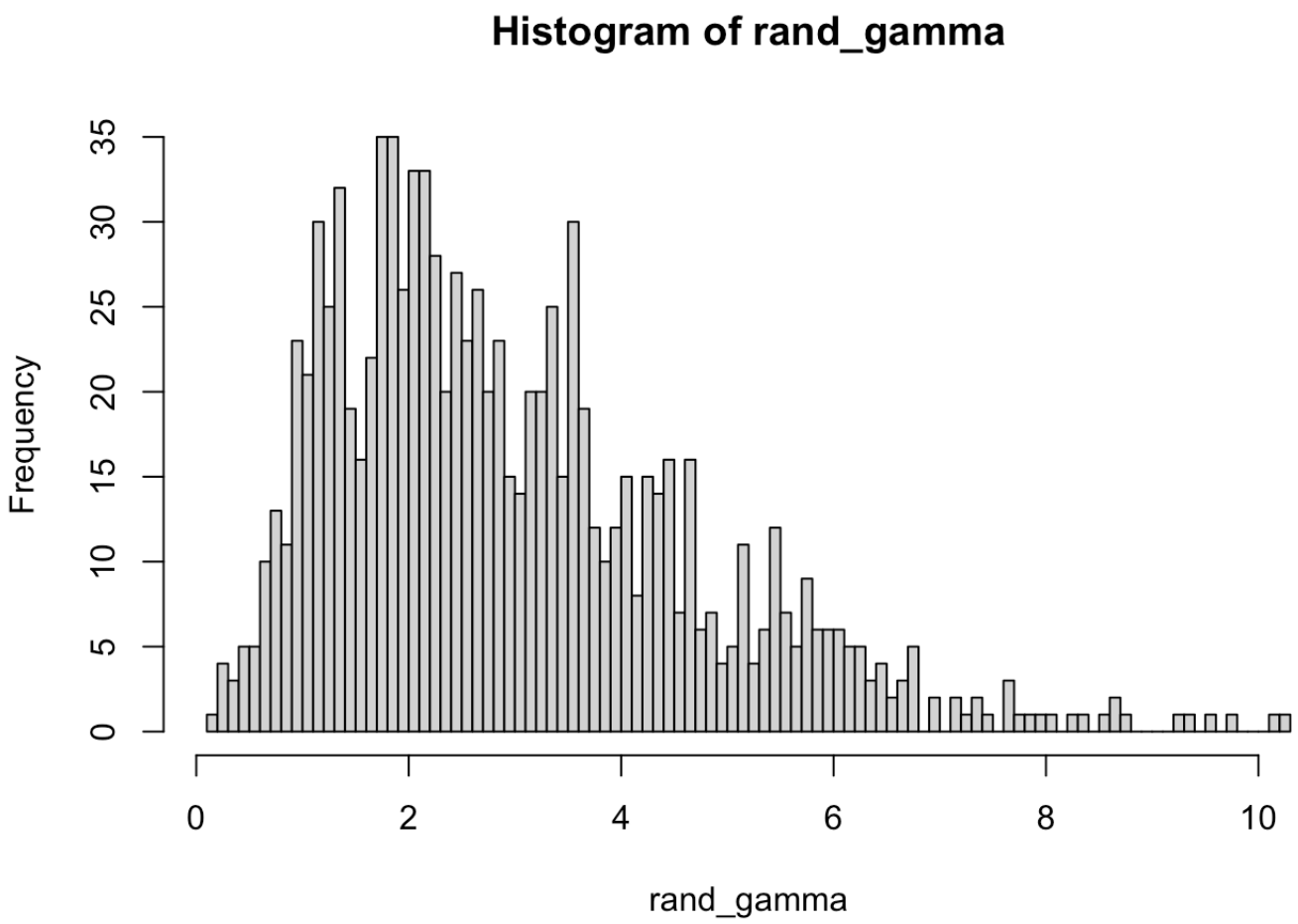



```
#CDF
dat1=pgamma(x,shape=3,scale=1)
plot(x,dat1,main="CDF of the Gamma Distribution",ylab="Cumulative Probability",type="
l")
abline(v=mean_gamma,col="red")
abline(v=median_gamma,col="blue")
```

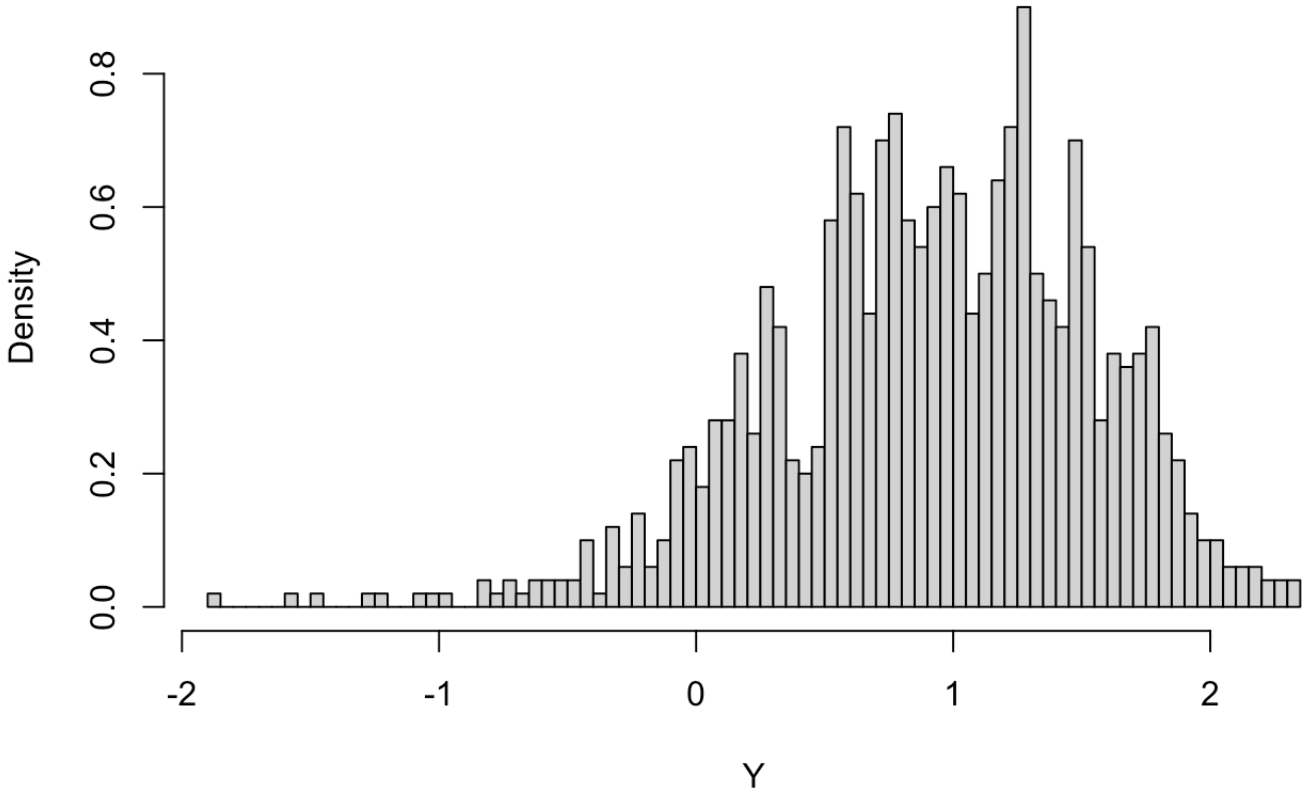


For each distribution below,generate a figure of the PDF and CDF of the transformation Y = log(X) random variable.Mark the mean and median in the figure.You may use simulation or analytic methods in order find the PDF and CDF of the transformation.

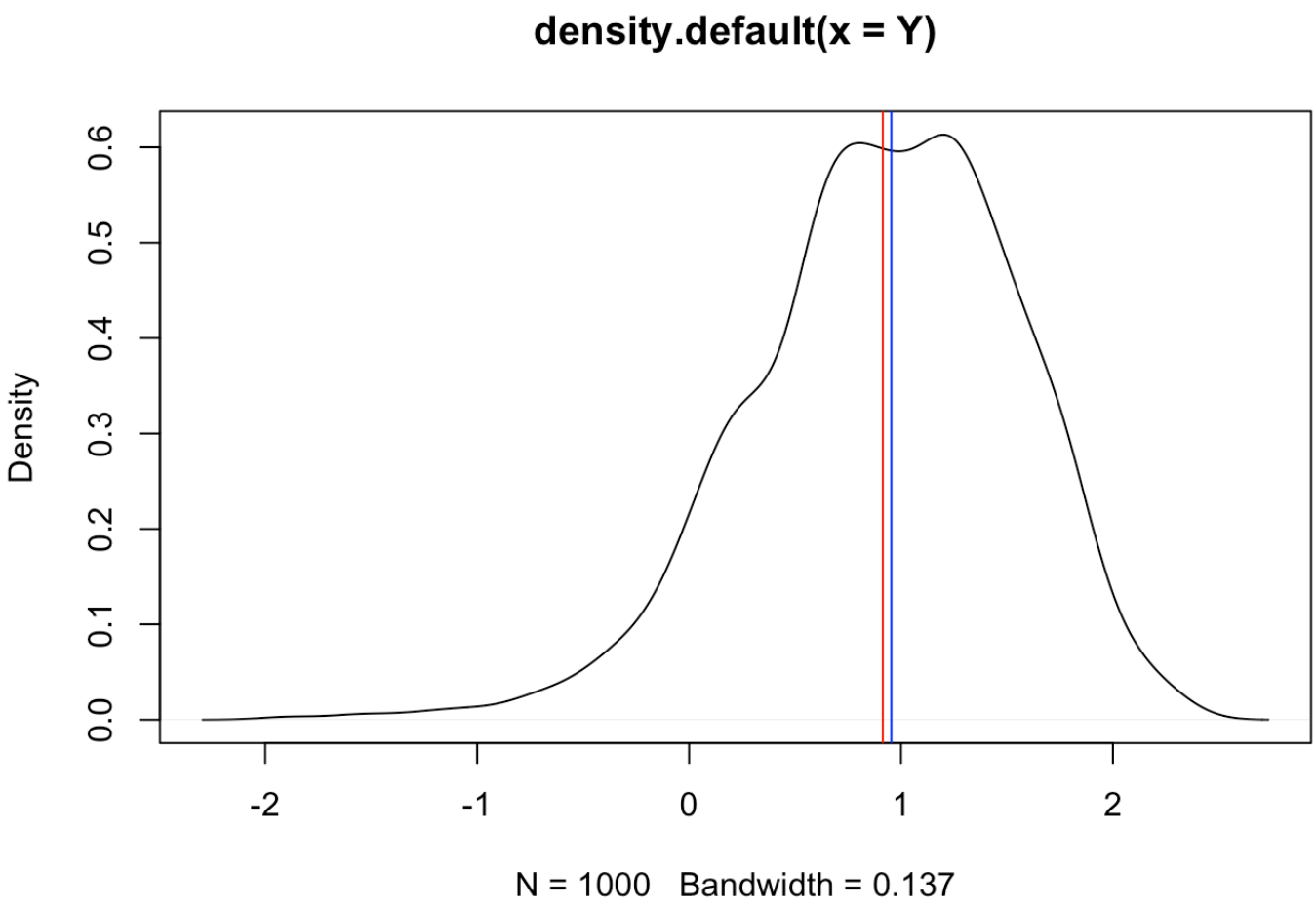
```
Y=log(rand_gamma) #new value

# for PDF, either hist() or density()
hist(rand_gamma,breaks=100)
```



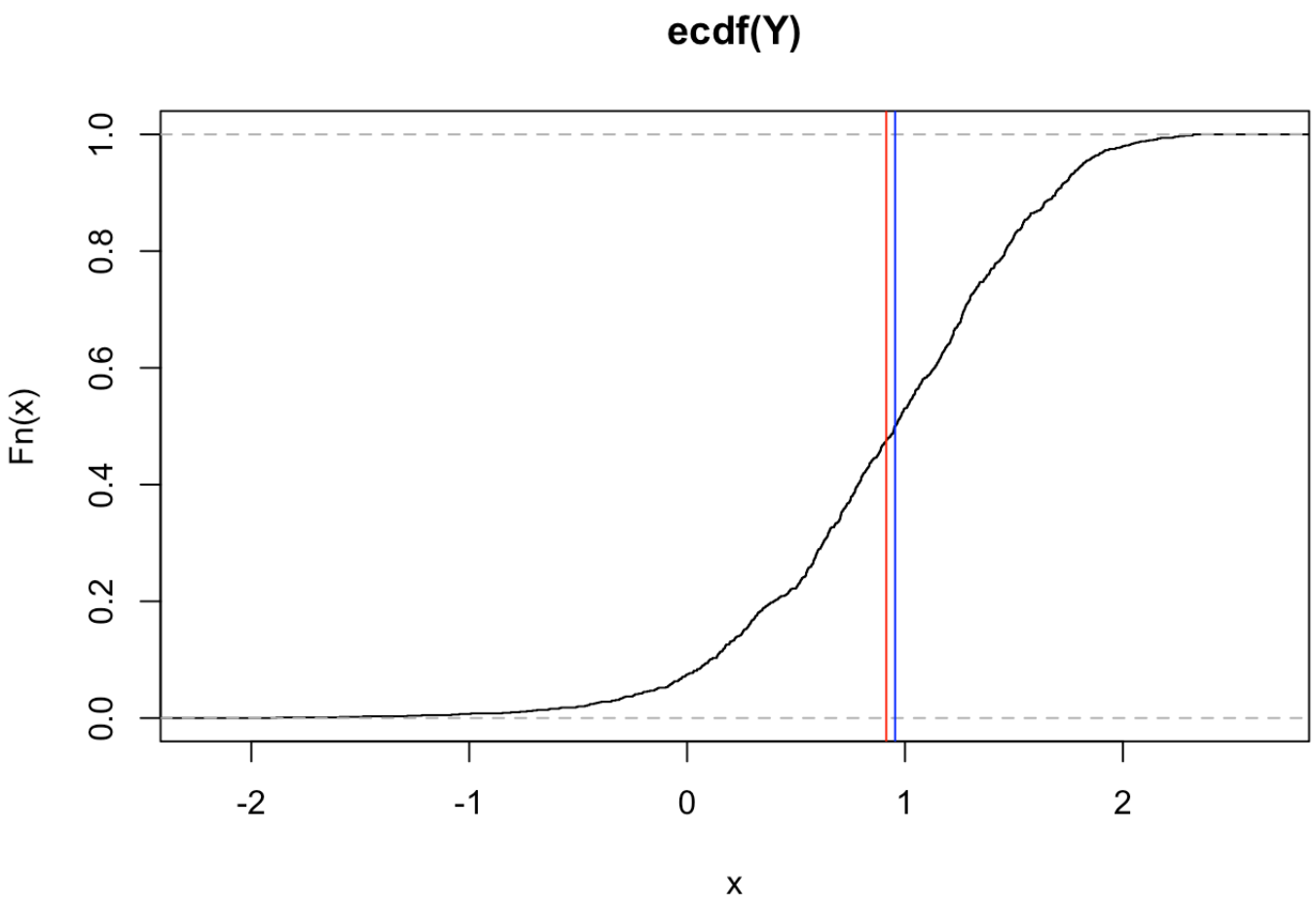

```
hist(Y,breaks=100,freq=FALSE) #lower transformation
```




```
mean_log_gamma=mean(Y)
median_log_gamma=median(Y)

Z=density(Y)
plot(Z) #the graph is very same as the Y
abline(v=mean_log_gamma,col="red")
abline(v=median_log_gamma,col="blue")
```




```
#CDF
y_cdf=ecdf(Y)
plot(y_cdf)
abline(v=mean_log_gamma,col="red")
abline(v=median_log_gamma,col="blue")
```

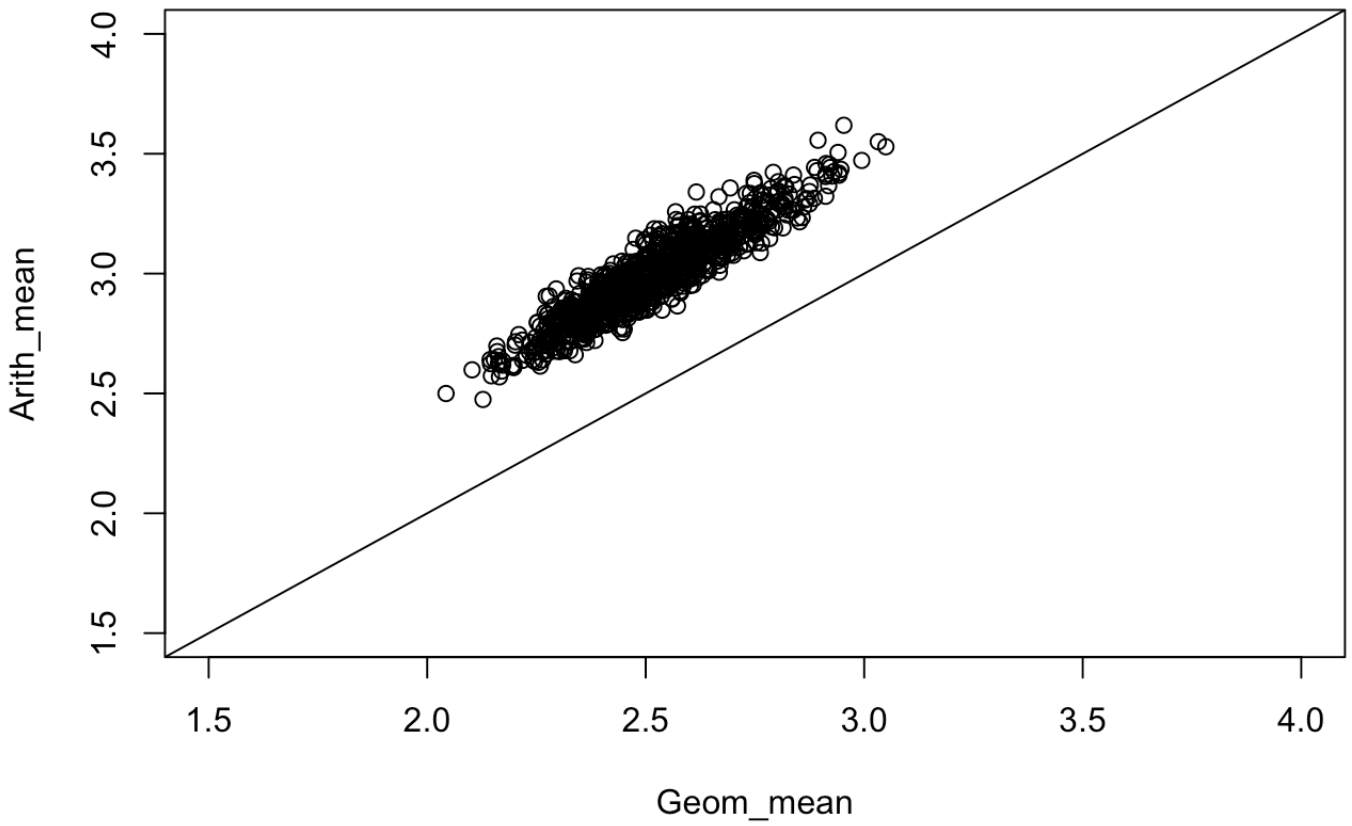


For each of the distributions below,generate 1000 samples of size 100.For each sample, calculate the geometric and arithmetic mean. Generate a scatter plot of the geometric and arithmetic sample means. Add the line of identify as a reference line.

```
#geometric mean= exp(mean(log(data))), used in computing rate of return in finance
#arithmetic mean= mean(log(data))
#arithmetic  mean>g

Gamma_sample=matrix(NA,ncol=100, nrow=1000)
for(k in 1:1000){
  Gamma_sample[k,]=rgamma(100,shape=3, scale=1)  #k row

}
dim(Gamma_sample)
```

```
## [1] 1000  100
```

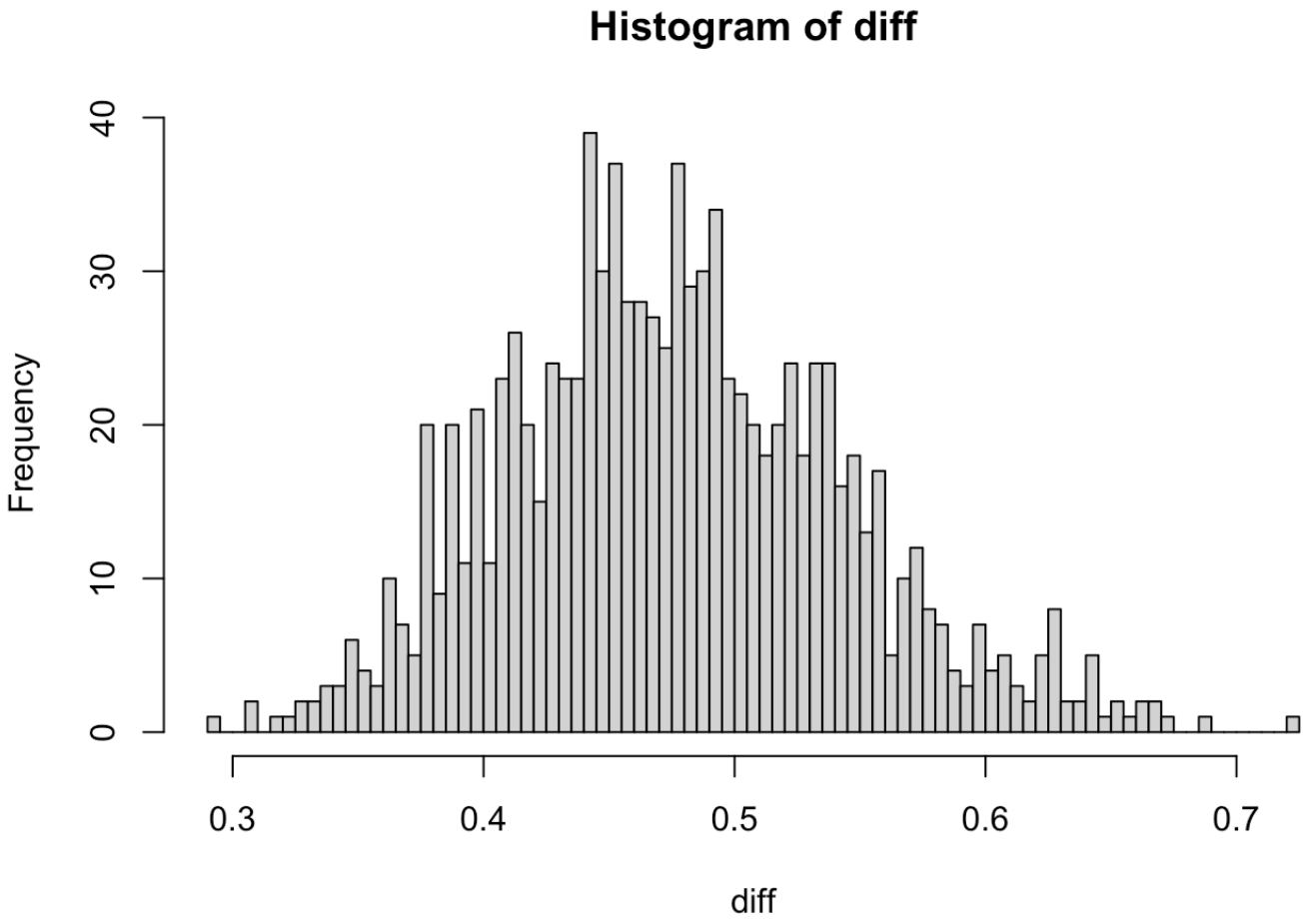
```
#Arith_mean and Geom_mean should have the same dimension

Arith_mean=rowMeans(Gamma_sample)

Geom_mean=exp(rowMeans(log(Gamma_sample)))

plot(Geom_mean,Arith_mean, xlim=c(1.5,4),ylim=c(1.5,4))
abline(c(0,1)) #identity line
```




```
#see from the high density plot : if Geom_mean is 2.5, the Arith_mean is greater than
2.5, because it is a wrong here.
#Arith_mean is all raise greater than Geom_mean
```

Generate a histogram of the difference between the arithmetic mean and the geometric mean.

```
diff= Arith_mean-Geom_mean
hist(diff, breaks=100)
```




```
#all the values are greater than 0, so means all the Arith_mean is greater than Geom_
mean for this example.
```

Analysis:

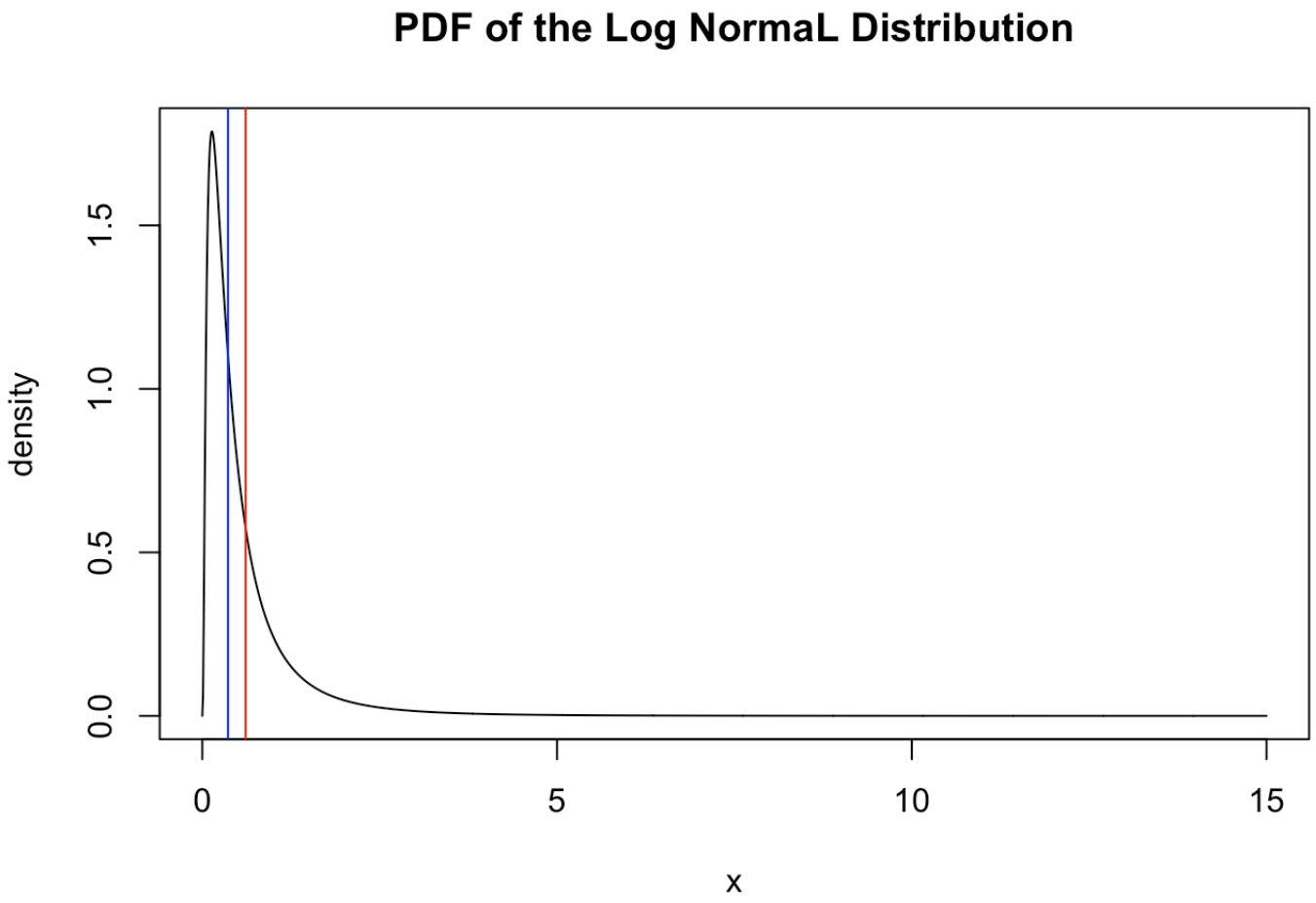
See from the high density plot : if Geom_mean is 2.5, the Arith_mean is greater than 2.5. So Arith_mean is all raise greater than Geom_mean in this example

Distribution 2

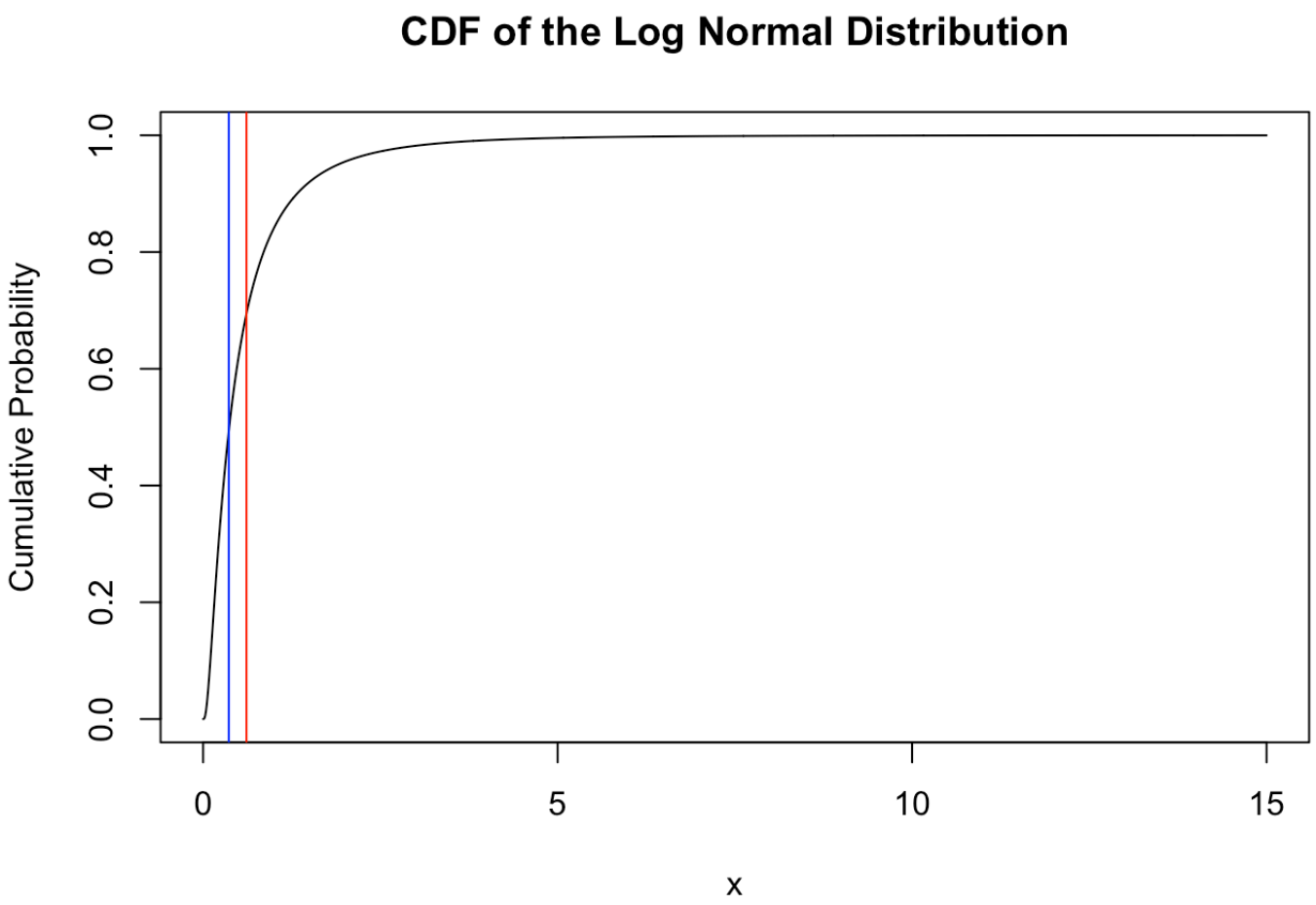
X~LOG NORMAL(μ= -1, σ= 1)

For each distribution below, generate a figure of the PDF and CDF.Mark the mean and median in the figure.

```
#PDF
#set the range of the x
x=seq(0,15,by=0.01)
dat = dlnorm(x,meanlog = -1, sdlog = 1, log = FALSE)
#data.frame(x,dat)
#plot(x,dat,type="l",main="PDF of the Log Normal Distribution",ylab="density")

#get mean and median
#method 2
rand_lognormal = rlnorm(1000, meanlog = -1, sdlog = 1)
mean_lognormal<-mean(rand_lognormal)
median_lognormal<-median(rand_lognormal)
plot(x,dat,type="l",main="PDF of the Log NormaL Distribution",ylab="density")
abline(v=mean_lognormal,col="red")
abline(v=median_lognormal,col="blue")
```

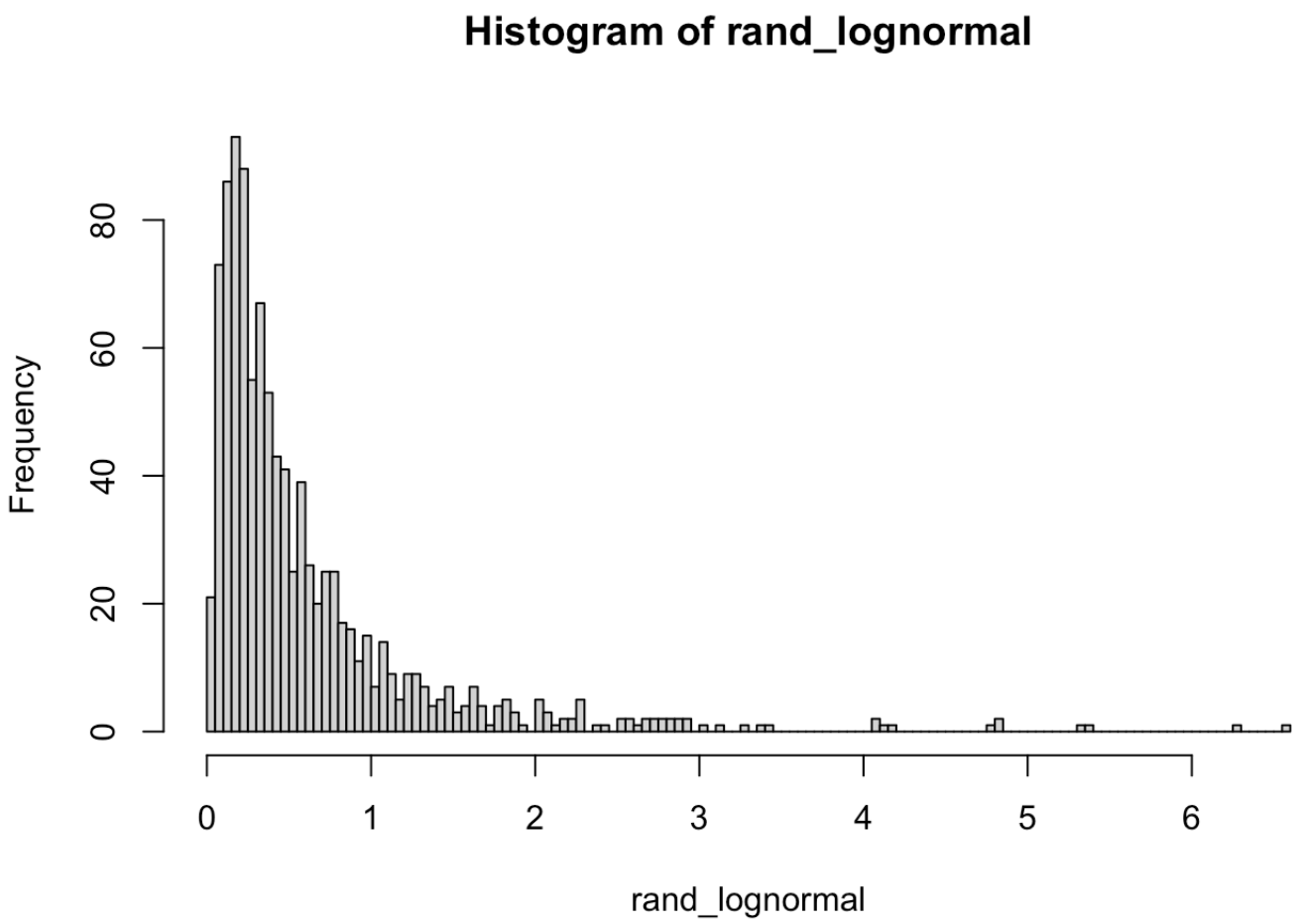


```
#CDF
dat1=plnorm(x,meanlog = -1, sdlog = 1)
plot(x,dat1,main="CDF of the Log Normal Distribution",ylab="Cumulative Probability",t
ype="l")
abline(v=mean_lognormal,col="red")
abline(v=median_lognormal,col="blue")
```

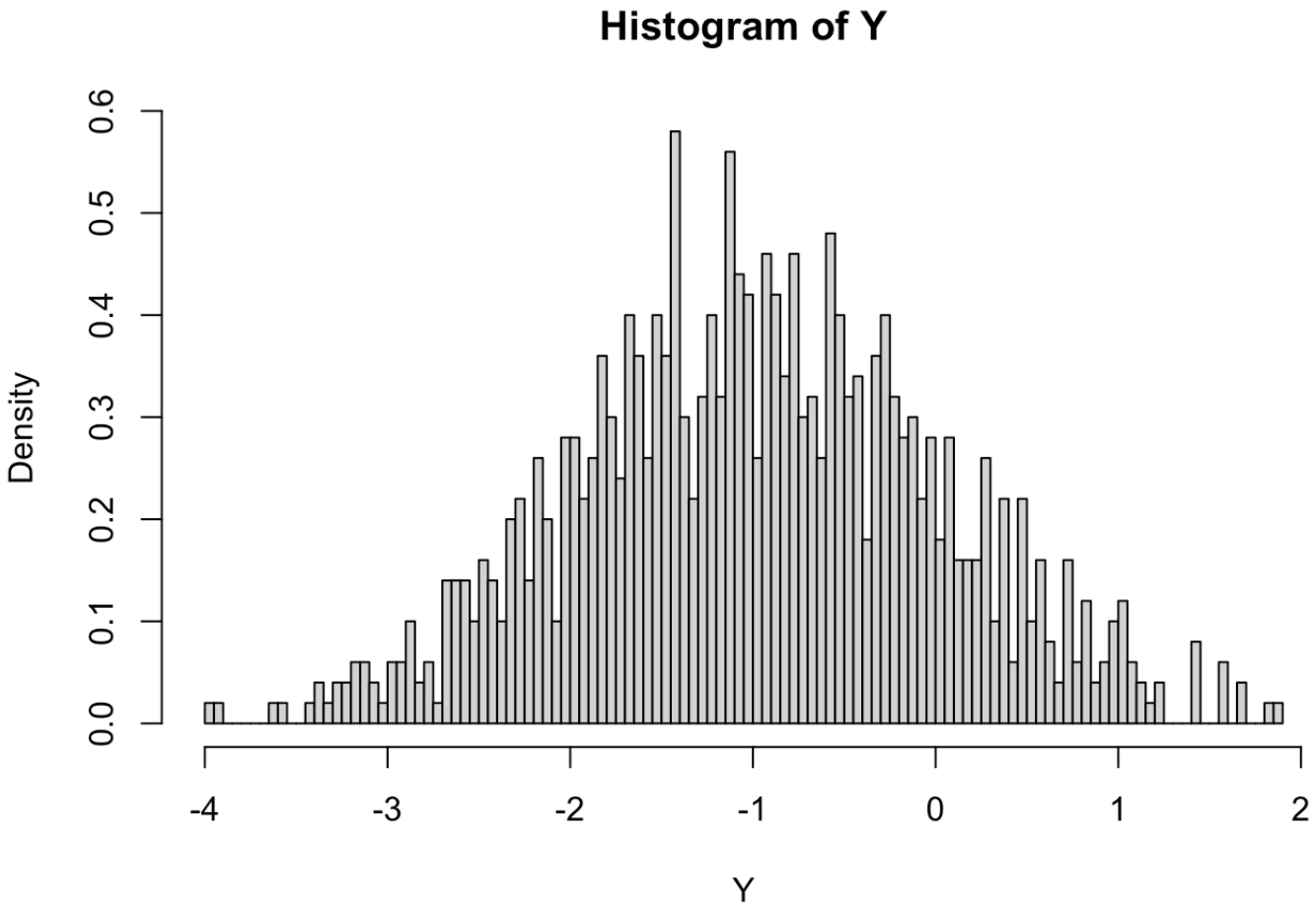


For each distribution below,generate a figure of the PDF and CDF of the transformation Y=log(X) random variable.Mark the mean and median in the figure.You may use simulation or analytic methods in order find the PDF and CDF of the transformation.

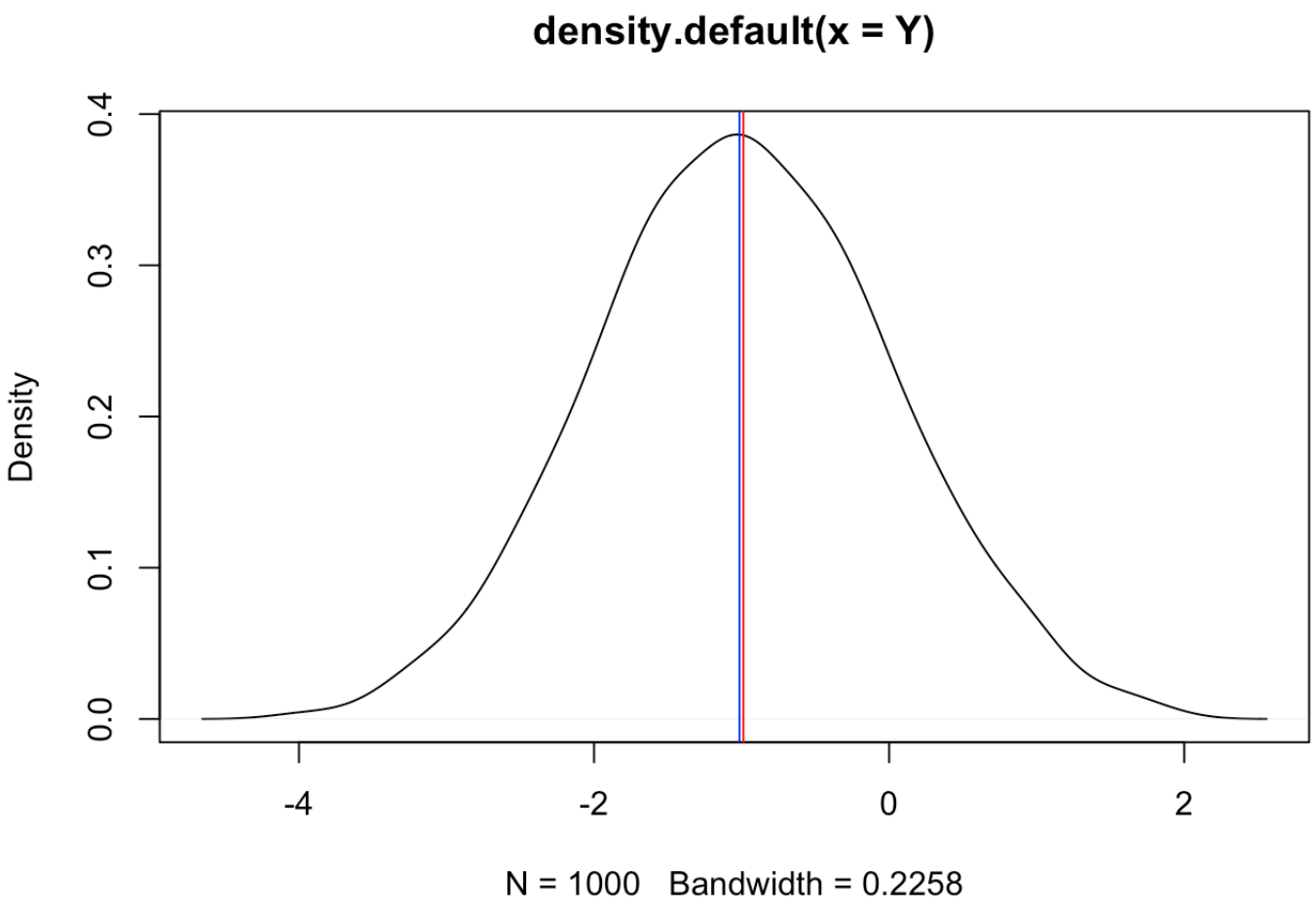
```
Y=log(rand_lognormal) #new value

# for PDF, either hist() or density()
hist(rand_lognormal,breaks=100)
```



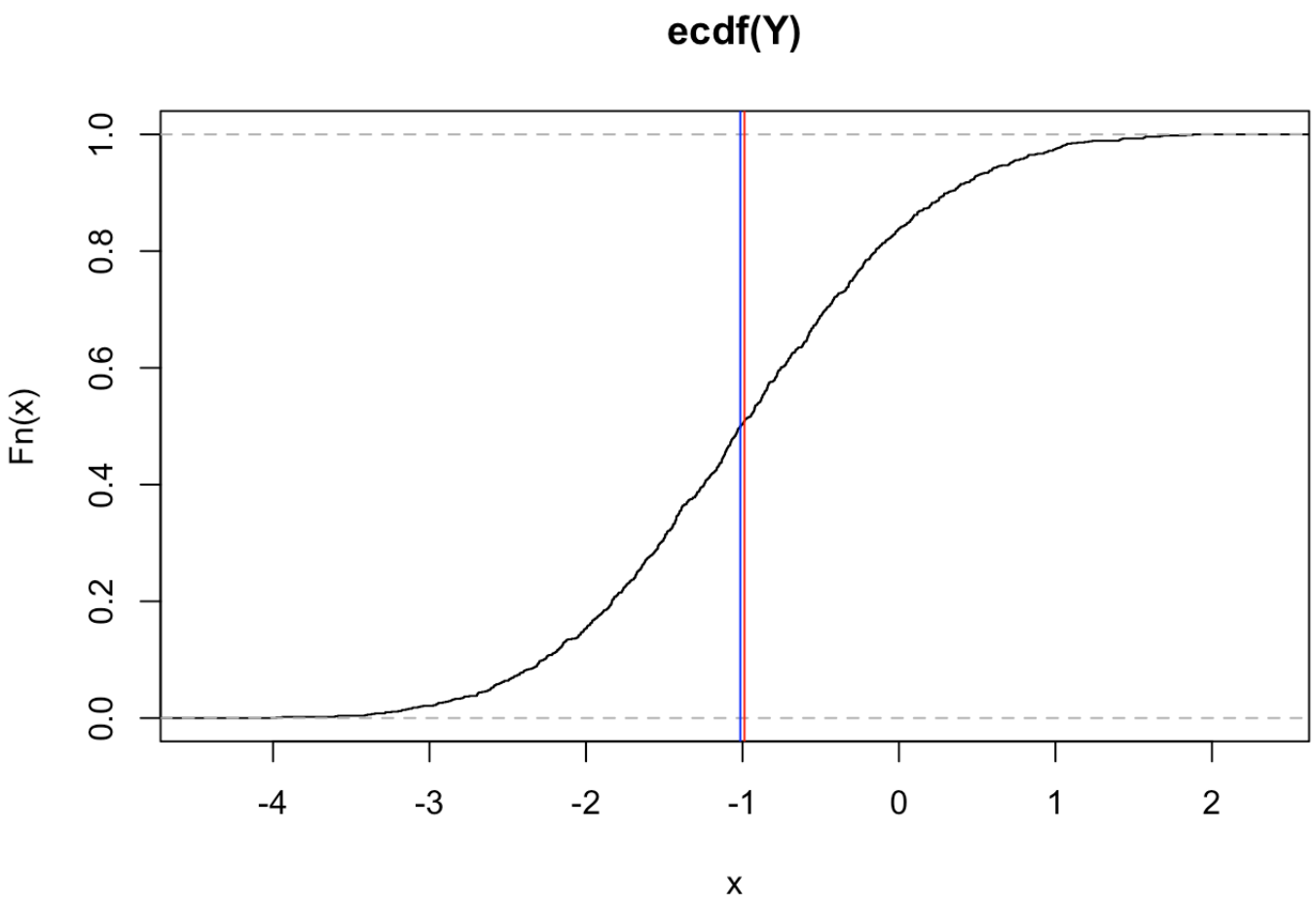

```
hist(Y,breaks=100,freq=FALSE) #lower transformation
```




```
mean_log_norm=mean(Y)
median_log_norm=median(Y)

Z=density(Y)
plot(Z) #the graph is very same as the Y
abline(v=mean_log_norm,col="red")
abline(v=median_log_norm,col="blue")
```




```
#CDF
y_cdf=ecdf(Y)
plot(y_cdf)
abline(v=mean_log_norm,col="red")
abline(v=median_log_norm,col="blue")
```

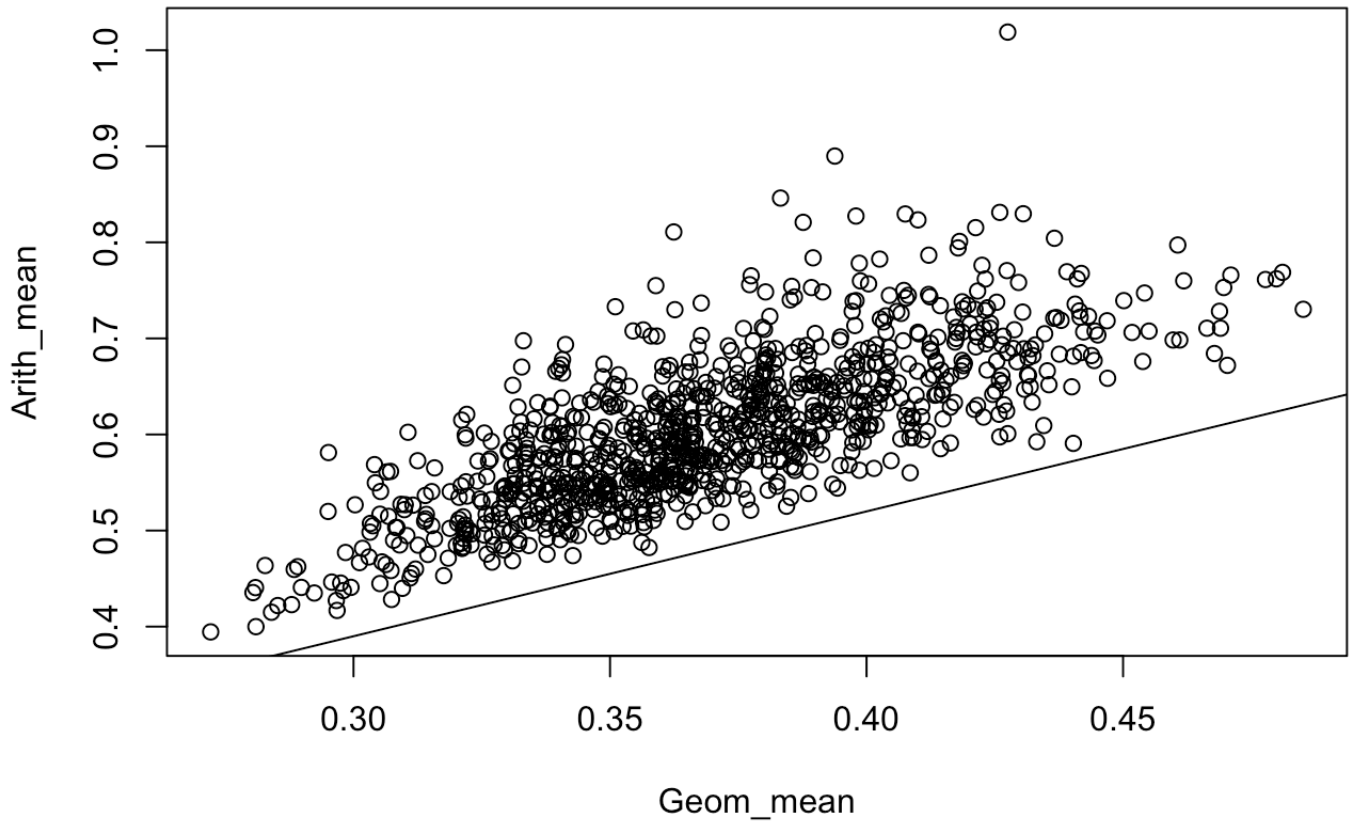


For each of the distributions below,generate 1000 samples of size 100.For each sample, calculate the geometric and arithmetic mean.Generate a scatter plot of the geometric and arithmetic sample means.Add the line of identify as a reference line.

```
#geometric mean= exp(mean(log(data))), used in
#geometric mean= exp(mean(log(data))), used in computing rate of return in finance
#arithmetic mean= mean(log(data))
#arithmetic  mean>g

lognormal_sample=matrix(NA,ncol=100, nrow=1000)
for(k in 1:1000){
  lognormal_sample[k,]=rlnorm(100, meanlog = -1, sdlog = 1)  #k row

}
dim(lognormal_sample)
```

```
## [1] 1000  100
```

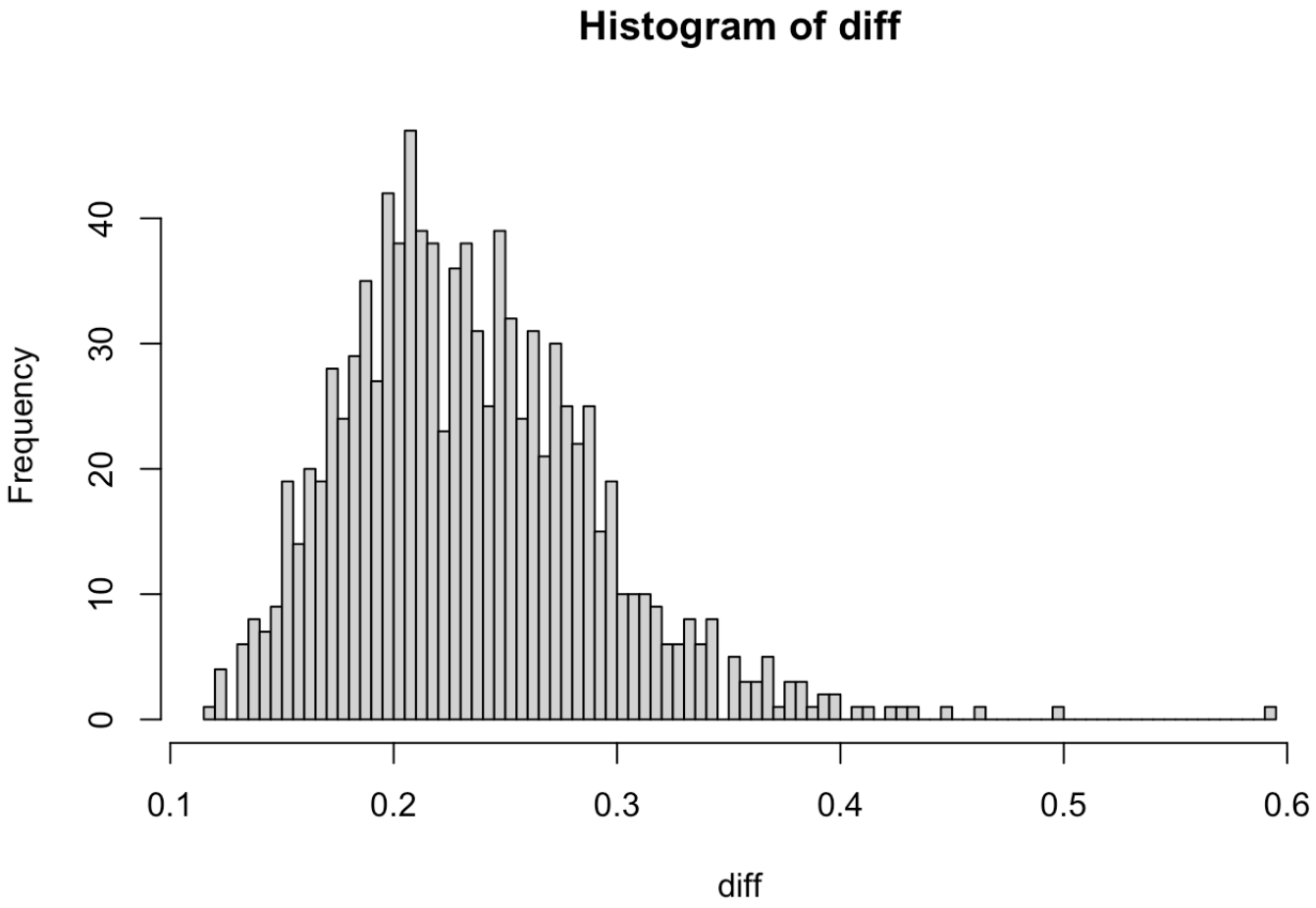
```
#Arith_mean and Geom_mean should have the same dimension

Arith_mean=rowMeans(lognormal_sample)

Geom_mean=exp(rowMeans(log(lognormal_sample)))

plot(Geom_mean,Arith_mean)

abline(c(0,1.3)) #identity line
```




```
#see from the high density plot : if Geom_mean is 2.5, the Arith_mean is greater than
2.5, because it is a wrong here.
#Arith_mean is all raise greater than Geom_mean
```

Generate a histogram of the difference between the arithmetic mean and the geometric mean.

```
diff= Arith_mean-Geom_mean
hist(diff, breaks=100)
```



Analysis:

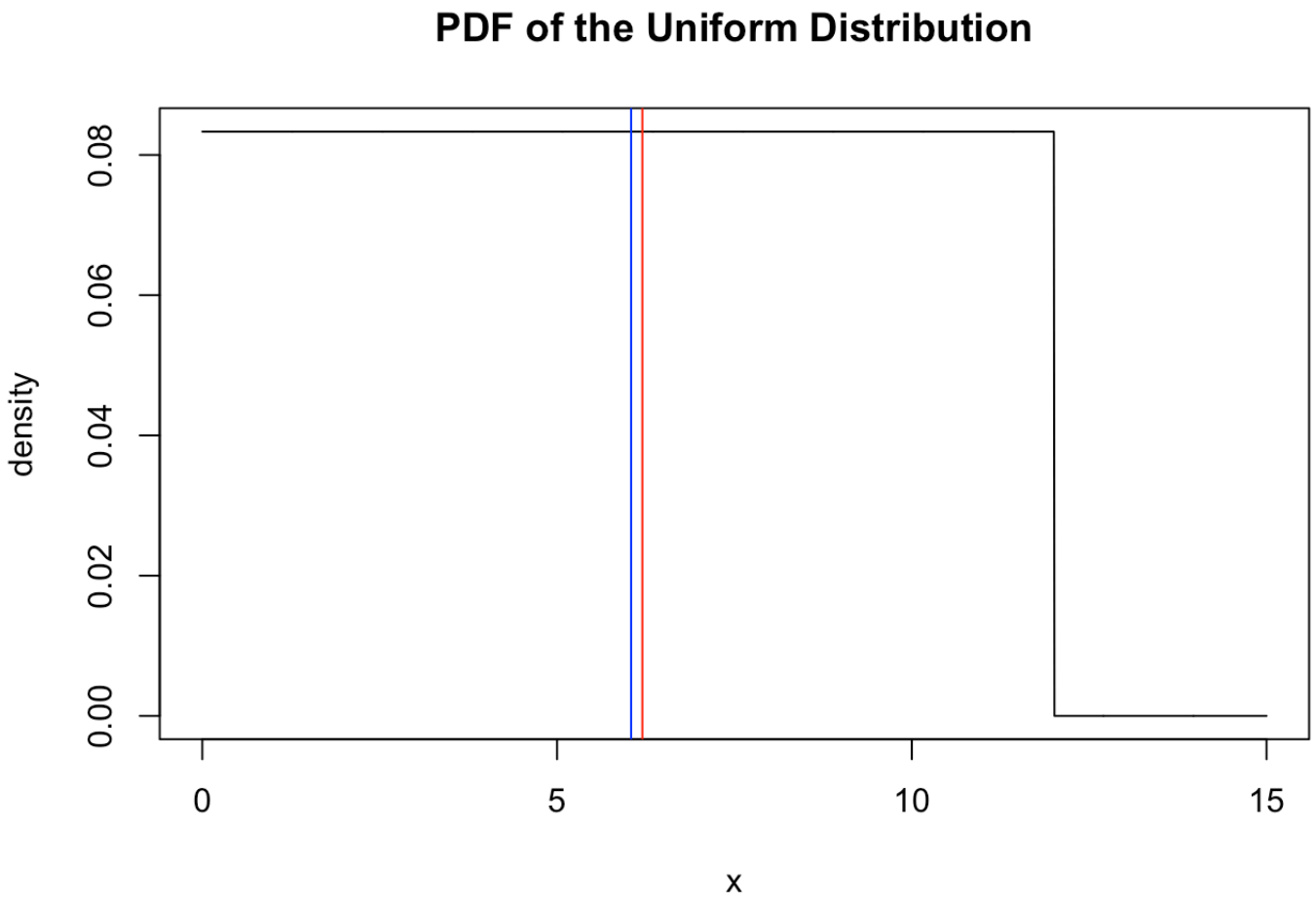
See from the high density plot : if Geom_mean is 2.5, the Arith_mean is greater than 2.5. So Arith_mean is all raise greater than Geom_mean in this example

Distribution 3

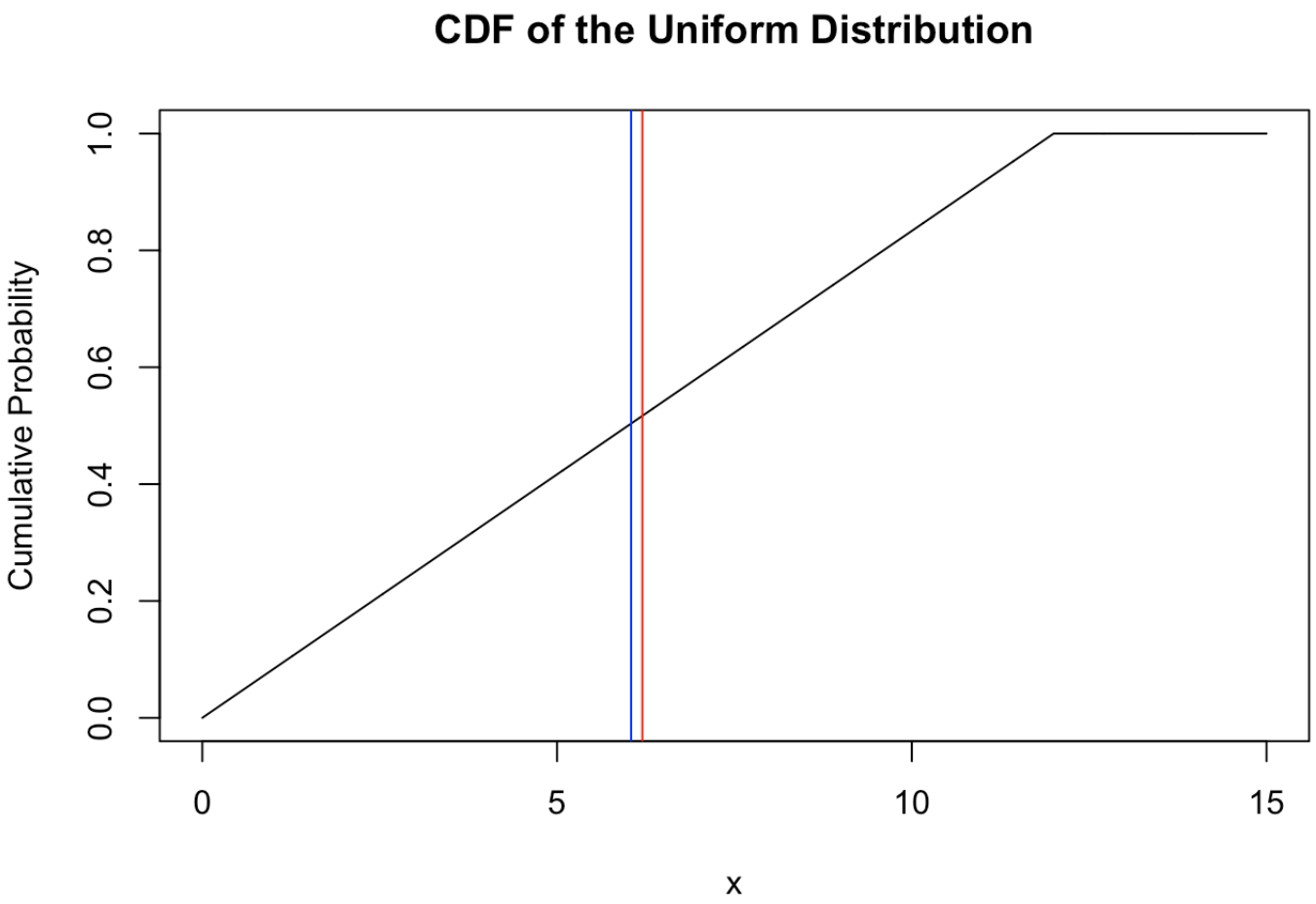
X~UNIFORM(0,12)

For each distribution below, generate a figure of the PDF and CDF.Mark the mean and median in the figure.

```
#PDF
#set the range of the x
x=seq(0,15,by=0.01)
dat = dunif(x,min = 0, max = 12)
#data.frame(x,dat)
#plot(x,dat,type="l",main="PDF of the Uniform Distribution",ylab="density")

#get mean and median
#method 2
rand_unif = runif(1000,min = 0, max = 12)
mean_unif<-mean(rand_unif)
median_unif<-median(rand_unif)
plot(x,dat,type="l",main="PDF of the Uniform Distribution",ylab="density")
abline(v=mean_unif,col="red")
abline(v=median_unif,col="blue")
```

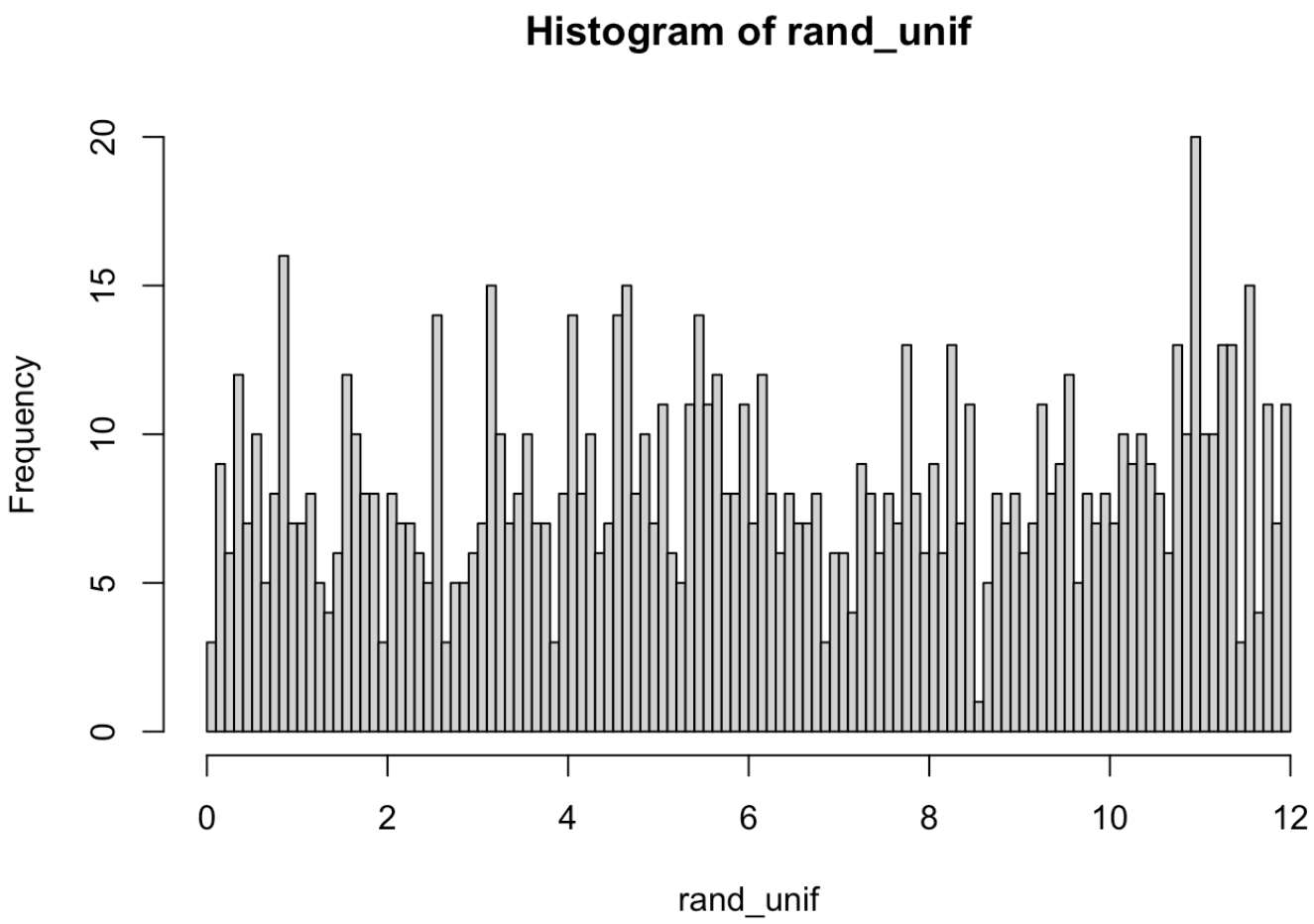


```
#CDF
dat1=punif(x,min = 0, max = 12)
plot(x,dat1,main="CDF of the Uniform Distribution",ylab="Cumulative Probability",type
="l")
abline(v=mean_unif,col="red")
abline(v=median_unif,col="blue")
```

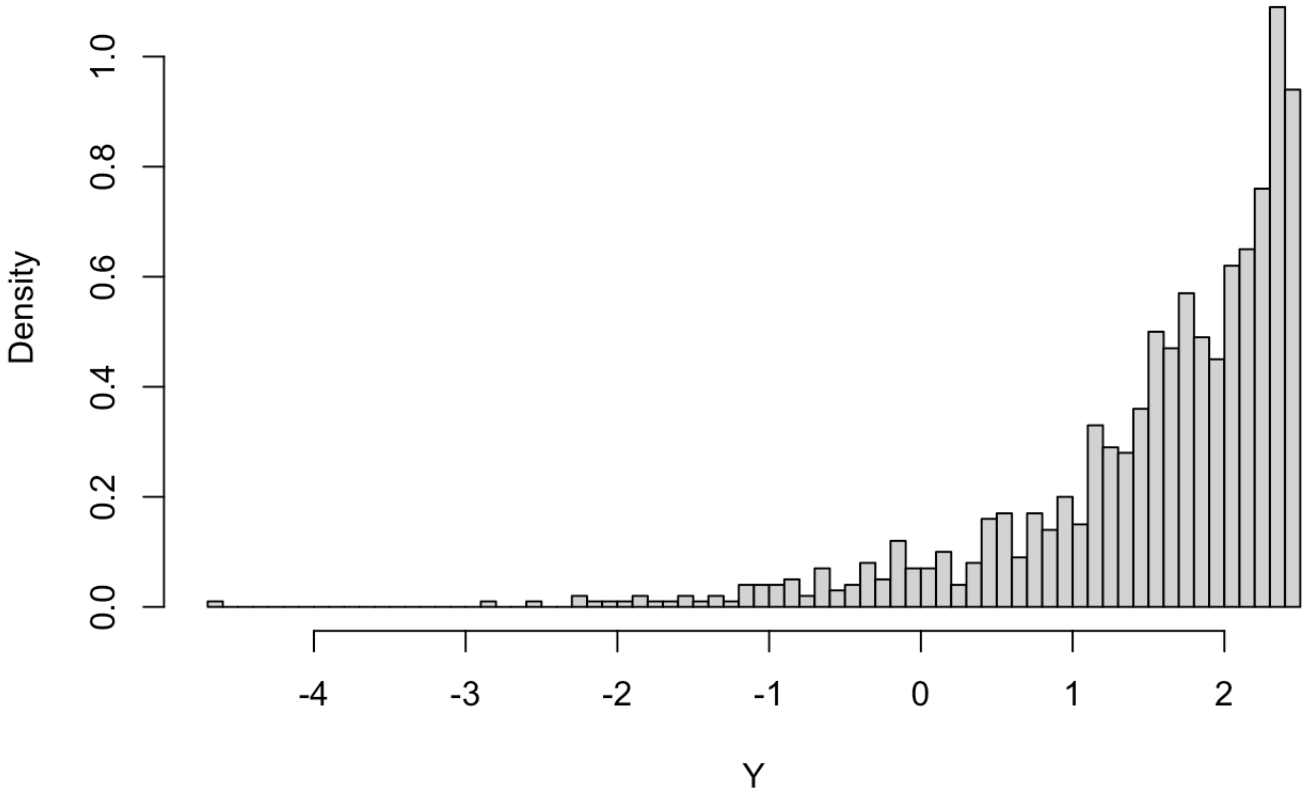


For each distribution below,generate a figure of the PDF and CDF of the transformation Y = log(X) random variable.Mark the mean and median in the figure.You may use simulation or analytic methods in order find the PDF and CDF of the transformation.

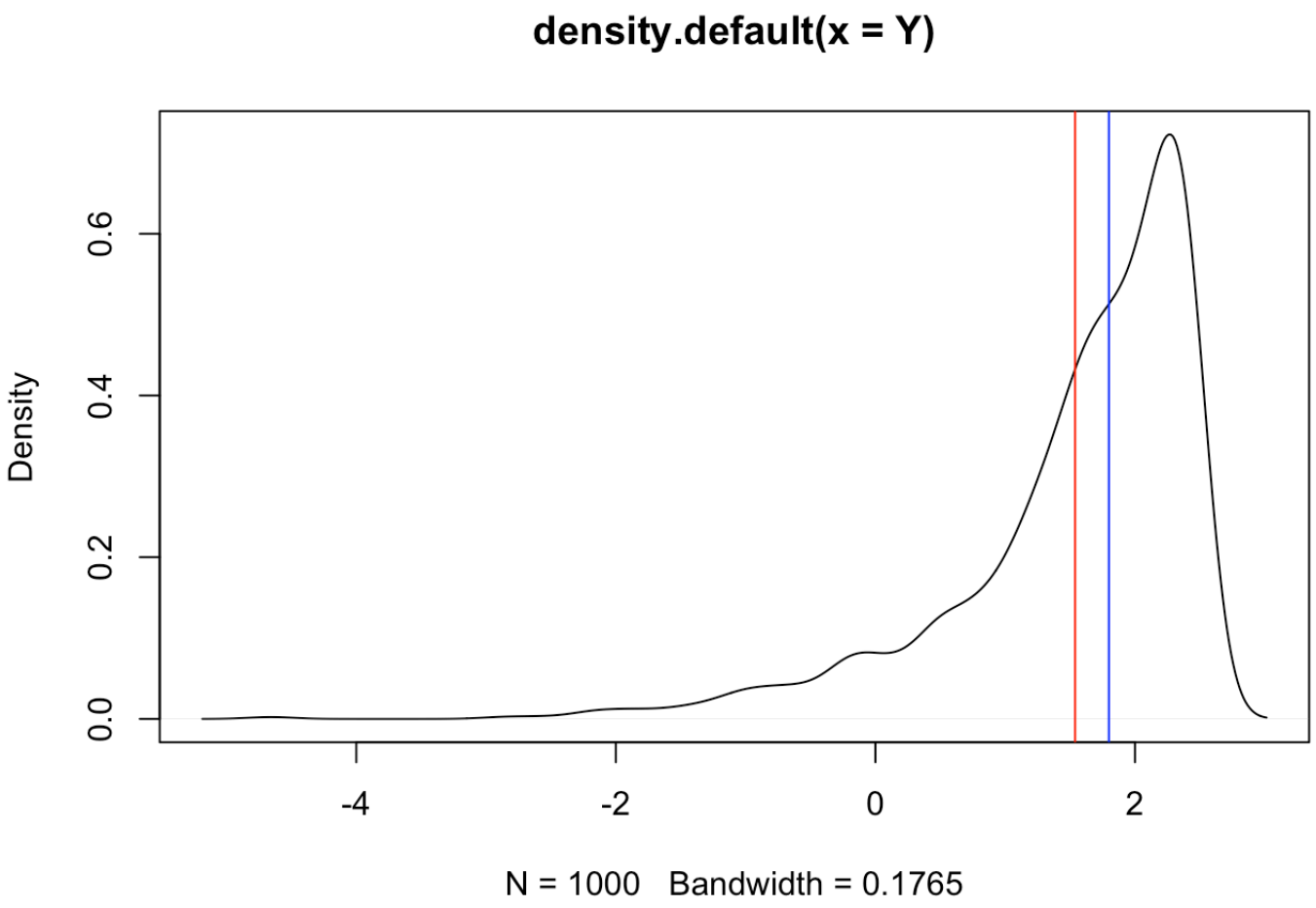
```
Y=log(rand_unif) #new value

# for PDF, either hist() or density()
hist(rand_unif,breaks=100)
```



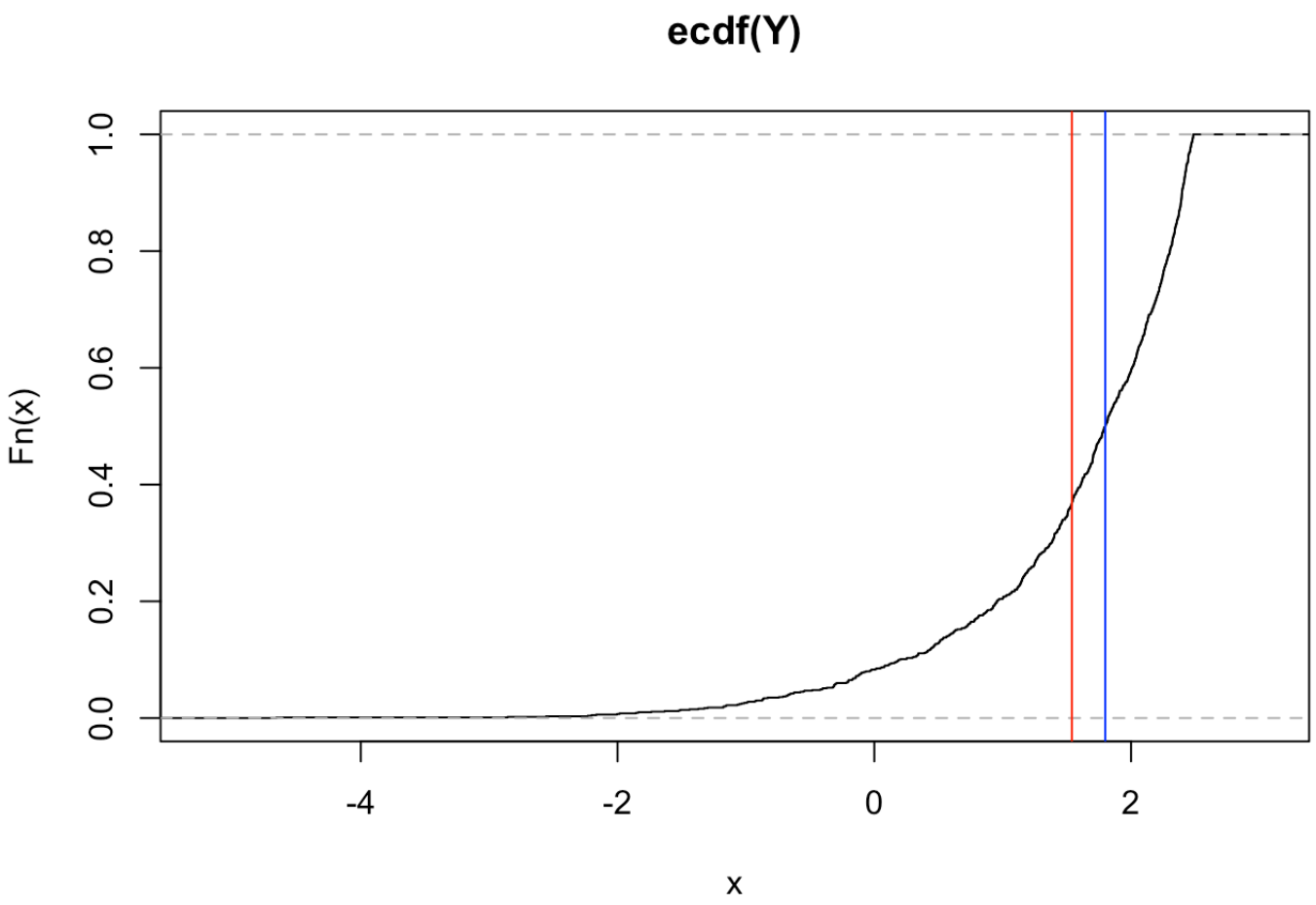

```
hist(Y,breaks=100,freq=FALSE) #lower transformation
```




```
mean_log_unif=mean(Y)
median_log_unif=median(Y)

Z=density(Y)
plot(Z) #the graph is very same as the Y
abline(v=mean_log_unif,col="red")
abline(v=median_log_unif,col="blue")
```




```
#CDF
y_cdf=ecdf(Y)
plot(y_cdf)
abline(v=mean_log_unif,col="red")
abline(v=median_log_unif,col="blue")
```

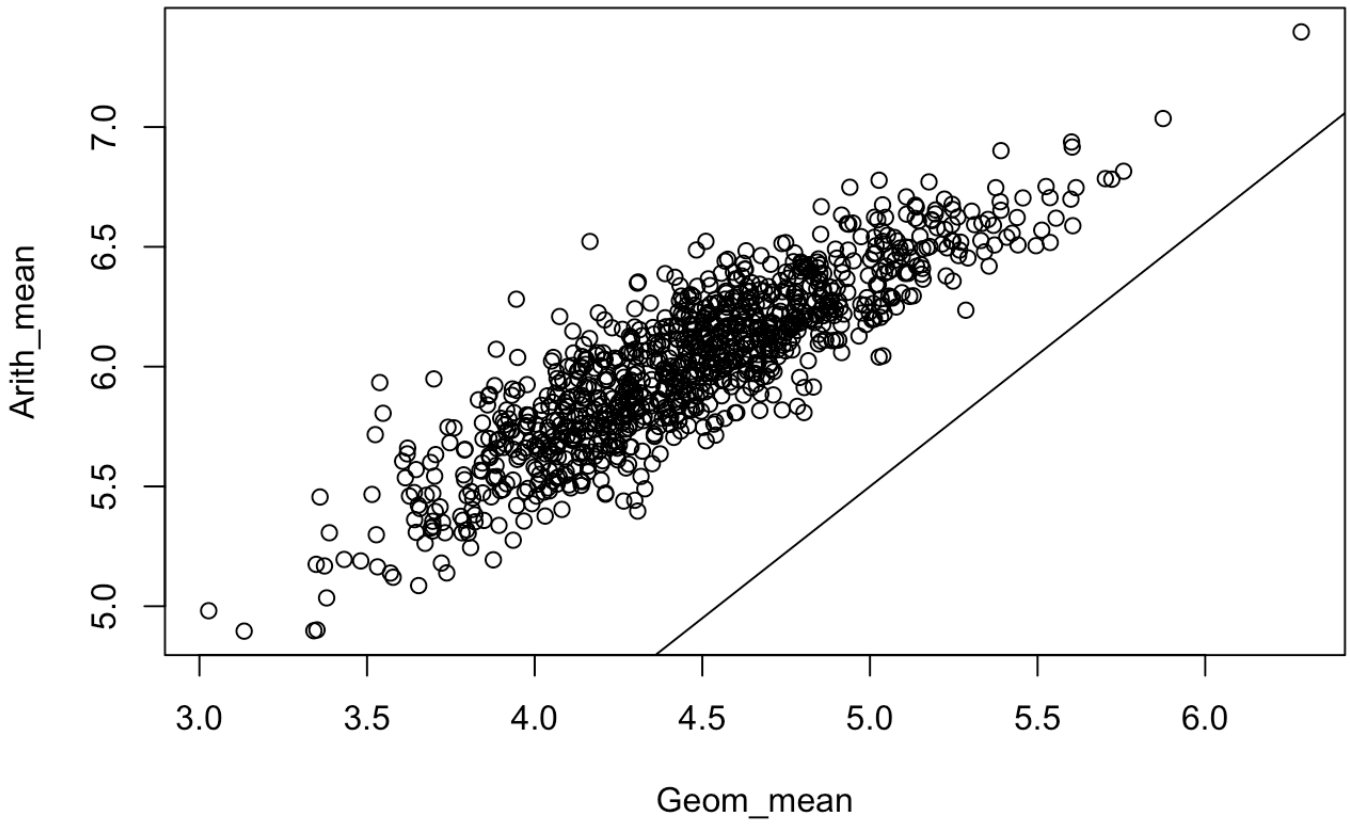


For each of the distributions below,generate 1000 samples of size 100.For each sample, calculate the geometric and arithmetic mean. Generate a scatter plot of the geometric and arithmetic sample means. Add the line of identify as a reference line.

```
#geometric mean= exp(mean(log(data))), used in computing rate of return in finance
#arithmetic mean= mean(log(data))
#arithmetic  mean>g

unif_sample=matrix(NA,ncol=100, nrow=1000)
for(k in 1:1000){
  unif_sample[k,]=runif(100,min = 0, max = 12)  #k row

}
dim(unif_sample)
```

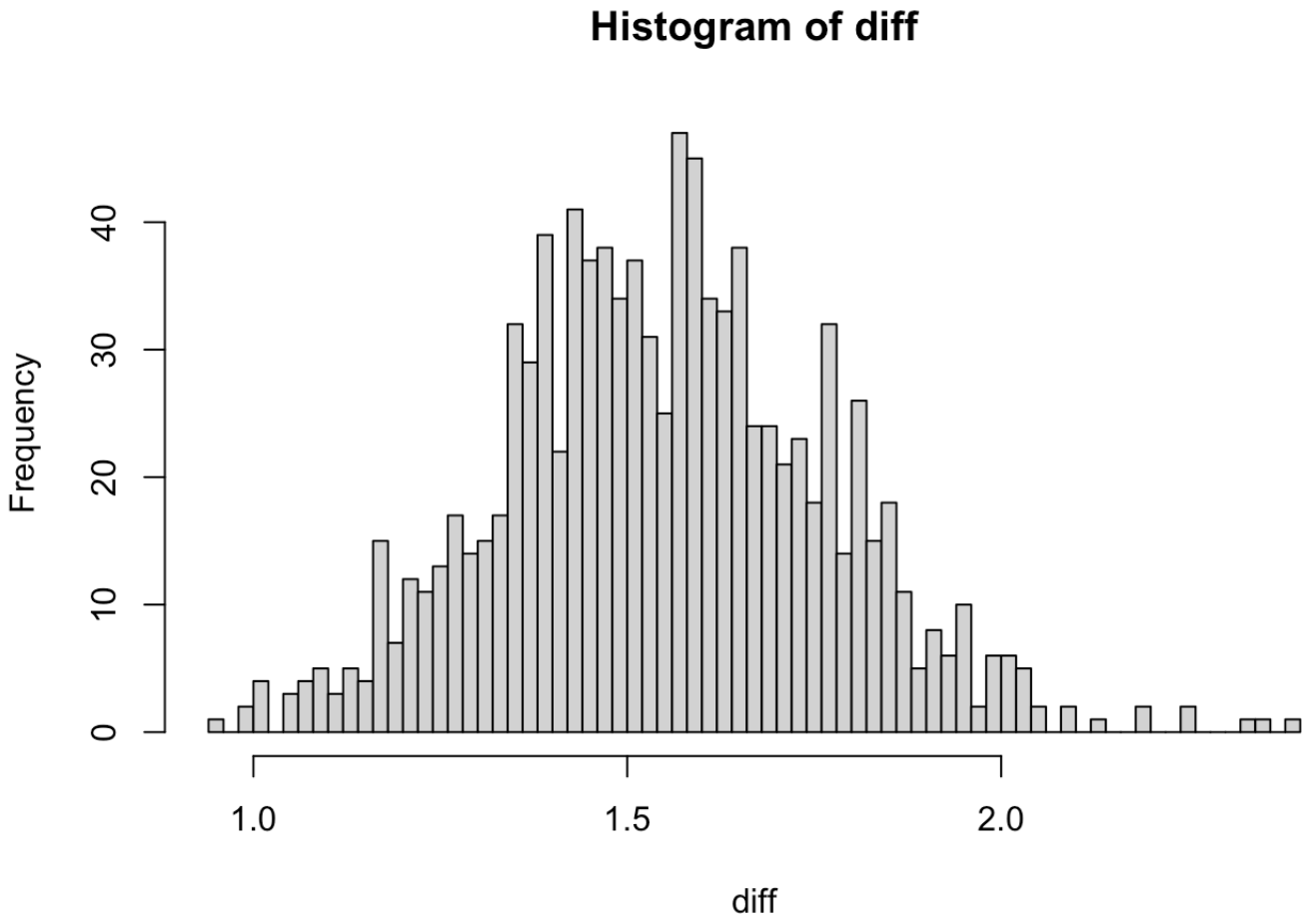
```
## [1] 1000  100
```

```
#Arith_mean and Geom_mean should have the same dimension

Arith_mean=rowMeans(unif_sample)

Geom_mean=exp(rowMeans(log(unif_sample)))

plot(Geom_mean,Arith_mean)
abline(c(0,1.1)) #identity line
```



Generate a histogram of the difference between the arithmetic mean and the geometric mean.

```
diff= Arith_mean-Geom_mean
hist(diff, breaks=100)
```




```
#all the values are greater than 0, so means all the Arith_mean is greater than Geom_
mean for this example.
```

Analysis:

See from the high density plot : if Geom_mean is 2.5, the Arith_mean is greater than 2.5. So Arith_mean is all raise greater than Geom_mean in this example

---

Part 2 (optional)

Show that if Xi>0 for all i, then the arithmetic mean is greater than or equal to the geometric mean.

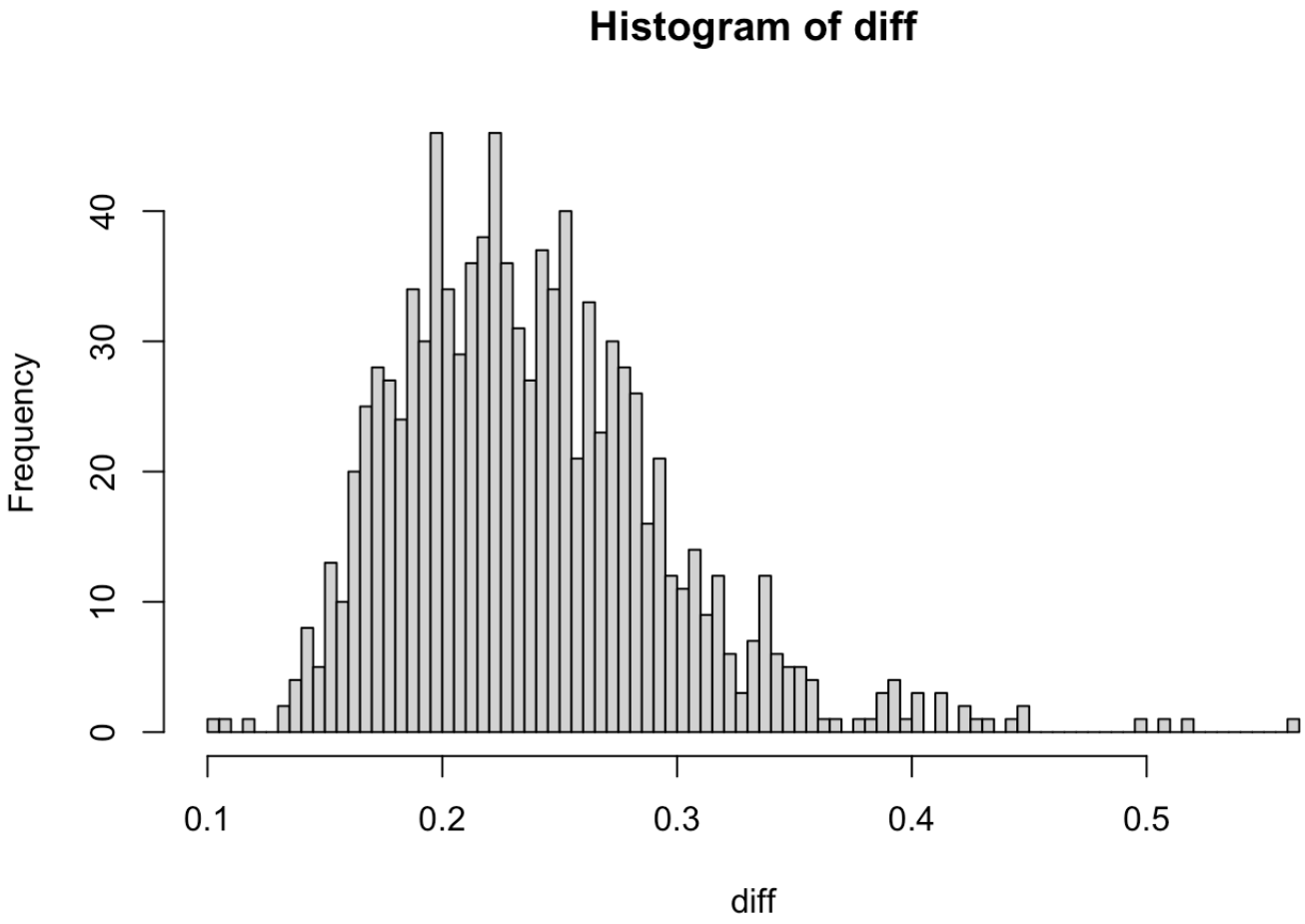
Hint: Start with the sample mean of the transformation Yi = log(Xi).

Assume: We use two different distributions: log normal and uniform. And make the diff= Arith_mean-Geom_mean.

```
#distribution 2: log_normal

lognormal_sample=matrix(NA,ncol=100, nrow=1000)
for(k in 1:1000){
  lognormal_sample[k,]=rlnorm(100, meanlog = -1, sdlog = 1)  #k row

}
dim(lognormal_sample)
```

```
## [1] 1000  100
```

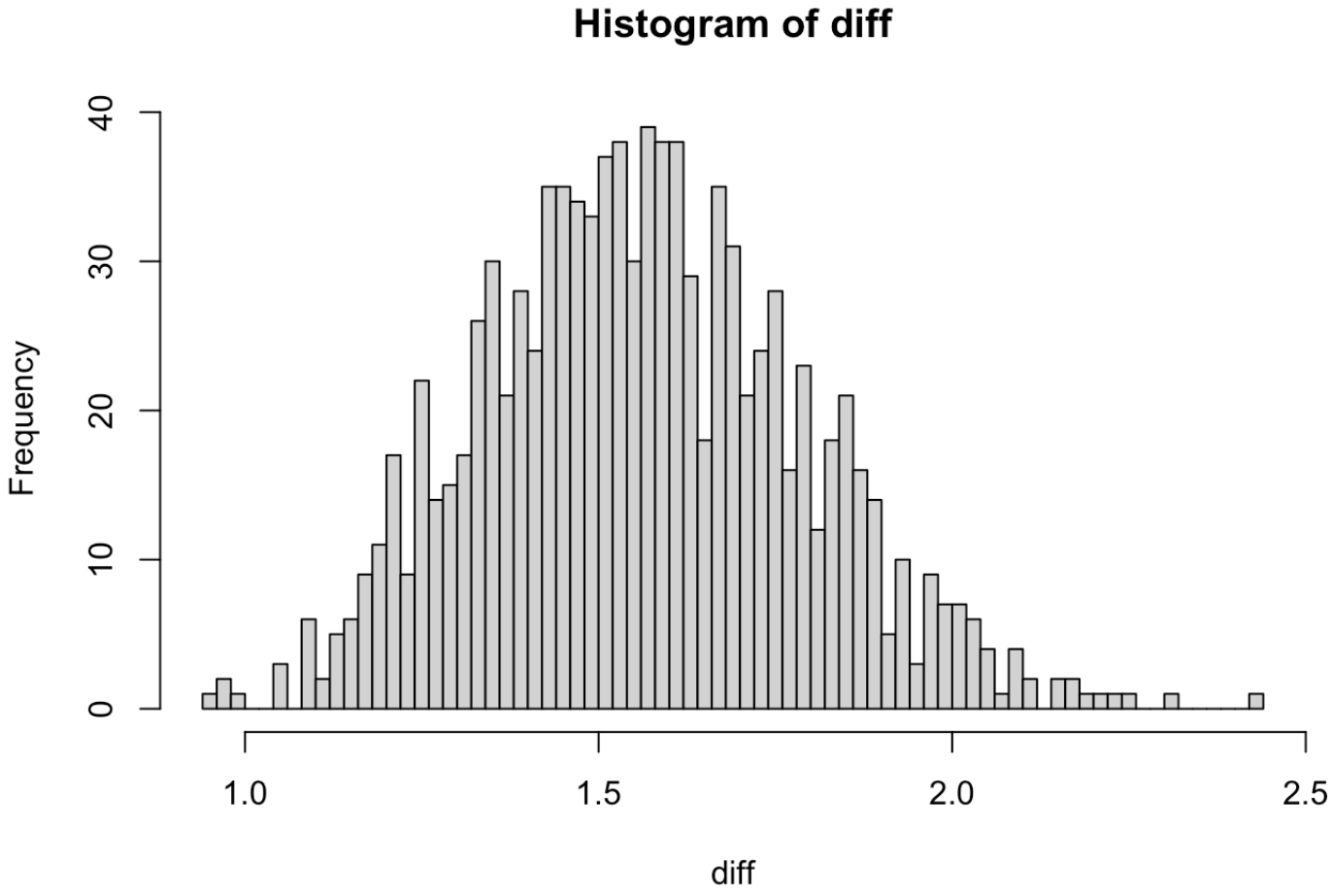
```
#Arith_mean and Geom_mean should have the same dimension

Arith_mean=rowMeans(lognormal_sample)

Geom_mean=exp(rowMeans(log(lognormal_sample)))

diff= Arith_mean-Geom_mean
hist(diff, breaks=100)
```




```
#distribution 3 : uniform
Y=log(rand_unif) #new value
unif_sample=matrix(NA,ncol=100, nrow=1000)
for(k in 1:1000){
  unif_sample[k,]=runif(100,min = 0, max = 12)  #k row

}
dim(unif_sample)
```

```
## [1] 1000  100
```

```
#Arith_mean and Geom_mean should have the same dimension

Arith_mean=rowMeans(unif_sample)

Geom_mean=exp(rowMeans(log(unif_sample)))

diff= Arith_mean-Geom_mean
hist(diff, breaks=100)
```



Analysis: Using two different distributions :log normal and uniform, we can see the two below graphs of diff = arithmetic mean - geometric mean >= 0. It can easily find out that if Xi>0 for all i, then the arithmetic mean is greater than or equal to the geometric mean.

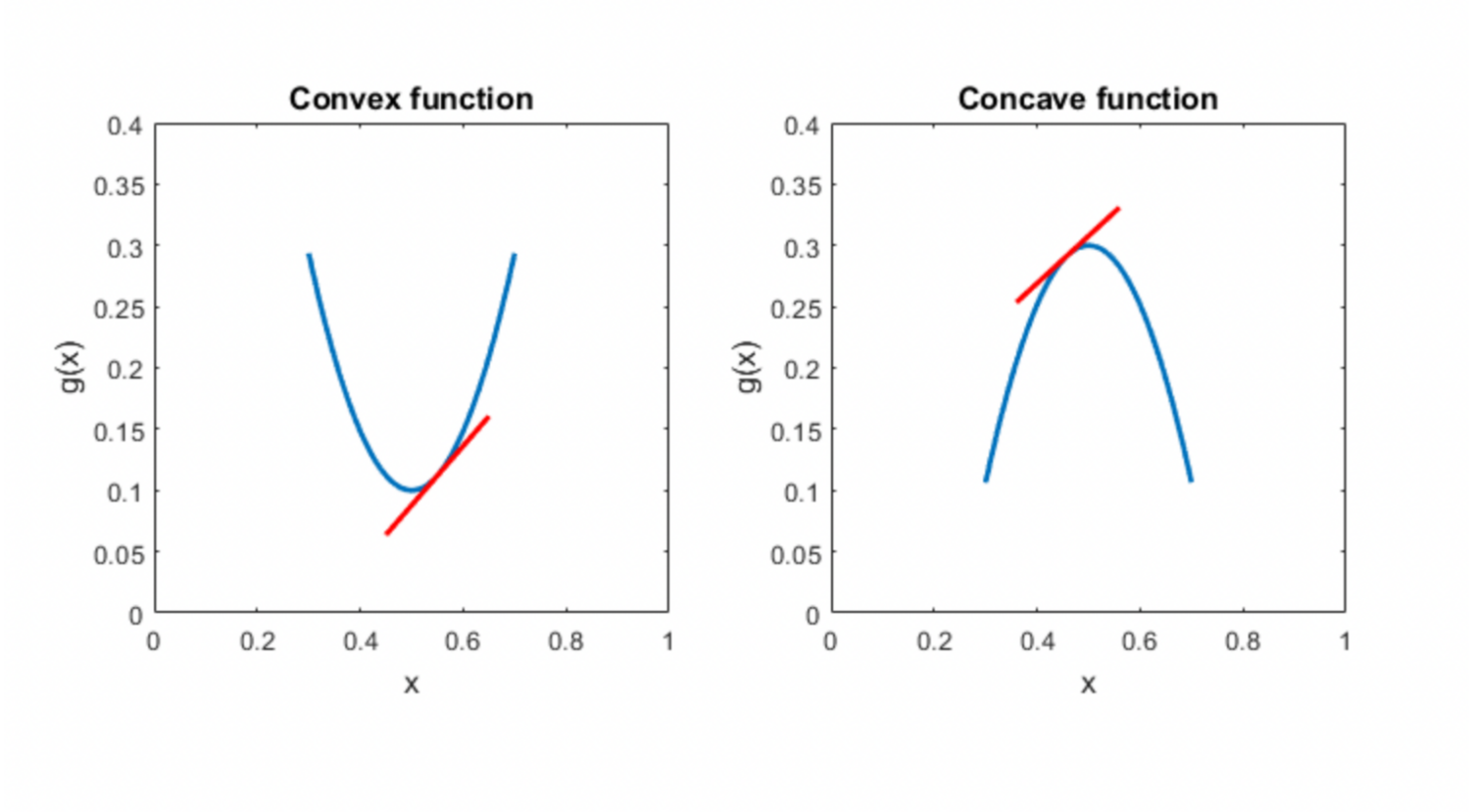
Part 3

What is them relationship between E[log(X)] and log[E(X)]? Is one always larger? Equal? Explain your answer.

Assume:

X is a positive value, i.e.,

We can see the different between convex function and convace function from the following graph:

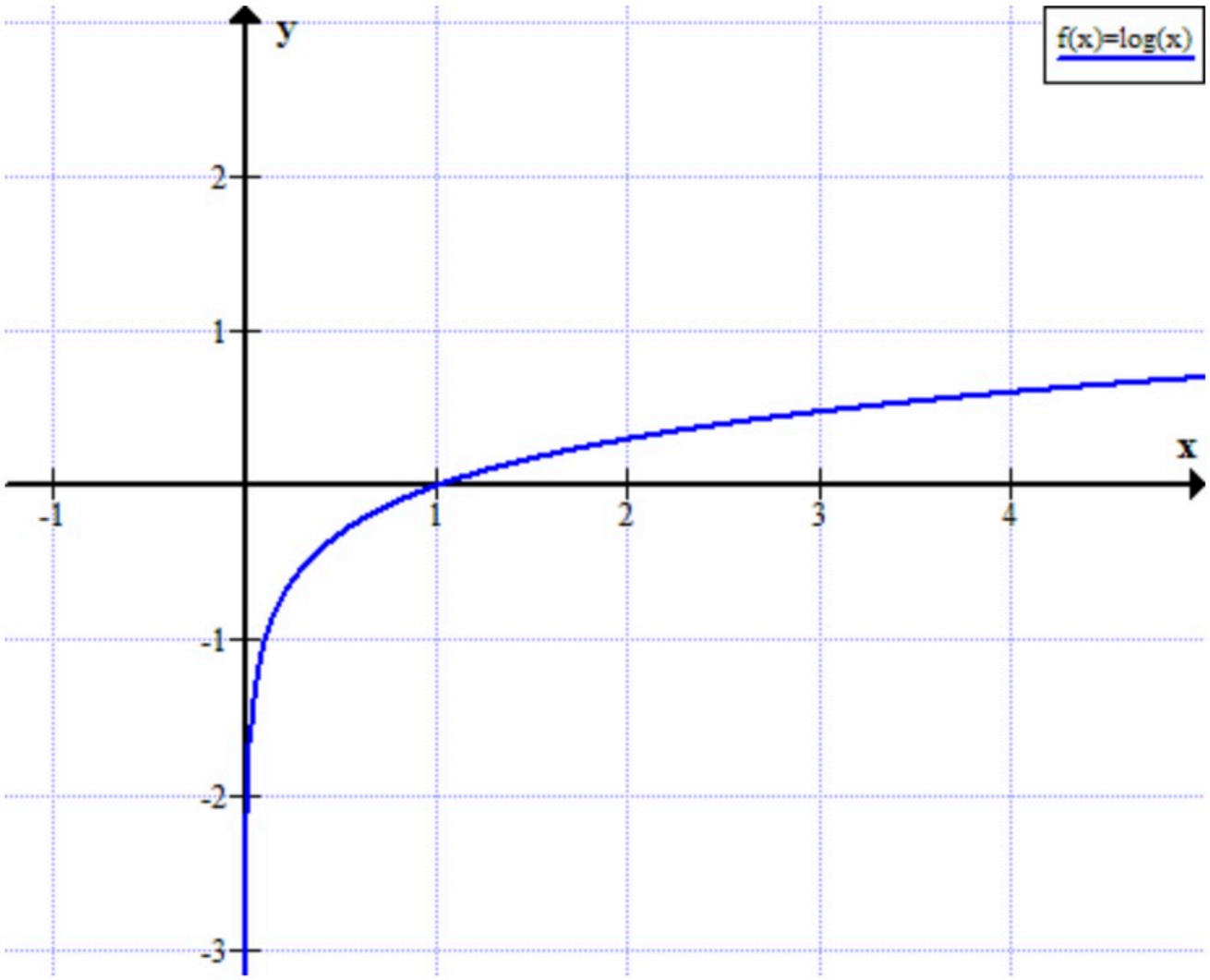


Convex & Concave functions

If the function g(x) is concave then : E(g(x)) <= g(E(x))

If the function g(x) is convex then : E(g(x)) >= g(E(x))

When we define g(x)=log(x), we know that log(x) is a concave function.



the graph of log(x)

So E[log(X)] <= log[E(X)] should be the relationship between the E[log(X)] and log[E(X)].

E[log(x)]= mean(log(x))

log(E[x])= log(mean(x))

```
#1, you can use Jensen's inequality if you know what it is.
#2, do simulation to answer this question

#Simulation
#(1)draw random samples from any distribution but x>0, eg:Gamma, uniform(0,10), or lo
g-Normal distribution
#(2)compute mean(log(x)) and log(mean(x)),and also the difference between them.
#(3) report (1) &(2) 10000 times, and count cases one is larger than another.

count = rep(NA, 10000)

for(k in 1:10000){

  x <- sample(1:1000, size=200)
  diff<-mean(log(x))-log(mean(x))
  if (diff >= 0){
    count[k] = 1
  }else if(diff<=0){
    count[k] = 0
  }
}

sum(count)
```

```
## [1] 0
```

Analysis:

For all the x>0, and g(x)=log(x), we define the count as the number of times that diff=1,0. We can find the answer: count =0, means all the diff<=0, which also means mean(log(x)) <= log(mean(x)). So we can make the conclusion that E[log(X)] <= log[E(X)]. But the conclusion is just for the condition of x>0 and g(x) is a concave function. When the g(x) function is a convex function, the conclusion will be opposite. That's what Jensen's inequality talks about.