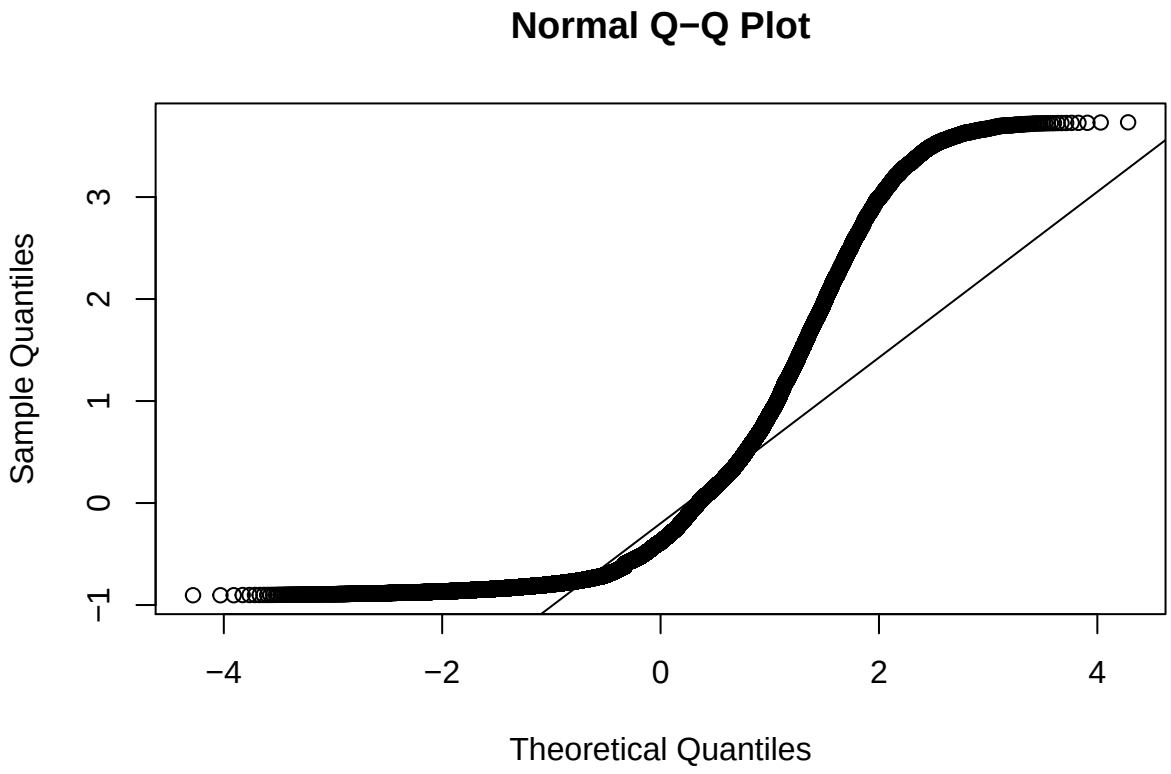
Homework 2

Hui (Sophie) Wang

1.

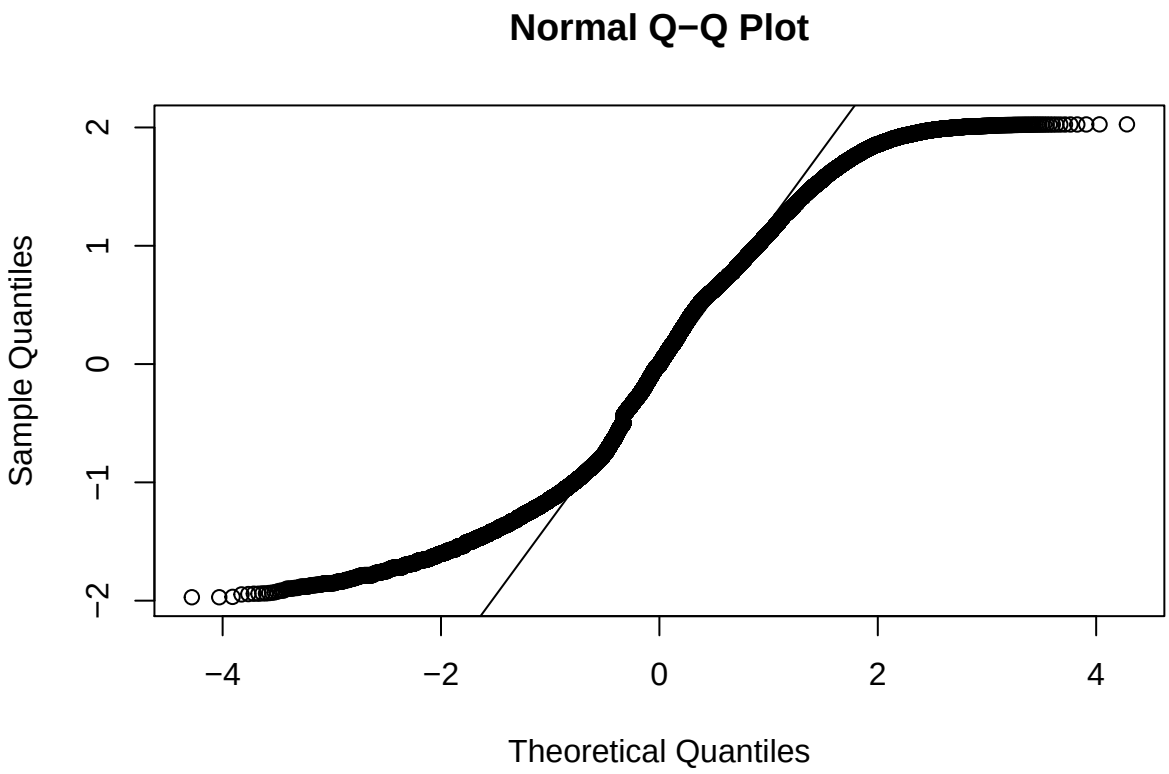
```
library(ggplot2)

x <- diamonds$price
x <- (x-mean(x))/sd(x)
qqnorm(x)
qqline(x)
```



(a)

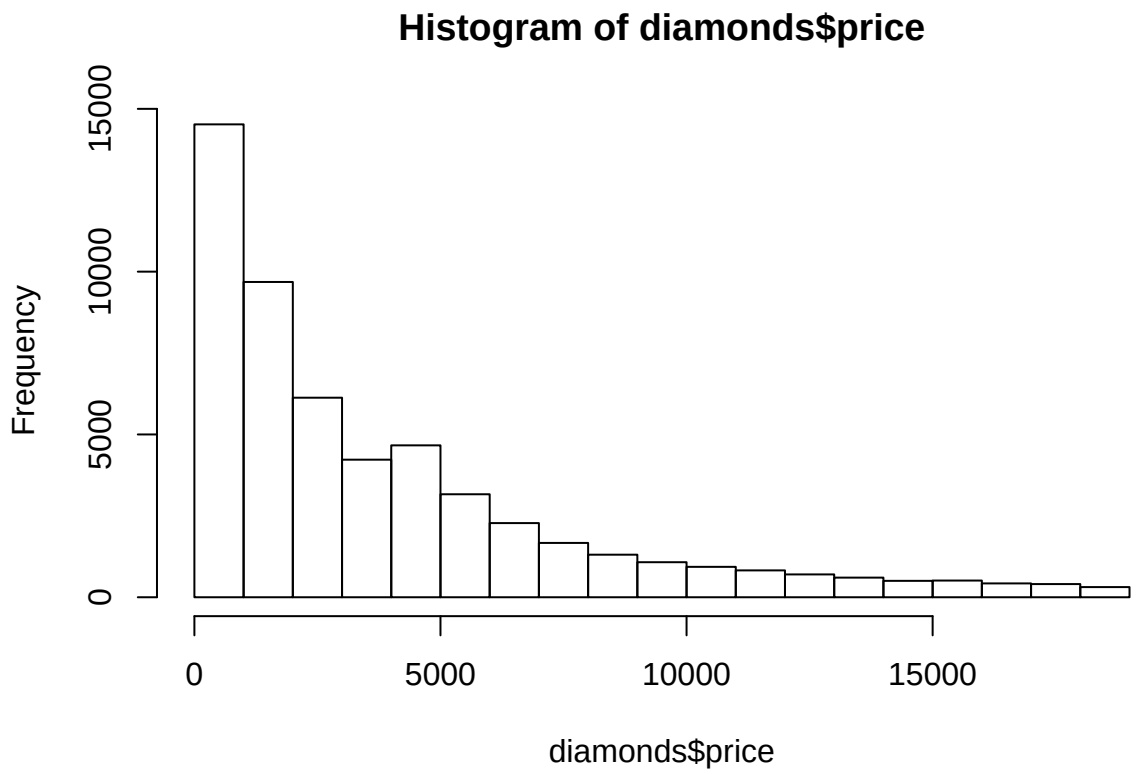
```
log_x <- log(diamonds$price)
log_x <- (log_x-mean(log_x))/sd(log_x)
qqnorm(log_x)
qqline(log_x)
```



The quantile-quantile plot of diamonds price (after standardlization) doesn't fit with the line y=x at all, which means the price of diamonds doesn't follow normal distribution. The quantile-quantile plot of natural log of diamonds price (after standardlization) doesn't fit with the line y=x perfectly, but at interval (-1, 1) they roughly coincide. Before log transformation, the smaller half of the data has a much smaller range than the larger half of the data. After log transformation, the difference of ranges between the two part of data becomes smaller. Therefore its distribution is closer to normal distribution.

(b)

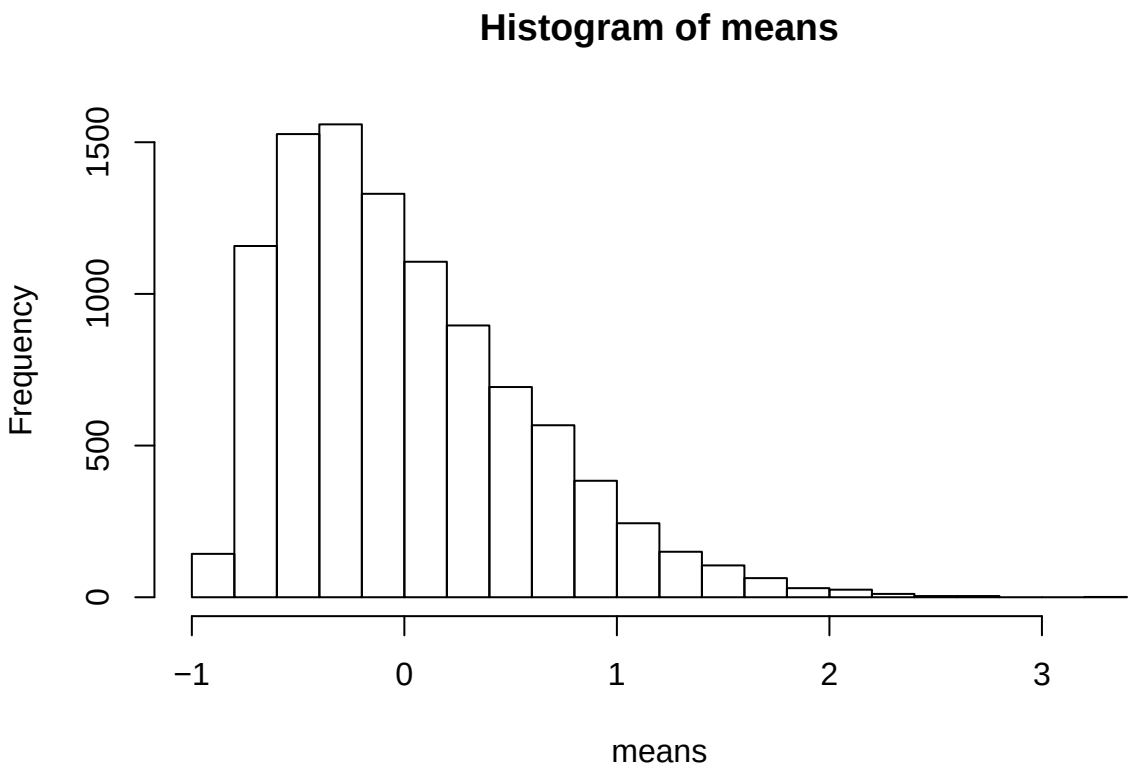
```
hist(diamonds$price)
```



i.

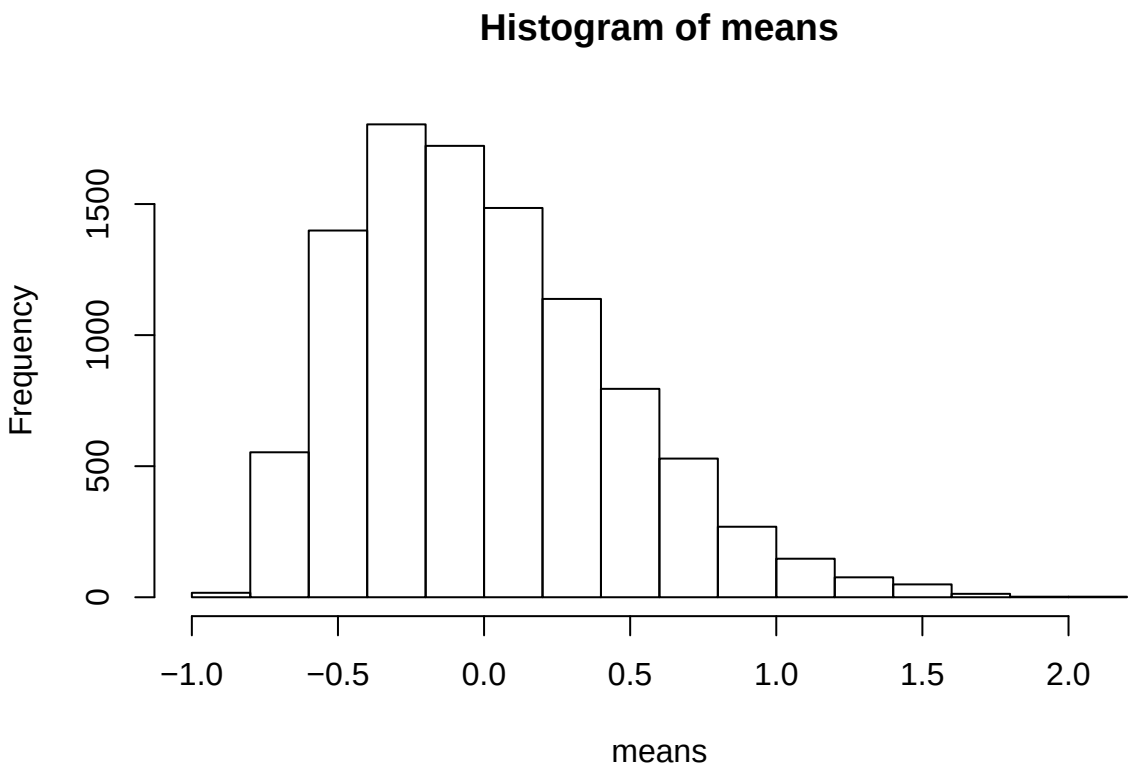
ii.

```
sample_mean <- function(data, size, n){
  means <- vector(mode='numeric',length=n)
  for(i in 1:n){
    sample1 <- sample(data, size)
    means[i] <- mean(sample1)
  }
  hist(means)
}
sample_mean(x, 3, 10000)
```

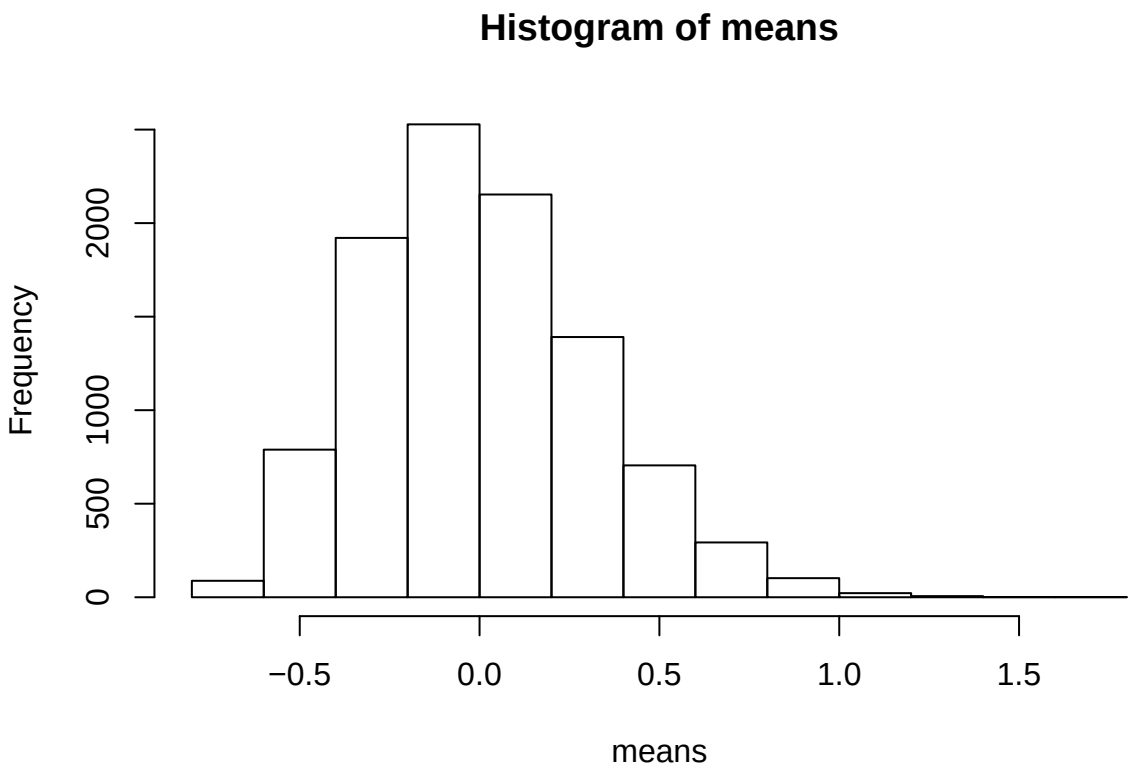


iii.

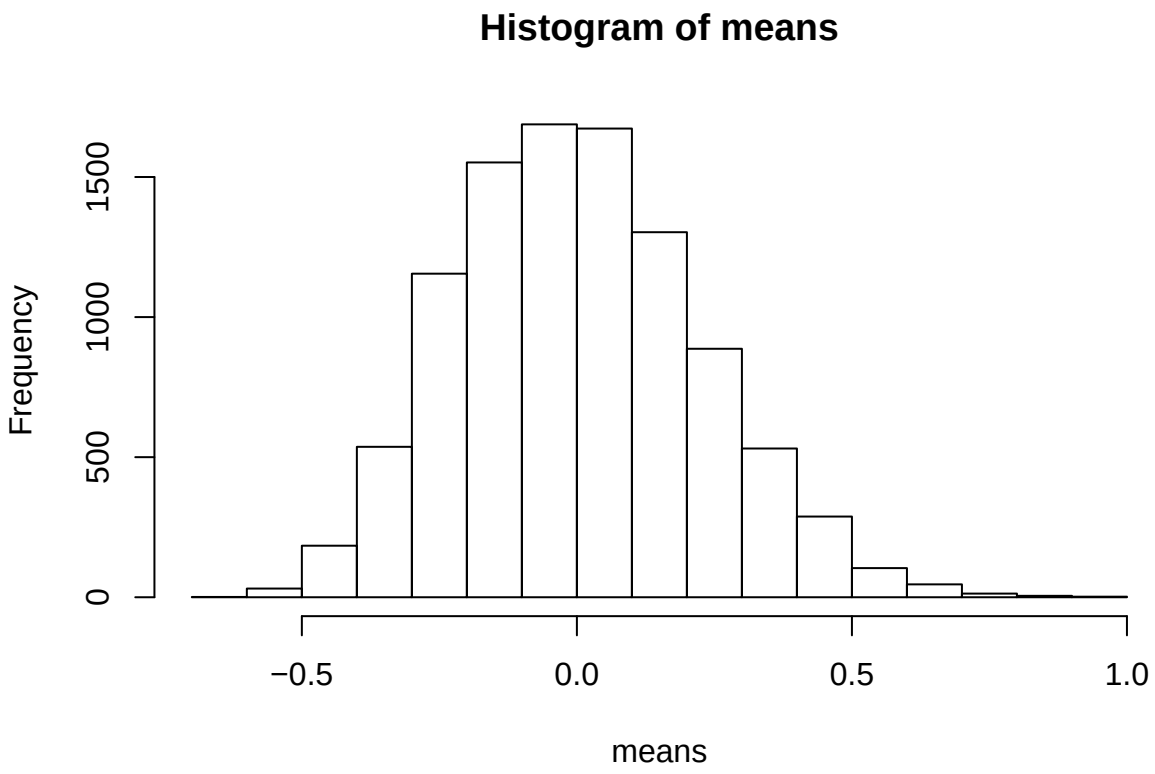
```
sample_mean(x, 5, 10000)
```




```
sample_mean(x, 10, 10000)
```

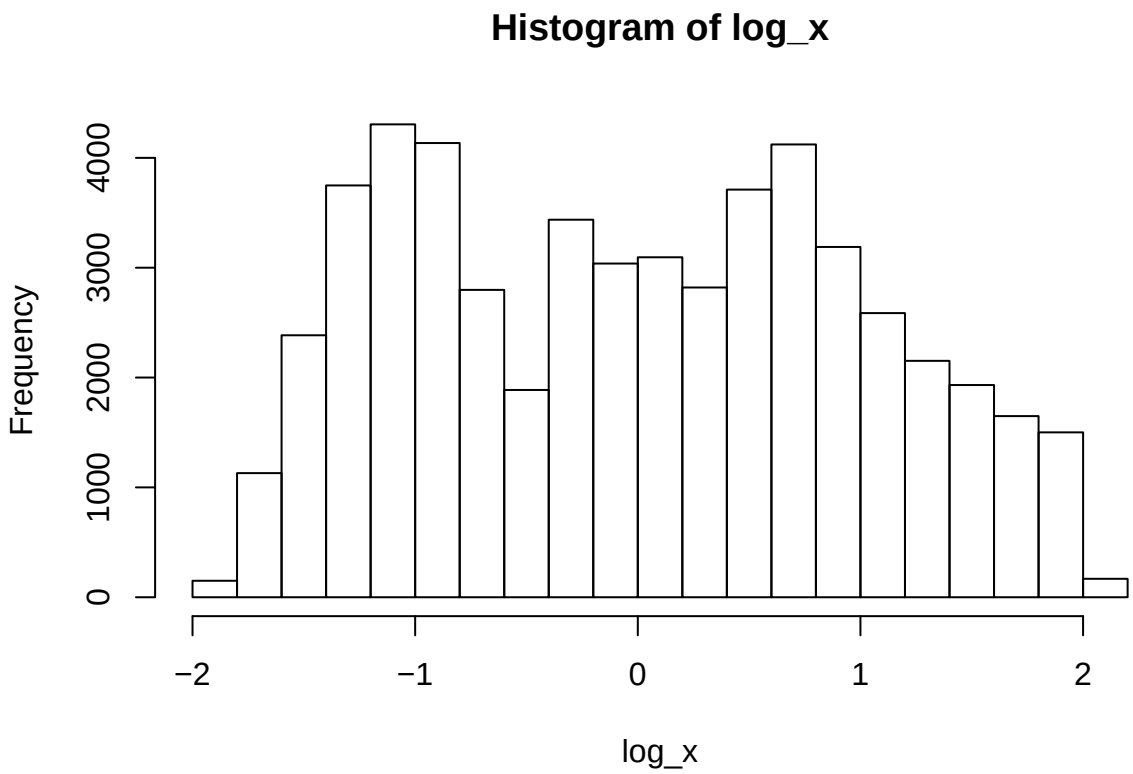



```
sample_mean(x, 20, 10000)
```



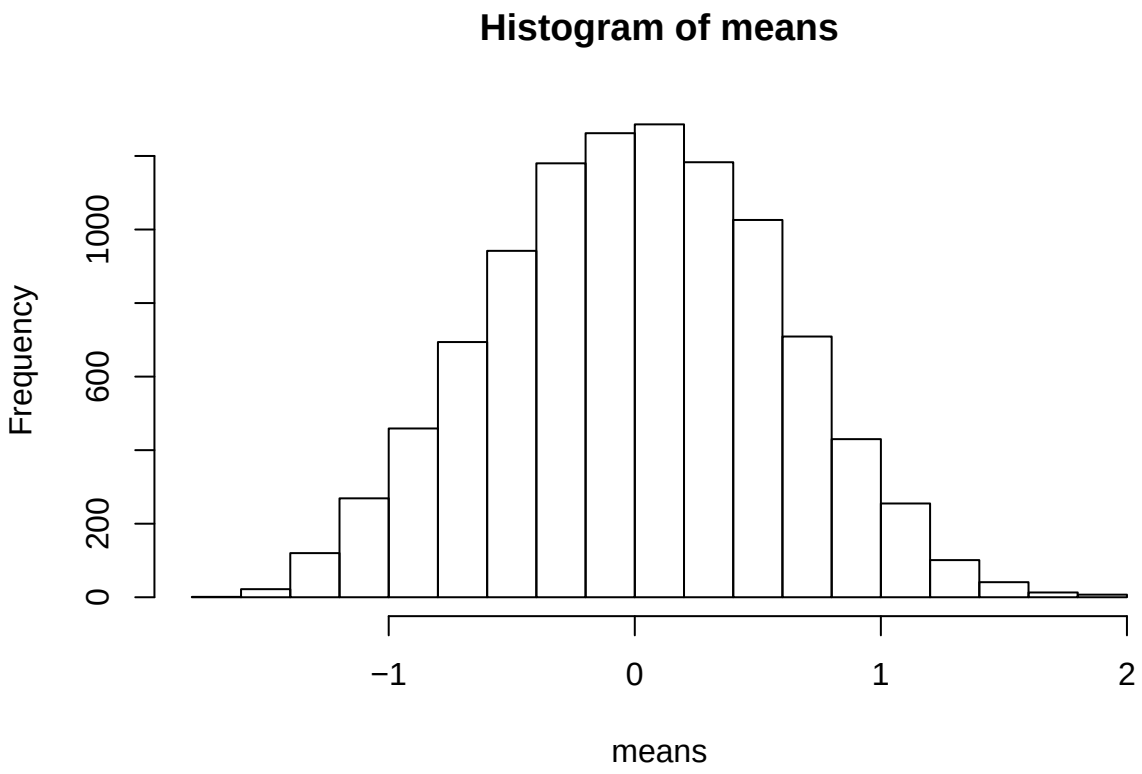
We can see that when sample size is 3, 5, 10 the distribution of sample means is still skewed like the population distribution, when sample size is 20, the distribution of sample means become normal.

```
hist(log_x)
```

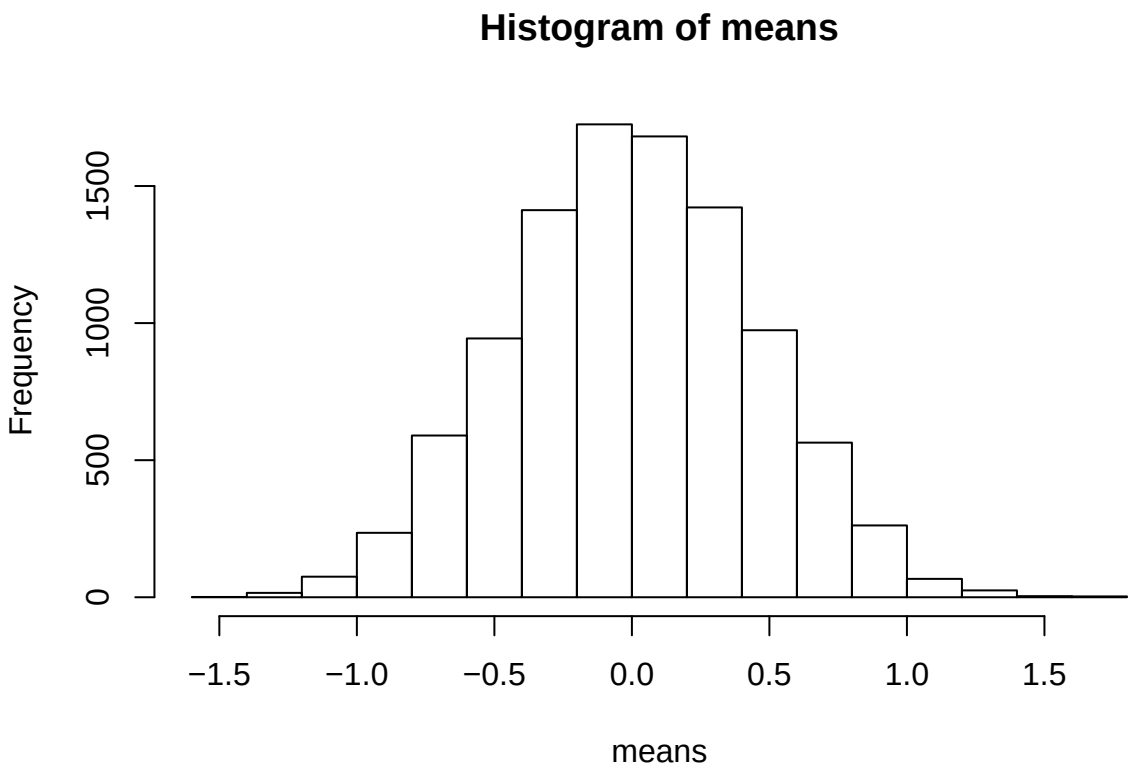


iv.

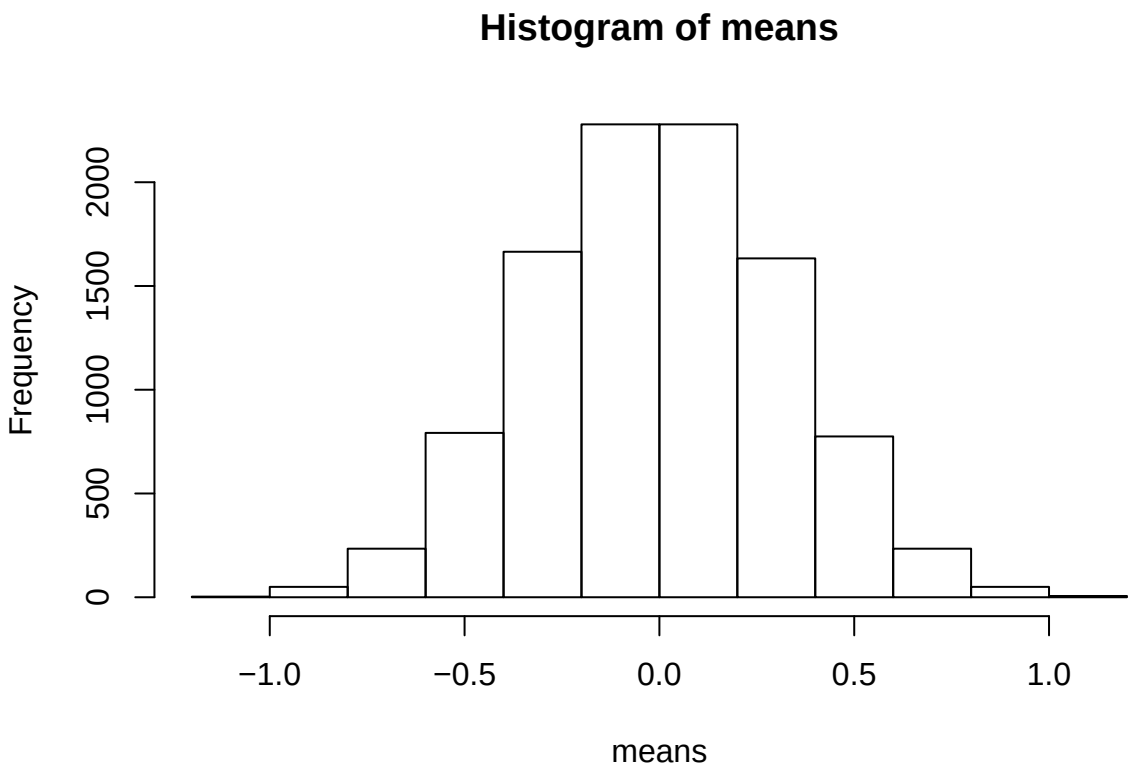
```
sample_mean(log_x, 3, 10000)
```



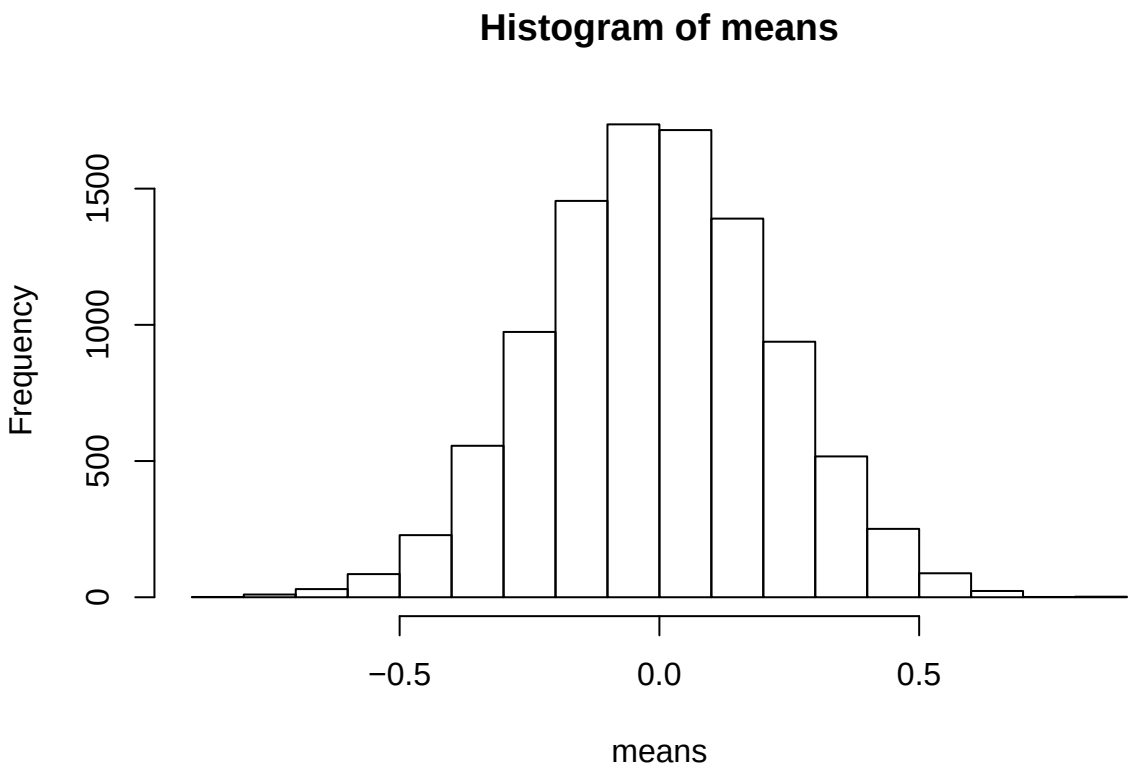

```
sample_mean(log_x, 5, 10000)
```




```
sample_mean(log_x, 10, 10000)
```




```
sample_mean(log_x, 20, 10000)
```



When we plot the natural log of price, sample means is normally distributed even if sample size is small. It is because the population distribution of log price is close to normal distribution.

2.

(a) According to Central Limit Theorem, the sampling distribution of sample mean is normal.

Its expected value is $\mu = 65in.$

Its standard error is $\sigma/\sqrt{n} = 0.5in.$

(b) Suppose the height of an adult American female is random variable H.

```
pnorm(66, mean=65, sd=0.5, lower.tail=FALSE)
```

i.

```
## [1] 0.02275013
```

$\mathrm{P}\{\mathrm{H} > 66\} = 0.0023$

```
pnorm(63, mean=65, sd=0.5)
```

ii.

```
## [1] 3.167124e-05
```

$P\{H < 63\} = 3.17 * 10^-5$

```
pnorm(66.5, mean=65, sd=0.5) - pnorm(63.5, mean=65, sd=0.5)
```

iii.

```
## [1] 0.9973002
```

$P\{H > 63.5 \&\& H < 66.5\} = P\{H < 66.5\}$ - $P\{H < 63.5\} = 0.997$

```
qnorm(0.975, mean=65, sd=0.5)
```

(c)

```
## [1] 65.97998
```

65 - (65.98 - 65)=64.02

Therefore, 95% confidence interval is [64.02, 65.98].

```
pnorm(63, mean=65, sd=2.5, lower.tail=FALSE)
```

(d)

```
## [1] 0.7881446
```

$P\{H > 66\} = 0.788$

The probability that the women is more than 66 in tall is 0.788.

(e) we use populatioin distribution iin (d), which is different from sampling distribution of sample mean used in (b) (i).

3.

(a) According to Central Limit Theorem, the sampling distribution of sample mean is normal.
Its expected value is equal to $\mu = 65in$
Its standard error is equal to $\sigma/\sqrt{n} = 0.25in$

(b)

```
pnorm(66, mean=65, sd=0.25, lower.tail=FALSE)
```

i.

```
## [1] 3.167124e-05
```

P{H > 66} = 3.17 * 10^-5

```
pnorm(63, mean=65, sd=0.25)
```

ii.

```
## [1] 6.220961e-16
```

P{H < 63} = 6.2 * 10^-16

```
pnorm(66.5, mean=65, sd=0.25) - pnorm(63.5, mean=65, sd=0.25)
```

iii.

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

P{H > 63.5 && H < 66.5} = P{H < 66.5} - P{H < 63.5} = 1

```
qnorm(0.975, mean=65, sd=0.25)
```

(c)

```
## [1] 65.48999
```

65 - (65.49 - 65)=64.51
Therefore, 95% confidence interval is [64.51, 65.49]

```r
pnorm(63, mean=65, sd=2.5, lower.tail=FALSE)
```

(d)

```
## [1] 0.7881446
```

$P\{H > 66\} = 0.788$

The probability that the women is more than 66 in tall is 0.788.

(e) we use populatioin distribution in (d), which is different from sampling distribution of sample mean used in (b) (i)

4.

Standard deviation of a random variable X is the square root of its variance, which is the average of the squared difference of the mean. The meaning of sample standard deviation is a description of how far the individuals within the sample differ from the sample mean. Sample standard deviation is calculated by $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{(n-1)}}$, where n is the size of the sample. Population standard deviatioin is calculated by $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}}$, where n is the size of the population.

The meaning of standard error of sample mean is an estimate of the standard deviation of sample mean, or how far the sample mean differ from the population mean, according to Central Limit Theorem. Standard deviation of sample mean is $\sigma/\sqrt{n}$. Standard error of the mean is calculated by $s/\sqrt{n}$, where s is sample standard deviation.

Therefore, sample standard deviation is larger than standard error of the mean for a given population.

5.

Standard error $SE = s/\sqrt{n}$. The larger the sample size n, the smaller the standard error.

6.

Central Limit Theorem states:

Assuming Y is a random variable, with population mean μ and population standard deviation σ

i) The sample mean of Y is $\mu_{\overline{y_n}} = \mu$.

ii) The standard deviation of sample mean is $\sigma_{\overline{y_n}} = \sigma/\sqrt{n}$.

iii) If Y is normally distributed $Y \sim N(\mu, \sigma)$, then the sample mean of Y is also normally distributed $\overline{y_n} \sim N(\mu, \sigma/\sqrt{n})$. If Y is not normally distributed, then the sample mean of Y approximate normal distribution $\overline{y_n} \sim N(\mu, \sigma/\sqrt{n})$

Central Limit Themrem is very important because it allows us to estimate population through samples, assuming the samples are representative of the population. It states that sample means follow normal distribution, with its mean equal to population mean, and its standard deviation equal to population standard deviation divided by the square root of sample size. And this works for any population distribution, whether it's normal distribution or not.