Random Samples and the Central Limit Theorem

Contents

In this project, we will work with the sampling distribution and the Central Limit Theorem (CLT).

The Central Limit Theorem is perhaps the most powerful concept/tool that we will learn in this course, it will send you back to the beginning of this class, require you to remind yourself and use the distributions and probability principles that we studied then, and yet at the end of this class now, it also sets the stage for new beginnings for many things to come. Make sure to have fun, and good luck!

Sampling Distribution of a Statistic.

Given sample data of the form:

$$X = \{x_1, x_2, x_3, \dots, x_n\},$$

consider the following statistic:

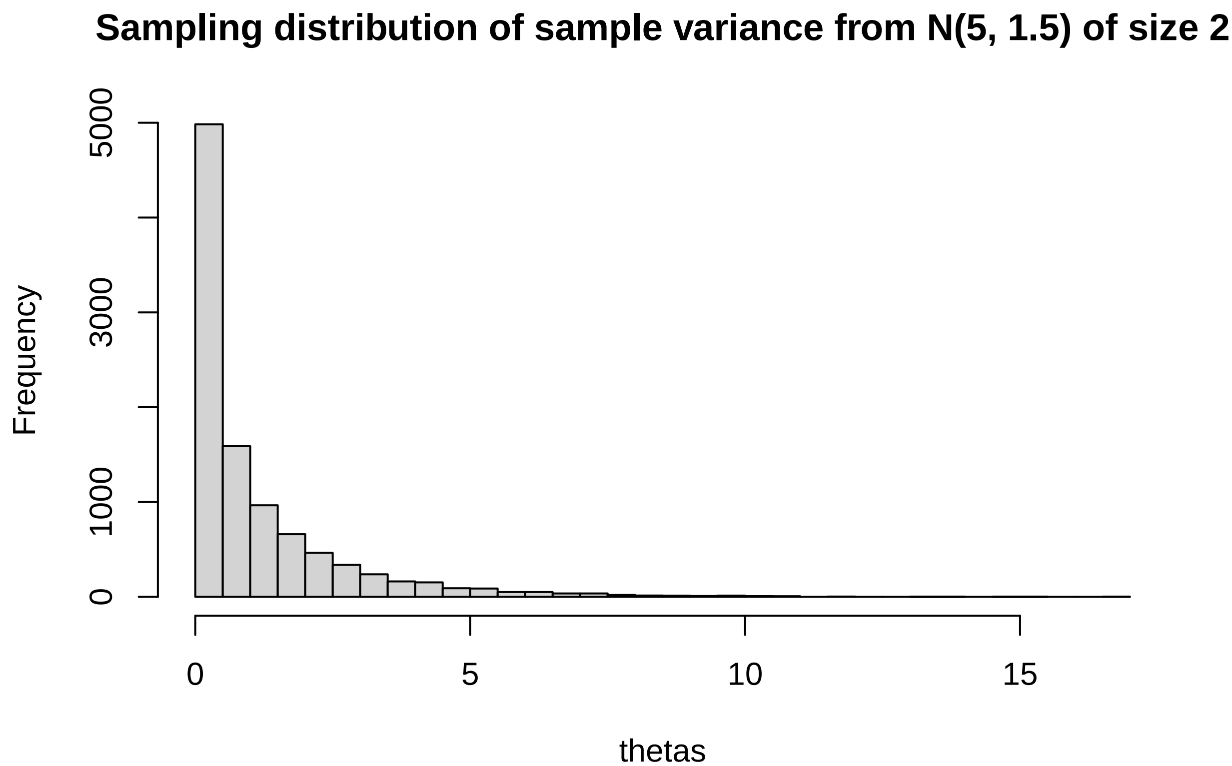
$$\hat{\theta}(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

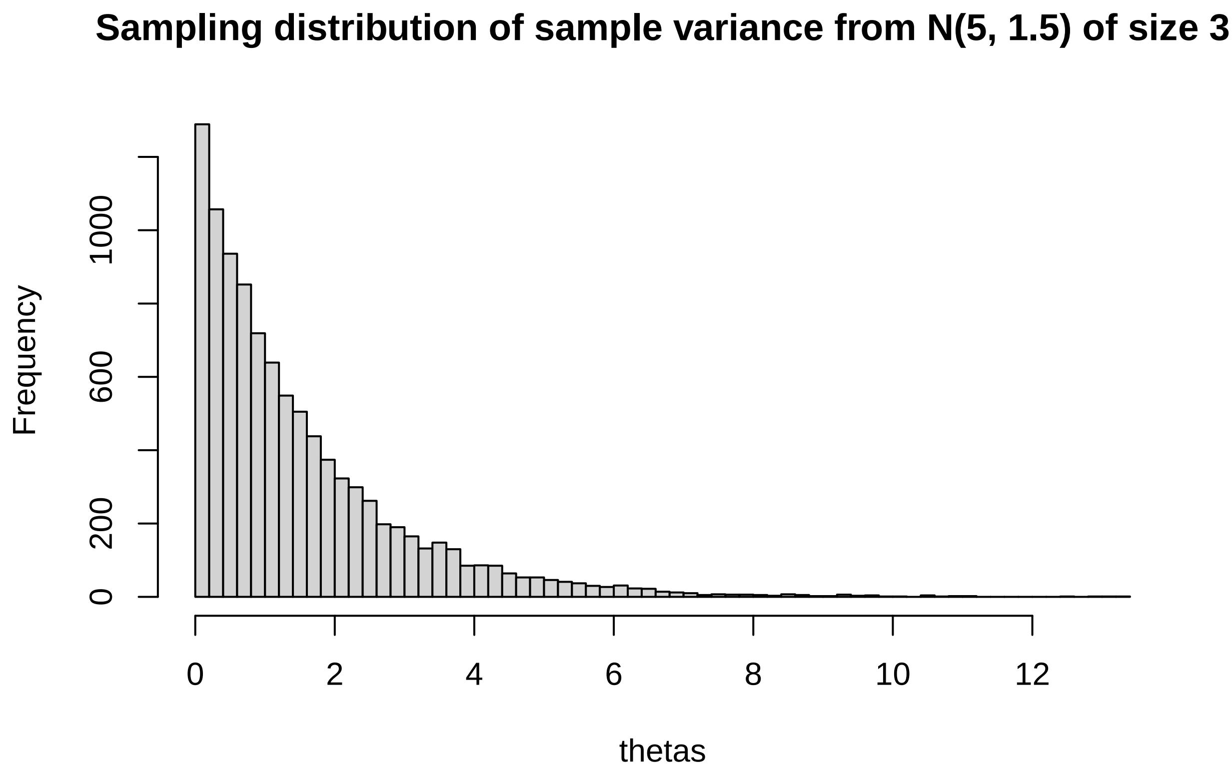
We will soon see that this statistic can be an "estimator" for the population variance σ^2. For now, write a function "theta_hat" that calculates the value of the statistic given sample data "samp".

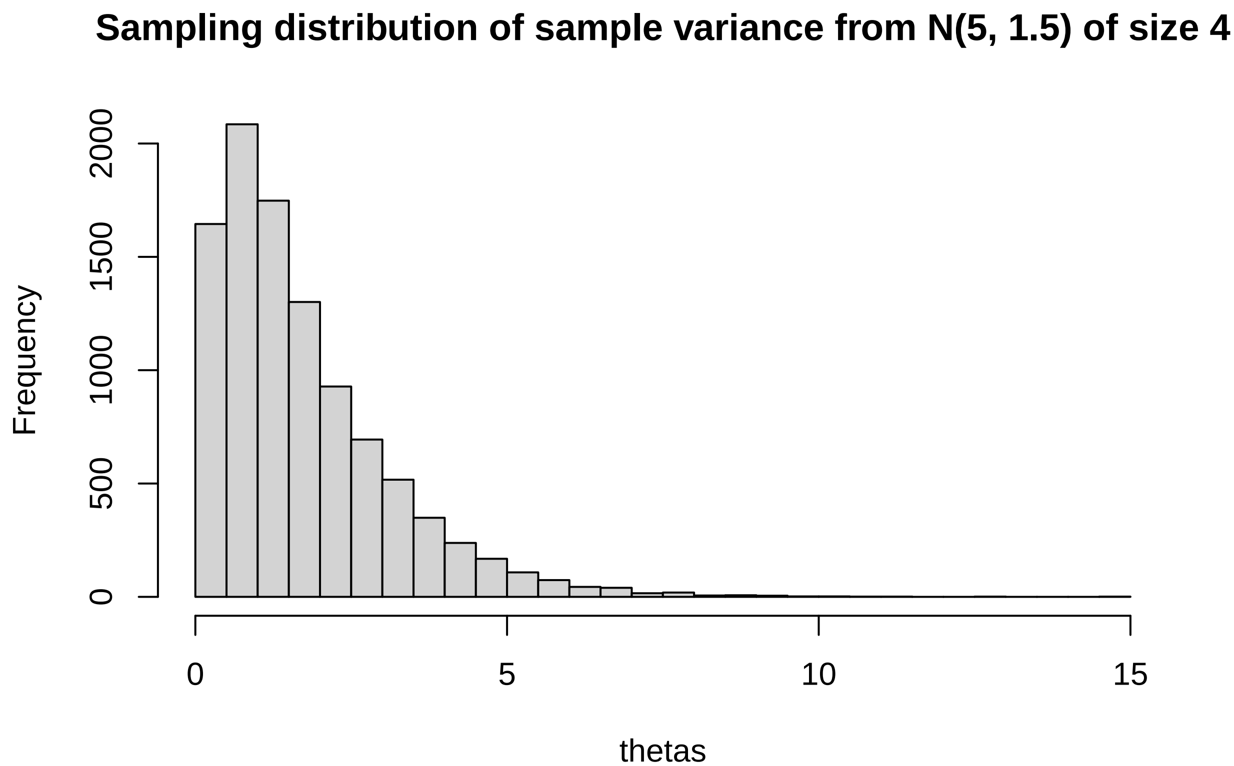
```r
theta_hat <- function(samp) {
  avg <- mean(samp)
  mean(sapply(samp, function(x) (x - avg)^2))
}
```

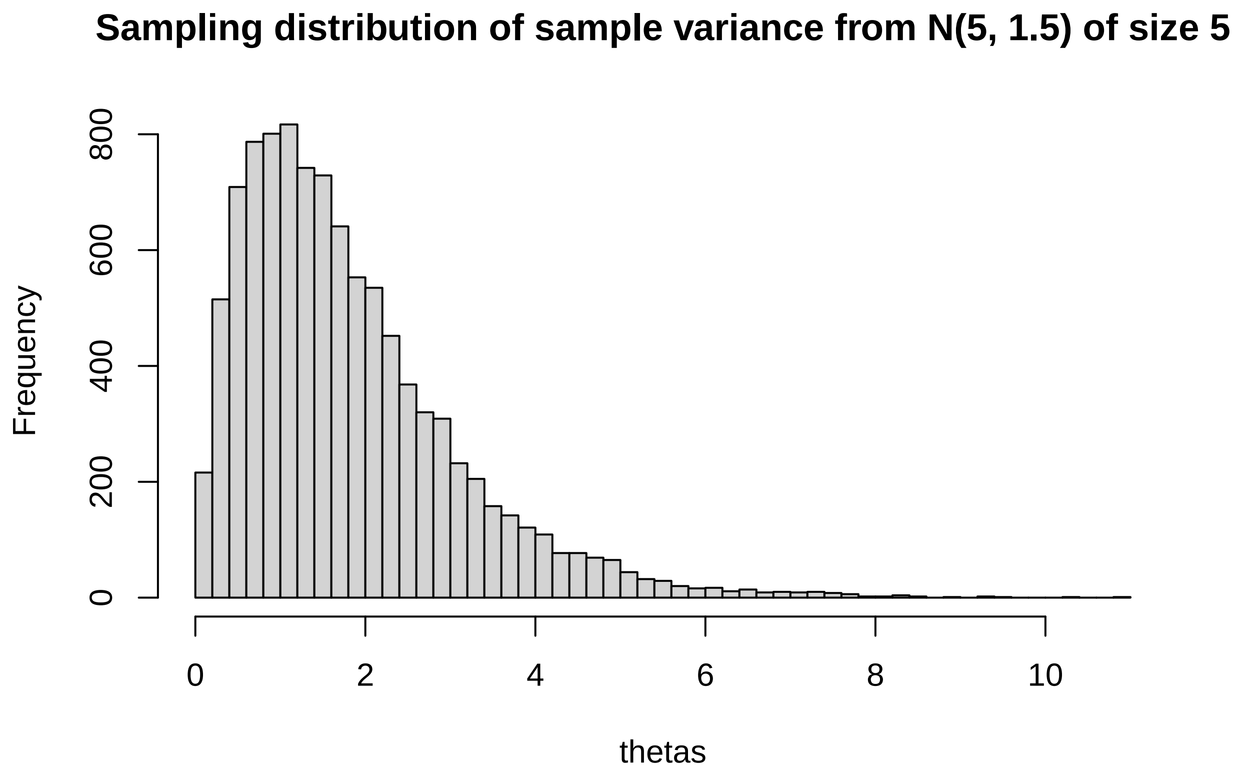
Use the replicate and hist function to calculate the sampling distribution of $\hat{\theta}$ when working with random samples coming from $N(\mu = 5, \sigma = 1.5)$ of sizes $n = 2, 3, 5, 10, 50, 500$.

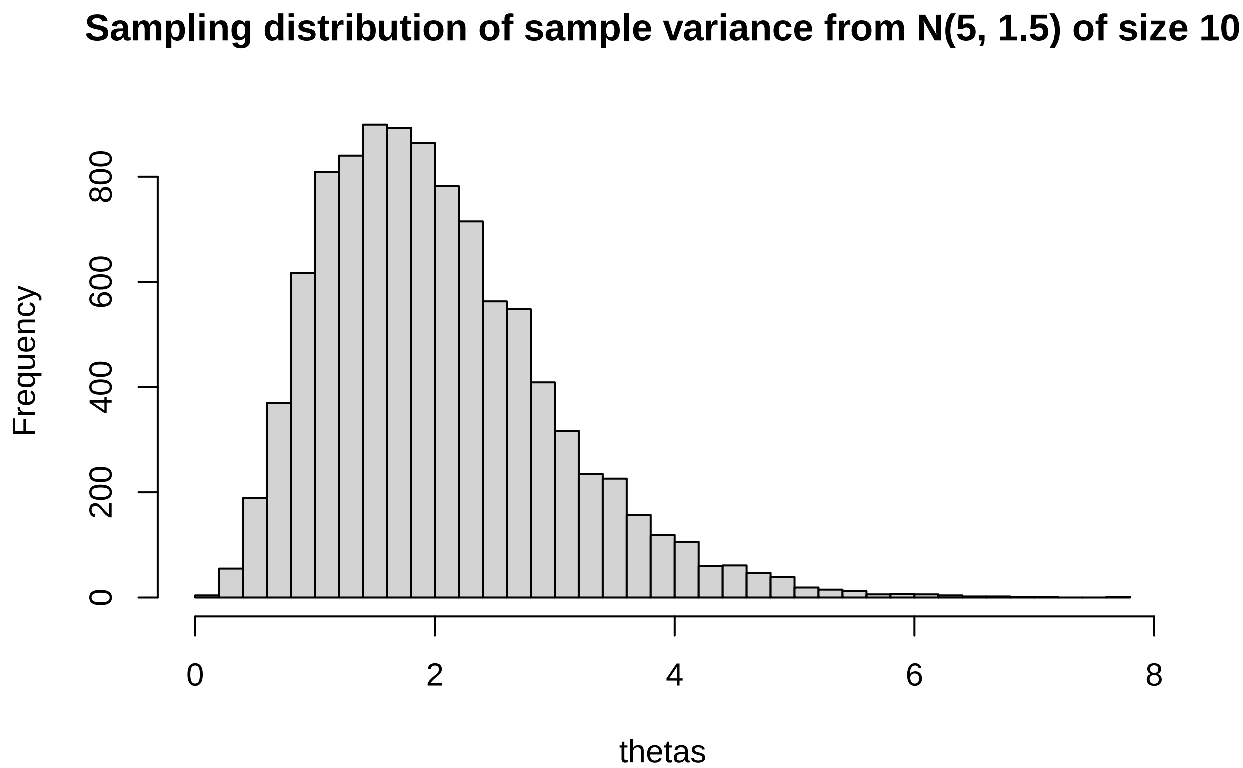
```r
B <- 10000
sizes <- c(2, 3, 4, 5, 10, 50 ,500)

for(n in sizes){
  thetas <- replicate(B, {
    samp <- rnorm(n, 5, 1.5)
    theta_hat(samp)
  })
  hist(thetas, breaks = 50,
       main = paste("Sampling distribution of sample variance from N(5, 1.5) of size", n))
}
```


Sampling distribution of sample variance from N(5, 1.5) of size 2
Frequency
0
1000
3000
5000
0
5
10
15
thetas

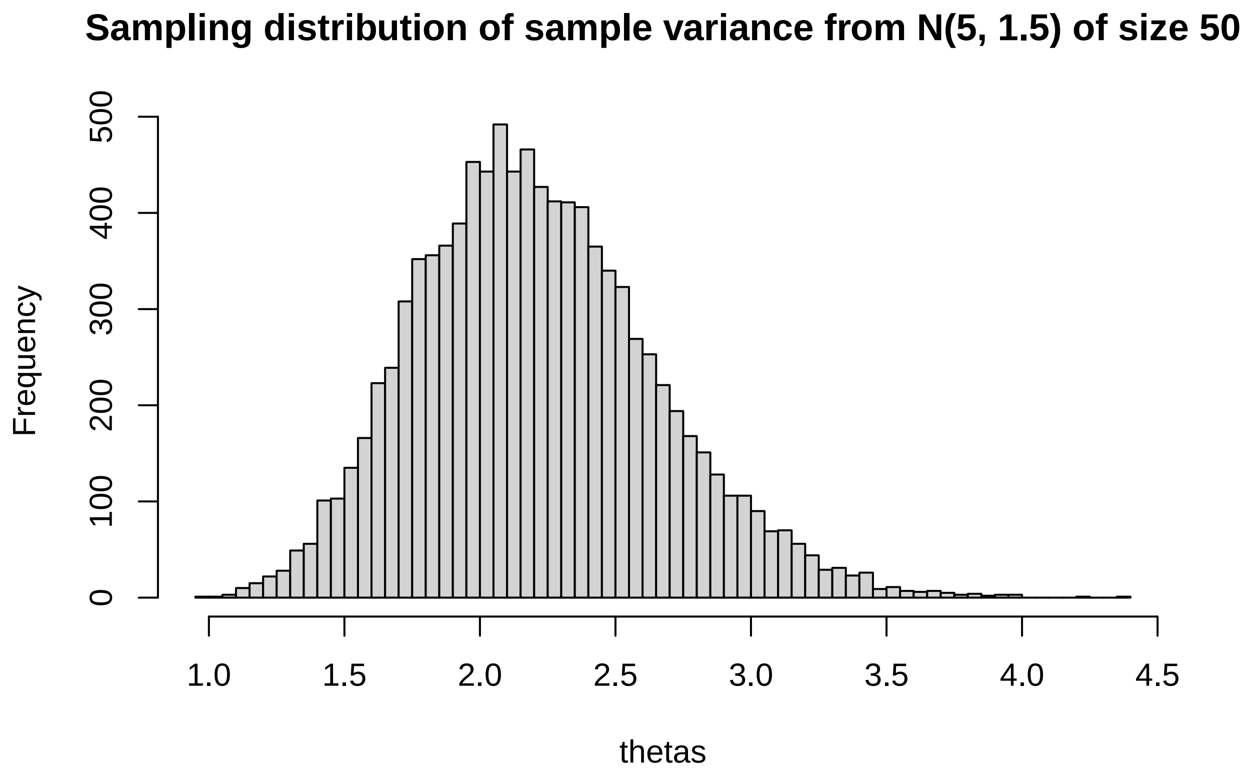

Sampling distribution of sample variance from N(5, 1.5) of size 3
Frequency
0
200
600
1000
0
2
4
6
8
10
12
thetas

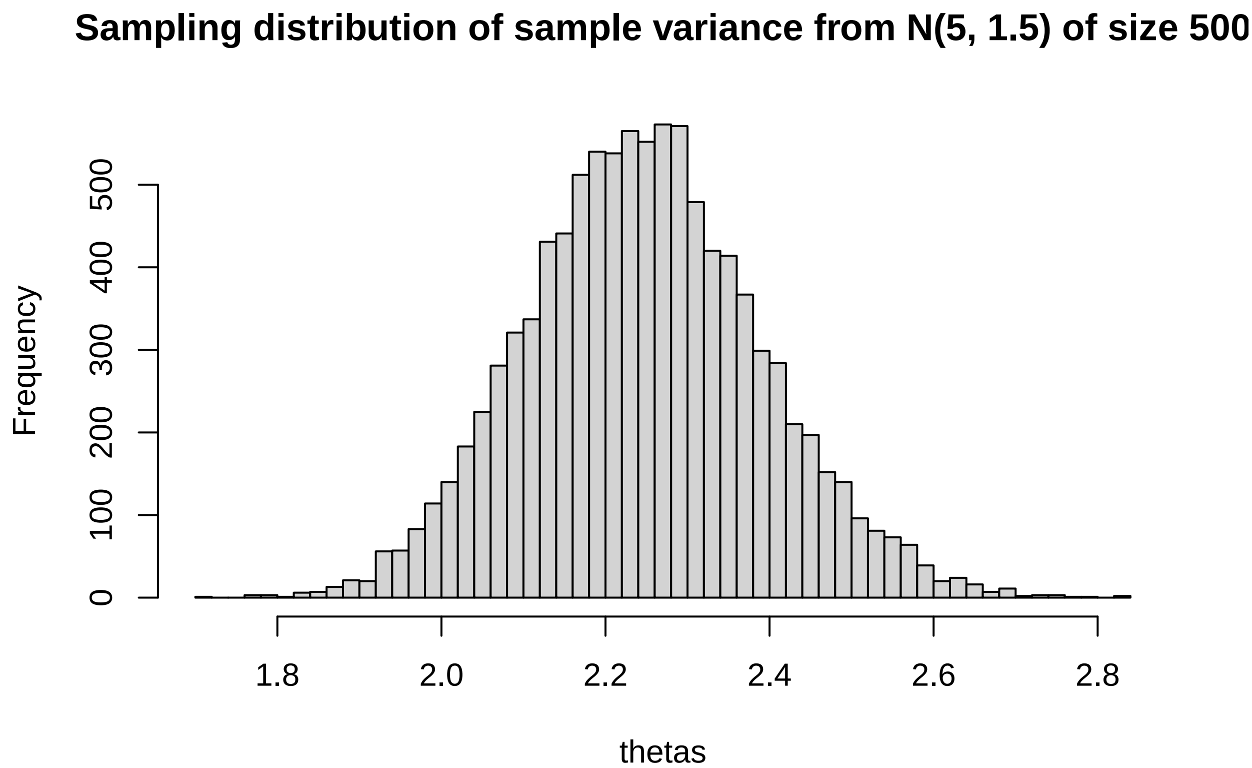

Sampling distribution of sample variance from N(5, 1.5) of size 4
Frequency
0
500
1000
1500
2000
0
5
10
15
thetas


Sampling distribution of sample variance from N(5, 1.5) of size 5
Frequency
0
200
400
600
800
0
2
4
6
8
10
thetas



Sampling distribution of sample variance from N(5, 1.5) of size 10
Frequency
0
200
400
600
800
0
2
4
6
8
thetas


Sampling distribution of sample variance from N(5, 1.5) of size 50
Frequency
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
thetas



For each of these cases of sample sizes, calculate the empirical expected value.

```
B <- 10000
sizes <- c(2, 3, 4, 5, 10, 50 ,500)

for(n in sizes){
  thetas <- replicate(B, {
    samp <- rnorm(n, 5, 1.5)
    theta_hat(samp)
  })
  print(paste("Sample size:",  n, "Empirical exp value of theta_hat ", round(mean(thetas), 5)))
}
```

```
## [1] "Sample size: 2 Empirical exp value of theta_hat  1.11225"
## [1] "Sample size: 3 Empirical exp value of theta_hat  1.52175"
## [1] "Sample size: 4 Empirical exp value of theta_hat  1.67492"
## [1] "Sample size: 5 Empirical exp value of theta_hat  1.81222"
## [1] "Sample size: 10 Empirical exp value of theta_hat  2.01975"
## [1] "Sample size: 50 Empirical exp value of theta_hat  2.20178"
## [1] "Sample size: 500 Empirical exp value of theta_hat  2.24609"
```

As sample size increases what is the relation between the empirical mean of $\hat{\theta}$ and σ^2? ANSWER: The empirical mean of $\hat{\theta}$ approaches σ^2 from the left. In this case, the value of $\hat{\theta}$ is closer and closer to $1.5^2 = 2.25$.

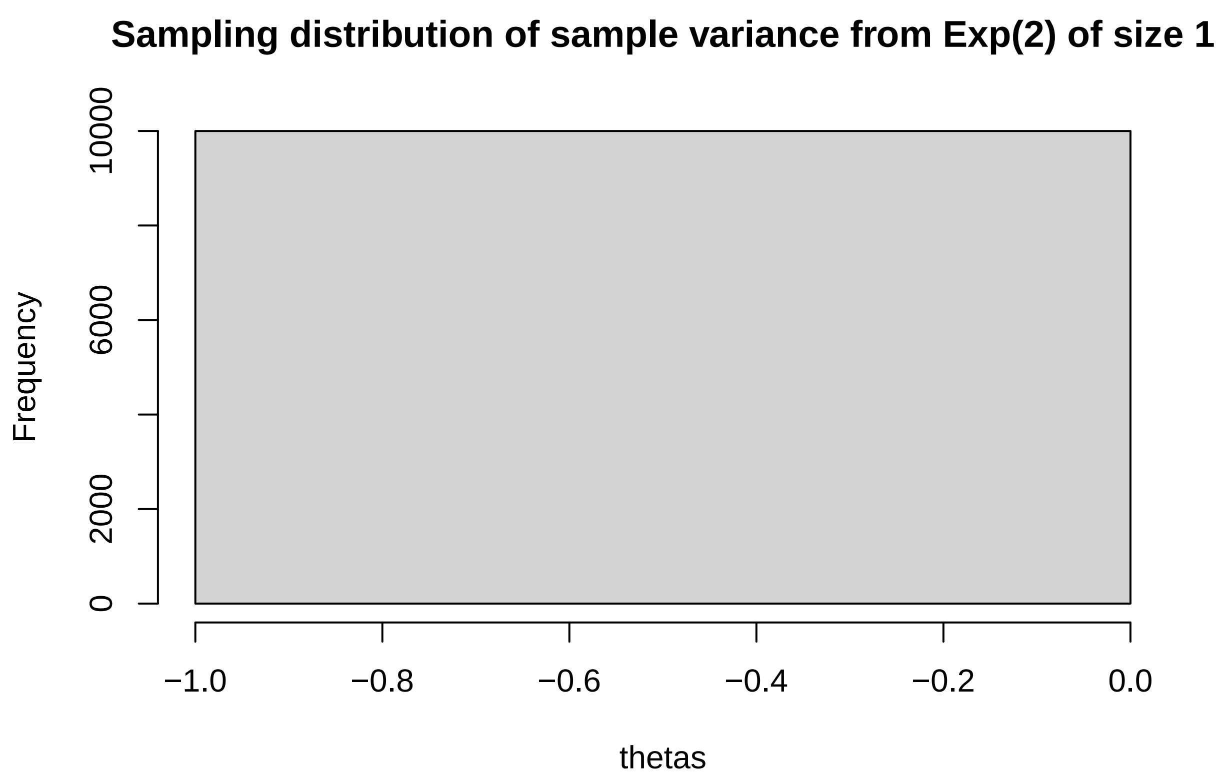
Central Limit Theorem

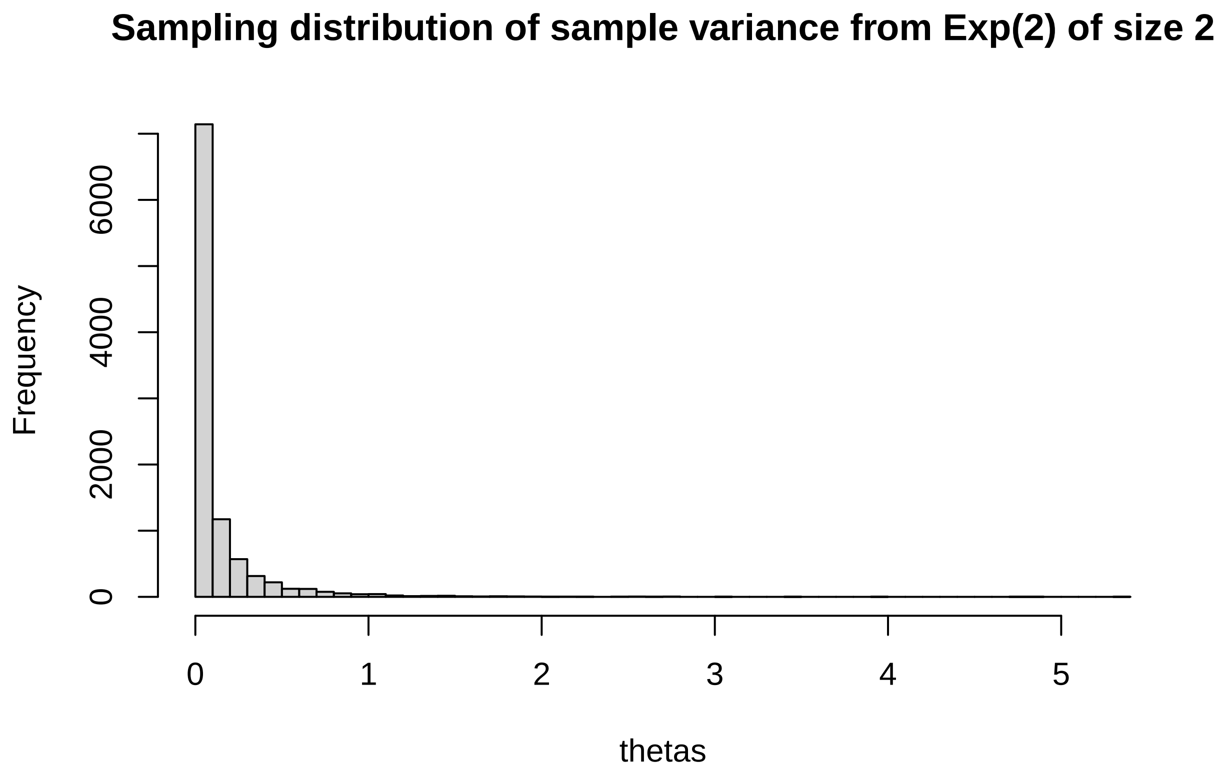
The Exponential Distribution

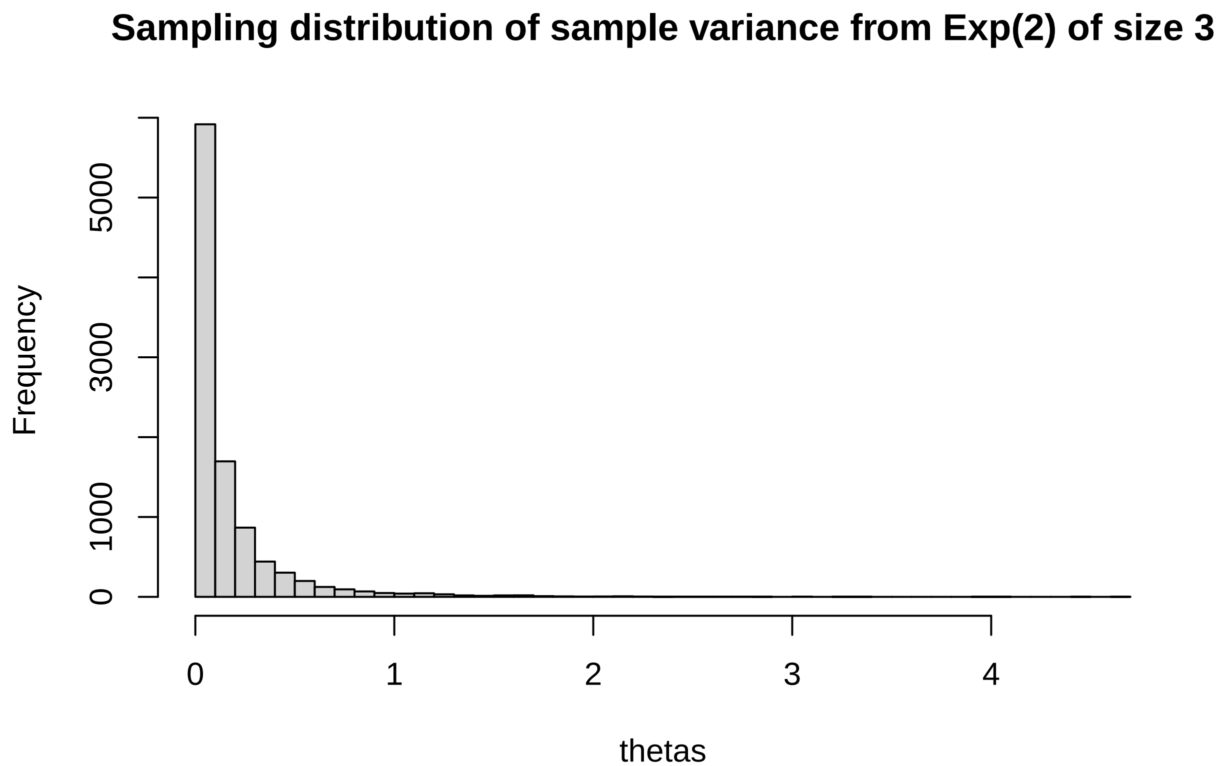
Suppose we are working with a population that has the exponential distibution with $\lambda = 2$.

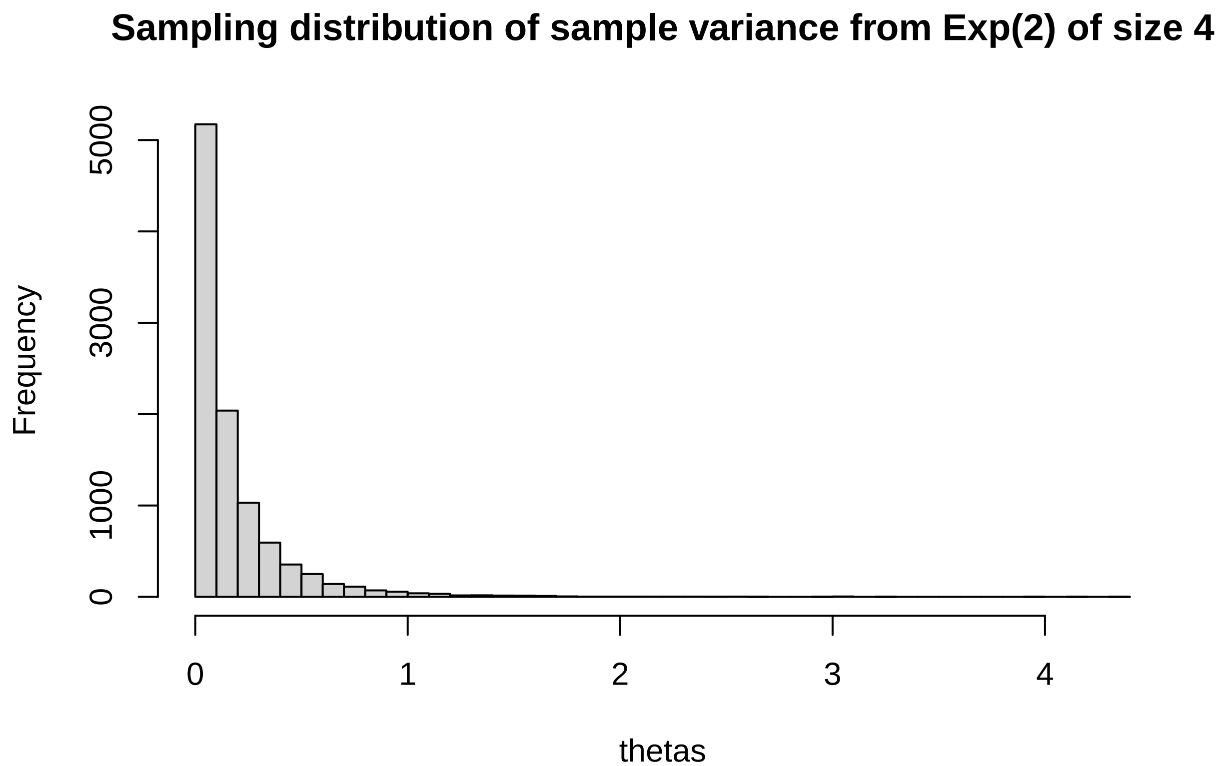
Use the replicate function to get the histograms for the sampling distribution of the sample mean when working with sample sizes $n = 1, 2, 3, 4, 15, 500$. Be sure to have appropriate titles for your histograms.

```r
B <- 10000
sizes <- c(1, 2, 3, 4, 15, 500)

for(n in sizes){
  thetas <- replicate(B, {
    samp <- rexp(n, 2)
    theta_hat(samp)
  })
  hist(thetas, breaks = 50,
       main = paste("Sampling distribution of sample variance from Exp(2) of size", n))
}
```

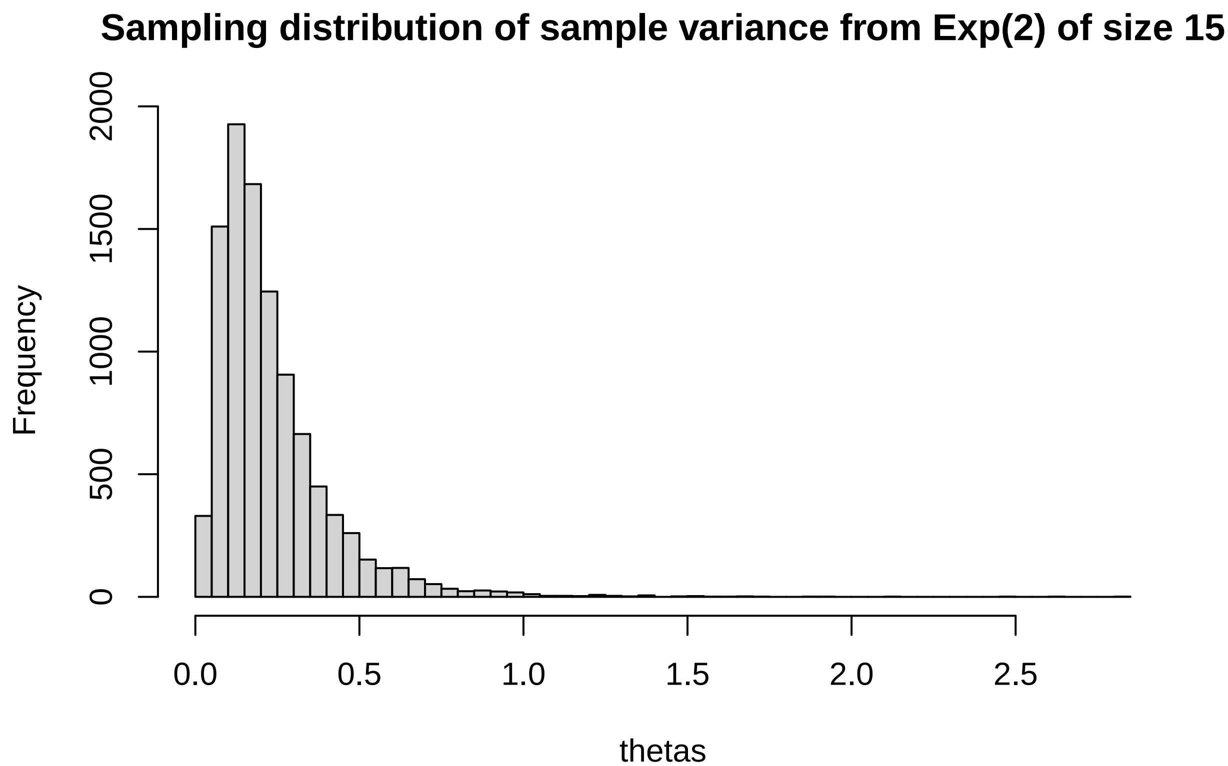


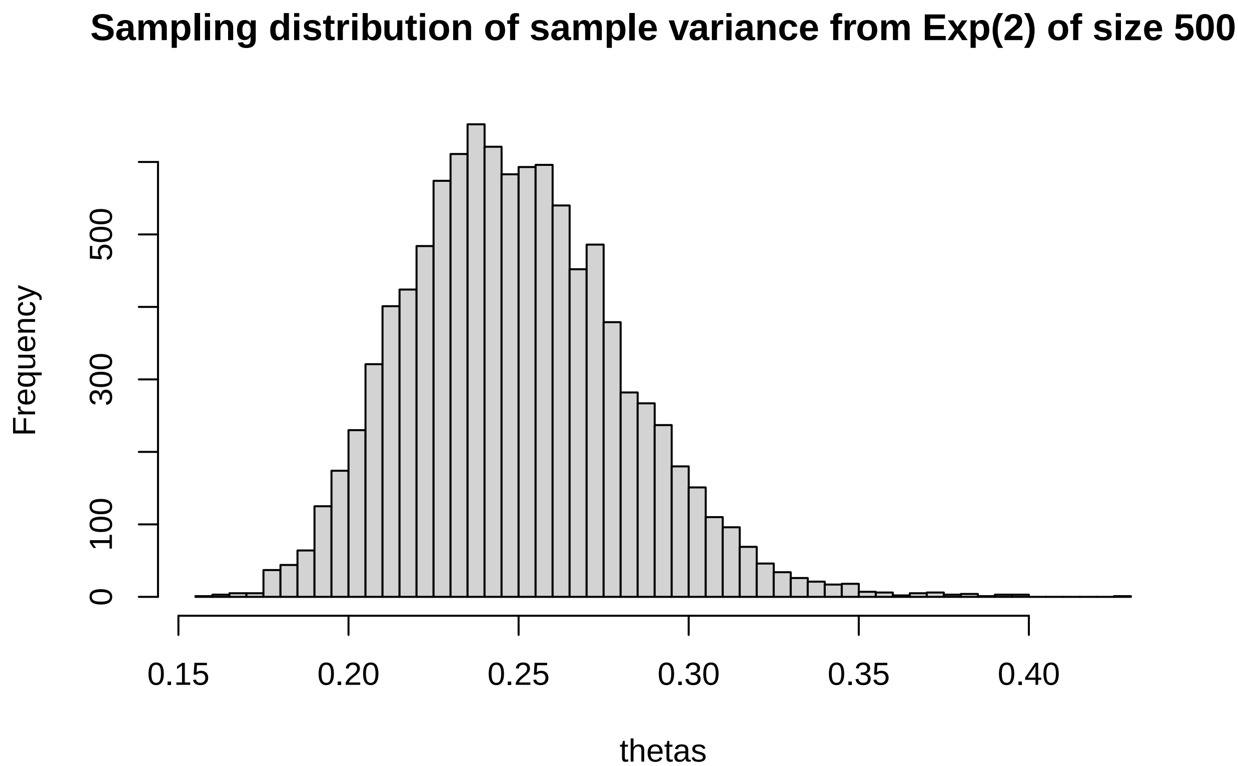

Sampling distribution of sample variance from Exp(2) of size 2
Frequency
0
2000
4000
6000
0
1
2
3
4
5
thetas


Sampling distribution of sample variance from Exp(2) of size 3
Frequency
0
1000
3000
5000
0
1
2
3
4
thetas



Sampling distribution of sample variance from Exp(2) of size 4
Frequency
0
1000
3000
5000
0
1
2
3
4
thetas


Sampling distribution of sample variance from Exp(2) of size 15
Frequency
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
2.0
2.5
thetas



What do you notice? ANSWER: Similarly to the previous experiment, the distribution of $\hat{\theta}$ approaches normal with $E(\hat{\theta}) = V(X) = 0.25$.

Discrete Uniform distibution

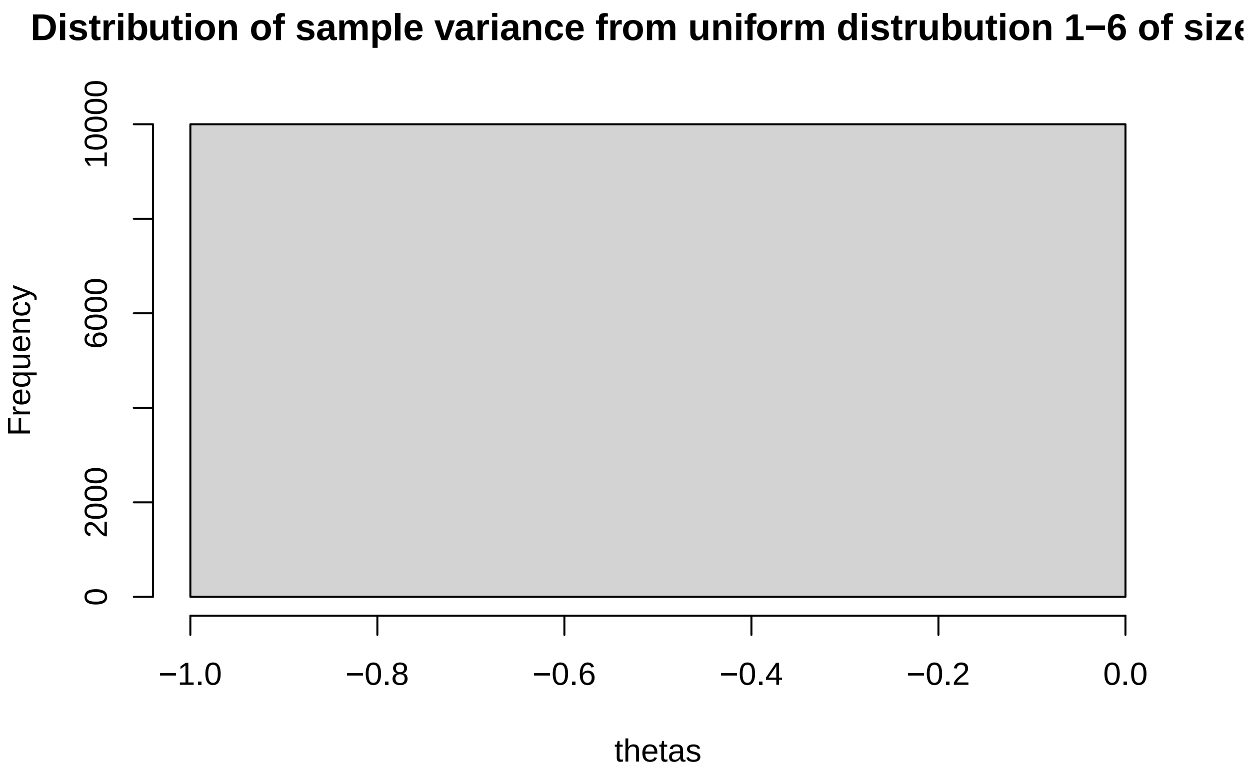
Suppose we are working with the discrete uniform random variable taking values $\{1, 2, 3, 4, 5, 6\}$.

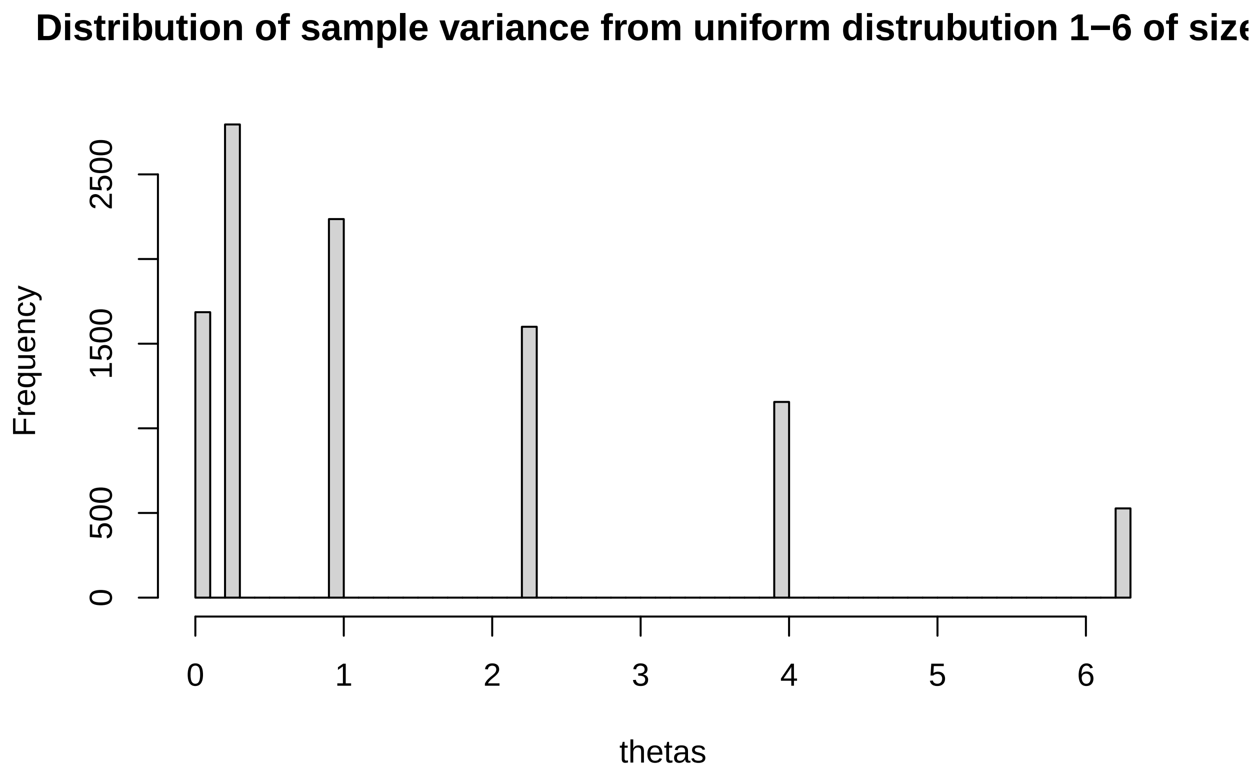
Define a function "disc_samp" that takes input "n" and returns a random sample of size "n" from this distribution.

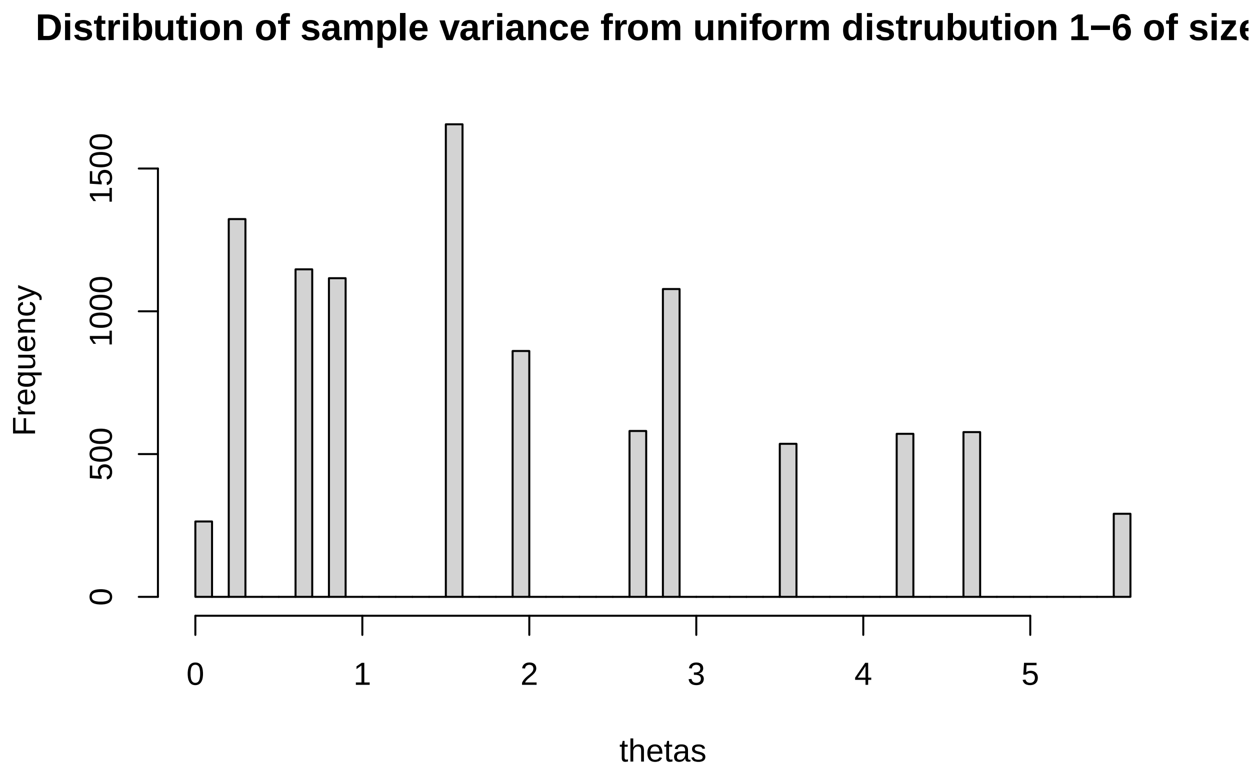
```
disc_samp <- function(n) sample(1:6, n, replace=TRUE)
```

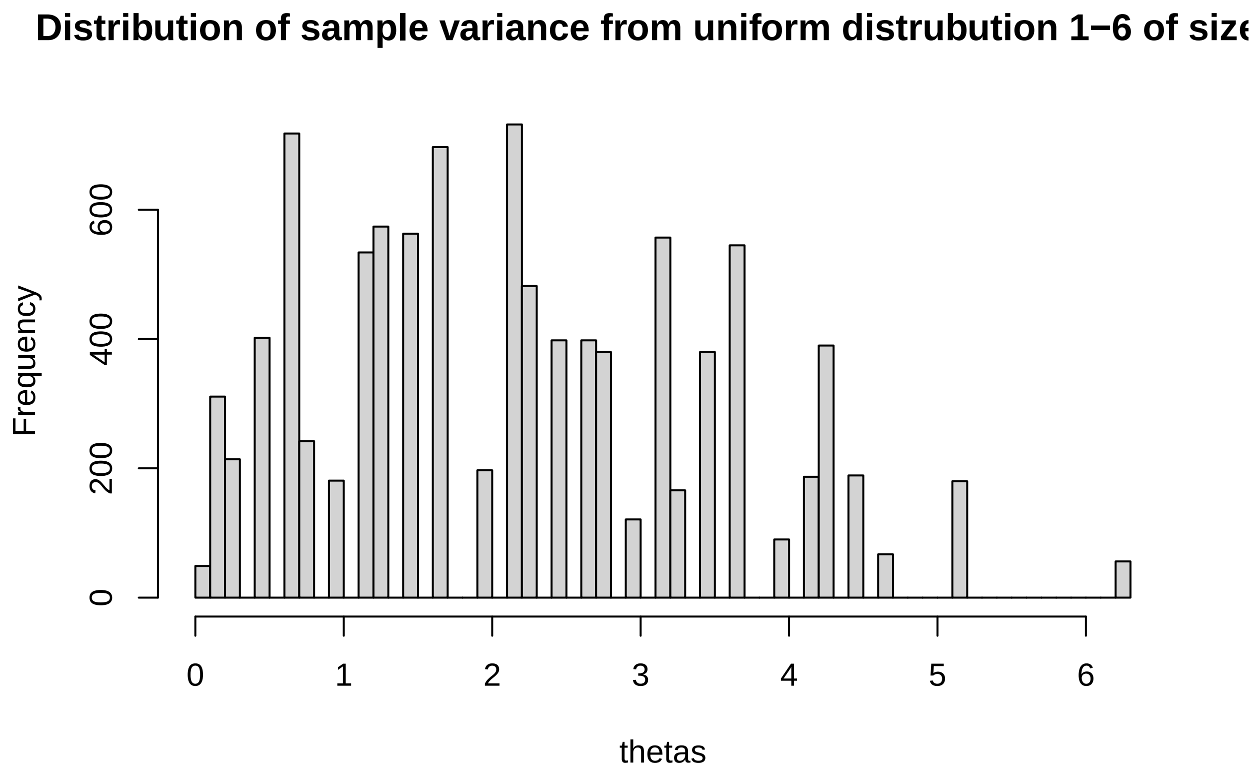
Use the "disc_samp" function and the replicate function to to get the histograms for the sampling distribution of the sample mean when working with sample sizes $n = 1, 2, 3, 4, 15, 500$. Be sure to have appropriate titles for your histograms.

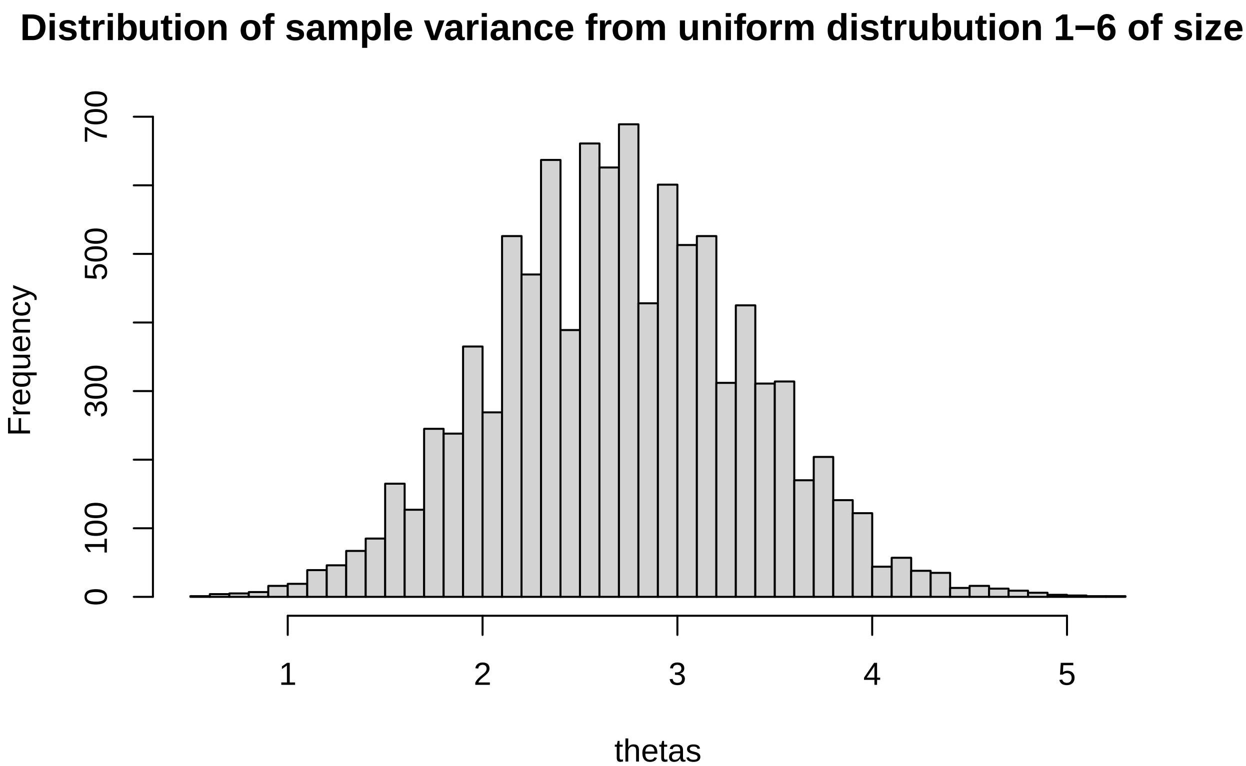
```
B <- 10000
sizes <- c(1, 2, 3, 4, 15, 500)

for(n in sizes){
  thetas <- replicate(B, {
    samp <- disc_samp(n)
    theta_hat(samp)
  })
  hist(thetas, breaks = 50,
       main = paste("Distribution of sample variance from uniform distrubution 1-6 of size", n))
}
```

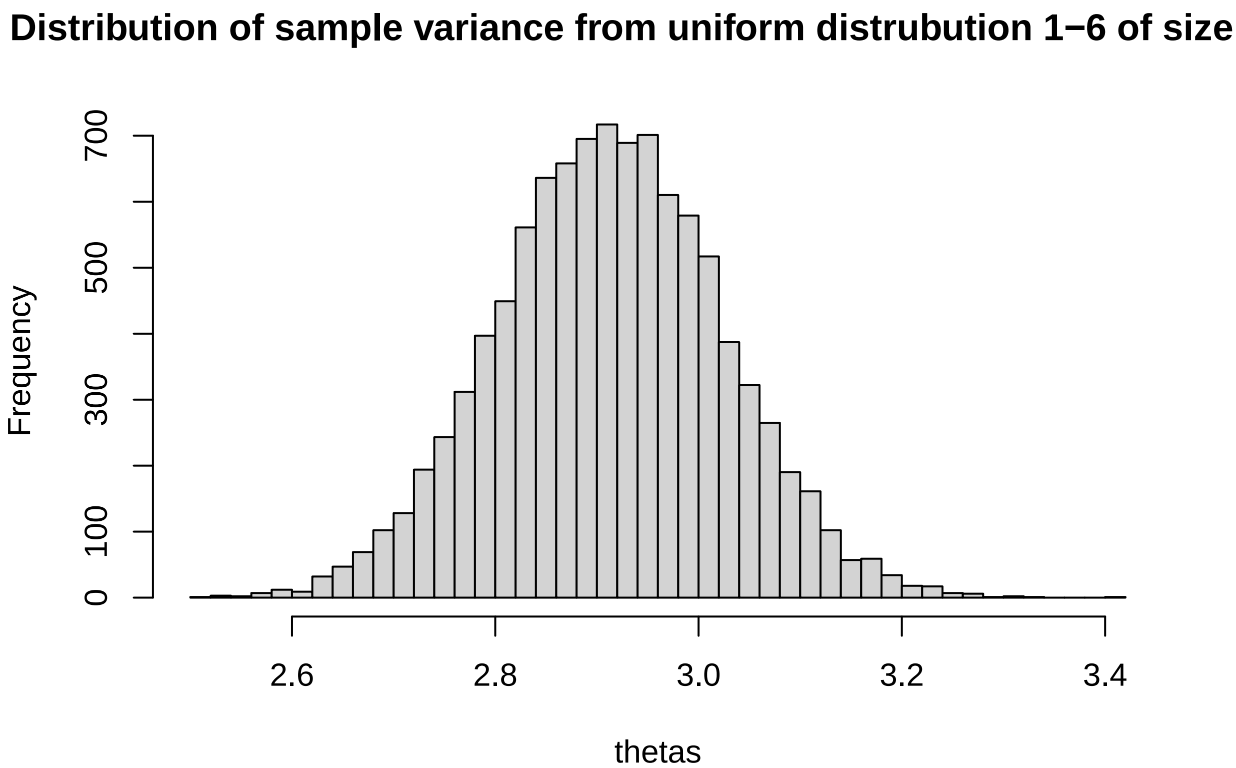

Distribution of sample variance from uniform distrubution 1−6 of size
Frequency
0
2000
6000
10000
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
thetas


Distribution of sample variance from uniform distrubution 1−6 of size
Frequency
0
500
1500
2500
0
1
2
3
4
5
6
thetas


Distribution of sample variance from uniform distrubution 1−6 of size
Frequency
0
500
1000
1500
0
1
2
3
4
5
thetas


Distribution of sample variance from uniform distrubution 1−6 of size
Frequency
0
200
400
600
thetas
0
1
2
3
4
5
6


Distribution of sample variance from uniform distrubution 1−6 of size
Frequency
0
100
300
500
700
1
2
3
4
5
thetas



What do you notice? ANSWER: Similarly to the previous experiment, the distribution of $\hat{\theta}$ approaches normal with $E(\hat{\theta}) = V(X) = \frac{6^2-1}{12} = 2.917$.