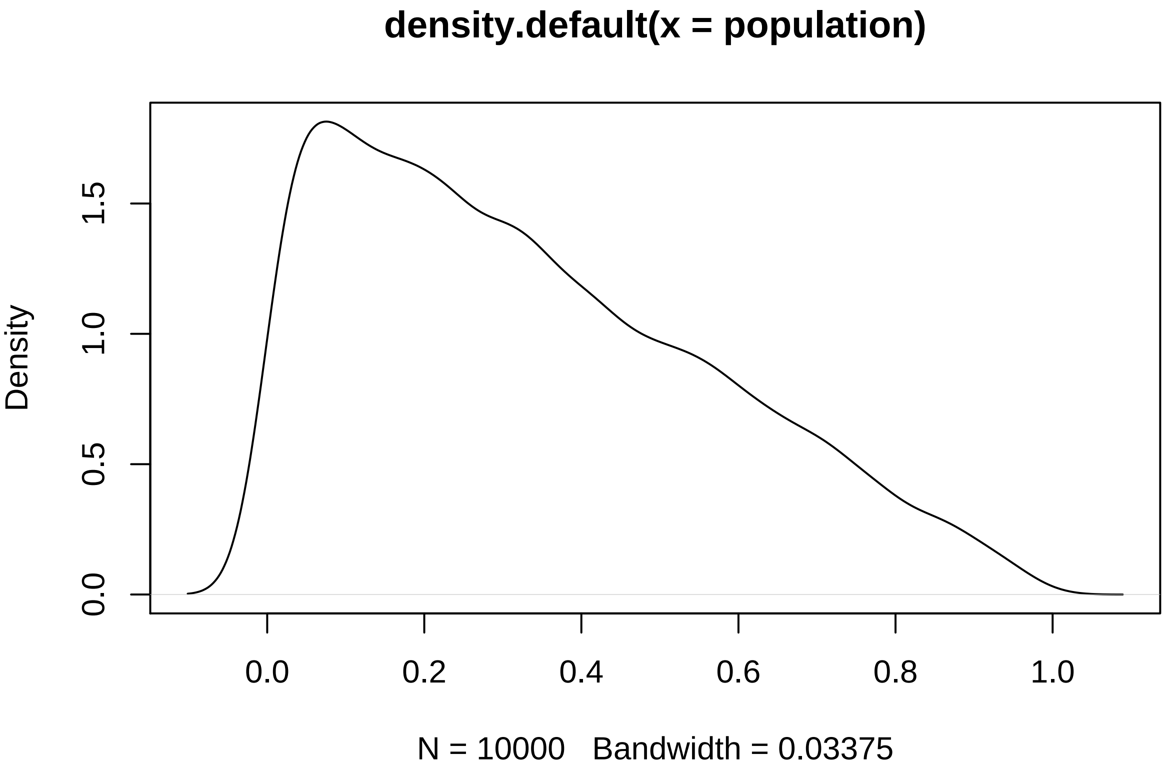
Lecture 1

Han Zhang

2/1/2021

Population is 10,000 units following Beta distribution.

```
population <- rbeta (10000, 1, 2)
plot (density (population))
```



```
# population mean
mean(population)
```

```
## [1] 0.3333757
```

```
# population variance
var(population)
```

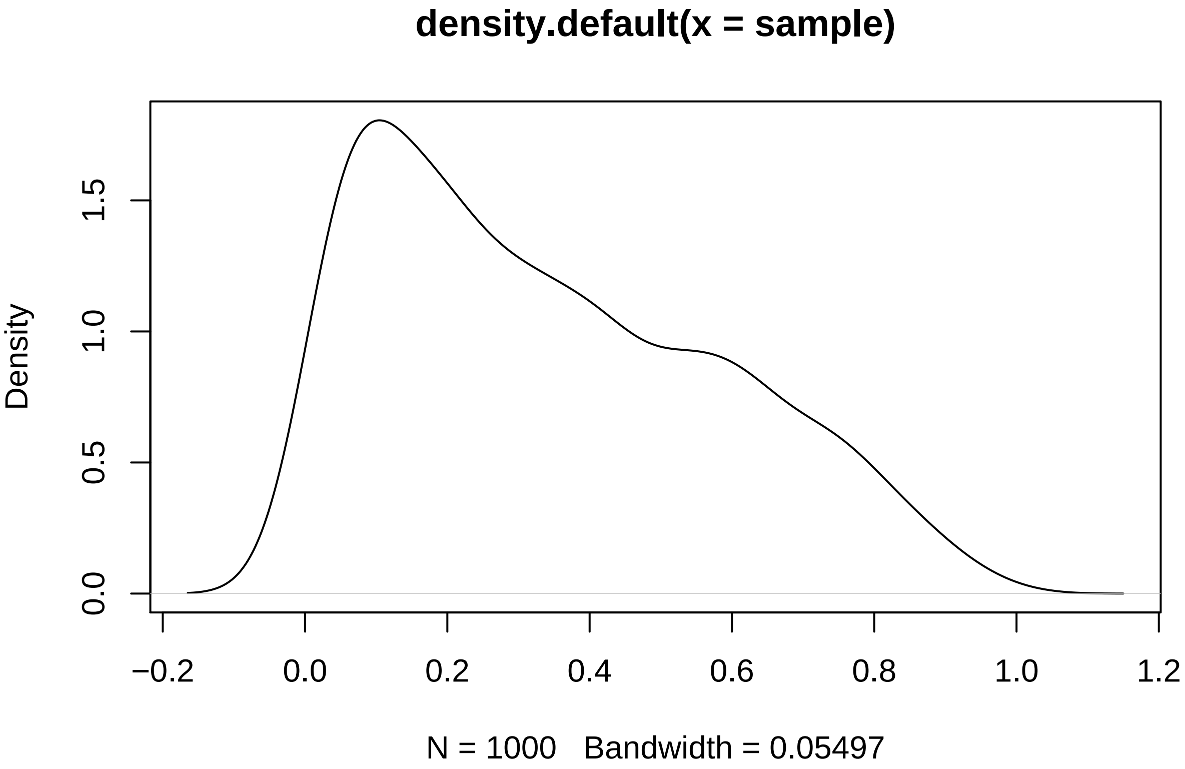
```
## [1] 0.05598226
```

Create samples from the population with size 1000

Still about X, original random variable.

We calculated $\bar{X}$ and $\hat{V}(X)$ Compare these estimates with population quantities. What do you find?

```
sample = sample(population, size = 1000)
plot (density (sample))
```



```
# sample mean as estimator of population mean
est.mean = mean(sample)

# unbiased estimator of population variance
est.var = 1000/(1000-1) * var(sample)
```

Normally-approximated confidence intervals

First we need to estimate standard error of the sample mean

We know it is

$\hat{V}(\bar{X}) = \frac{\hat{V}(X)}{n}$. We have calculatd $\hat{V}(X)$ in the previous code chunk.

```
sample_mean = mean(sample)

#standard error of the sample mean
ss = sqrt(est.var/1000)

print ("point estimate of mean")
```

```
## [1] "point estimate of mean"
```

```
print (sample_mean)
```

```
## [1] 0.3418856
```

```
print ("95% normal-approximated confidence interval of mean")
```

```
## [1] "95% normal-approximated confidence interval of mean"
```

```
c(sample_mean - 1.96 * ss, sample_mean + 1.96 * ss)
```

```
## [1] 0.3268072 0.3569640
```

bootstrap confidence interval for sample mean

Bootstrap is slower, but needs no Central Limit Theorem.

```
bootstrap_means = c() # store bootstrapped medians
for (i in 1:10000){
  # resample from the sample with replacement
  boot_data <- sample(sample, 1000, replace = T)
  boot_mean <- mean(boot_data)
  bootstrap_means <- c(bootstrap_means, boot_mean)
}

# then simple quantile function to
print ("point estimate of mean")
```

```
## [1] "point estimate of mean"
```

```
print (sample_mean)
```

```
## [1] 0.3418856
```

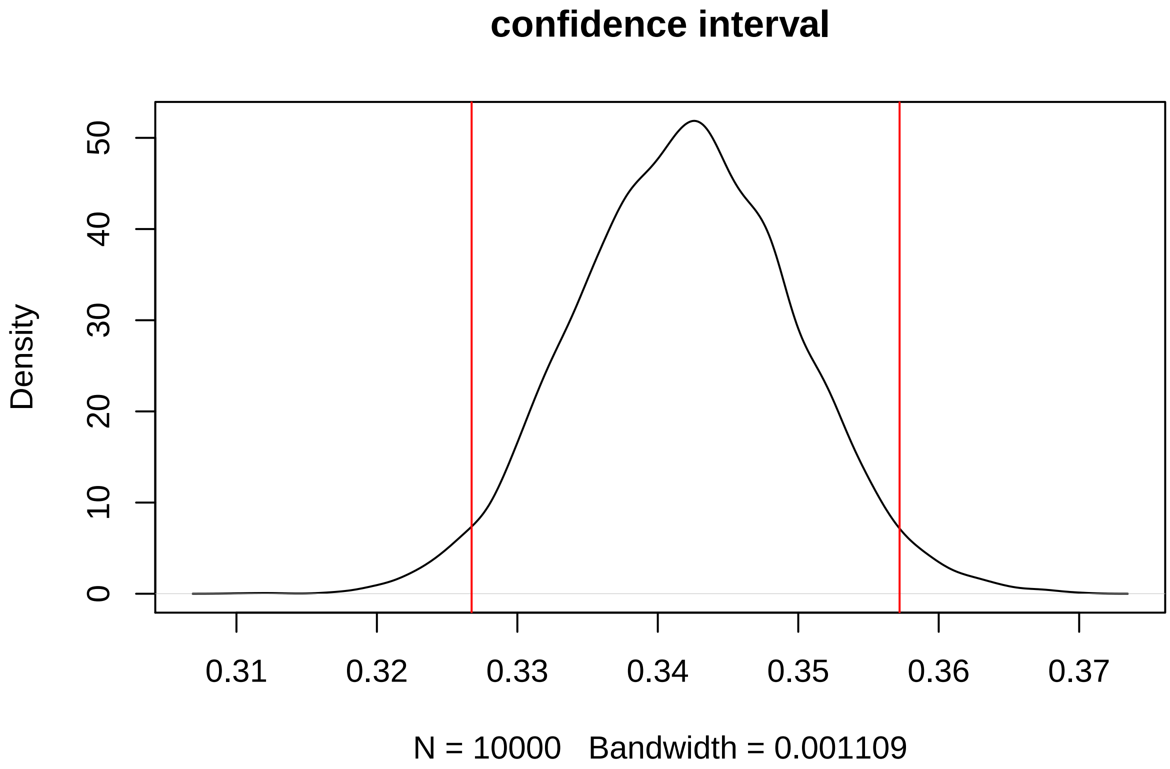
```
print ("95% bootstrap confidence interval of mean")
```

```
## [1] "95% bootstrap confidence interval of mean"
```

```
conf = quantile(bootstrap_means, c(0.025, 0.975))
print (conf)
```

```
##      2.5%     97.5%
## 0.3267431 0.3572145
```

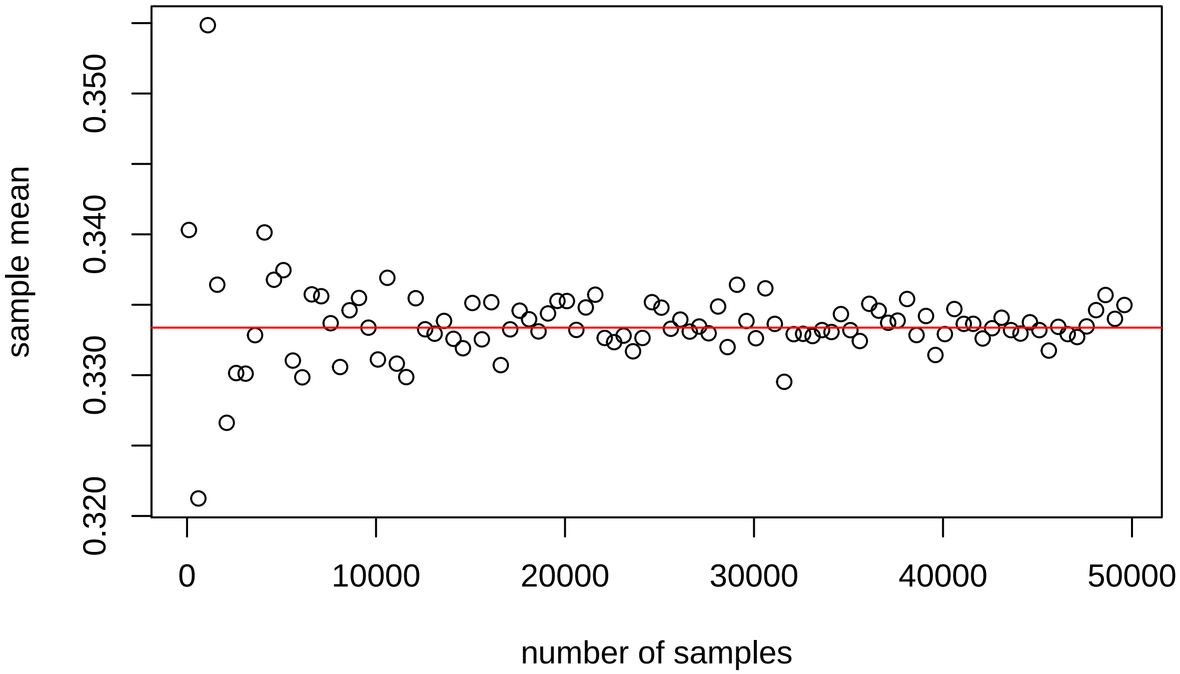
```
plot(density(bootstrap_means), main = "confidence interval")
abline(v = conf[1], col = "red")
abline(v = conf[2], col = "red")
```



Law of Large Numbers

As n increases, sample mean approaches population mean.

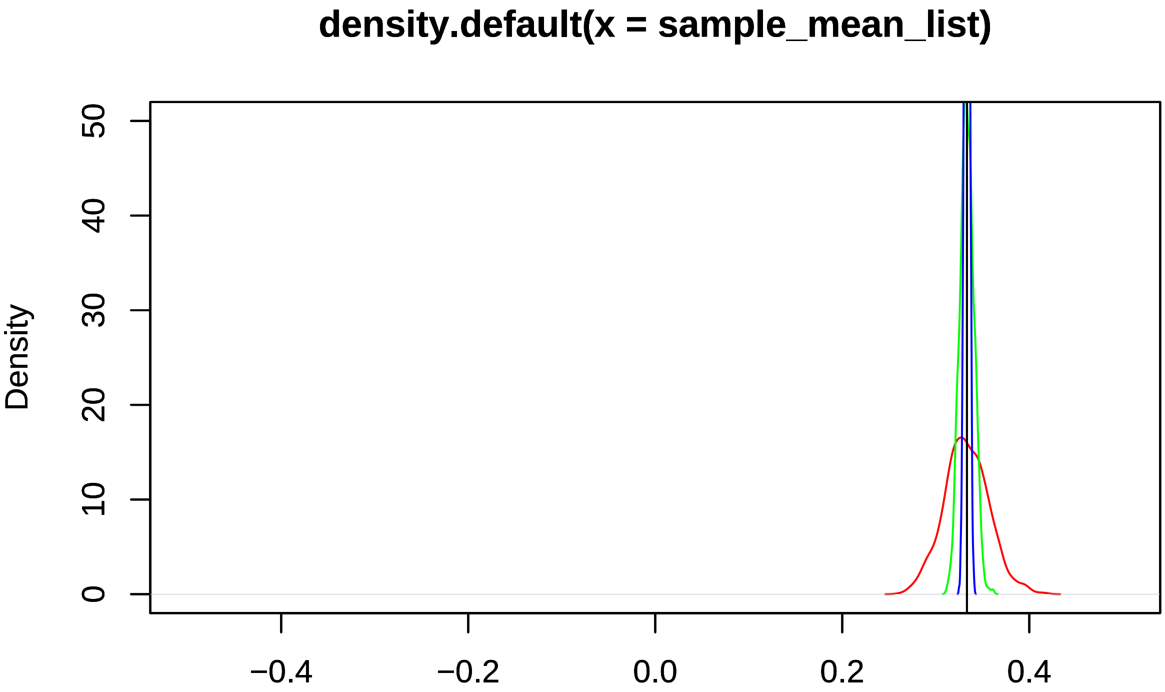
```
sample_mean_list <- c()
sample_times <- seq(100, 50000, 500)
for (n in sample_times ){
  sample <- sample(population, size = n, replace = TRUE)
  sample_mean_list <- c(sample_mean_list, mean(sample))
}

plot(sample_times, sample_mean_list, xlab = "number of samples", ylab = "sample mean")
abline(h = mean(population), col = "red")
```



Central limit theorem

- As n increases, distribution of sample means approaches normal

```r
# Now we show the central limit theorem
sample_times <- c(100, 1000, 10000)
col <- c("red", "green", "blue")
i = 1
for (n in sample_times ){
  # central limit theorem talks about distribution of the sample mean
  # we cannot calculate the distribution for a single sample, so we draw sample multiple times
  sample_mean_list <- c()
  for (m in 1:1000)
  {
    sample <- sample(population, size = n, replace = TRUE)
    sample_mean_list <- c(sample_mean_list, mean(sample))
  }
  # sample_mean_list_standard <- sample_mean_list - mean(population)
  plot(density (sample_mean_list),  col = col[i], xlim = c(-0.5, 0.5), ylim  = c(0,50), xlab = "")
  abline(v = mean(population), col = "black")
  par(new = T)
  i = i  + 1
}
```



bootstrap

For instance, we care about the *median* of population. Sample median is clearly an estimate of population median. But how can we obtain the *95% confidence interval* of estimated median?

```
sample = sample(population, size = 1000)
```

```
sample_median = median(sample)


bootstrap_medians = c() # store bootstrapped medians
for (i in 1:10000){
  # resample from the sample with replacement
  boot_data <- sample(sample, 1000, replace = T)
  boot_median <- median(boot_data)
  bootstrap_medians <- c(bootstrap_medians, boot_median)
}

# then simple quantile function to
print ("point estimate of median")
```

```
## [1] "point estimate of median"
```

```
print (sample_median)
```

```
## [1] 0.2817869
```

```
print ("95% confidence interval of median")
```

```
## [1] "95% confidence interval of median"
```

```
quantile(bootstrap_medians, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2576151 0.2980866
```

As you may see, the bootstrapped confidence interval may not be symmetric; normal approximated confidence interval, on the other hand, is by definition symmetric.