Law of Large Numbers - Simulation

Yichen Han

Goal Setting: we want to visualize LLN so that we can prove:

- that the sample mean converges to the population expectation, and
- that the variance of sample mean converges to 0.

Idea: we simulate a random variable with a known expectation and variance, draw many samples of different sizes from it, and then we plot the sample mean and its variance for different sample sizes.

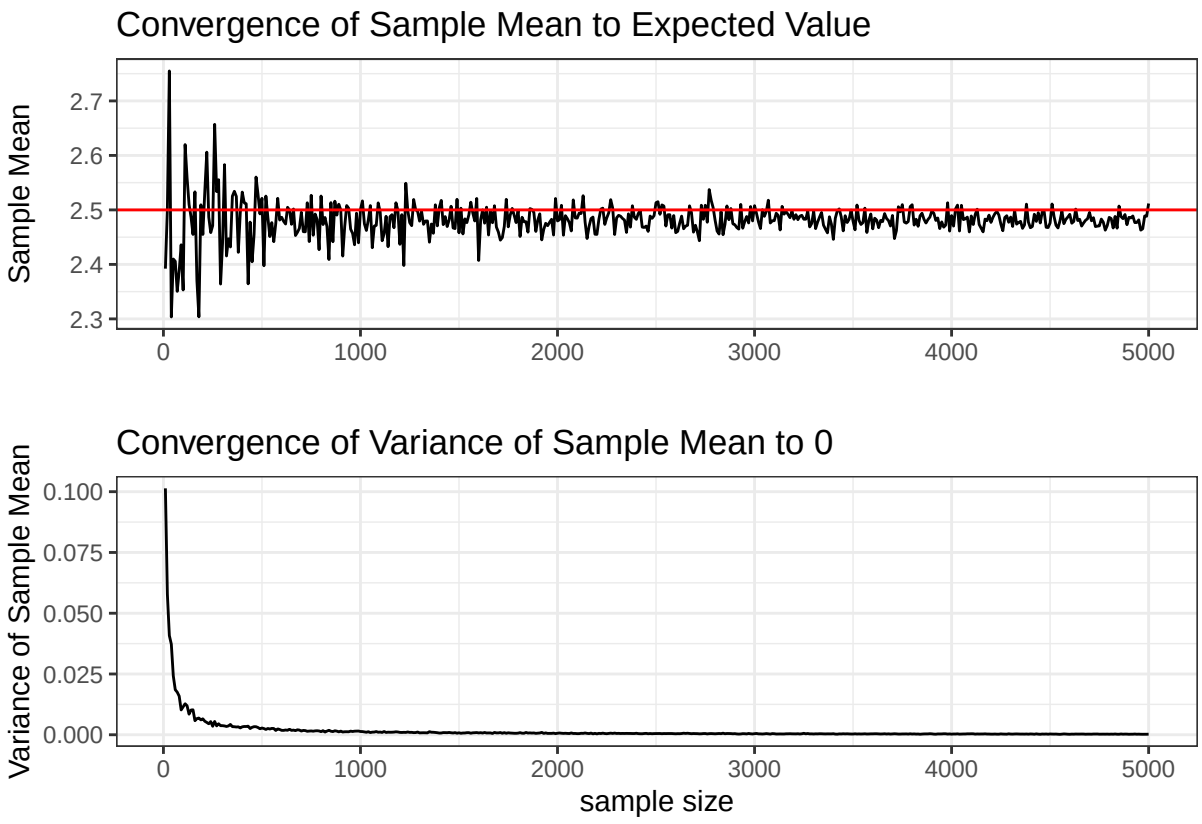
```
library(ggplot2)
library(dplyr)
library(tidyr)
library(patchwork)
library(gridExtra)
```

Graphical Simulation

We choose gamma distribution with $\alpha = 5$ and $\beta = 2$, with expected value $= \frac{\alpha}{\beta} = 2.5$.

```
# set seed for reproducibility
set.seed(123)
# a large population
population <- rgamma(n = 10000, shape = 5, rate = 2)
# define sample sizes
sample_sizes <- seq(10, 5000, by = 10)
# a fixed number of simulations for each sample size
m <- 100
# Initialize
variance_sample_mean <- numeric(length(sample_sizes))
sample_mean <- numeric(length(sample_sizes))
probs <- numeric(length(sample_sizes))
for (i in 1:length(sample_sizes)) {
  n <- sample_sizes[i]
  sample <- sample(population, size = n)
  sample_means <- numeric(m)
  sample_mean[i] <- mean(sample)
  # vairance of sample mean is based on m simulations of same sample size
  for (j in 1:m) {
    sample <- rgamma(n = n, shape = 5, rate = 2)
    sample_means[j] <- mean(sample)
  }
  probs[i] <- mean(abs(sample_means - 2.5) > 0.01)
  variance_sample_mean[i] <- var(sample_means)
```

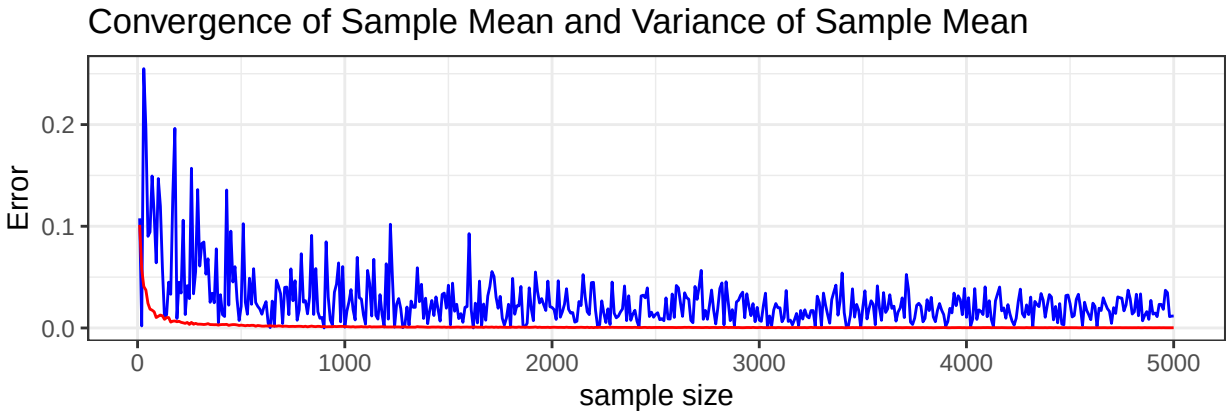
```r
}
# define expected value
expected_value <- 2.5
# data frame for plotting
df <- data.frame(
  sample_size = sample_sizes,
  sample_mean = sample_mean,
  variance_mean = variance_sample_mean,
  expected_value = expected_value,
  probs = probs
)
# plot sample mean
meanplot <- ggplot(df, aes(x = sample_size, y = sample_mean)) +
  geom_line() +
  geom_hline(yintercept = expected_value, col = "red") +
  labs(title = "Convergence of Sample Mean to Expected Value", x = "",
       y = "Sample Mean") + theme_bw()
# plot variance of sample mean
varplot <- ggplot(df, aes(x = sample_size, y = variance_mean)) +
  geom_line() +
  labs(title = "Convergence of Variance of Sample Mean to 0", x = "sample size",
       y = "Variance of Sample Mean") + theme_bw()
# display stacked plots
meanplot / varplot
```



Numerical Simulation

It is also possible to approach the problem numerically. We do this by calculating the error of the sample mean and its variance at some different levels of sample sizes, and observe their convergence.

```r
# calculate error
error_mean <- abs(sample_mean - expected_value)
error_var <- abs(variance_sample_mean)

# turn error to df
df <- df %>%
  mutate(
    error_mean = error_mean,
    error_var = error_var
  )

# get error_mean and error_var at 1%, 25%, 50%, 75%, 99% of observations
# for example, with first 5% of sample, the error_mean is ...
percentile_indices <- c(0.01, 0.25, 0.50, 0.75, 0.99) * nrow(df)

# Order the data frame
ordered_df <- df %>%
  arrange(sample_size)

# Function to calculate mean and variance up to a given index
calc_stats <- function(data, index) {
  subset_data <- data[index, ]
  return(data.frame(
    error_mean = subset_data$error_mean,
    error_var = subset_data$error_var,
    probs = subset_data$probs
  ))
}

# Apply the function for each percentile
qs <- lapply(percentile_indices, function(index) calc_stats(ordered_df, index))

# Combine the results into a data frame
qs <- do.call(rbind, qs)
rownames(qs) <- c("1%", "25%", "50%", "75%", "99%")
qs <- qs %>%
  mutate(
    error_mean = round(error_mean, 4),
    error_var = round(error_var, 6),
    probs = round(probs, 4)
  )

errorplot <- ggplot(df, aes(x = sample_size)) +
  geom_line(aes(y = error_mean), col = "blue") +
  geom_line(aes(y = error_var), col = "red") +
  labs(title = "Convergence of Sample Mean and Variance of Sample Mean",
       x = "sample size", y = "Error") + theme_bw()
table <- tableGrob(as.data.frame(qs))
grid.arrange(errorplot, table, ncol = 1)
```



	error_mean	error_var	probs
1%	0.0901	0.024222	0.93
25%	0.0234	0.000975	0.73
50%	0.0133	0.000494	0.72
75%	0.0096	0.000328	0.6
99%	0.0183	0.000258	0.56

Rate of Convergence

According to the numerical simulation, the sample mean converges to the expected value at a slower rate than its variance converges to 0.

This is not an isolated case. In general, the rate of convergence of the sample mean IS slower than the rate of convergence of its variance.

It is not mathematically challenging to prove this statement, which we will do in the following out of interest.

The sample mean is calculated through:

$$\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$$

According to central limit theorem, the sample mean is normally distributed with:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Thus, the variance is euqal to:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = O(\frac{1}{n}) \to 0$$

i.e., the variance of the sample mean converges to 0 at a rate of $O(\frac{1}{n})$.

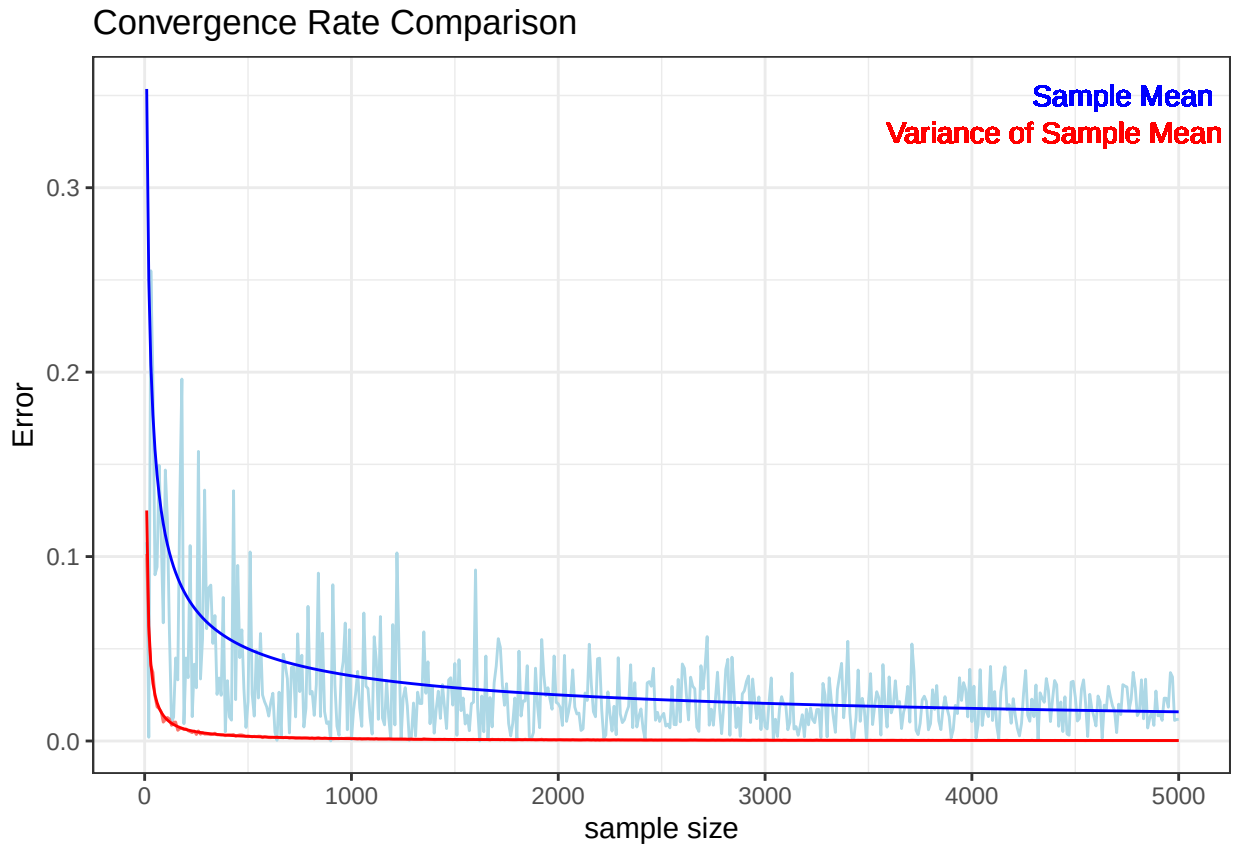
The distribution has the standard error (standard deviation from its expected value):

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

It is not hard to see that the standard error converges to 0 at a rate of $O(\frac{1}{\sqrt{n}})$. Thus, the sample mean converges to the expected value at a rate of $O(\frac{1}{\sqrt{n}})$.

Conclusion: With a rate of $O(\frac{1}{\sqrt{n}})$, the sample mean converges to the expected value at a slower rate than the sample variance converges to 0, which has a rate of $O(\frac{1}{n})$.

```
ggplot(df, aes(x = sample_size)) +
  geom_line(aes(y = error_mean), col = "lightblue") +
  geom_line(aes(y = error_var), col = "#F8766D") +
  labs(title = "Convergence Rate Comparison", x = "sample size",
       y = "Error") + theme_bw() +
  coord_cartesian(xlim = c(0, 5000)) +
  geom_line(aes(y = 1.118/sqrt(sample_size)), col = "blue") +
  geom_line(aes(y = 1.25/sample_size), col = "red") +
  geom_text(aes(x = 4400, y = 0.33, label = "Variance of Sample Mean"), col = "red") +
  geom_text(aes(x = 4730, y = 0.35, label = "Sample Mean"), col = "blue") +
  theme(legend.position = "none")
```



We see that $O(\frac{1}{\sqrt{n}})$ only roughly describes the rate of convergence of the sample mean, with considerable variance in data, this is due to the fact that the sample mean is asymptotically normally distributed, and we only estimated its deviation from the expected value.

Meanwhile, $O(\frac{1}{n})$ is almost a perfect approximation for the rate of convergence of the variance of sample mean.