Central Limit Theorem - An Illustration

Introduction

This document is a rudimentary illustration of the Central Limit Theorem. We start with an exponential distribution, and observe the distribution of sample means.

Theory

Let X follow an exponential distribution with rate λ, thus

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \frac{1}{\lambda^2}, \text{ and} \\
\sigma_X &= \sqrt{\mathrm{Var}(X)} = \frac{1}{\lambda}
\end{aligned}
$$

Let $\{X_i\}_{i=1}^n$ be a series of n i.i.d. variables s.t. $X_i \sim \exp(\lambda)$. We define the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. We can see that

$$
\begin{aligned}
\mathrm{E}(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{\lambda} \\
\mathrm{Var}(\bar{X}) &= \sum_{i=1}^{n}\frac{1}{n^2}\mathrm{Var}(X_i) = \frac{1}{n\lambda^2}, \text{ and} \\
\sigma_{\bar{X}} &= \sqrt{\mathrm{Var}(\bar{X})} = \frac{1}{\sqrt{n}\lambda}
\end{aligned}
$$

Thus, according to the Central Limit Theorem, when n is sufficiently large, $\bar{X}$ should converge towards a normal distribution with mean $\frac{1}{\lambda}$ and standard deviation $\frac{1}{\sqrt{n}\lambda}$.

Simulation

For the purpose of simulation, we set $\lambda = 0.20$, $n = 40$, we run the simulation $1,000$ times.

```r
lambda <- 0.2
n_sims <- 1000
n <- 40


theo_x <- seq(2.5, 7.5, length = 1000)

theo_df<- data.frame(
        x = theo_x,
        y = dnorm(theo_x, mean = 1/lambda, sd = 1/(lambda*sqrt(n)))
)

rm(theo_x)
```

```r
set.seed(100)
sim_data <- matrix(rexp(n_sims * n, rate = lambda), nrow = n_sims)
```

```r
lambda <- 0.2
n_sims <- 1000
n <- 40

sim_data <- matrix(rexp(n_sims * n, rate = lambda), nrow = n_sims)

means_df <- data.frame(
        x = apply(sim_data, 1, mean)
)

x_bar <- apply(sim_data, 1, mean)
x_bar_mean <- mean(x_bar)
x_bar_sd <- sd(x_bar)
```

In theory, $\bar{X}$ should have a mean of $\frac{1}{\lambda} = 5$, and standard deviation of $\frac{1}{\sqrt{n}\lambda} = 0.7905694$. We observe a sample mean of 5.0105574 and standard deviation of 0.8069805.

Plotting

Visually, the sample means follow a distribution which is approximately normal. There are more sophisticated ways to illustrate the convergence (e.g., Q-Q plots, comparing moments, etc.) but the following chart would suffice for our purposes.

```r
g <- ggplot(means_df, aes(x = x, y = ..density.., fill = "Simulated")) +
        geom_histogram(bins = 50) +
        geom_line(data = theo_df,
                  aes(y = y, color = "Theoretical")) +
        scale_color_manual(name = NULL,
                           values = c("Theoretical" = "blue")) +
        scale_fill_manual(name = NULL,
                          values = c("Simulated" = "grey")) +
        theme(legend.position = "top") +
        ggtitle(TeX("Simulated v. Theoretical Distribution of $\\bar{X}$"),
                subtitle = TeX("where \\bar{X}=\\frac{1}{40}\\sum_{i=1}^{40}X_i, and X_1, X_2, ..., X_{4
g
```

Simulated v. Theoretical Distribution of $\overline{X}$

where $\overline{X} = \frac{1}{40}\sum_{i=1}^{40} X_i$, and $X_1, X_2, \ldots, X_{40}$ are iid and follow an exponential distribution with $\lambda = 0.2$