Title: Statistical Inference Course Project: Part 1

Overview

This project investigates the exponential distribution in R and compare it with the Central Limit Theorem.

Simulations

As part of our exploration of the CLT we will simulate 1000 samples of size 40.

1. we simulate all 1000*40 draws from an exponential distribution, $Exp(\lambda = 0.2)$, and calculate the mean and standard deviation of the entire sample data.

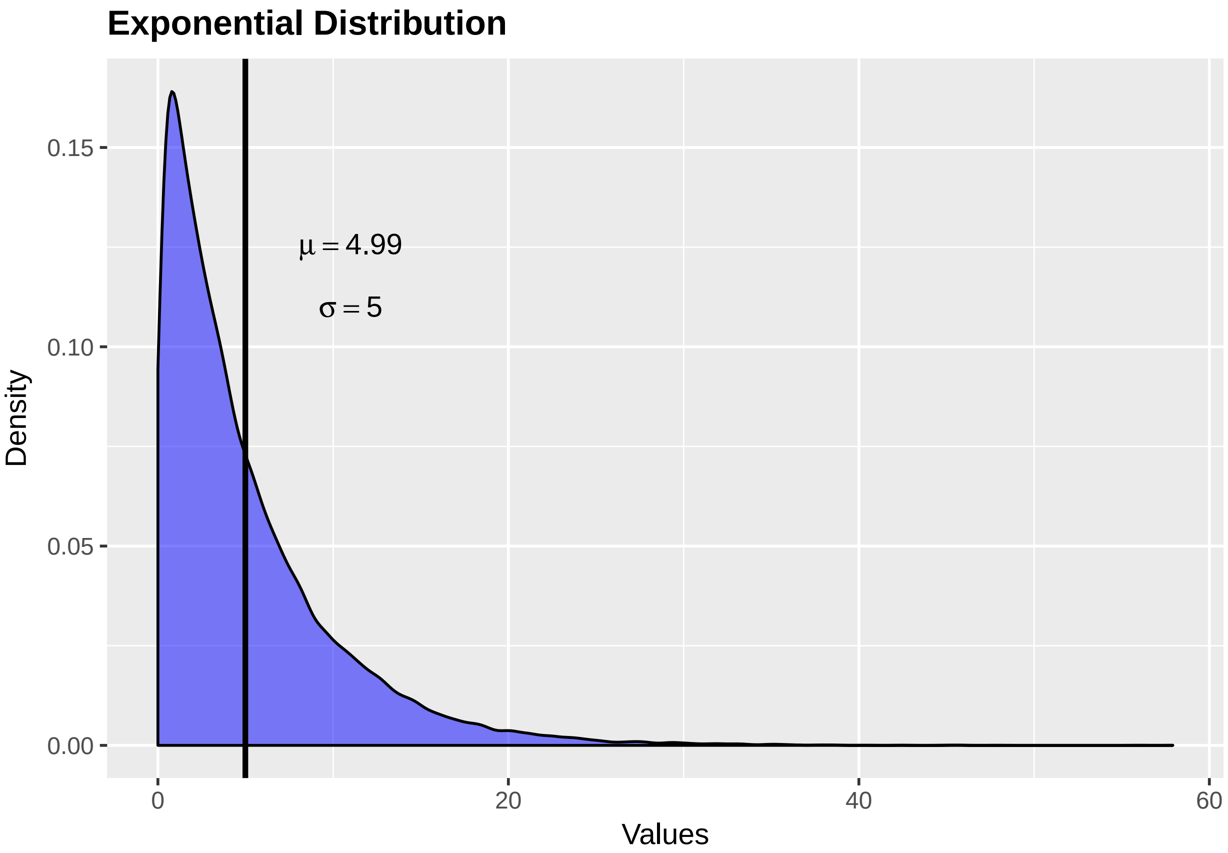
```
lambda <- 0.2
n <- 40
sims <- 1000

set.seed(1)
expData <- data.frame(rexp(n*sims, lambda))
colnames(expData) <- "v"
```

2. Summary of the simulated sample data, including mean, standard deviation and the histogram.

```
summary(expData$v)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00018  1.42869  3.44672  4.99003  6.91539 57.91790
```

```
exp_mean <- mean(expData$v)
exp_sd <- sd(expData$v)
#hist(expData$v)
exp_plot <- ggplot(expData, aes(x=v)) + geom_density(alpha=0.5, fill="blue") +
  ggtitle("Exponential Distribution") + xlab("Values") + ylab("Density") +
  geom_vline(aes(xintercept=exp_mean), size=1, color="black") + annotate("text", x=11, y=0.125, label=pa
  annotate("text", x=11, y=0.11, label=paste("sigma ==", round(exp_sd, 2)), parse=T) + theme(plot.title
exp_plot
```



3. Simulation of sample means from the above distribution we reshape the draws into 1000 samples of size 40. We then calculate the mean of each of the 1000 samples and the overall mean of the sample means.

```
set.seed(1)
samples <- matrix(expData$v, sims, n)
mns <- data.frame(apply(samples, 1, mean))
colnames(mns) <- "v"
mns_mean <- mean(mns$v)
mns_std <- sd(mns$v)
```

4. According to the CLT, the distribution of the sample means should follow a normal distribution with a mean equal to the population mean (μ) and a standard deviation equal to the population standard deviation divided by the square root of the sample size ($\frac{\sigma}{\sqrt{n}}$).

For theoretical mean and standard deviation of the exponential distribution , $Exp(\lambda)$, both the mean and standard deviation are equal to $\frac{1}{\lambda}$.

Using these properties we calculate the mean and standard deviation(referred to as the standard error) of the theoretical CLT normal distribution.

```
clt_mu <- 1.0/lambda
clt_sd <- 1.0/lambda
clt_se <- clt_sd/sqrt(n)
```

5. With these paramaters of the normal distribution calculated, we can careate a table show the difference between theoretical data and the sampled data, as below:

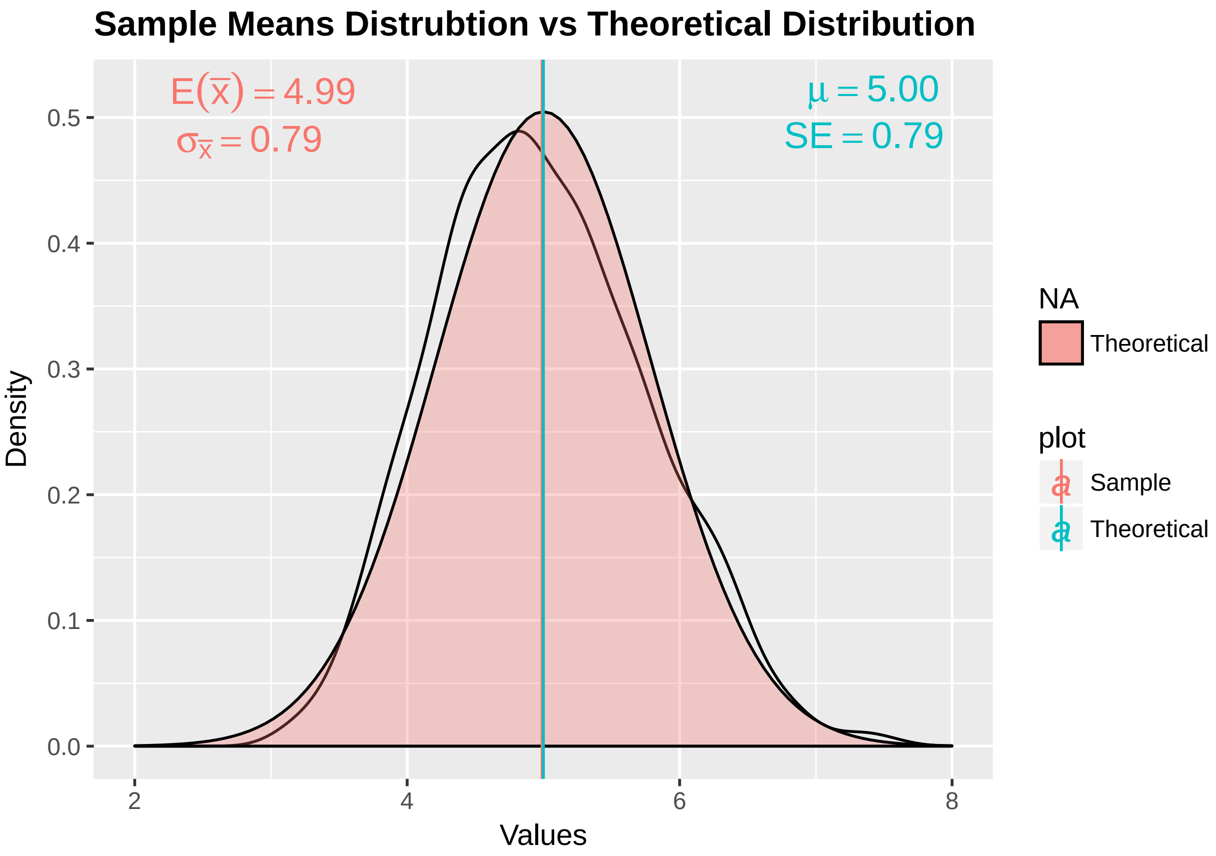
```
sampled_distr <- data.frame( mean = c(round(mns_mean,2), clt_mu),
  std = c(round(mns_std,2), clt_se),
  meanlab=c(deparse(formatC(round(mns_mean,2),digits=2,format="f")),deparse(formatC(clt_mu, digits=2, f
  plot=c("Sample", "Theoretical"), hjust=c(2, -2), label=c("E(bar(x)) ==", "mu =="),
  sd=c(deparse(formatC(round(mns_std,2),digits=2,format="f")), deparse(formatC(round(clt_se,2), digits=
  sdlab=c("sigma[bar(x)] ==", "SE =="))

sampled_distr[, c("mean", "std")]
```

```
##   mean       std
## 1 4.99 0.7900000
## 2 5.00 0.7905694
```

6. We can see that the sampled mean and standard deviation are very close to the theoretical value, and we can use the follow plot to confirm CLT normal distribution.

```
clt_plot <- ggplot(mns, aes(x=v)) +
    geom_density(alpha=0.5) +
    stat_function(fun=dnorm, color="black", geom="ribbon", mapping=aes(ymin=0, ymax=..y.., fill="Theore
    geom_vline(data=sampled_distr, aes(xintercept=mean, color=plot)) +
    geom_text(data=sampled_distr, aes(mean, .52, label=paste(label, meanlab), color=plot, hjust=hjust),
   geom_text(data=sampled_distr, aes(mean, .52, label=paste(sdlab, sd), color=plot, hjust=hjust+0.5, vj
   scale_x_continuous(limits=c(2,8)) +
   ggtitle("Sample Means Distrubtion vs Theoretical Distribution") +
   theme(plot.title = element_text(face="bold")) +
   xlab("Values") +
   ylab("Density")

print(clt_plot)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```



Result

The table and the plot above illustrates the accuracy of the CLT with regard to the mean and standard deviation of the distribution of sample means.

The sample means' distritubion is very close to a normal distribuiton, with normal mean close to underlying original distribution's mean, its sample standard deviation is approximately equal to the standard deviation of the underlying exponential distribution divided by the square root of the sample size.