Statistical Inference Course Project

Maury Miller

Introduction

This project is a demonstration of the **Central Limit Theorem** (CLT) with an exponential distribution. The CLT states that the distribution of averages of iid variables (properly normalized) has a normal distribution.

Experiment

The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also also 1/lambda. Set lambda = 0.2 for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

The experiment will try to answer the following questions.

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.
4. Evaluate the coverage of the confidence interval for 1/lambda: $\bar{X} \pm 1.96S$

Calculations for Distribution of Averages

```
library(ggplot2)

# Generate random data
set.seed(123)
lambda <- .2
nosim <- 1000
n <- 40
dat <- data.frame(x = c(replicate(nosim, mean(rexp(n,lambda)))),
    size = factor(rep(c(n), rep(nosim, 1))))
```

The sample mean, sample variance, and confidence intervals.

```
mean(dat$x)
```

```
## [1] 5.012
```

```
(1/(lambda * lambda))/n
```

```
## [1] 0.625
```

```
intervals <- mean(dat$x) + c(-1, 1) * qnorm(.975) * 5 / sqrt(n)
intervals
```

```
## [1] 3.462 6.561
```

The plot shows the distribution, a normal distribution curve, the population mean in green and the sample mean in blue.

```
# Plot distribution of averages
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(binwidth=.2, colour = "black", aes(y = ..dens
g <- g + geom_vline(aes(xintercept=mean(x)), color="blue", linetype="longdash", size=.75) # sample mean
g <- g + geom_vline(xintercept = 1/lambda, size = .75, linetype="longdash", colour="green")
g <- g  + stat_function(fun=dnorm, args=list(mean=5, sd=sqrt(5)/sqrt(n)), size=1)

g
```

Coverage

The plot displays the coverage for a set of lambda values centered around the initial lambda value of .2

```
lambdavalues <- seq(0.06, 0.4, by = 0.02)
nosim <- 1000
coverage <- sapply(lambdavalues, function(lambda) {
lhats <- replicate(nosim, mean(rexp(n,lambda)))
ll <- lhats - qnorm(0.975) * (1/lambda)/sqrt(n)
ul <- lhats + qnorm(0.975) * (1/lambda)/sqrt(n)
mean(ll < 1/lambda & ul > 1/lambda)
})

ggplot(data.frame(lambdavalues, coverage), aes(x = lambdavalues, y = coverage)) + geom_line(size = 2) +
```

Conclusion

The experiment showed the sample averages of an exponential distribution are approximately normal. And the averages are centered around the population mean with a sample deviation of

$$\frac{1/\lambda}{\sqrt{n}}$$

As I increased n, the density became narrower and narrower. This was expected. The coverage was very close to the 95th percentile for all the lambda values that I tested.