Statistical Inference Course Project, Part I

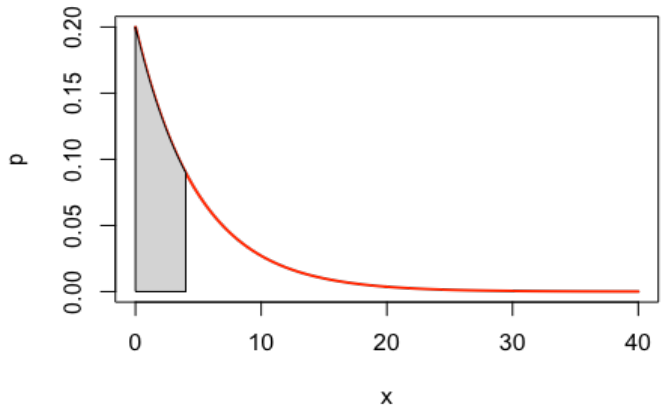
The Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}$$

The exponential distribution has its name because its probability distribution function has the shape of the exponential function. It crosses the Y-axis at some positive value called λ and slopes downwards to the right, decreasing towards zero as the values of random variable X increase, never reaching zero. The amount of slope in the curve is determined by the value of λ, also called *the rate*.

Definition and graph adapted from *http://www.ynegve.info/Post/169/a-tutorial-on-probability-and-exponential-distribution*

```
x = seq(0, 40, length = 200)
y = dexp(x, rate = 0.2)
plot(x, y, type="l", lwd = 2, col = "red", ylab = "p")
x = seq(0, 4, length = 200)
y = dexp(x, rate = 0.2)
polygon(c(0, x, 4), c(0, y, 0), col = "lightgray")
```



The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean or expected value of the exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also also $\frac{1}{\lambda}$.

Theoretical Center

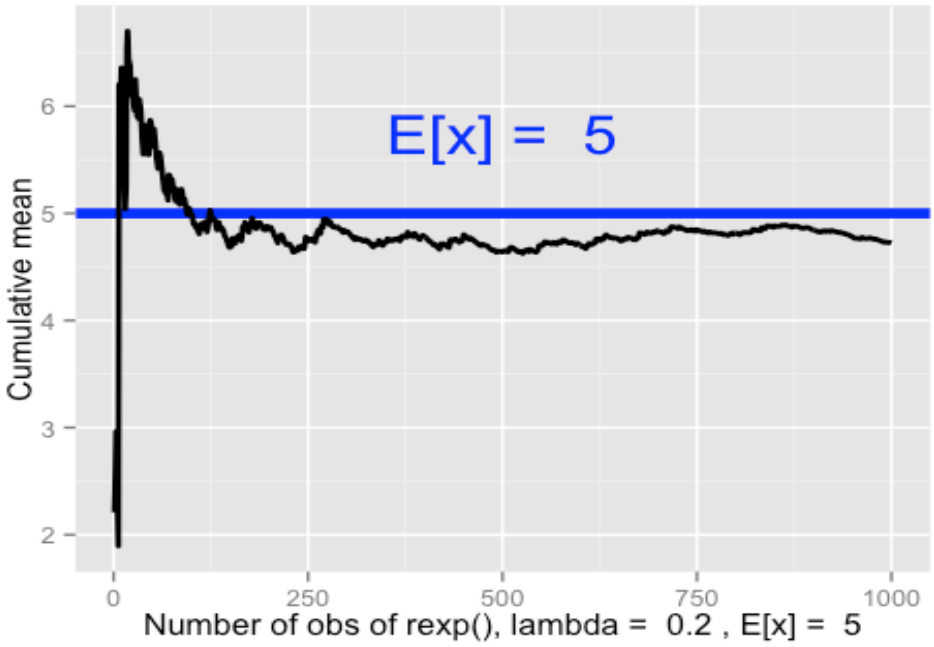
First, given values for the rate λ, we show where the distribution is centered at and compare it to the theoretical center of the distribution. In this case, where $\lambda = 0.2$, the central tendency or mean is expected to be $E[x] = \frac{1}{\lambda} = \frac{1}{0.2} = 5$.

```
lambda <- 0.2       # rate of descent
mu <- 1/lambda      # expected value
sigma <- 1/lambda   # standard deviation
```

Here, we illustrate through simulation that `rexp(n, lambda)` follows the law of large numbers, with cumulative means approaching $\frac{1}{\lambda}$ as the number of observations $n \to \infty$.

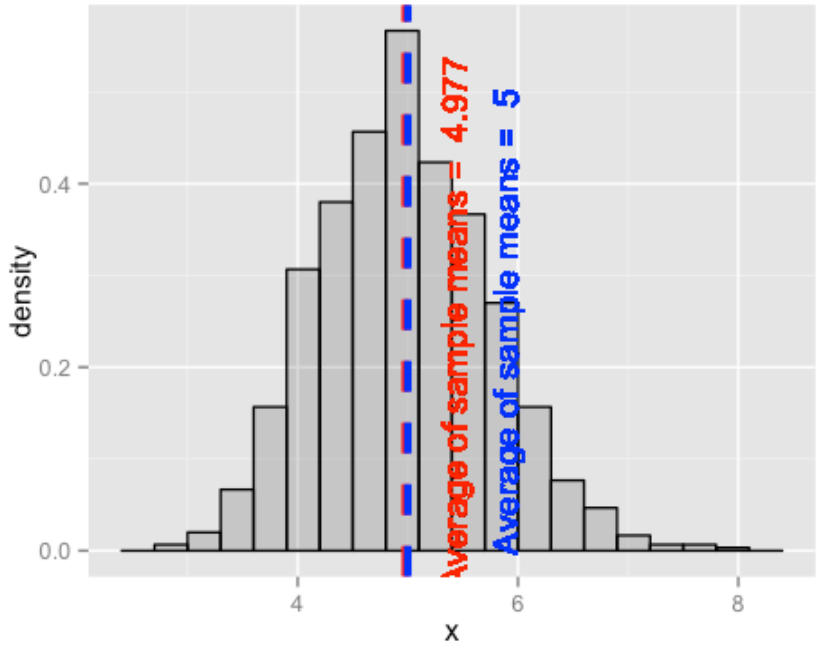
```
set.seed(12345)
nosim <- 1000

means <- cumsum(rexp(nosim, lambda)) / (1  : nosim)
g <- ggplot(data.frame(x = 1 : nosim, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = mu, size = 2, color = "blue") +
geom_line(size = 1)
g <- g + annotate("text", label = paste("E[x] = ", mu), x = 500, y =
1.15*mu, size = 8, colour = "blue")
g + labs(x = paste("Number of obs of rexp(), lambda = ", lambda, ",
E[x] = ", round(mu, 3)), y = "Cumulative mean")
```



Now, we look specifically at the central tendency of the sample means from 1,000 trials of 40 observations each.

```
n <- 40
x <- apply(matrix(rexp(n*nosim, lambda), nosim, n), 1, mean)
g <- ggplot(data.frame(x = x), aes(x = x)) + geom_histogram(alpha =
.20, binwidth=.3, colour = "black", aes(y = ..density..))
g + geom_vline(aes(xintercept = mean(x)), color = "red", linetype =
"dashed", size = 1.5) +
geom_text(aes(x = mean(x), label=paste("\n\nAverage of sample means =
", round(mean(x), 3)), y=0.25),
color="red", angle=90, text=element_text(size=6)) +
geom_vline(aes(xintercept = mu), color = "blue", linetype = "dashed",
size = 1.5) +
geom_text(aes(x = mu, label=paste("\n\n\n\nAverage of sample means = ",
round(mu, 3)), y=0.25),
color="blue", angle=90, text=element_text(size=6))
```



Variability

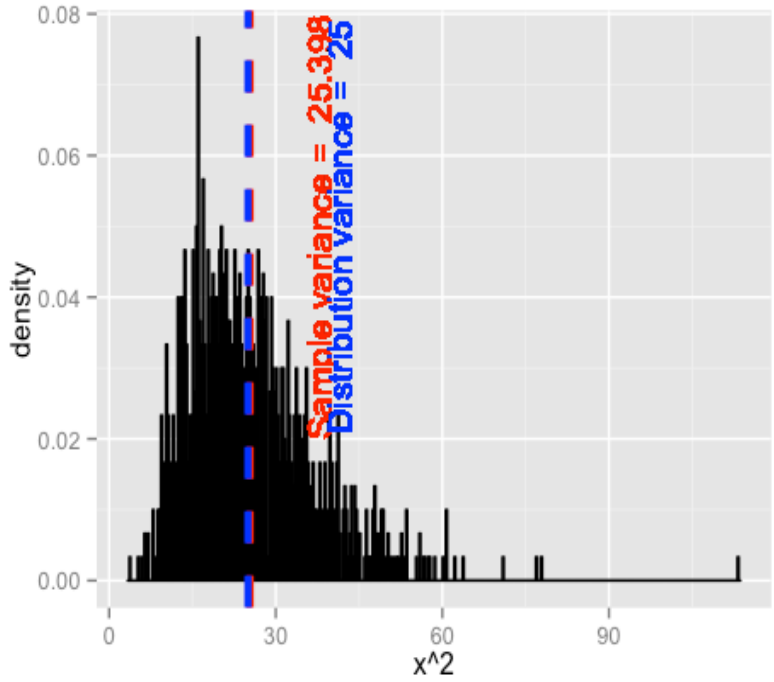
Now, we show how variable the distribution is and we compare it to the theoretical variance of the distribution. We know that the population variance for the exponential function is given by $(\frac{1}{\lambda})^2$. That gives us:

$$\sigma^2 = (\frac{1}{\lambda})^2 = (\frac{1}{0.2})^2 = 25$$

```
n <- 40
x <- apply(matrix(rexp(n*nosim, lambda), nosim, n), 1, sd)
g <- ggplot(data.frame(x = x), aes(x = x^2)) + geom_histogram(alpha =
0.20, binwidth = 0.3, colour = "black", aes(y = ..density..))
g + geom_vline(aes(xintercept = mean(x^2)), color = "red", linetype =
```

```
"dashed", size = 1.5) +
geom_text(aes(x = mean(x^2), label=paste("\n\n\nSample variance = ",
round(mean(x^2), 3)), y = 0.05),
color="red", angle=90, text=element_text(size=6)) +
geom_vline(aes(xintercept = sigma^2), color = "blue", linetype =
"dashed", size = 1.5) +
geom_text(aes(x = sigma^2, label=paste("\n\n\n\nDistribution variance =
", round(sigma^2, 3)), y = 0.05),
color="blue", angle=90, text=element_text(size=6))

## Warning: position_stack requires constant width: output may be
incorrect
```

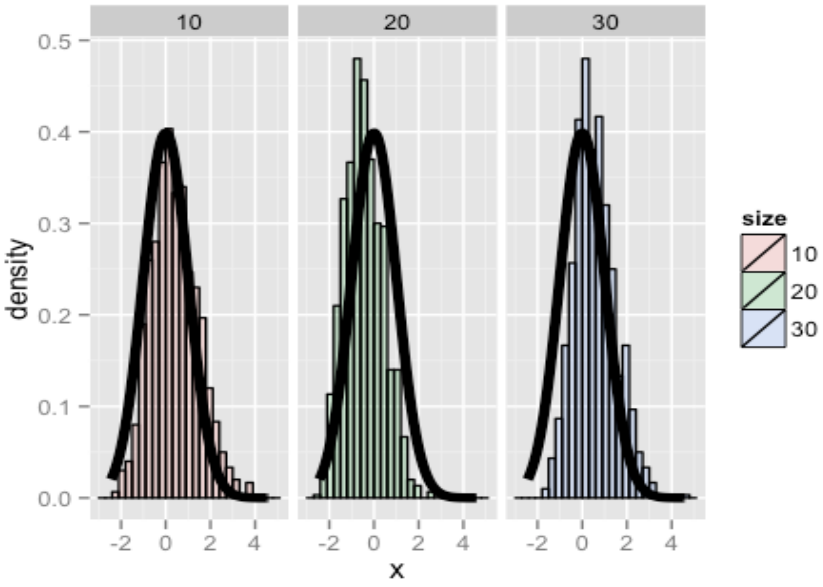


Normality

Through an experiment, we illustrate the approximate normality of the distribution by simulating of multiples of **1000 trials** of sample means of `rexp(n, lambda)`, where $\lambda = 0.2$, where each trial contains **40 observations** each.

```
n <- 40
cfunc <- function(x, n) sqrt(n) * (mean(x) - mu) / sigma
dat <- data.frame(
  x = c(apply(matrix(sample(rexp(n, lambda), nosim * 10, replace =
TRUE),
                     nosim), 1, cfunc, 10),
        apply(matrix(sample(rexp(n, lambda), nosim * 20, replace =
TRUE),
                     nosim), 1, cfunc, 20),
```

```
        apply(matrix(sample(rexp(n, lambda), nosim * 30, replace =
TRUE),
                     nosim), 1, cfunc, 30)
        ),
  size = factor(rep(c(10, 20, 30), rep(nosim, 3))))
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(alpha = .20,
binwidth=.3, colour = "black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g + facet_grid(. ~ size)
```



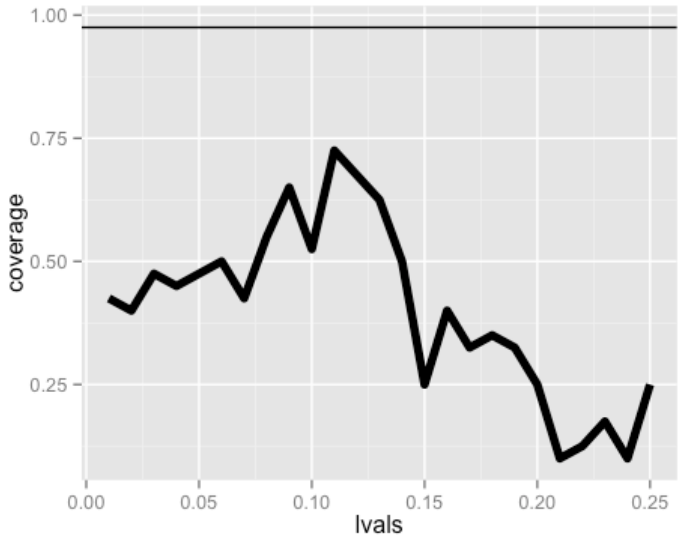
This experiment demonstrates thta the distribution is approximately normal, as expected under the C.L.T.

Coverage

Finally, we valuate the coverage of the confidence interval for $\frac{1}{\lambda}$: $\overline{X} \pm 1.96 \frac{s}{\sqrt{n}}$. We know that a standard deviation of 1.96 corresponds to the 97.5 percentile. We look at variations in $\lambda = 0.01, 0.02, \ldots, 0.25$.

```
n <- 40
lvals <- seq(0.01, 0.25, by = 0.01)
coverage <- sapply(lvals, function(l){
  phats <- rexp(n, l) / n
  ll <- phats - qexp(0.975) * sd(phats) / sqrt(n)
  ul <- phats + qexp(0.975) * sd(phats) / sqrt(n)
  mean(ll < l & ul > l)
})

ggplot(data.frame(lvals, coverage), aes(x = lvals, y = coverage)) +
geom_line(size = 2) + geom_hline(yintercept = 0.975)
```



The data suggests that in general we are seeing very poor coverage for the confidence interval at 97.5th percentile ($\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$).