# Statistical Inference- Course Project

## The exponential distribution

(Note: we need load the `dplyr` package for data manipulation and the `matrixStats` package for the `rowSds` function.)

## Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s.

We set the seed, the rate parameter `lambda`, the number of observations `nobs`, and the number of simulations `sims`.

```
set.seed(12345); lambda = 0.2; nobs = 40; sims = 1000
```

Now we take 40000 random draws from the exponential distribution. We will organize this into a matrix with 1000 rows and 40 columns and view this as simulating 1000 repetitions of a draw of 40 random exponentials with rate $\lambda = 0.2$.

```
exp <- matrix(rexp(sims*nobs, lambda), nrow = sims, ncol = nobs)
```

Now we take the mean of each row. (We'll need standard deviations and confidence interval endpoints later, so we might as well compute them now. Partway through, we convert `exp` to a data frame to be able to use the handy `mutate` function.)

```
exp_mean <- rowMeans(exp)
exp_sd <- rowSds(exp)
exp <- cbind(exp, exp_mean = exp_mean, exp_sd = exp_sd)
exp <- data.frame(exp)
exp <- exp %>%
    mutate(CI_left = exp_mean - 1.96 * exp_sd / sqrt(nobs),
           CI_right = exp_mean + 1.96 * exp_sd / sqrt(nobs))
```

### 1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

The mean of our sample means is 4.972 versus the theoretical mean, which is

$$\frac{1}{\lambda} = \frac{1}{0.2} = 5.$$

### 2. Show how variable it is and compare it to the theoretical variance of the distribution.

The standard deviation of our sample means is 0.7847 versus the theoretical standard deviation, which is

$$\frac{\sigma}{\sqrt{n}} = \frac{1/\lambda}{\sqrt{40}} = 0.7906.$$

Equivalently, the variance of our sample means is 0.6158 versus the theoretical variance, which is

$$\frac{\sigma^2}{n} = \frac{1/\lambda^2}{40} = 0.625.$$

**3. Show that the distribution is approximately normal.**

Here is a histogram of the sample means with the normal distribution $N\left(1/\lambda, \frac{1/\lambda}{\sqrt{40}}\right)$ superimposed.

```r
h <- hist(exp$exp_mean,
          main = "Histogram of sample means",
          xlab = "Mean")
xfit<-seq(min(exp$exp_mean),max(exp$exp_mean),length=100)
yfit<-dnorm(xfit,mean = 1/lambda, sd = (1/lambda)/sqrt(nobs))
yfit <- yfit*diff(h$mids[1:2])*length(exp$exp_mean)
lines(xfit, yfit, col="blue", lwd=2)
```



**Histogram of sample means**
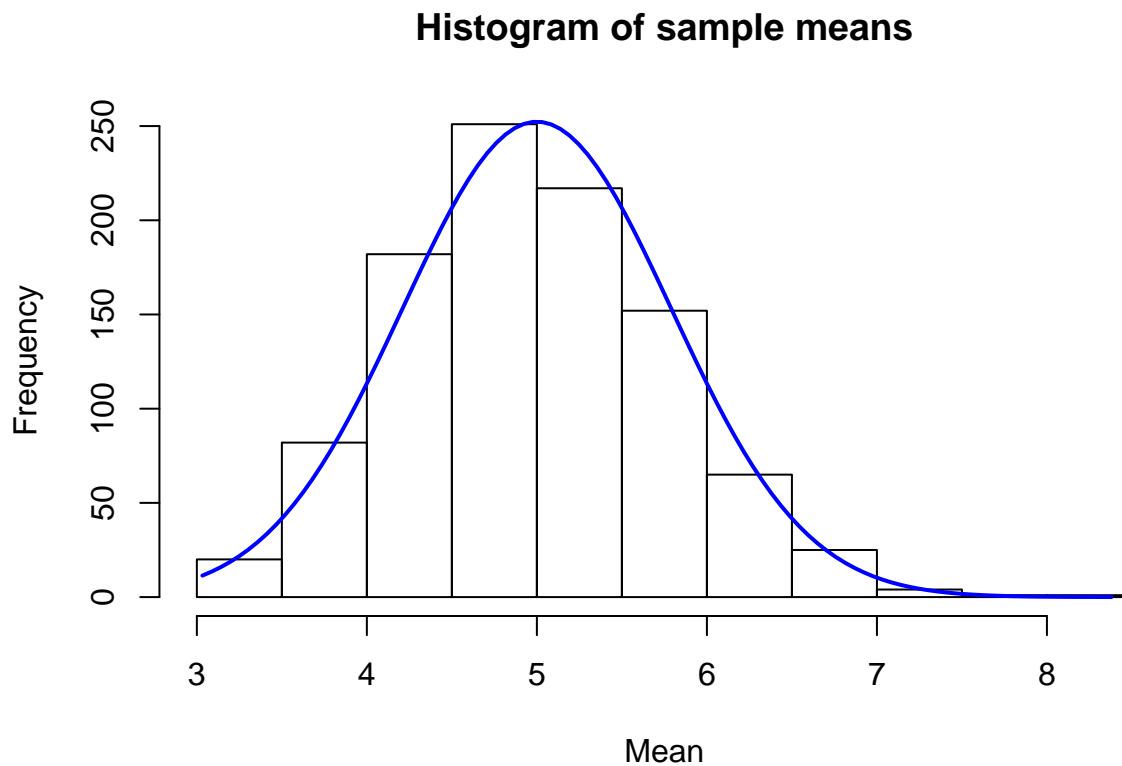
**4. Evaluate the coverage of the confidence interval for $1/\lambda$:**

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$$

The following code calculates the percentage of confidence intervals that capture the true (theoretical) mean.

```r
exp <- exp %>%
    mutate(CI_good = (CI_left < 1/lambda) & (1/lambda < CI_right))
percent_good <- mean(exp$CI_good)
```

So 92% of the simulated intervals capture the true mean $\frac{1}{\lambda} = 5$.