ST501: Group K - R Project

Yun Ji, Sarah McLaughlin, Kevin Carmody

12/1/2019

Part I: Visualizing Convergence in Probability

1.1) Show that the minimum order statistic converges in probability to 0.

Let $Y_i \sim exp(1)$ for $i = 1, 2, \ldots, n$, since $\lambda = 1$, the PDFs are defined as:

$$f_{Y_i}(y) = f_Y(y) = e^{-y} \text{ for } 0 < y < \infty.$$

The CDF would be defined as:

$$F_Y(y) = \begin{cases} 1 - e^{-y} & 0 < y < \infty \\ 0 & \text{o.w.} \end{cases}$$

With the minimum order statistic:

$$F_{Y_{(1)}}(y) = \begin{cases} 1 - (1 - F_Y(y))^n = 1 - e^{-ny} & 0 < y < \infty \\ 0 & \text{o.w.} \end{cases}$$

The support of $Y_{(1)}$ is the same as $Y: 0 < y < \infty$, thus $|Y_{(1)} - 0| = Y_{(1)} - 0$, and $P(|Y_{(1)} - 0| < \epsilon) = P(Y_{(1)} - 0 < \epsilon) = P(Y_{(1)} < \epsilon) = F_{Y_{(1)}}(\epsilon)$.

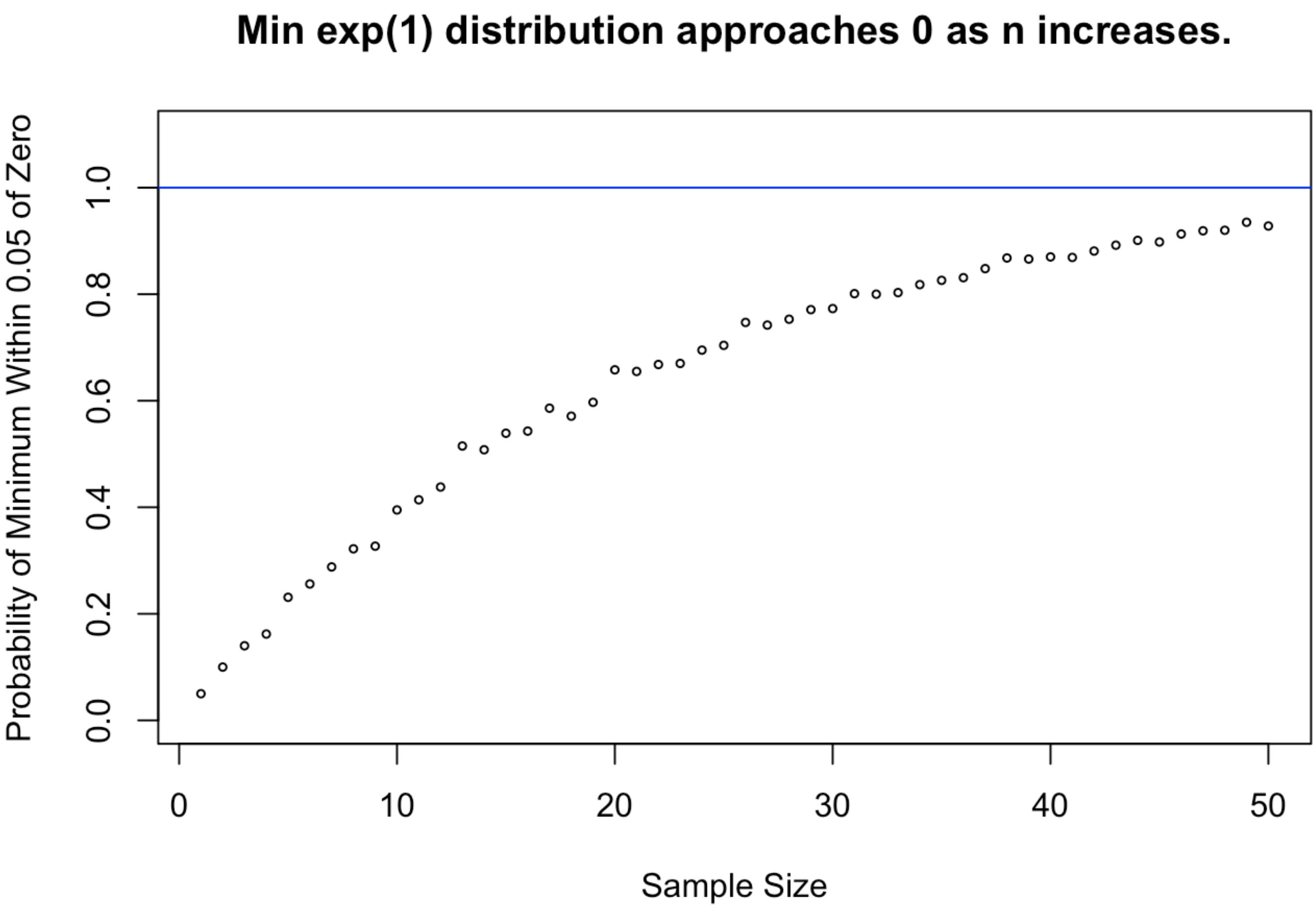
Since $\epsilon > 0$ by definition, then $F_{Y_{(1)}}(\epsilon) = 1 - e^{-n\epsilon}$ for all values of ϵ. The $\lim_{n\to\infty} F_{Y_{(1)}}(\epsilon) = \lim_{n\to\infty}(1 - e^{-n\epsilon}) = 1$. Therefore, $\lim_{n\to\infty} P(|Y_{(1)} - 0| < \epsilon) = 1$ and $Y_{(1)}$ converges in probability to 0.

1.2-5) Visualize and approximate the probability statement above.

```
set.seed(10)

#Part 1: Convergence in Probability
#Initiate set of sample sizes (n) and number of data sets to generate for each sample
size (N)
n = c(1:50)
N = 1000

#Initiate data structures for empirical data, minimums, and convergence probabilities
p1data = list()
p1min = list()
p1convprob = list()

for (i in 1:length(n)) {
  p1data[[i]] = matrix(0, nrow = N, ncol = n[i])
  p1min[[i]] = matrix(0, nrow = N, ncol = 1)
  p1convprob[[i]] = 0
}

#Generate values of the exp(1) distribution and find the minimum for each sample
for (i in 1:length(n)) {
  for (j in 1:N) {
    p1data[[i]][j,] = rexp(n = n[i], rate = 1)
    p1min[[i]][j,] = min(p1data[[i]][j,])
  }
}

#Set epsilon as 0.05 and find the proportion of abs(minimum minus 0) within this epsi
lon for sample sizes 1 thru 50
eps = 0.05
for (i in 1:length(n)) {
  p1convprob[[i]] = sum(abs(p1min[[i]]-0) < eps)/N
}
```

1.5) Create a plot with the sample size on the x-axis and the probability of interest on the y-axis.

```
#Plot sample size vs. probability computed above
plot(x = 1:length(n), y = p1convprob, main = "Min exp(1) distribution approaches 0 as
n increases.",
     xlab = "Sample Size", ylab = "Probability of Minimum Within 0.05 of Zero", ylim
= c(0.0, 1.1), cex = 0.5)
abline(h=1, col = "blue")
```



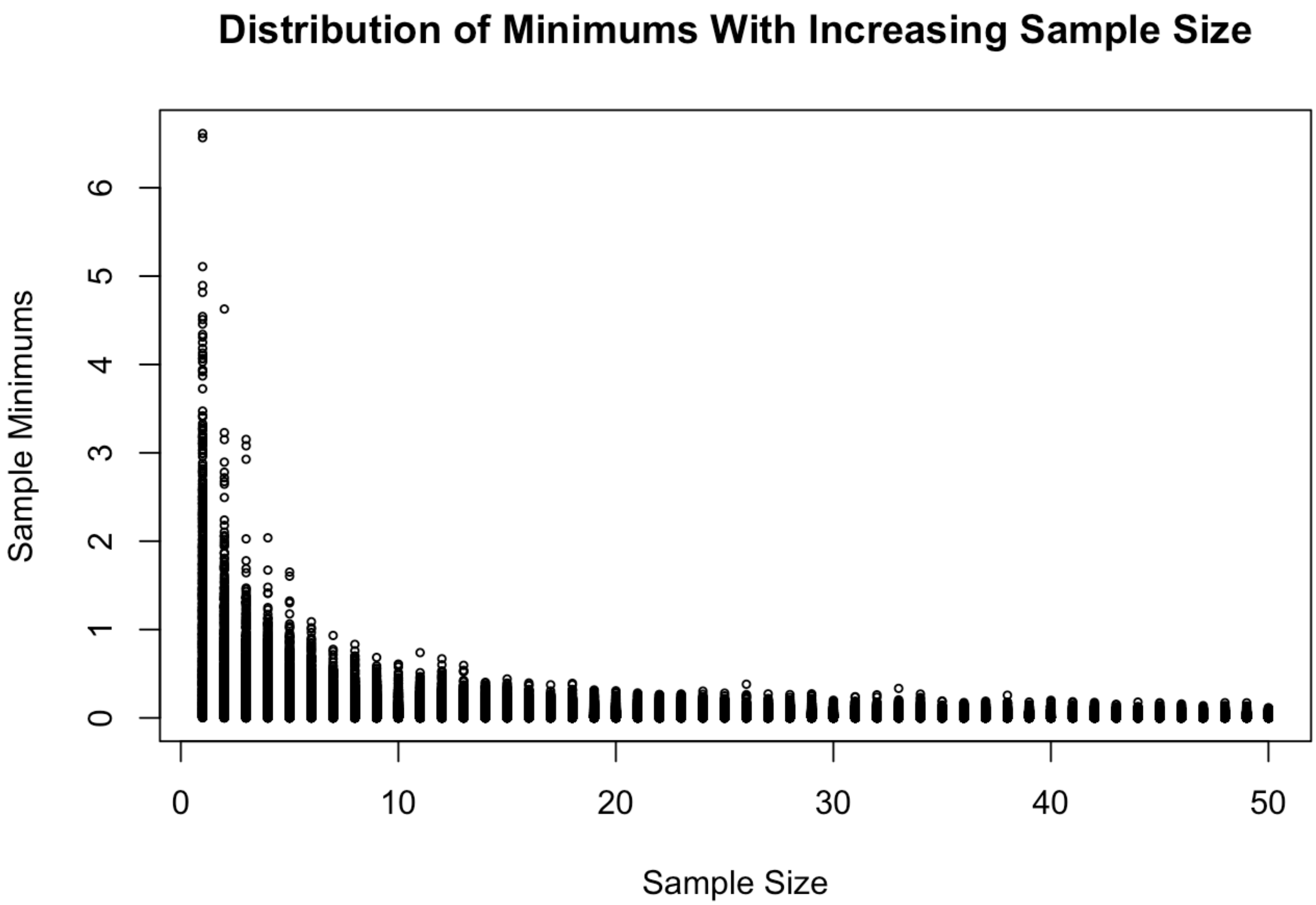
1.6) Explain how the plot above can help someone understand convergence in probability to a constant.

The convergence in probability definition states that a sequence of random variables (in this case, the minimums from random samples of the exp(1) distribution) converges to some random variable as the sample size grows. In this simulation, we show that as the sample size of an exp(1) distribution grows, the minimum converges to 0. As the sample size increases, the distribution of the minimum of the sample becomes more and more concentrated about 0. The graph created displays how the probability that the minimum of the sample is “close” (within the epsilon value of 0.05) to zero converges to 1.

```
(This is the old answer. I'll remove after any edits to Draft #2) Per the plot, we se
e that as the sample size increases, the probability of being within the 0.05 boundar
y increases. If the sample size were to continue to increase, we'd get closer and clo
ser values to 1. Thus, it illustrates that as our sample size increases, so does the
probability and if we were to increase to infinity the probability would converge to
1.
```

1.7) Create a plot with the sample size on the x-axis and the values of the minimum on the y-axis.

```
#Plot distribution of minimums for each sample size
plotdata = matrix(0, nrow = 0, ncol = 2)
for (x in 1:length(n)) {
  plotdata = rbind(plotdata, cbind(rep(x,N), p1min[[x]]))
}

plot(plotdata, main = "Distribution of Minimums With Increasing Sample Size",
     xlab = "Sample Size", ylab = "Sample Minimums", cex = 0.5)
```



1.8) Explain how the plot above can help someone understand convergence in probability to a constant.

In the graph, as n increases, the distribution of the sample minimum becomes more concentrated towards 0. As n approaches infinity, the minimum of each random sample will converge to 0.

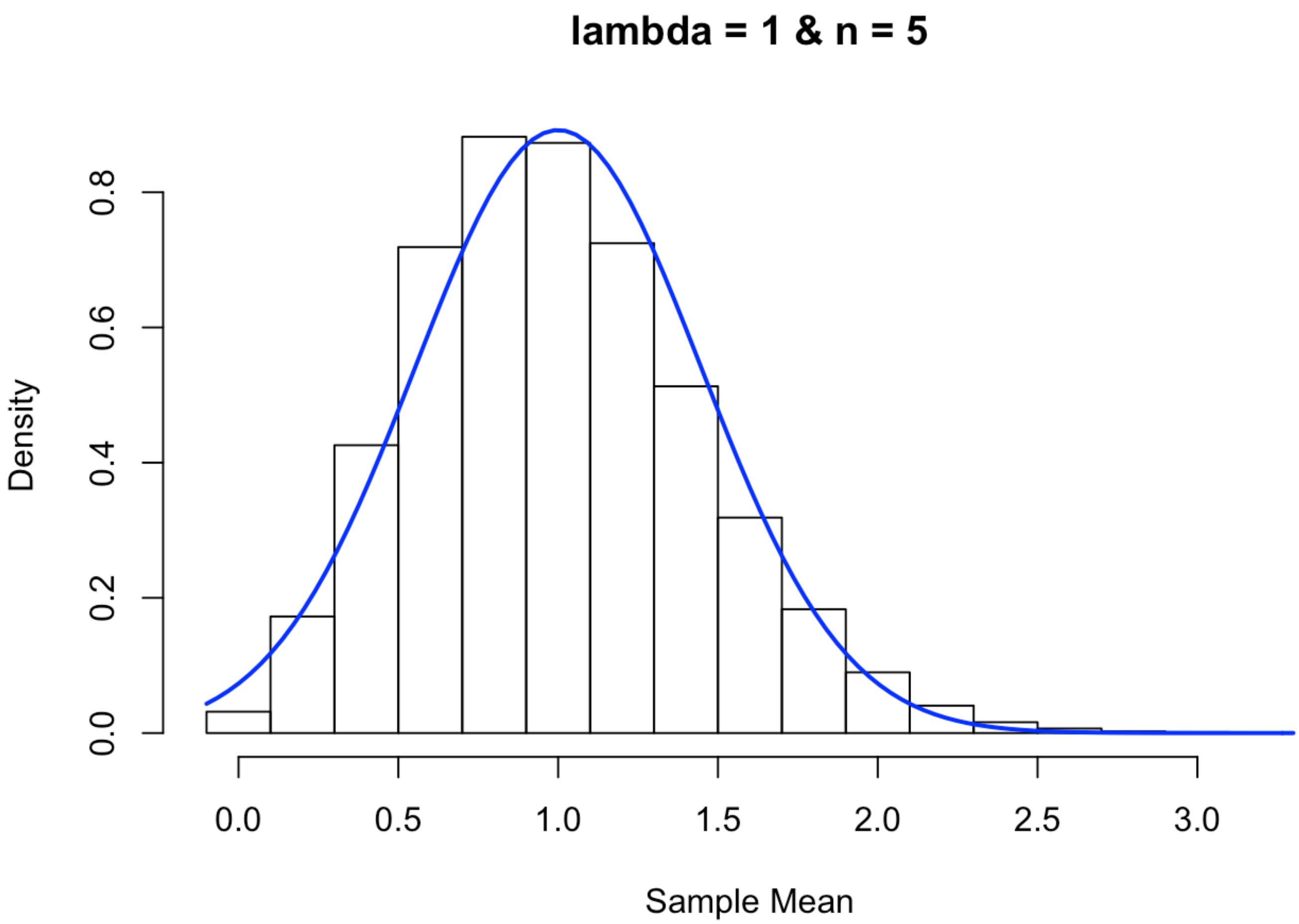
```
(This is the old answer. I'll remove after any edits to Draft #2) Per the plot, we se
e that as the sample size increases, the our minimums become less and less dispersed.
As we continue increasing the sample size the upper bound continues to decrease, ulti
mately converging to 0, should n increase to infinity.
```

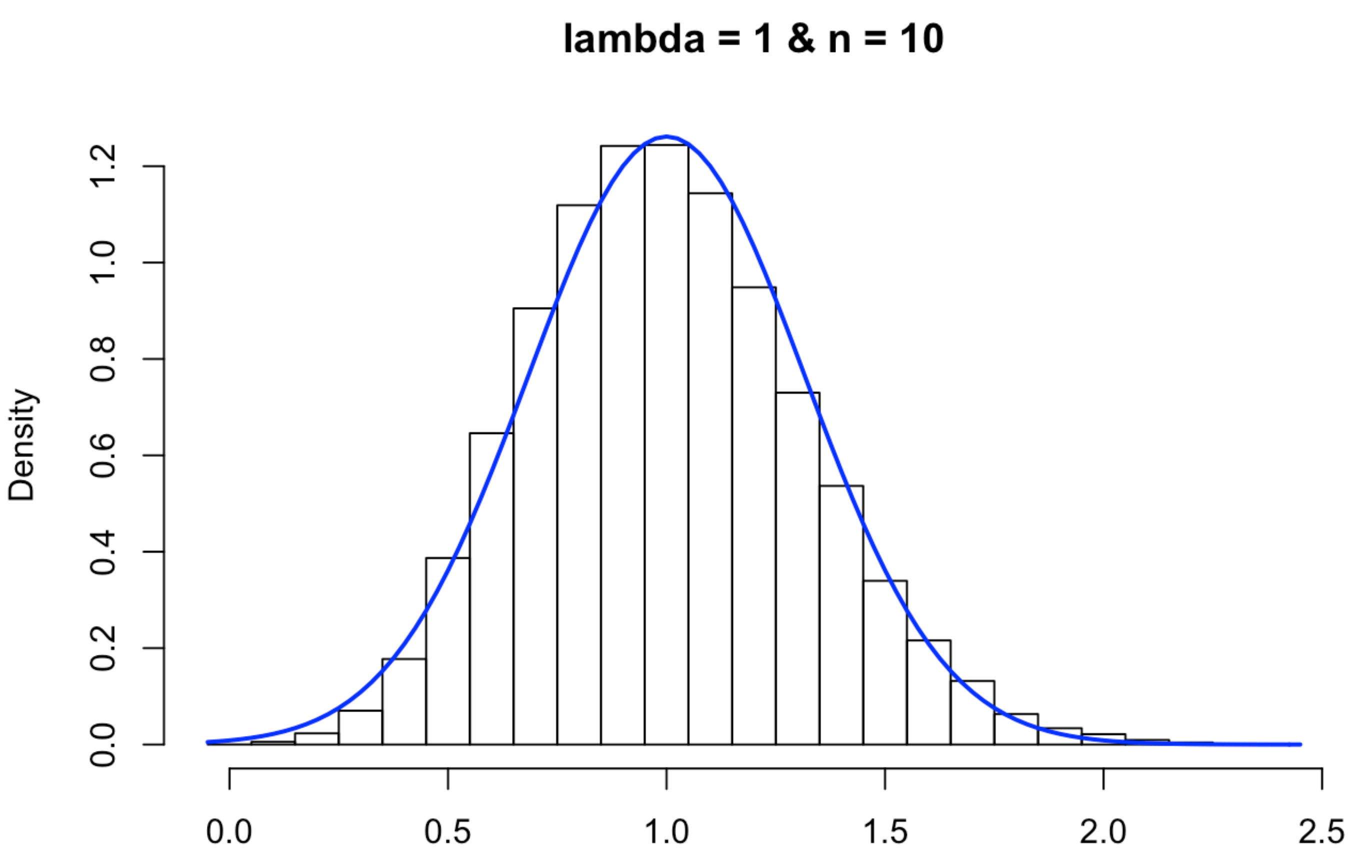
Part II: Visualizing Convergence in Distribution

2.1-3) Visualizing the CLT.

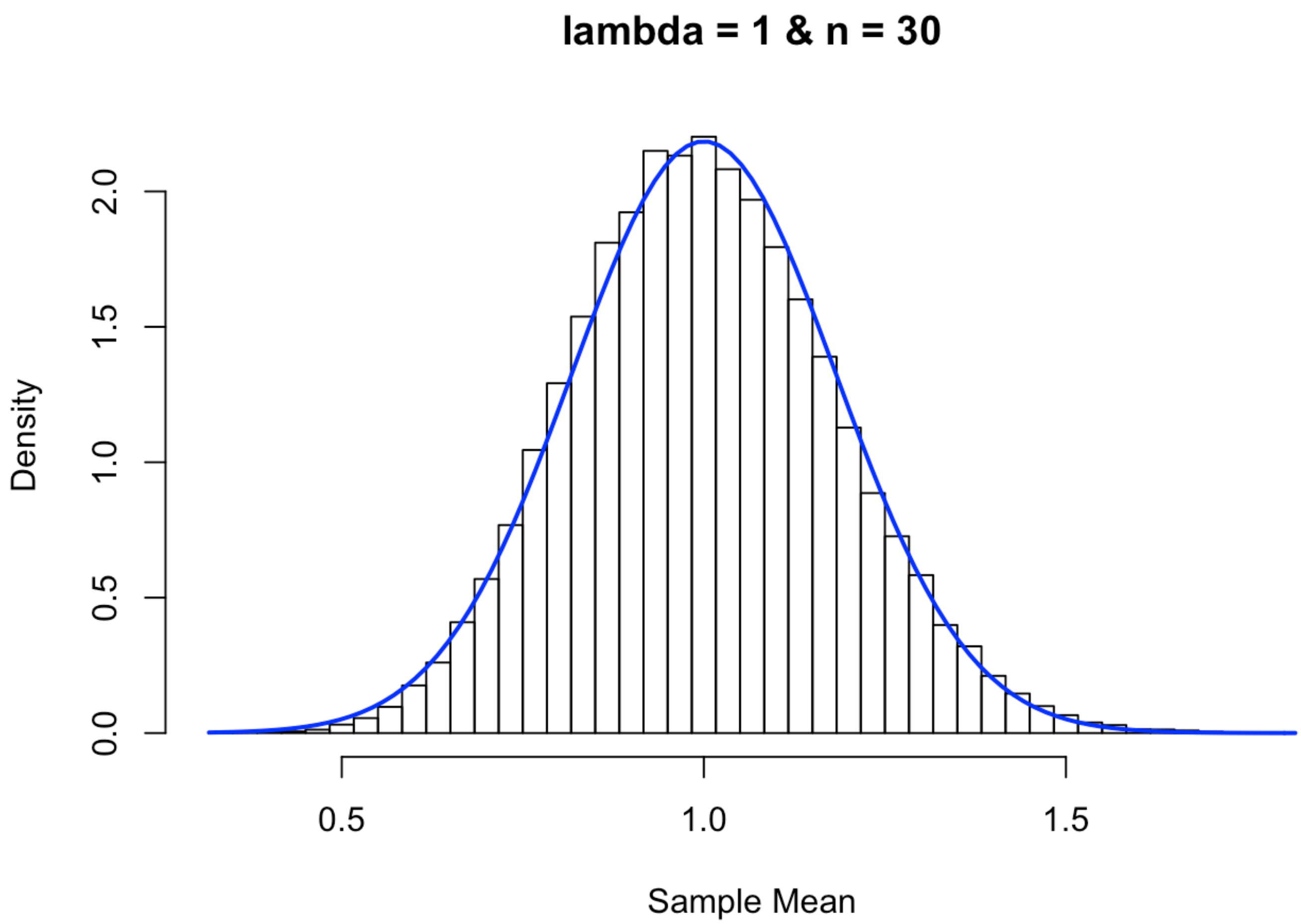
```
set.seed(10)

#Part 2: Convergence in Distribution
#Initiate set of lambda parameter values, set of sample sizes (n) and number of data
sets to generate for each combination of lambda and sample size (N)
lambda = c(1,5,25)
n = c(5,10,30,100)
N = 50000

#Initiate data structures for empirical data, means, empirical & derived probabilitie
s
p2data = list()
p2mean = list()
p2empprob = list()
p2derprob = list()

for (l in 1:length(lambda)) {
  p2data[[l]] = list()
  p2mean[[l]] = list()
  p2empprob[[l]] = list()
  p2derprob[[l]] = list()
  for (i in 1:length(n)) {
    p2data[[l]][[i]] = matrix(0, nrow = N, ncol = n[i])
    p2mean[[l]][[i]] = matrix(0, nrow = N, ncol = 1)
    p2empprob[[l]][[i]] = 0
    p2derprob[[l]][[i]] = 0
  }
}

#Generate values of the Poi(lambda) distributions and compute their means for every c
ombination of lambda and sample size
for (l in 1:length(lambda)) {
  for (i in 1:length(n)) {
    for (j in 1:N) {
      p2data[[l]][[i]][j,] = rpois(n = n[i], lambda = lambda[l])
    }
    p2mean[[l]][[i]] = rowMeans(p2data[[l]][[i]])
  }
}

#Plot empirical vs. derived (normal) distributions for means
for (l in 1:length(lambda)) {
  for (i in 1:length(n)) {
    mu = lambda[l]
    sigma = sqrt(lambda[l]/n[i])
    start = min(p2mean[[l]][[i]])
```

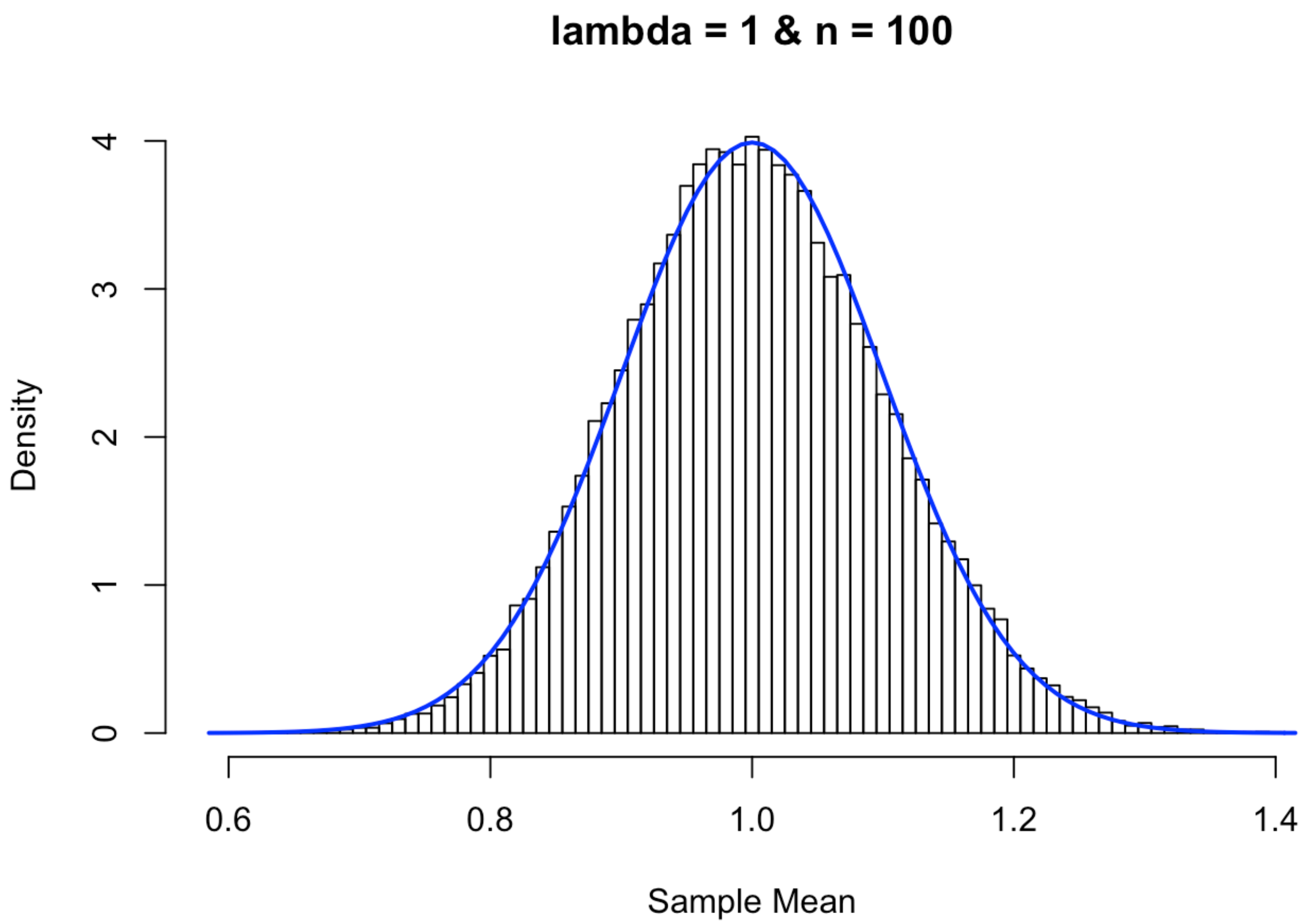
```
    end = max(p2mean[[l]][[i]])
    stepsize = 1/n[i]
    hist(p2mean[[l]][[i]], breaks = seq(from = start-stepsize/2, to = end+stepsize/2,
by = stepsize),
         main = sprintf("lambda = %d & n = %d", lambda[l], n[i]), freq = FALSE, xlab
= "Sample Mean")
    curve(1/(sigma*sqrt(2*pi))*exp(-(x-mu)^2/(2*sigma^2)), from = start-stepsize/2, t
o = end+stepsize/2,
          add = TRUE, lwd = 2, col = "blue")
  }
}
```

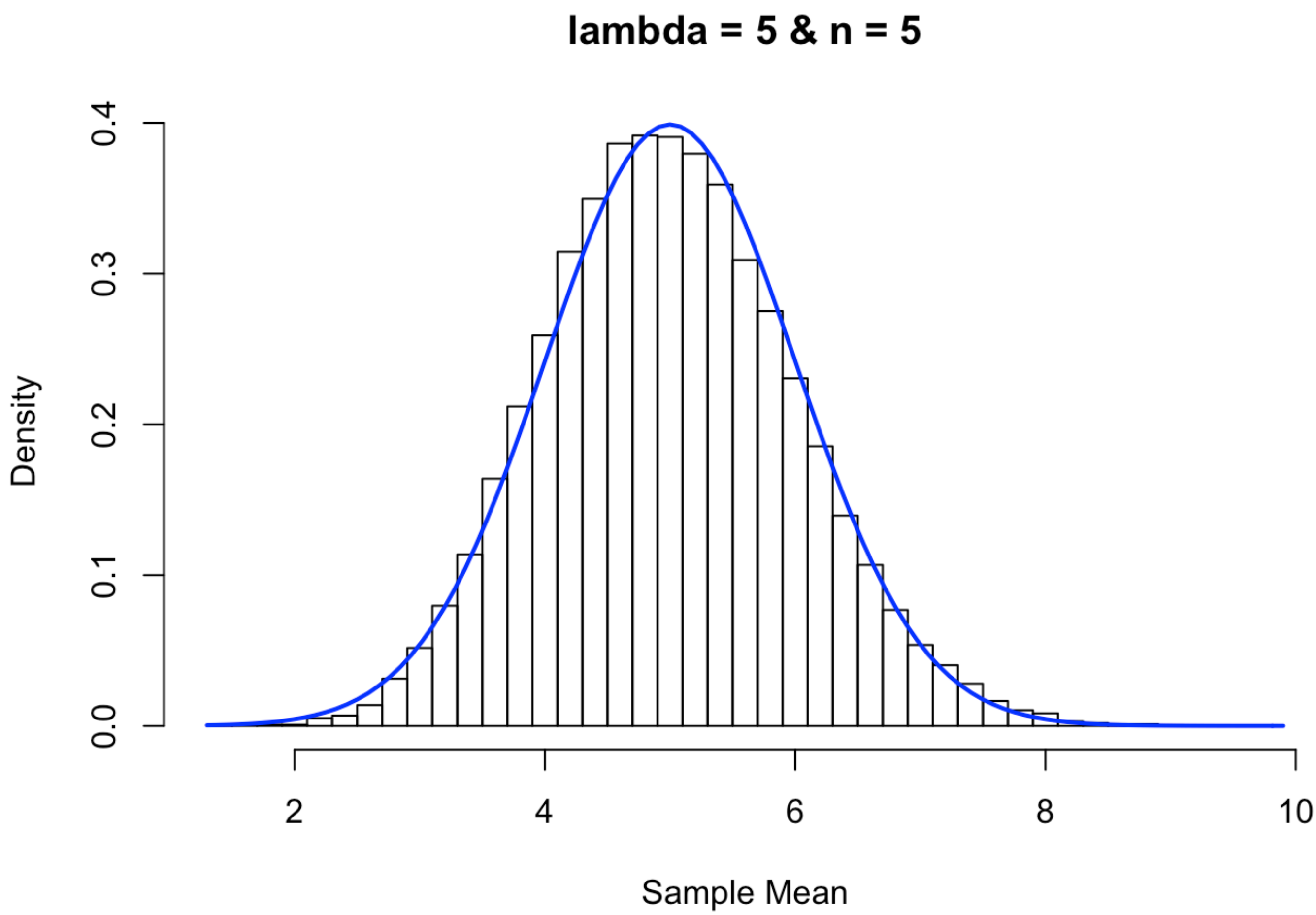

lambda = 1 & n = 5
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sample Mean

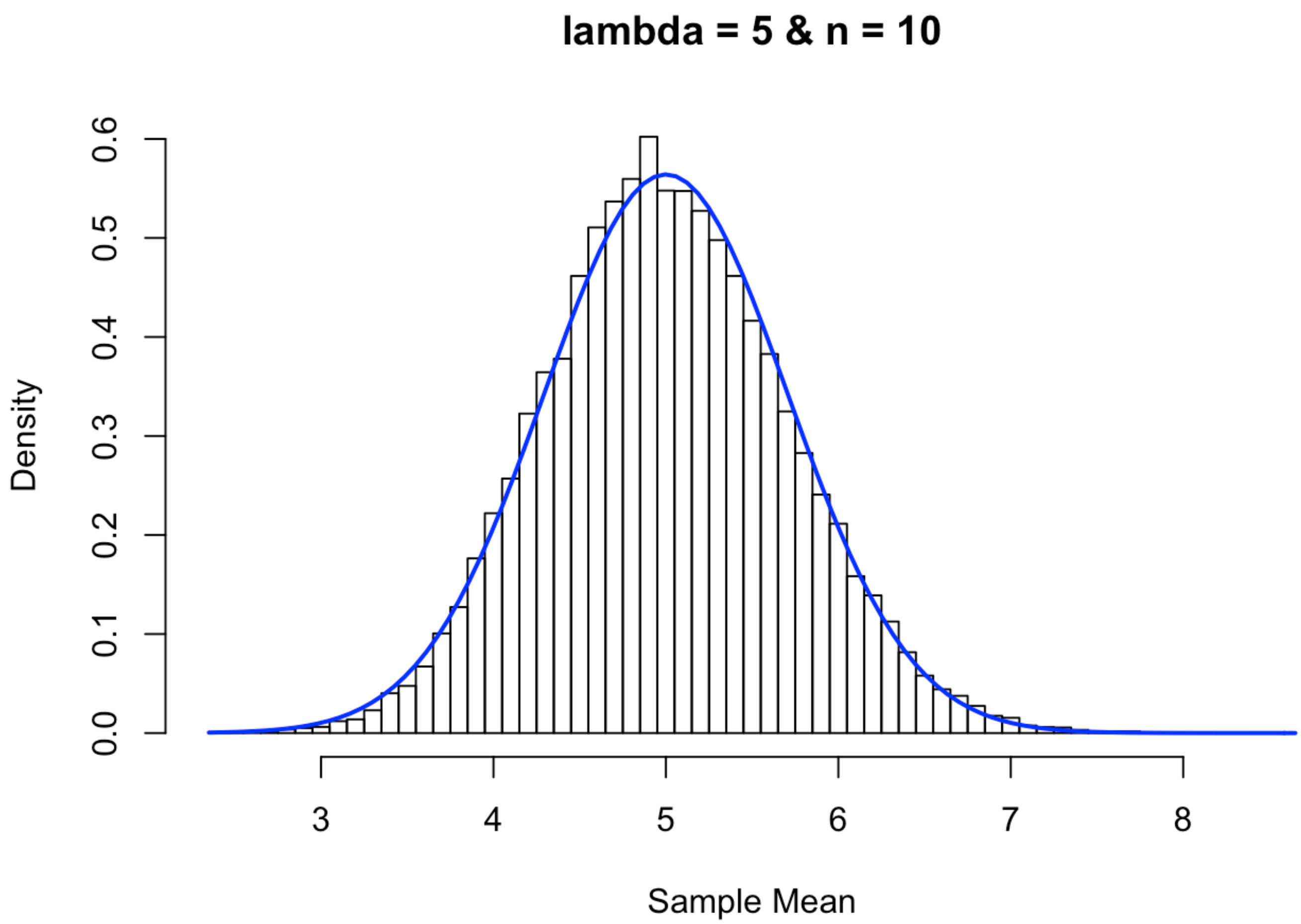

lambda = 1 & n = 10
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5

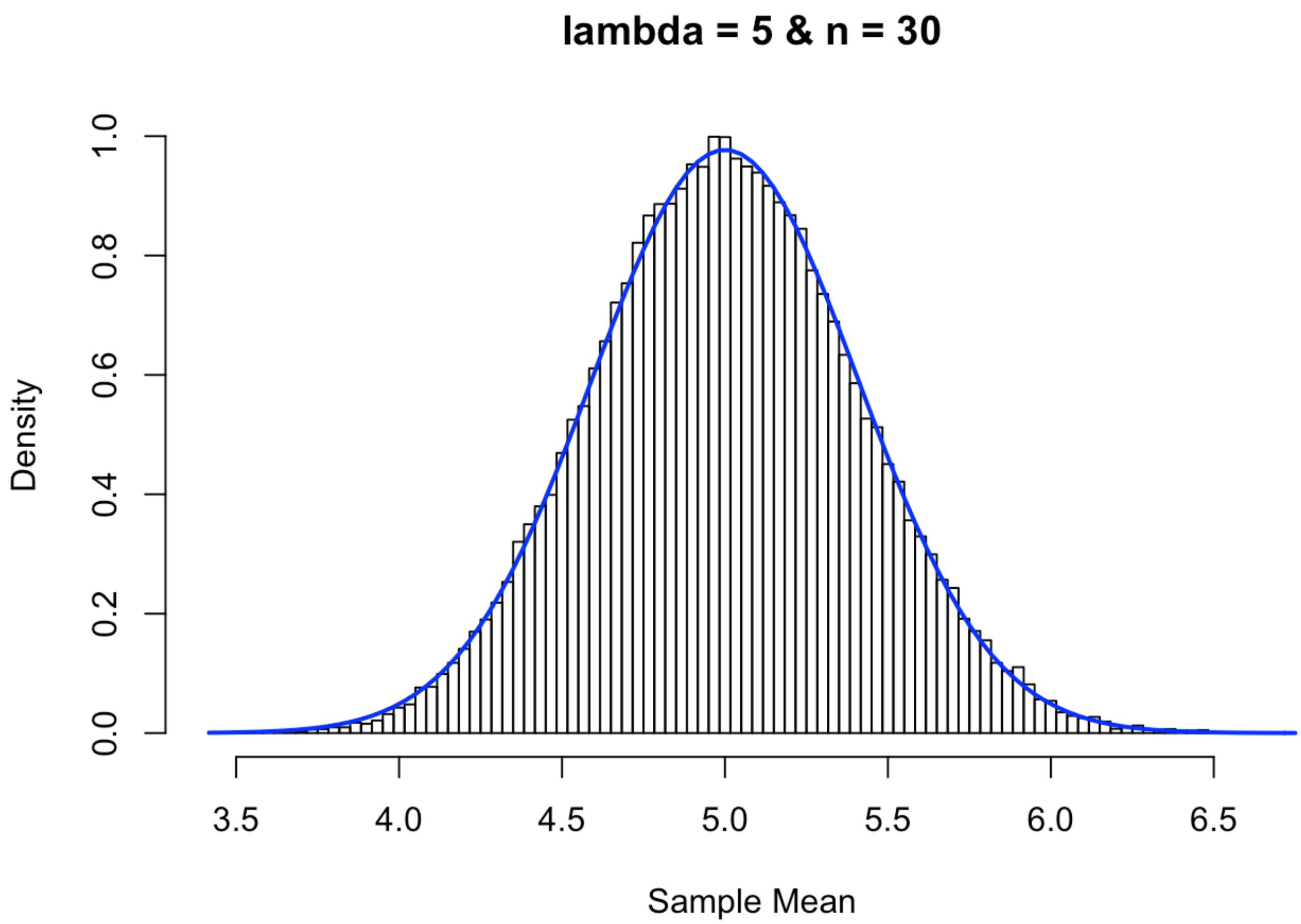
Sample Mean

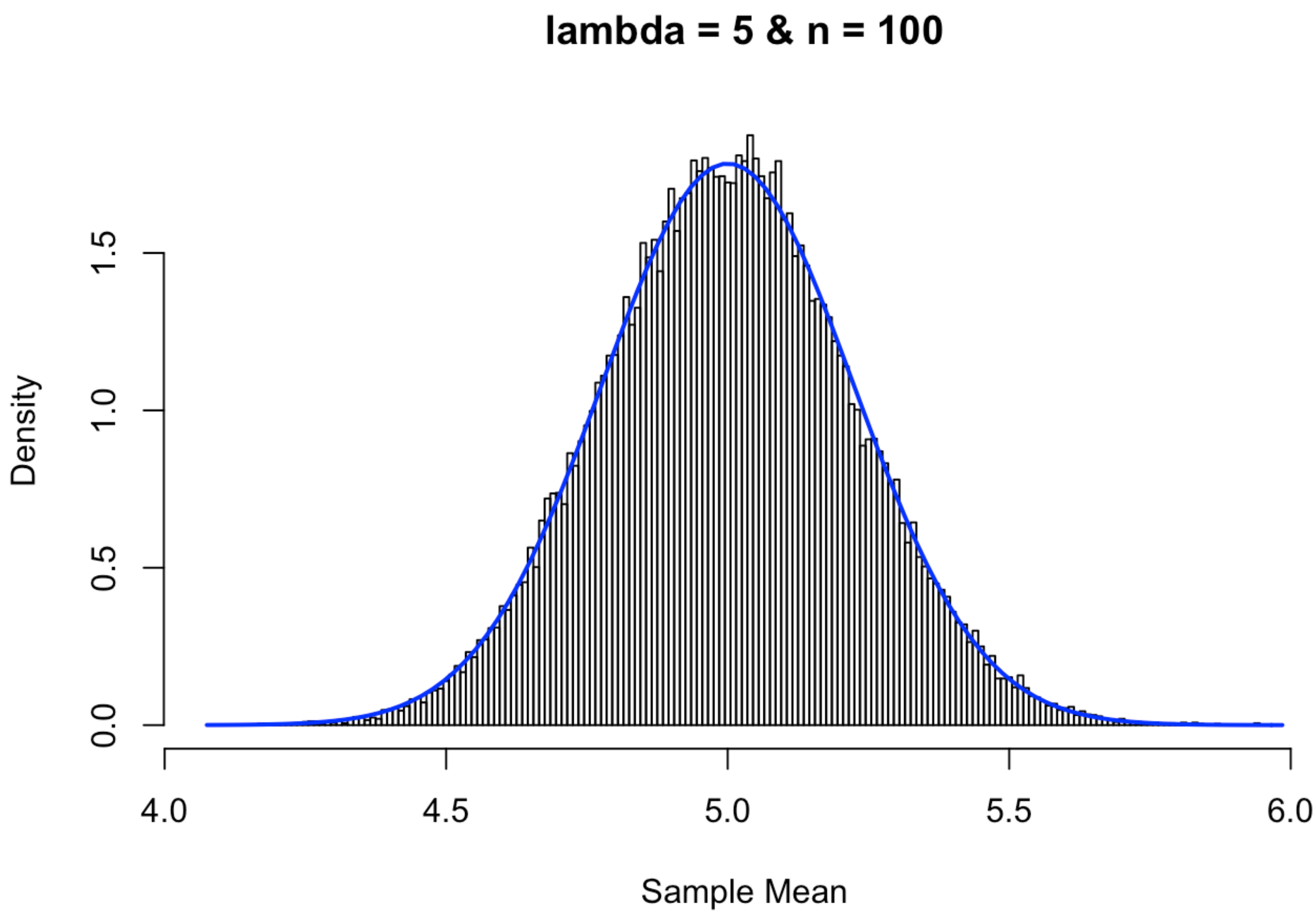


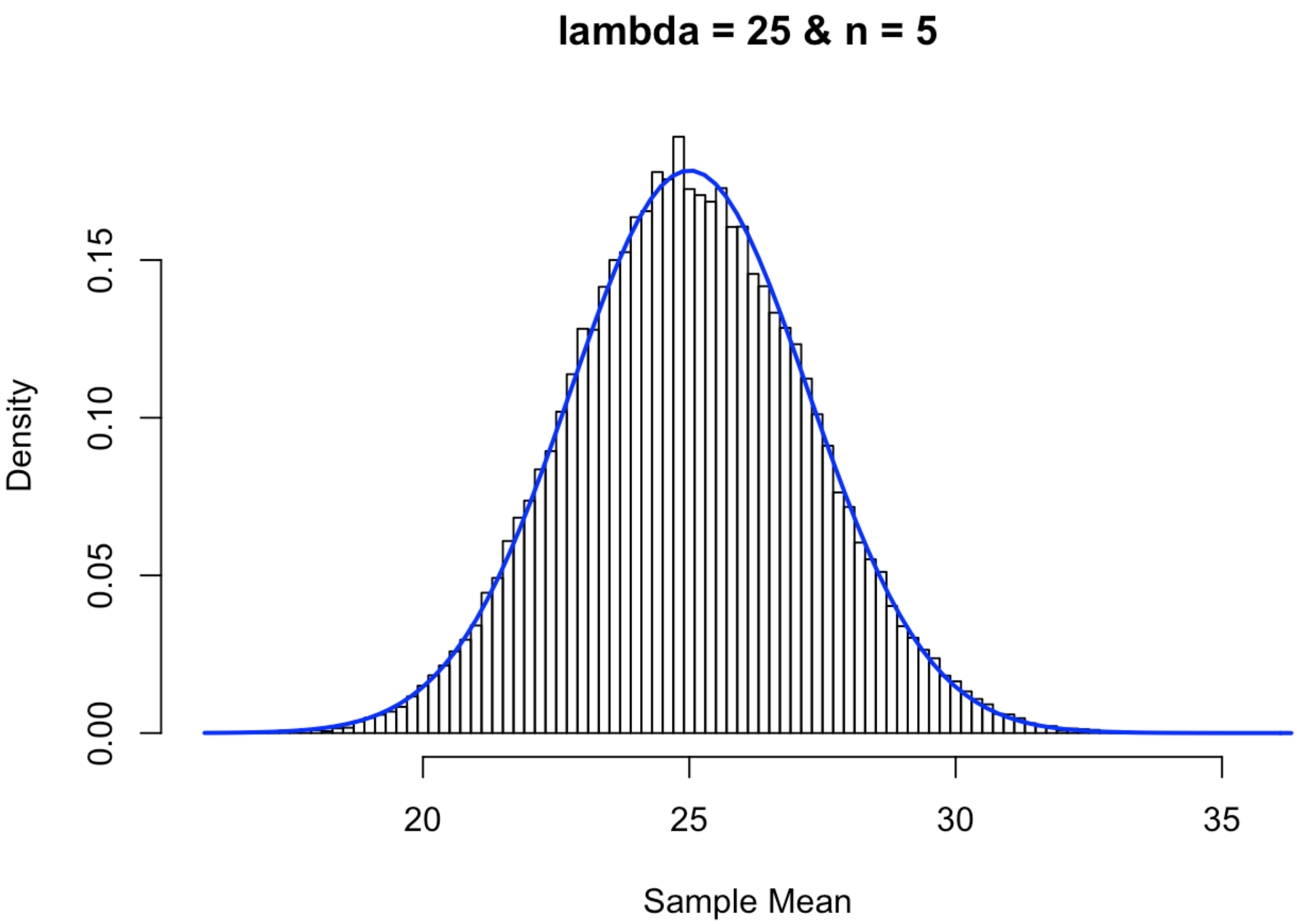

lambda = 1 & n = 100
Density
0
1
2
3
4
0.6
0.8
1.0
1.2
1.4
Sample Mean

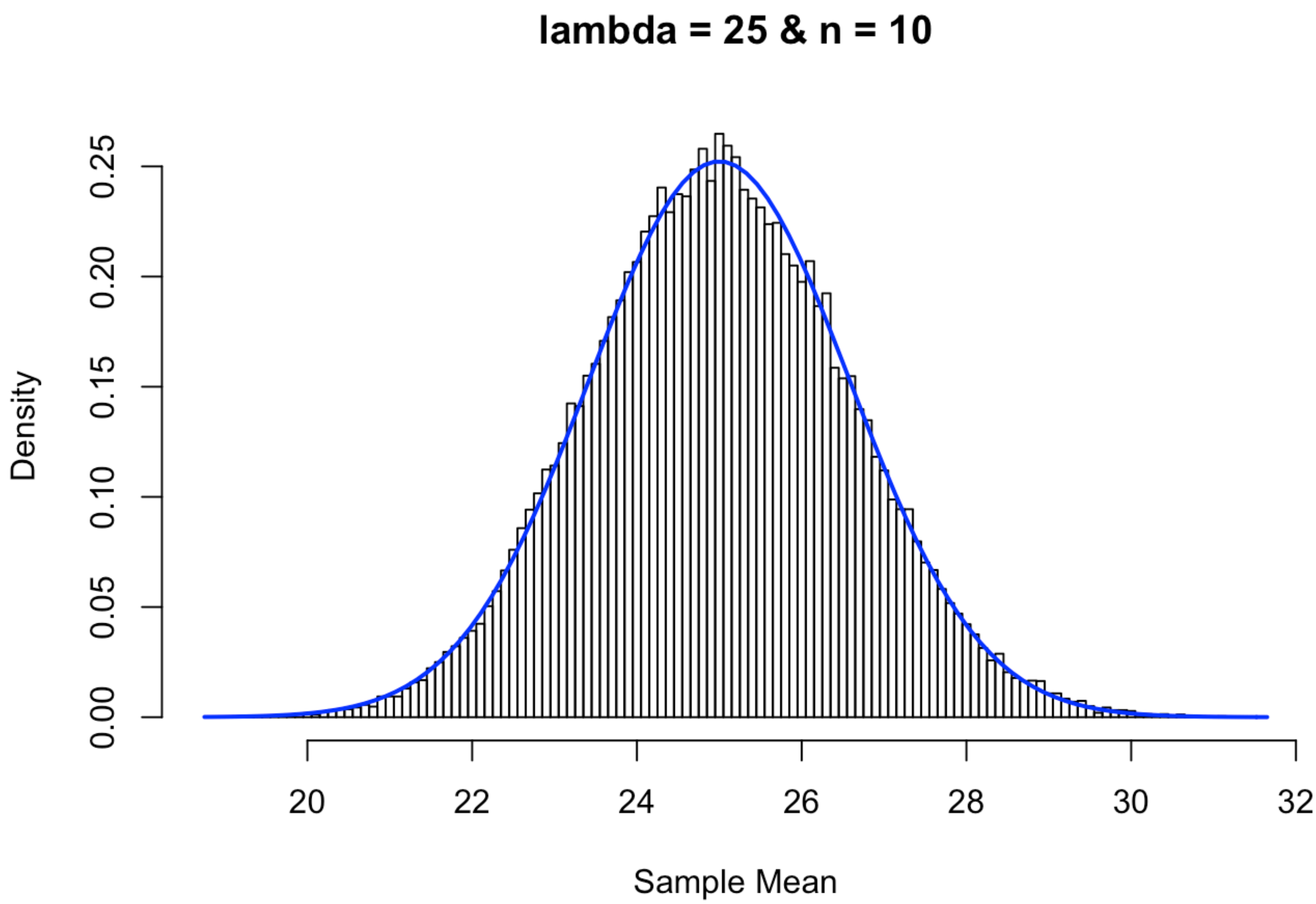

lambda = 5 & n = 5
Density
0.0
0.1
0.2
0.3
0.4
2
4
6
8
10
Sample Mean


lambda = 5 & n = 10
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
3
4
5
6
7
8
Sample Mean

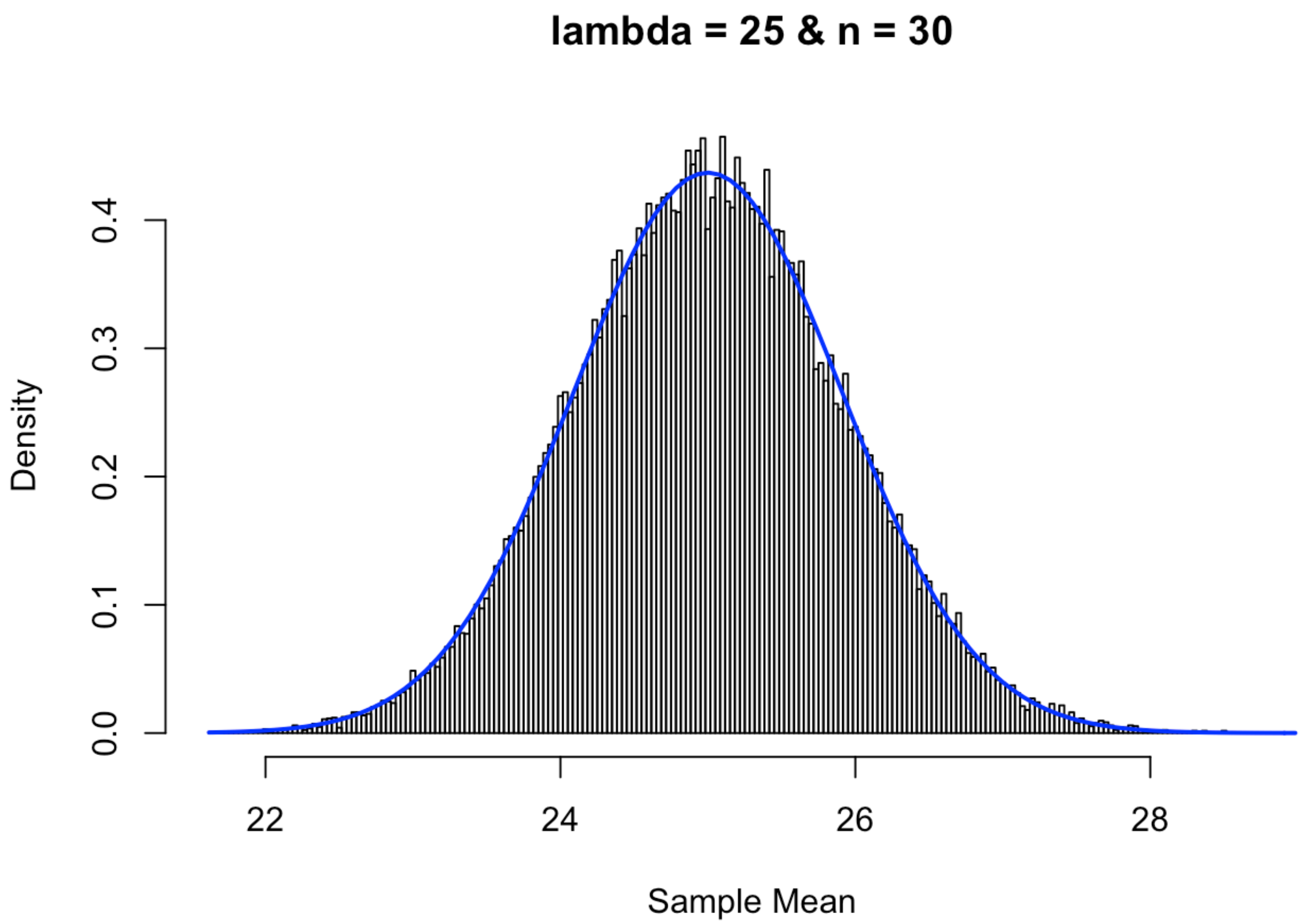

lambda = 5 & n = 30
Density
0.0
0.2
0.4
0.6
0.8
1.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Sample Mean

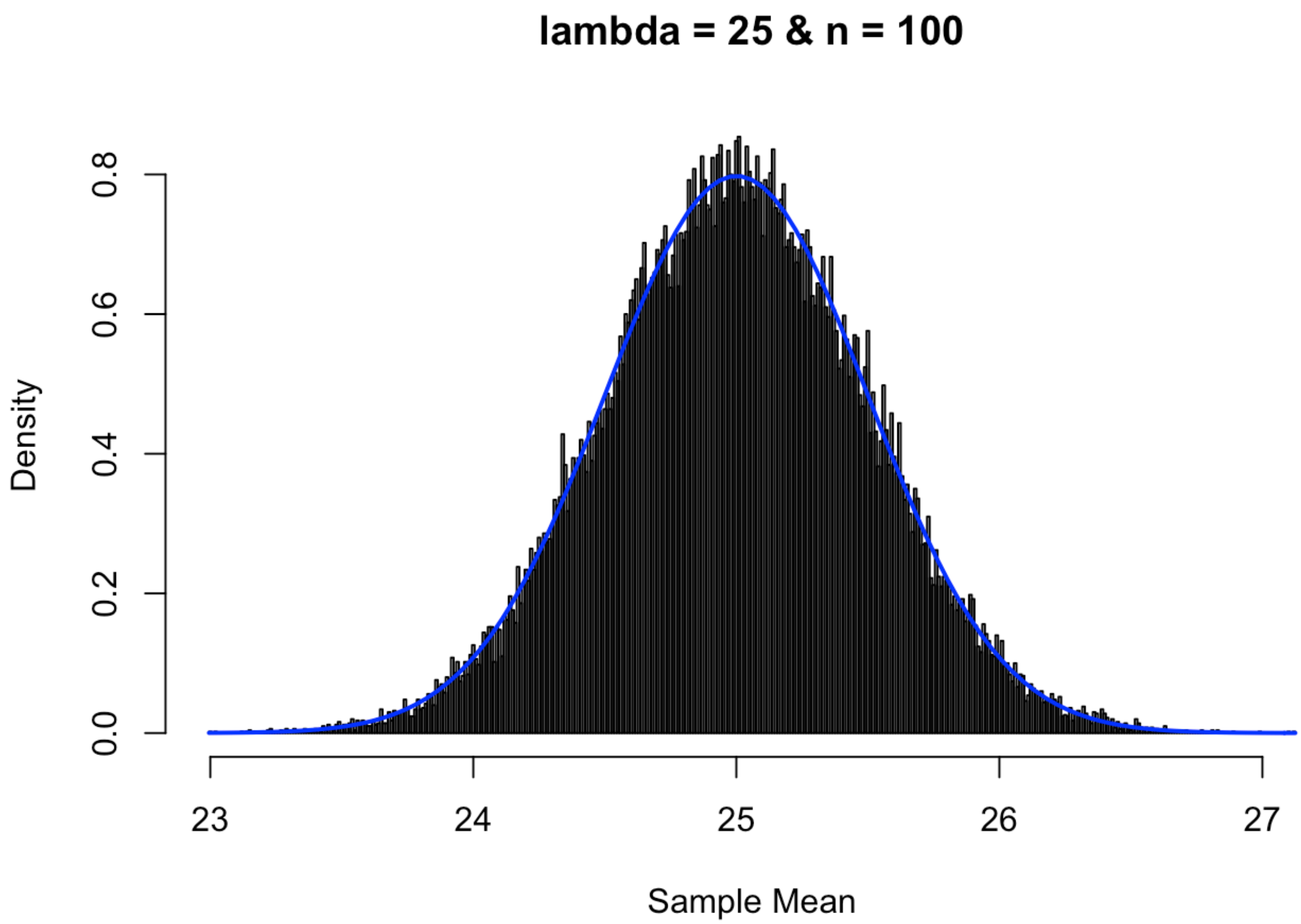

lambda = 5 & n = 100
Density
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
Sample Mean


lambda = 25 & n = 5
Density
0.00
0.05
0.10
0.15
20
25
30
35
Sample Mean



lambda = 25 & n = 10
Density
0.00
0.05
0.10
0.15
0.20
0.25
20
22
24
26
28
30
32
Sample Mean


lambda = 25 & n = 30
Density
0.0
0.1
0.2
0.3
0.4
22
24
26
28
Sample Mean




```
#For every lambda/sample size combination, compute proportion of empirical means grea
ter than or equal to lambda+2*sqrt(lambda/n) and compare to normal probability
for (l in 1:length(lambda)) {
  for (i in 1:length(n)) {
    mu = lambda[l]
    sigma = sqrt(lambda[l]/n[i])
    steps = 1/n[i]
    p2empprob[[l]][[i]] = sum(p2mean[[l]][[i]] >= mu+2*sigma-steps/2)/N
    p2derprob[[l]][[i]] = pnorm(q = mu+2*sigma, mean = mu, sd = sigma, lower.tail = F
ALSE)
    print(sprintf("For lambda = %-2d and n = %-3d the empirical probability is %-1.5f
and the normal probability is %-1.5f", lambda[l], n[i], p2empprob[[l]][[i]], p2derpro
b[[l]][[i]]))
  }
}
```

```
## [1] "For lambda = 1  and n = 5   the empirical probability is 0.06800 and the norm
al probability is 0.02275"
## [1] "For lambda = 1  and n = 10  the empirical probability is 0.04818 and the norm
al probability is 0.02275"
## [1] "For lambda = 1  and n = 30  the empirical probability is 0.03186 and the norm
al probability is 0.02275"
## [1] "For lambda = 1  and n = 100 the empirical probability is 0.02814 and the norm
al probability is 0.02275"
## [1] "For lambda = 5  and n = 5   the empirical probability is 0.03310 and the norm
al probability is 0.02275"
## [1] "For lambda = 5  and n = 10  the empirical probability is 0.03092 and the norm
al probability is 0.02275"
## [1] "For lambda = 5  and n = 30  the empirical probability is 0.02936 and the norm
al probability is 0.02275"
## [1] "For lambda = 5  and n = 100 the empirical probability is 0.02396 and the norm
al probability is 0.02275"
## [1] "For lambda = 25 and n = 5   the empirical probability is 0.02970 and the norm
al probability is 0.02275"
## [1] "For lambda = 25 and n = 10  the empirical probability is 0.02462 and the norm
al probability is 0.02275"
## [1] "For lambda = 25 and n = 30  the empirical probability is 0.02434 and the norm
al probability is 0.02275"
## [1] "For lambda = 25 and n = 100 the empirical probability is 0.02406 and the norm
al probability is 0.02275"
```

2.4) How do the plots and probabilities above help someone understand convergence in distribution?

The 12 plots above exhibit the Central Limit Theorem which states the distribution of the mean of a Random Variable sample ($\bar{Y}$) is well approximated by the Normal Distribution with mean $= \lambda$ and variance $= \lambda/n$. We ran the simulation with three different lambdas (1, 5, 25) with four different sample sizes (5, 10, 30, 100). For each of the lambdas, as n increased, the distribution of ($\bar{Y}$) was better approximated by the Normal(λ, λ/n).

2.5) Why do you think the large-sample approximation works better for larger λ values?

For the Poisson Distribution, as λ increases, the distribution becomes more Normal. If the random samples are being taken from a parent population that can be better approximated by the Normal Distribution, the distribution of the sample means will also be Normal. This is why the $\lambda = 25$ sample mean distribution looks Normal at a sample size of 5 while the $\lambda = 1$ does not look Normal at the same sample size.