In this handout, we describe the use of **Monte Carlo simulation** to illustrate how the Central Limit Theorem (CLT) works. Recall what the CLT says:

Result 2: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a population distribution with mean μ and variance σ^2 (not necessarily a normal distribution). When the sample size n is large, the sample mean

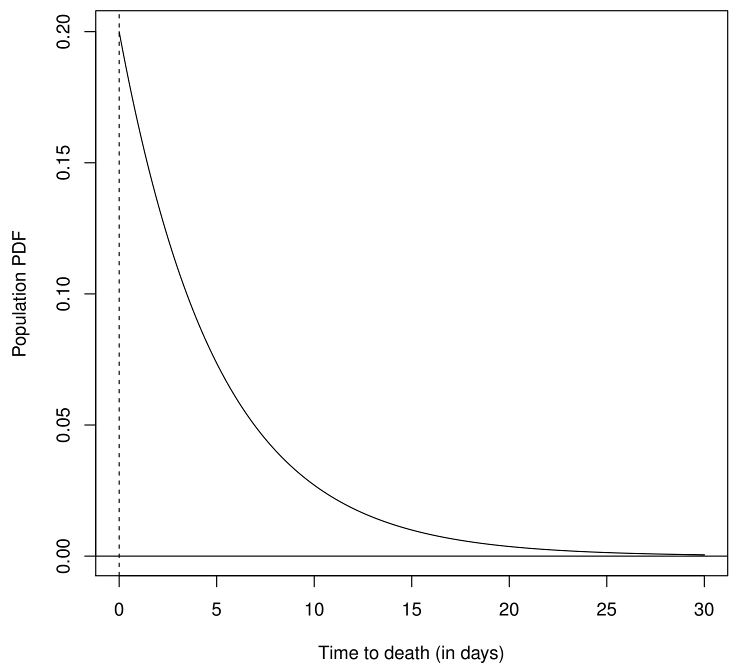
$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right).$$

The symbol $\mathcal{AN}$ is read "approximately normal."

To fix our ideas, suppose we consider Example 6.3, where the death time Y (in days) was modeled using

$$Y \sim \text{exponential}(\lambda = 1/5).$$

This is the population distribution. It describes the time to death for all individual rats in the population.



Consider observing a random sample of $n = 10$ rats and their death times:

$$Y_1, Y_2, ..., Y_{10} \quad \longrightarrow \quad \text{calculate } \overline{Y}$$

R can automate this process:

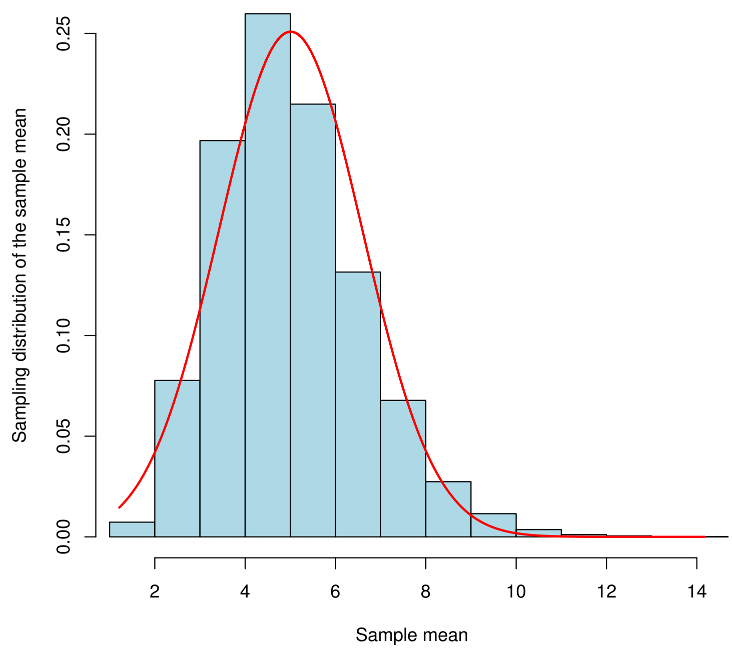
```
n = 10 # sample size
lambda = 1/5 # exponential parameter
exp.data = rexp(n,lambda) # simulate exponential random sample
mean(exp.data) # calculate sample mean
```

We can repeat this process a large number of times

$$\begin{array}{rcll}
\text{Sample 1:} & Y_1, Y_2, ..., Y_{10} & \longrightarrow & \text{calculate } \overline{Y} \\
\text{Sample 2:} & Y_1, Y_2, ..., Y_{10} & \longrightarrow & \text{calculate } \overline{Y} \\
\text{Sample 3:} & Y_1, Y_2, ..., Y_{10} & \longrightarrow & \text{calculate } \overline{Y} \\
 & \vdots & & \\
\text{Sample } B\text{:} & Y_1, Y_2, ..., Y_{10} & \longrightarrow & \text{calculate } \overline{Y}
\end{array}$$

and then look at the empirical distribution formed by plotting all of the sample means in a histogram.

Here is what I got with $B = 10000$; i.e., simulate 10,000 random samples, each of size $n = 10$:



The smooth curve is the normal probability density function calculated at the overall mean and the standard deviation (of the $B = 10000$ sample means).

Interpretation:

- The histogram offers an empirical look at the sampling distribution of $\overline{Y}$, when the sample size is $n = 10$ and the population distribution is exponential with $\lambda = 1/5$.
- The smooth curve is the normal distribution that most closely agrees with the histogram.
- We can see that the normal approximation to the sampling distribution of $\overline{Y}$ (when the sample size $n = 10$) is not that good.

Let's explore what happens when we increase the sample size. I repeated this simulation when $n = 25$, $n = 50$, and $n = 100$.

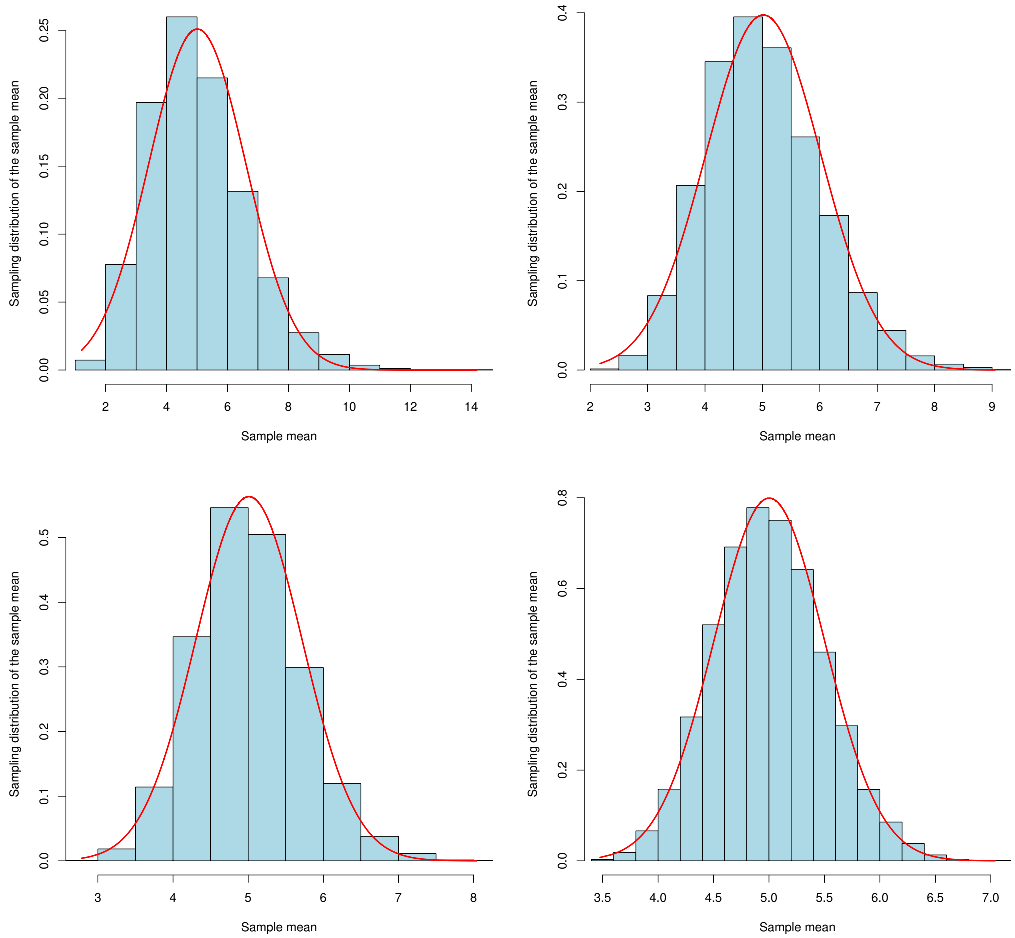


Figure 1: CLT simulation exercise. Sampling distribution of the sample mean $\overline{Y}$ when the population distribution is exponential with $\lambda = 1/5$. Upper left: $n = 10$. Upper right: $n = 25$. Lower left: $n = 50$. Lower right: $n = 100$.

Interpretation:

- As the sample size increases, the normal approximation (smooth curve) to the empirical distribution of the sample mean $\overline{Y}$ (histogram) gets better and better.
- This is precisely what the CLT says should happen.

R code for this simulation exercise is on the next page.

R CODE:

```
n = 10 # sample size
lambda = 1/5 # exponential parameter
B = 10000 # number of Monte Carlo samples

# Generate B samples of exponential(lambda) data, each of size n
# Rows hold the samples (10000 rows)
exp.data = matrix(rexp(n*B,lambda),nrow=B,ncol=n)

# Calculate sample mean for each row (sample)
sample.mean = apply(exp.data,1,mean)

# Make histogram of 10000 sample means (one calculated from each row)
# This is the Monte Carlo distribution
hist(sample.mean,xlab="Sample mean",prob=TRUE,
    xlim=c(min(sample.mean),max(sample.mean)),
    ylab="Sampling distribution of the sample mean",
    main="",col="lightblue")
# Overlay normal density to assess the approximation
lines(sort(sample.mean),
    dnorm(sort(sample.mean),mean(sample.mean),sd(sample.mean)),
    col="red",lwd=2)
```