In this handout, we describe the use of **Monte Carlo simulation** to illustrate how the Central Limit Theorem (CLT) works for the sample proportion $\widehat{p}$. Recall the relevant results below.

If the individual success/failure statuses in the sample adhere to the Bernoulli trial assumptions, then

$$Y = \text{ the number of successes out of } n \text{ sampled individuals}$$

follows a binomial distribution, that is, $Y \sim b(n, p)$. The sample proportion is

$$\widehat{p} = \frac{Y}{n}.$$

Sampling distribution: The Central Limit Theorem says that

$$\widehat{p} \sim \mathcal{AN}\left(p,\ \frac{p(1-p)}{n}\right),$$

when the sample size n is large.

Consider observing a binomial random variable

$$Y \sim b(n, p) \quad \longrightarrow \quad \text{calculate } \widehat{p}$$

R can automate this process:

```
n = 100 # sample size
p = 0.50 # population proportion; pr("success")
binomial.data = rbinom(1,n,p)
sample.prop = binomial.data/n # sample proportion
```

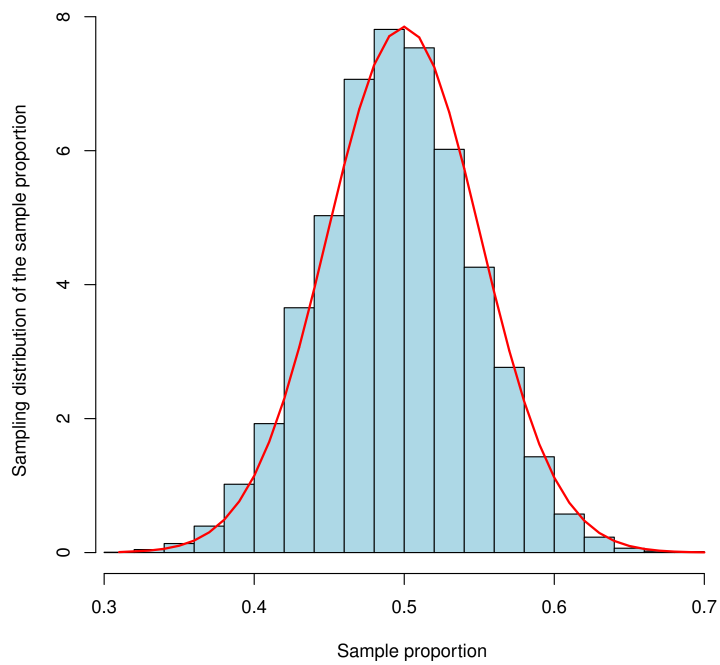
We can repeat this process a large number of times

$$\begin{array}{rlll}
\text{Sample 1:} & Y \sim b(n,p) & \longrightarrow & \text{calculate } \widehat{p} \\
\text{Sample 2:} & Y \sim b(n,p) & \longrightarrow & \text{calculate } \widehat{p} \\
\text{Sample 3:} & Y \sim b(n,p) & \longrightarrow & \text{calculate } \widehat{p} \\
\vdots & & & \\
\text{Sample } B\text{:} & Y \sim b(n,p) & \longrightarrow & \text{calculate } \widehat{p}
\end{array}$$

and then look at the empirical distribution formed by plotting all of the sample proportions $\widehat{p}$ in a histogram.

The figure at the top of the next page shows what I obtained when I did this with $B = 10000$ samples, each with $n = 100$ (sample size) and $p = 0.5$ (population proportion). The smooth curve is the normal probability density function calculated at the overall mean and the standard deviation (of the $B = 10000$ sample proportions).

The histogram offers an empirical look at the sampling distribution of $\widehat{p}$, when the sample size is $n = 100$ and the population proportion is $p = 0.5$. From the figure, we can see that the normal approximation to the sampling distribution of $\widehat{p}$ is quite good.



R CODE:

```
n = 100 # sample size
B = 10000 # number of Monte Carlo samples
p = 0.50 # population proportion; pr("success")
binomial.data = rbinom(B,n,p)
sample.prop = binomial.data/n # sample proportions

hist(sample.prop,xlab="Sample proportion",prob=TRUE,
    xlim=c(min(sample.prop),max(sample.prop)),
    ylab="Sampling distribution of the sample proportion",
    main="",col="lightblue")

# Overlay normal density to assess the approximation
lines(sort(sample.prop),
    dnorm(sort(sample.prop),mean(sample.prop),sd(sample.prop)),
    col="red",lwd=2)
```