Central Limit Theorem

Dr. Robert Buscaglia

January 31, 2021

```
#Libraries I use for this work
library(tidyverse)
library(knitr)
```

Central Limit Theorem

For a random sample of n observations from a population distribution with finite mean μ and variance σ^2, the distribution of Z_n converges in distribution to a $Normal(0,1)$. That is,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} N(0,1)$$

It then follows that the distribution of $\bar{X}_n$ is well approximated by a $Normal(\mu, \sigma^2/n)$. This approximation does not specify a distributional assumption of the population from which the random sample is taken making it generally applicable when the mean and variance are finite.

CLT and sample size

How can we assess how the CLT does under different sample sizes? We will explore two general methods.

1. If the theory provides an exact distribution for $\bar{X}_n$, then we can compare directly our CLT estimate and the true probability. This is not always possible.
2. When the exact distribution of $\bar{X}_n$ is unknown, we can use simulation to compare the estimated distribution of $\bar{X}_n$ to the theoretical Normal distribution suggested by the CLT.

Assessing CLT Estimates

We will use both graphical investigation and calculation of errors to compare the approximations from the CLT to the exact or simulated distribution of $\bar{X}_n$. Graphically we can compare the probability mass function or probability density function of the estimated or exact distribution of $\bar{X}$ to that of the approximating Normal distribution. Assessment of the error in our calculations can be tabulated. Below I will use the percent relative error

$$PRE = \frac{\hat{p} - p}{p} \times 100\%$$

where $\hat{p}$ is the estimated probability from CLT and p is the exact probability determined from an exact distribution. When the exact distribution cannot be used, it is useful to also visually inspect the distribution rather than using only tabulated errors that also include simulation error.

Example 1 - Comparing Exact and CLT-based Approximations

Consider a random sample of size n from an Exponential distribution with parameter $\lambda > 0$. That is, let $X_1, ..., X_n \overset{\text{iid}}{\sim} Exp(\lambda)$. Then for each fixed value of n, it can shown that $\bar{X}_n \sim Gamma(n, n\lambda)$. Be sure you could prove this result. This allows us to show that

$$E[\bar{X}_n] = \frac{1}{\lambda} \qquad \text{and} \qquad V[\bar{X}_n] = \frac{1}{n\lambda^2}$$

Now consider the central limit theorem which suggests that

$$\bar{X}_n \dot{\sim} Normal\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right)$$

Let us compare the CLT approximation to the known distribution for different sample sizes.

Setting up the comparison

We must first choose parameters for the approximation. We will work with a different range of sample sizes, $n = 2, 10, 25, 100, 1000$ and let us fix the rate parameters to be $\lambda = 1/4$. Notice with this that we can write out

$$\begin{aligned} X_1, ..., X_n &\overset{\text{iid}}{\sim} Exp(1/4) \\ \bar{X}_n &\sim Gamma(n, n/4) \\ E[\bar{X}_n] = 4 \quad &\text{and} \quad V[\bar{X}_n] = \frac{16}{n} \\ \bar{X}_n &\dot{\sim} Normal(4, 16/n) \qquad \text{by CLT} \end{aligned}$$

Let us initialize this for obtaining the probabilities from R. Here I enter in the values above and must also set a sample size n, which I start defaulted at $n = 2$.

```
n <- 2 #sample size
lambda <- 1/4 #rate parameters
mu_xbar <- 1/lambda #mean of X-bar
var_xbar <- 1/(n*lambda^2) #variance of X-bar
sd_xbar <- sqrt(var_xbar) #sd of X-bar
```

A starting example

We will start with $n = 2$ and consider $P(\bar{X}_n < 4)$. We can calculate the exact probability in this case by knowing the distribution of $\bar{X}_n$. We use R for this calculation

```
pgamma(q = 3, shape = n, rate = n*lambda)
```

```
## [1] 0.4421746
```

Be clear here the parameters of the distribution. The shape is what Chihara labels as r and the rate λ. We observe that $P(\bar{X}_n < 3) = 0.4422$ using the known distribution of $\bar{X}_n$. Now, let us calculate the approximation

```
### mean and sd were calculated above in an earlier
pnorm(3, mean = mu_xbar, sd = sd_xbar)
```

```
## [1] 0.3618368
```

Based on the central limit theorem we estimate $P(\bar{X}_n < 4) = 0.3618$. This gives a $PRE = 18.2\%$.

```
abs(0.3618 - 0.4422)/(0.4422)
```

```
## [1] 0.1818182
```

This looks like a poor estimate, and is, especially given this is not in the tail. Let us try a tail probability. Consider $P(\bar{X}_n < 1)$

```
gam2 <- pgamma(q = 1, shape = n, rate = n*lambda)
gam2
```

```
## [1] 0.09020401
```

```
norm2 <- pnorm(q = 1, mean = mu_xbar, sd = sd_xbar)
norm2
```

```
## [1] 0.1444222
```

```
(norm2 - gam2)/(gam2)
```

```
## [1] 0.6010617
```

This gives $P(\bar{X}_n < 4) = 0.0902$ from the known distribution and $P(\bar{X}_n < 1) = 0.1444$ from the CLT estimate, producing a $PRE = 60.11\%$. Yikes!

Tabulation

Let us use the power of R to produce a large tabulation comparing the known and CLT estimated probabilities. We will use a diverse set of known probabilities $0.01, 0.01, 0.1, 0.4, 0.5, 0.6, 0.75, 0.8, 0.99$.

```
set.probs <- c(0.001, 0.01, 0.1, 0.4, 0.5, 0.6, 0.75, 0.8, 0.99)
set.probs
```

```
## [1] 0.001 0.010 0.100 0.400 0.500 0.600 0.750 0.800 0.990
```

We can obtain the value at which the gamma distributions will obtain this probability from the quantiles.

```
quant.gamma <- qgamma(set.probs, shape = n, rate = n*lambda)
quant.gamma
```

```
## [1]  0.09080404  0.29710948  1.06362322  2.75284268  3.35669398  4.04462649
## [7]  5.38526906  5.98861669 13.27670414
```

These values represent x^* which is the chosen quantile of the gamma distribution. For example, $P(\bar{X}_n < x^*) = 0.001$ gives $x^* = 0.09080404$. We can then find what the estimated probability was by evaluating at the determined quantiles for the known gamma.

```
est.probs <- pnorm(quant.gamma, mean = mu_xbar, sd = sd_xbar)
est.probs
```

```
## [1] 0.08346904 0.09523847 0.14959640 0.32962941 0.41003963 0.50629419 0.68785046
## [8] 0.75899773 0.99948062
```

Let us establish a table to compare the probability from the known distribution to the CLT estimate.

```
CLT.1 <- data.frame(x.star = quant.gamma, Known.Prob = set.probs,
                    CLT.Prob = est.probs %>% round(4),
                    PRE = 100*abs(est.probs - set.probs)/(set.probs) )
CLT.1$PRE <- CLT.1$PRE %>% round(2)
CLT.1 %>% kable()
```

x.star	Known.Prob	CLT.Prob	PRE
0.0908040	0.001	0.0835	8246.90
0.2971095	0.010	0.0952	852.38
1.0636232	0.100	0.1496	49.60
2.7528427	0.400	0.3296	17.59
3.3566940	0.500	0.4100	17.99
4.0446265	0.600	0.5063	15.62
5.3852691	0.750	0.6879	8.29
5.9886167	0.800	0.7590	5.13
13.2767041	0.990	0.9995	0.96

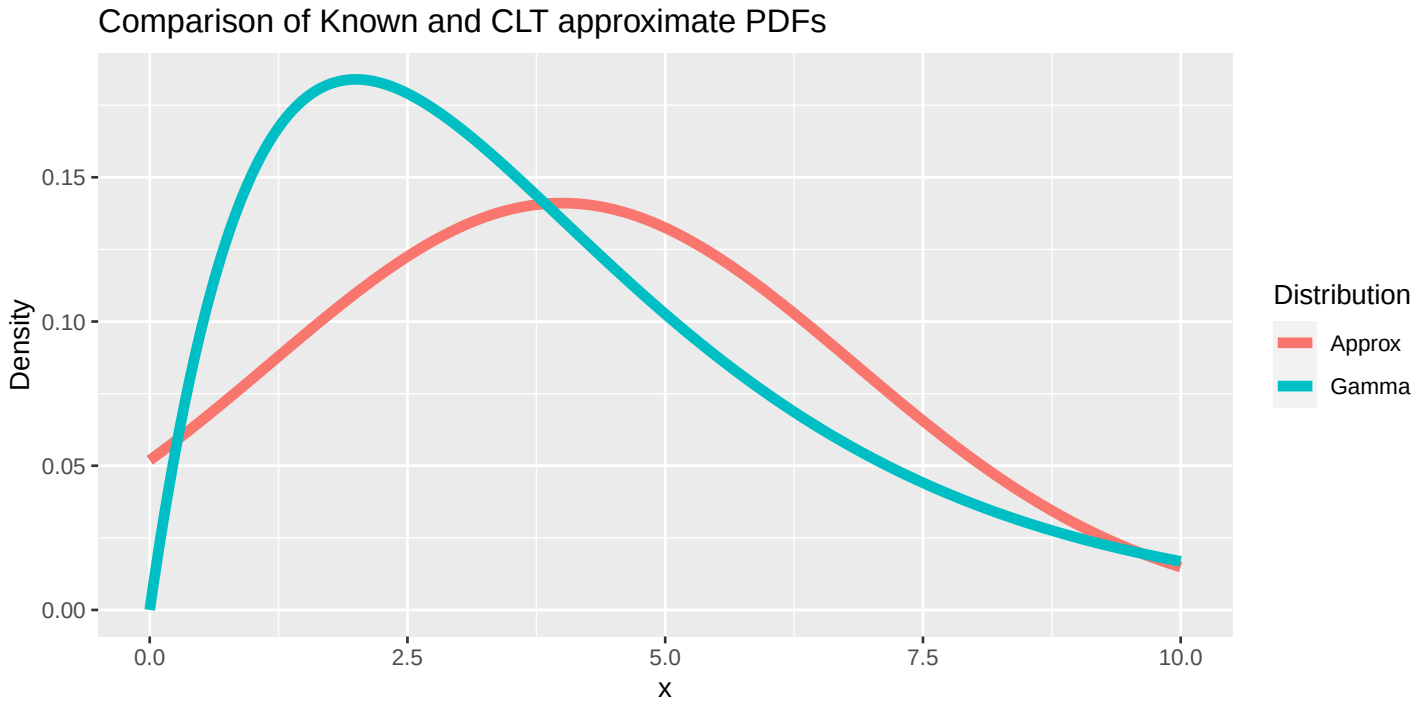
We can now start to see the comparison, and it is clear that the CLT estimate is poor, with very high error in the extreme tails. Recall this is all done at $n = 2$ so far.

Visualize

Before moving on and seeing how the sample size will change this error, let us prepare a visual of the known distribution and its CLT estimate.

```
n2.df <- data.frame(x = seq(0, 10, 0.01))
n2.df <- n2.df %>% mutate(Gamma = dgamma(x, shape = n, rate = n*lambda),
                          Approx = dnorm(x, mean = mu_xbar, sd = sd_xbar))
n2.df.long <- n2.df %>% pivot_longer(cols = 2:3, names_to = 'Distribution',
                                     values_to = 'Density')

ggplot(n2.df.long) +
   geom_line(aes(x = x, y = Density, color = Distribution), size = 2) +
   labs(title = 'Comparison of Known and CLT approximate PDFs')
```



The graphic makes it clear the normal distribution is a poor approximate of the true distribution when $n = 2$.

A survey of sample size

Let us produce a table that gives a summary of information across sample sizes. We will produce two new tables. One that tabulates the estimated probability in comparison with the true values, a second that allows us to see the PRE.

```r
survey.df <- data.frame(True.Prob = c(0.001, 0.01, 0.1, 0.4, 0.5, 0.6, 0.75, 0.8, 0.99))
pre.df <- data.frame(True.Prob = c(0.001, 0.01, 0.1, 0.4, 0.5, 0.6, 0.75, 0.8, 0.99))
n.size <- c(2, 10, 25, 50, 100, 500, 1000)
for(n in n.size)
{
   lambda <- 1/4 #rate parameters
   mu_xbar <- 1/lambda #mean of X-bar
   var_xbar <- 1/(n*lambda^2) #variance of X-bar
   sd_xbar <- sqrt(var_xbar) #sd of X-bar

   set.probs <- c(0.001, 0.01, 0.1, 0.4, 0.5, 0.6, 0.75, 0.8, 0.99)
   quant.gamma <- qgamma(set.probs, shape = n, rate = n*lambda)
   est.probs <- pnorm(quant.gamma, mean = mu_xbar, sd = sd_xbar)
   pre <- 100*abs(est.probs - set.probs)/(set.probs)

   survey.df <- cbind(survey.df, est.probs %>% round(4))
   pre.df <- cbind(pre.df, pre %>% round(2))
}

colnames(survey.df) <- c('True.Prob', paste0('n=',n.size))
colnames(pre.df) <- c('True.Prob', paste0('n=',n.size))
```

First a table showing the comparison of the probabilities.

```r
survey.df %>%
   kable(align='c',
         caption='Comparison of Known and CLT estimated probabilities.')
```

Table 2: Comparison of Known and CLT estimated probabilities.

True.Prob	n=2	n=10	n=25	n=50	n=100	n=500	n=1000
0.001	0.0835	0.0130	0.0057	0.0035	0.0025	0.0015	0.0013
0.010	0.0952	0.0317	0.0212	0.0171	0.0147	0.0119	0.0113
0.100	0.1496	0.1161	0.1091	0.1061	0.1042	0.1018	0.1012
0.400	0.3296	0.3645	0.3769	0.3835	0.3882	0.3947	0.3962
0.500	0.4100	0.4583	0.4735	0.4812	0.4867	0.4941	0.4958
0.600	0.5063	0.5598	0.5750	0.5825	0.5877	0.5946	0.5962
0.750	0.6879	0.7275	0.7368	0.7410	0.7438	0.7473	0.7481
0.800	0.7590	0.7871	0.7929	0.7953	0.7969	0.7987	0.7991
0.990	0.9995	0.9973	0.9955	0.9943	0.9933	0.9916	0.9912

Here is a table summarizing the percent relative errors with increasing sample size.

```
pre.df %>%
   kable(align='c',
         caption='Percent Relative Errors for increasing sample sizes.')
```

Table 3: Percent Relative Errors for increasing sample sizes.

True.Prob	n=2	n=10	n=25	n=50	n=100	n=500	n=1000
0.001	8246.90	1200.44	466.09	254.26	149.36	52.20	34.84
0.010	852.38	217.13	112.12	71.41	46.88	18.99	13.11
0.100	49.60	16.06	9.14	6.11	4.15	1.76	1.23
0.400	17.59	8.87	5.77	4.14	2.95	1.34	0.95
0.500	17.99	8.34	5.30	3.76	2.66	1.19	0.84
0.600	15.62	6.70	4.16	2.92	2.05	0.91	0.64
0.750	8.29	3.00	1.77	1.20	0.83	0.36	0.25
0.800	5.13	1.61	0.89	0.59	0.39	0.16	0.11
0.990	0.96	0.73	0.56	0.44	0.33	0.16	0.12

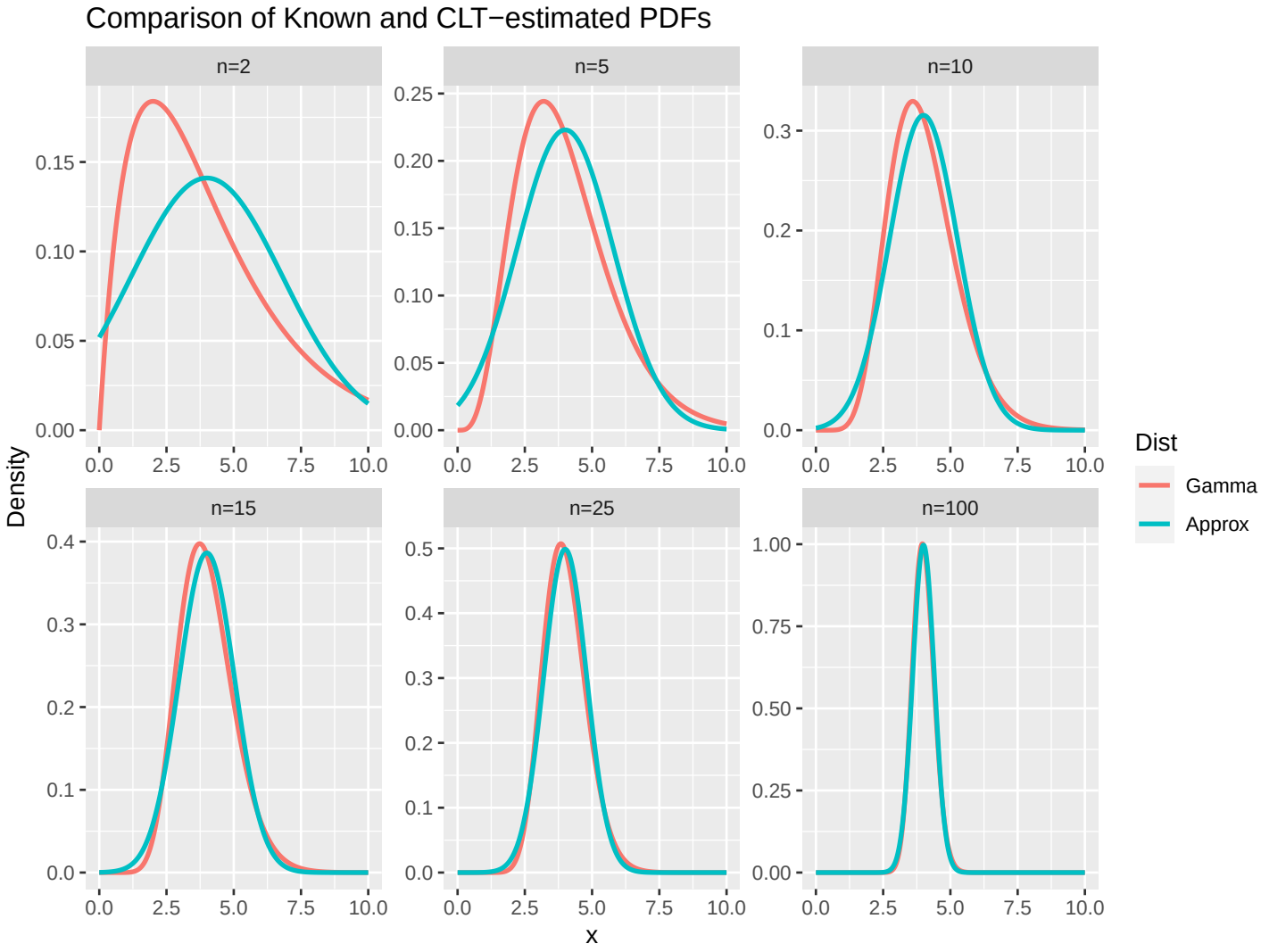
Visualize

A comparison of the known and approximate PDFs for each sample size. I changed the sample sizes a bit to show a bit more of the changes that occur as n increases.

```
full.df <- NULL
n.size <- c(2, 5, 10, 15, 25, 100)
for(n in n.size)
{
temp <- data.frame(x = seq(0, 10, 0.01))
lambda <- 1/4 #rate parameters
mu_xbar <- 1/lambda #mean of X-bar
var_xbar <- 1/(n*lambda^2) #variance of X-bar
sd_xbar <- sqrt(var_xbar) #sd of X-bar
temp.gamma <- temp %>% mutate(Size = factor(rep(paste0('n=',n), length(x))),
                              Dist = factor(rep('Gamma', length(x))),
                              Density = dgamma(x, shape = n, rate = n*lambda))
temp.approx <- temp %>% mutate(Size = factor(rep(paste0('n=',n), length(x))),
                              Dist = factor(rep('Approx', length(x))),
                              Density = dnorm(x, mean = mu_xbar, sd = sd_xbar))

full.df <- rbind(full.df, temp.gamma, temp.approx)
}

ggplot(full.df) + geom_line(aes(x = x, y = Density, col=Dist), size=1) +
   facet_wrap(~Size, scales='free') +
   labs(title='Comparison of Known and CLT-estimated PDFs')
```


Comparison of Known and CLT−estimated PDFs
n=2
n=5
n=10
n=15
n=25
n=100
Density
x
Dist
Gamma
Approx

Example 2 - Assessing CLT-based Approximations without a Known Distribution

For this example we will consider a random sample from a population with a $Beta(\alpha, \beta)$ distribution. If $X_1, ...X_n \overset{\text{iid}}{\sim} Beta(\alpha, \beta)$, then the distribution of $\bar{X}_n$ is difficult to work with analytically. Instead, for this example, we will use the what knowledge we do have of this distribution to make estimates based on the CLT. To compare, we will use numerical simulation of drawing samples from this population and evaluating the observed mean. Repeating this enough times will allow us to compare the sampling distribution of $\bar{X}_n$ to the estimate suggested by the CLT.

Recall Beta distribution.

Recall for a random variable X that follows a Beta distribution with parameters α and β both non-negative,

$$E[X] = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad V[X] = \frac{\alpha\beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)}$$

The beta-distribution allows for many unique 'shapes' of curves defined on the interval $x \in (0, 1)$. For a review of the PDF, see the Chihara appendix.

To evaluate the estimate provided by the central limit theorem, these are the only two values we need to know! We then know that for large enough sample sizes, that

$$\bar{X}_n \dot{\sim} Normal\left(\mu = \frac{\alpha}{\alpha + \beta}, \quad \sigma^2 = \frac{1}{n} \cdot \frac{\alpha\beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)}\right)$$

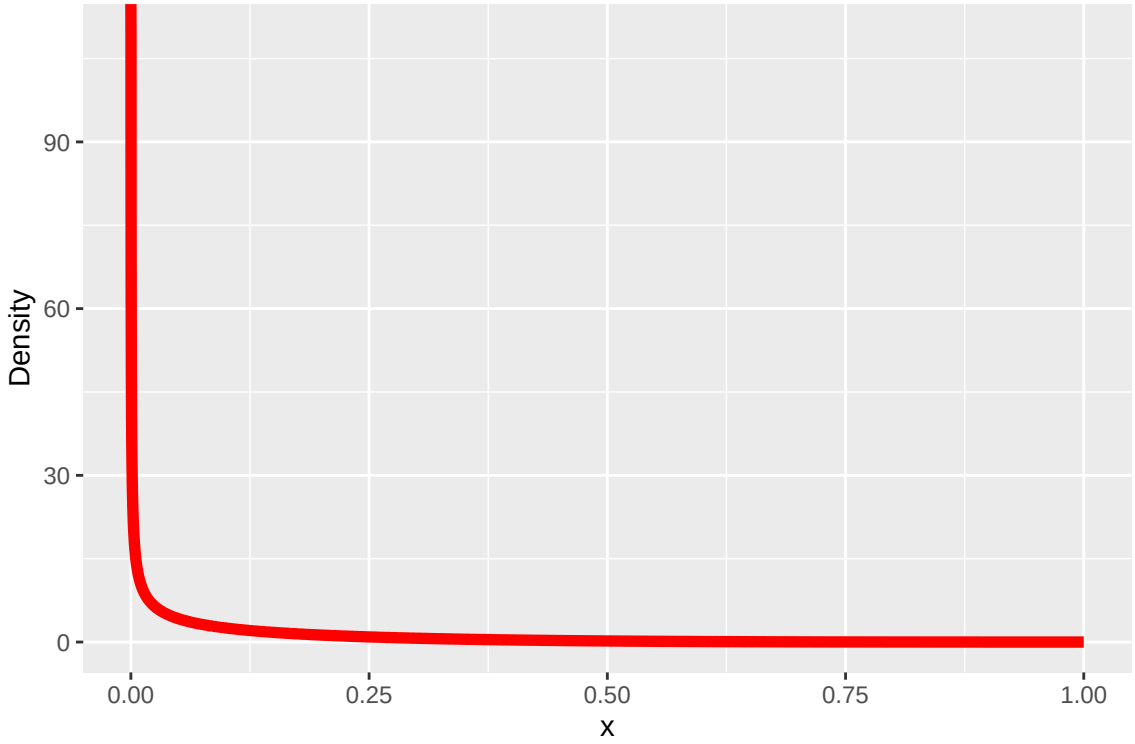
Setting up the approximation

We will need parameters to make the approximation. Let us assume that

$$X_i \sim Beta\left(\alpha = \frac{1}{2}, \beta = 4\right) \quad \text{for } i = 1, ..., n$$

We will use this for the approximations made below. Here is what the PDF of the distribution for visualization.

```
df.e2 <- data.frame( x = seq(0, 1, length.out = 1e4))
df.e2 <- df.e2 %>% mutate( y = dbeta(x, 1/2, 4))
ggplot(df.e2, aes(x = x, y = y)) + geom_line(size = 2, col='red') +
   labs(y = 'Density')
```



Next,let us find the mean and variance of this distribution.

$$E[X] = \frac{\frac{1}{2}}{\frac{1}{2}+4} = \frac{1}{9} \quad \text{and} \quad V[X] = \frac{\frac{1}{2}\cdot 4}{(\frac{1}{2}+4)^2\cdot(\frac{1}{2}+4+1)} = \frac{2}{\frac{81}{4}\cdot\frac{11}{2}} = \frac{16}{891}$$

We can then state for a random sample of size n that

$$\bar{X}_n \dot{\sim} Normal\left(\mu = \frac{1}{9}, \sigma^2 = \frac{16}{891n}\right)$$

A sample problem

Find $P(\bar{X}_n < 0.1)$ based on the CLT approximation for a sample of size $n = 32$. Based on this sample size we have that

$$\bar{X}_n \dot{\sim} Normal\left(\mu = \frac{1}{9}, \sigma^2 = \frac{16}{891\cdot 32} = \frac{1}{1782}\right)$$

Then based on the CLT we can find an estimate of the probability using R

```
pnorm(0.1, 1/9, sqrt(1/1782))
```

```
## [1] 0.31952
```

Thus, based on the CLT, $P(\bar{X}_n < 0.1) = 0.3195$. As we saw above though, the error rate of the approximation relays heavily on the sample size and position of the approximation (tails were harder to estimate).

Assessing the CLT-approximation

Without being able to calculate the known probability based on a distributional assumption, we will have to relay on computation. We will draw samples from the beta distribution of size n and calculate the mean. If we repeat ths process enough (how does $1e5$ sound for today), we will get a good look at the sampling distribution of $\bar{X}_n$. We can then compare our simulated densities to that of the estimated CLT density.

An example

Let us start with $n = 2$, which draws two samples from the population and takes the mean. Start by drawing two samples randomly (this is the r-functions, which stand for `random` as in random draws). I save my random draw as the object `rand.sample`.

```
rand.sample <- rbeta(n = 2, shape1 = 1/2, shape2 = 4)
rand.sample
```

```
## [1] 0.01096662 0.12390856
```

Now all that is left is to calculate the mean and store it somewhere. Lets make a vector to store the means we calculate called `means.out`. Then the next (and final) calculation is

```
means.out <- numeric()
means.out <- mean(rand.sample)
means.out
```

```
## [1] 0.06743759
```

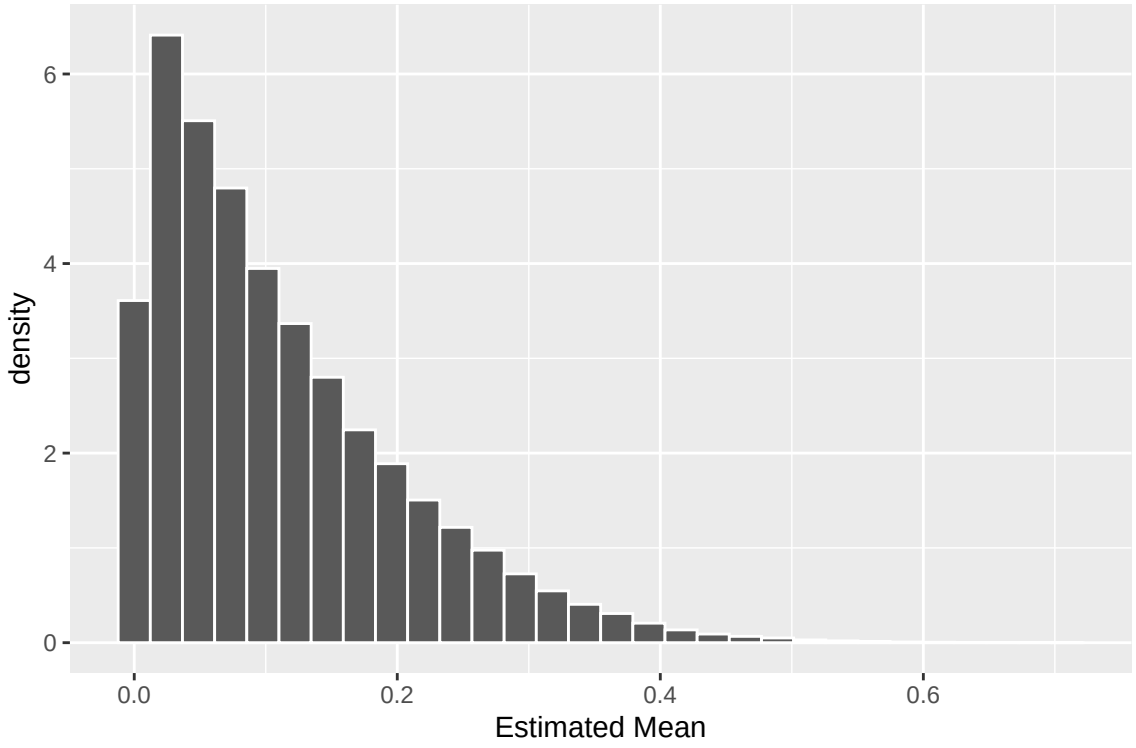
A simulation

Let us put this together and simulate this process $1e5$ times. Here is the simulation in full. Notice I make the vector `means.out` before running the loop, this ensures I start with an empty vector. I have used `set.seed` to make the simulation reproducible.

```
set.seed(2021)
means.out <- numeric()
n.set = 2 ## sample size of random sample
for(i in 1:1e5)
{
   rand.sample <- rbeta(n = n.set, shape1 = 1/2, shape2 = 4)
   means.out[i] <- mean(rand.sample)
}
```

With the means simulated for a sample of size $n = 2$, we can now visualize an esimated density curve for $\bar{X}_n$.

```
means.df <- data.frame( means.out = means.out )
ggplot(means.df, aes(x = means.out)) + geom_histogram(aes(y=..density..), col='white') +
   labs(x = 'Estimated Mean')
```

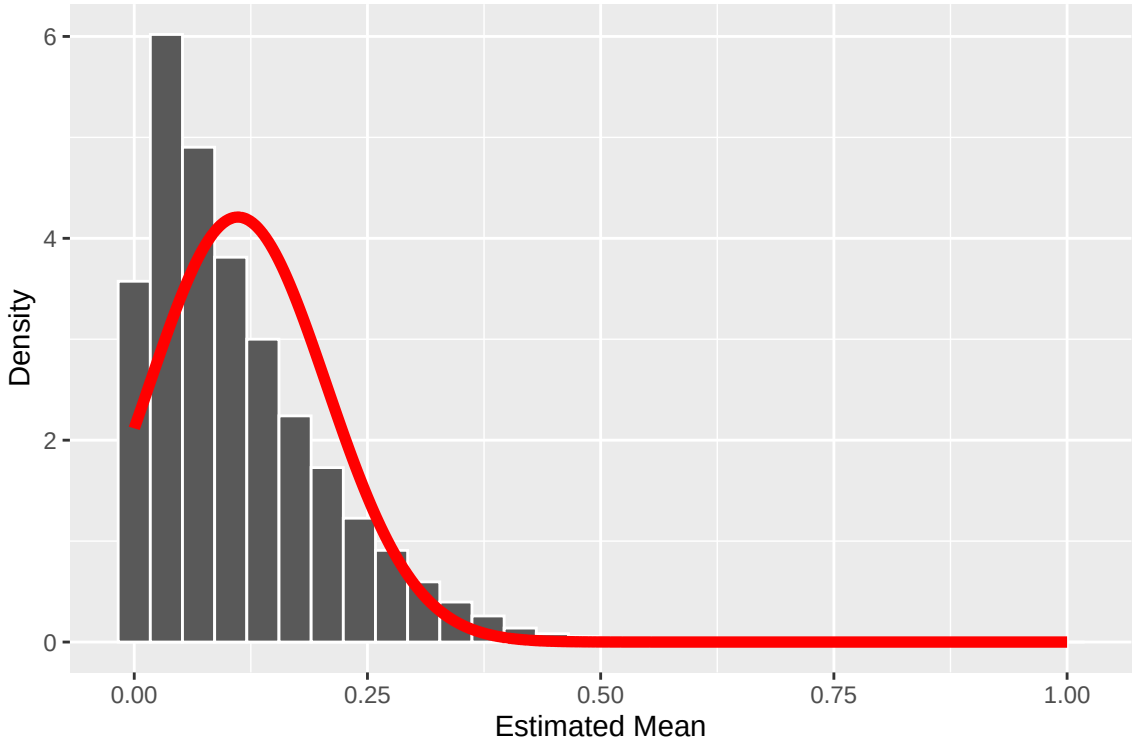


Let us overlay the known density curve for the normal approximation based on the CLT. First a data.frame to store the normal density curve.

```
n.set = 2
approx.df <- data.frame(x = seq(0, 1, length.out = 1e4)) %>%
   mutate(Approx = dnorm(x, mean = 1/9, sd = sqrt(16/(891*n.set))))
```

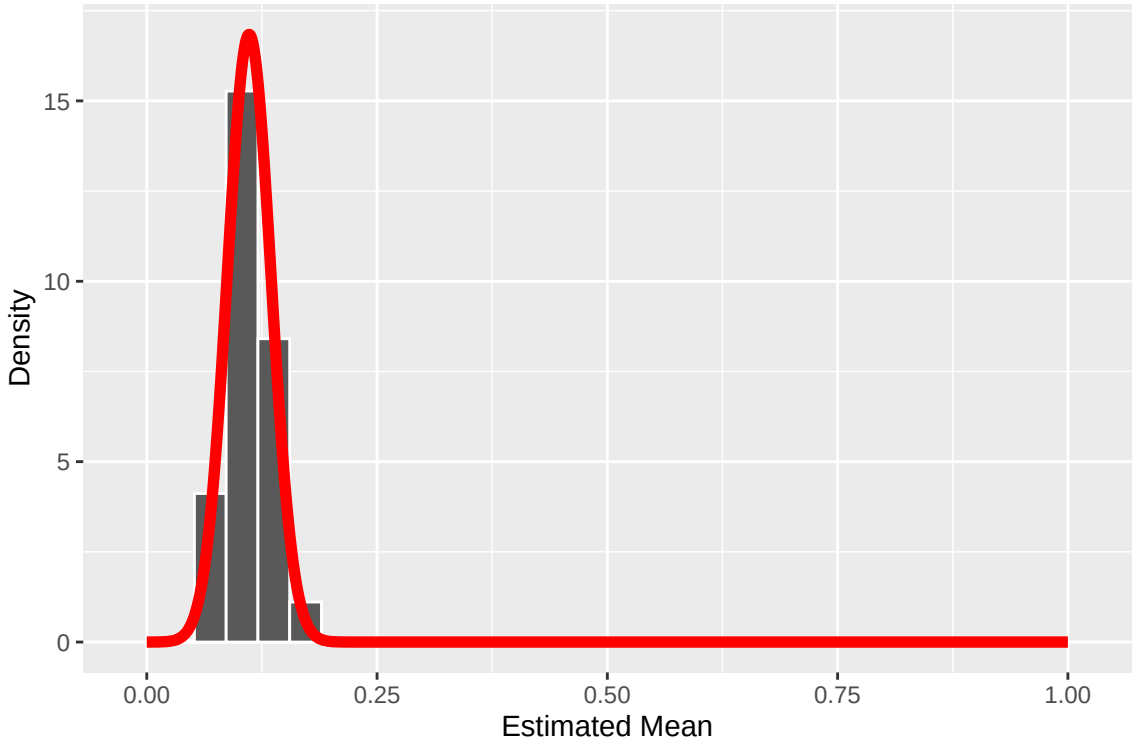
Now the graph

```
ggplot(means.df, aes(x = means.out)) +
   geom_histogram(aes(y=..density..), col='white') +
   labs(x = 'Estimated Mean', y='Density') +
   geom_line(data = approx.df, aes(x = x, y = Approx), col='red', size=2)
```



We can observe that even with $n = 2$, there is starting to be a similar shape for the estimated distribution. We could also compare the simulated probabilities to the CLT approximation. Let us do so using our example from above that used $n = 32$. First, lets run the simulation and visualize again.

```
set.seed(2021) ## reproducible
means.out <- numeric() ## vector to store answers
n.set = 32 ## sample size of random sample
for(i in 1:1e5) ## simulation
{
   rand.sample <- rbeta(n = n.set, shape1 = 1/2, shape2 = 4)
   means.out[i] <- mean(rand.sample)
}
## data frame for simulation
means.df <- data.frame( means.out = means.out )
## data frame for approximation
approx.df <- data.frame(x = seq(0, 1, length.out = 1e4)) %>%
   mutate(Approx = dnorm(x, mean = 1/9, sd = sqrt(16/(891*n.set))))
### make the graph
ggplot(means.df, aes(x = means.out)) +
   geom_histogram(aes(y=..density..), col='white') +
   labs(x = 'Estimated Mean', y='Density') +
   geom_line(data = approx.df, aes(x = x, y = Approx), col='red', size=2)
```



We can make an estimate of the probability from the simulated sampling. Recall we were interested $P(X < 0.1)$. By accounting for how many times a value less than 0.1 was observed, and dividing by the total number of iterations, we obtain the simulation estimate.

```
length(which(means.df$means.out < 0.1))/length(means.df$means.out)
```

```
## [1] 0.33304
```

Recall the value based on CLT was

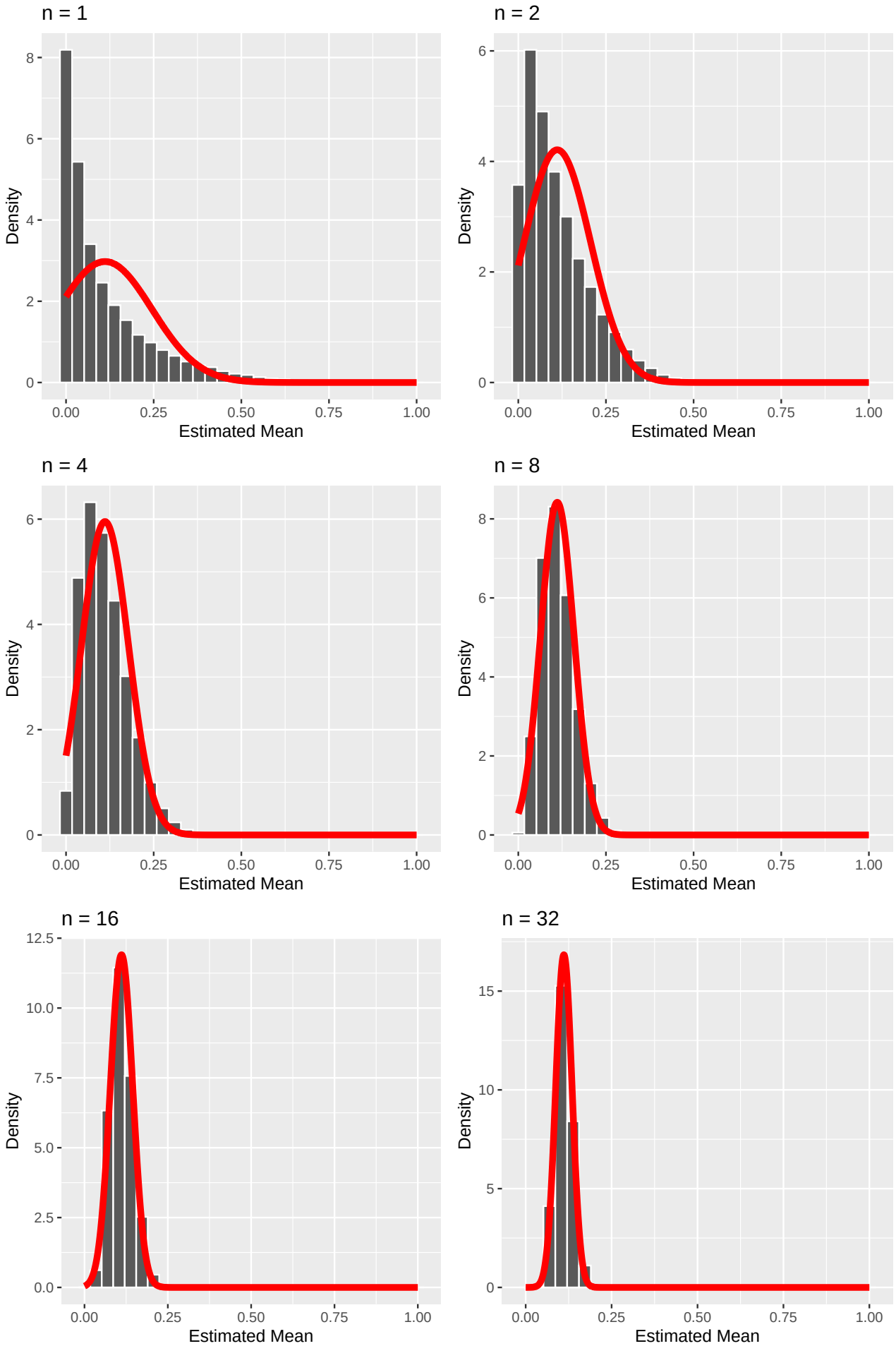
```
pnorm(0.1, 1/9, sqrt(1/1782))
```

```
## [1] 0.31952
```

The difficulty here is we now have two estimates, neither is from a proven distribution assumption. We could increase the number of iterations in the simulation to improve the precision of the approximation, but this requires more computing. Thus, the CLT provides a much less intensive approximation. This may not be as nice of an assessment as we could get from working out the theoretical details of the distribution of $\bar{X}_n$. However, there are cases such as this one where doing so is difficult to impossible. Computations and using the power of CLT has provided quick estimates without the need for a distributional assumption on $\bar{X}_n$.

Visualizing sample size and CLT

Let us finish up by preparing a visual comparison of the simulated density curves for $\bar{X}_n$ and the overlaid estimated density curve from the CLT approximation. We will do so for a variety of random sample sizes. I hide my code because I made these a bit faster without preparing one large data.frame. Instead I made 6 different graphs and used `cowplot`. Ask questions if interested!


n = 1
n = 2
n = 4
n = 8
n = 16
n = 32
Density
Estimated Mean
0.00
0.25
0.50
0.75
1.00