Introduction to Central Limit Theorems

```r
library(fastR2)
```

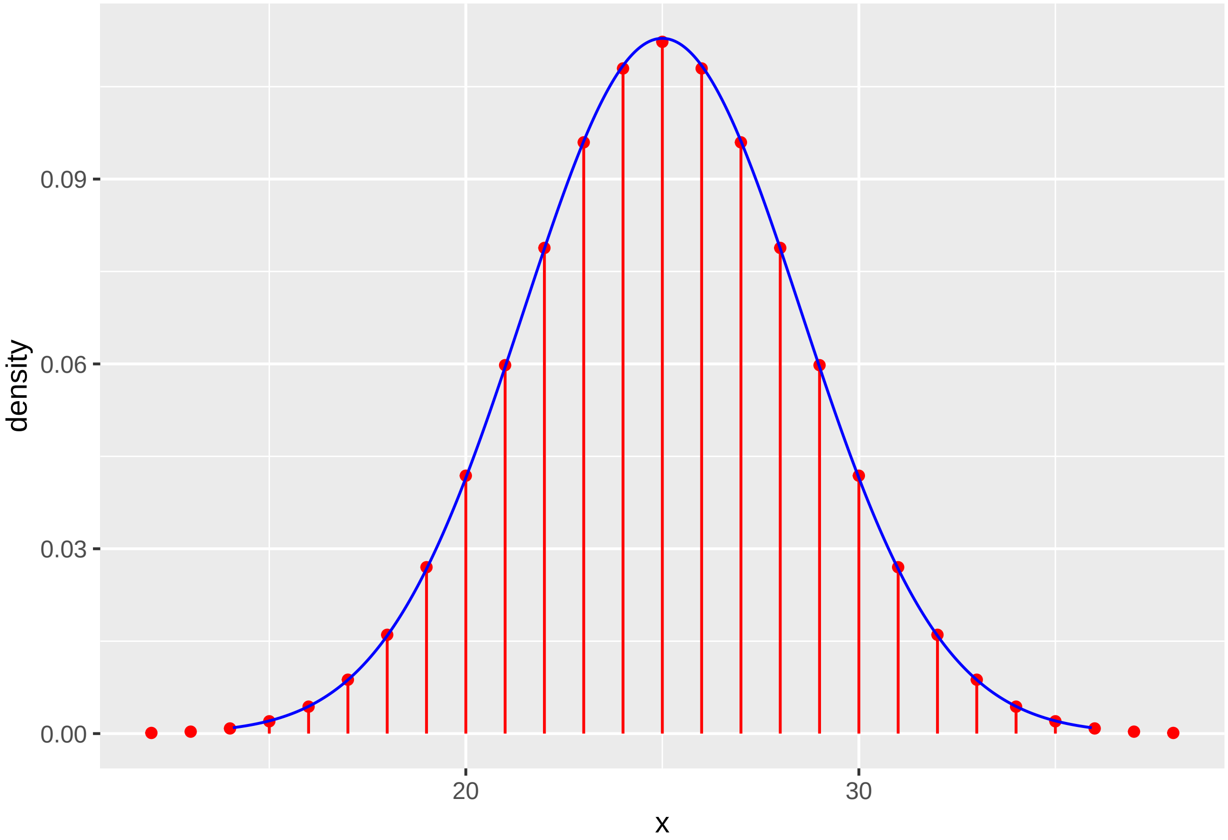
Introduction

We have introduced random variables and their distributions. Specifically, we have learned about binomial random variables and normal random variables. The family of normal distributions is central to probability and mathematical statistics and part of the reason why is because the distribution of many random variables are well-approximated by a normal distribution. This is a result (or family of results) known as the central limit theorem (CLT). In this course we do not have time to fully develop the CLT but we can pretty easily develop some intuition for this important result via computing and simulation.

An Example

A simplified version of the CLT says that if X is a binomial random variable with size n and probability of success ρ, then if n is sufficiently large, the distribution of X can be well-arppoximated by a normal distribution with mean $\mu = n\rho$ and variance $\sigma^2 = n\rho(1-\rho)$. One way to illustrate this is to plot the pmf for X together with the pdf for a normal distribution with$\mu = n\rho$ and variance $\sigma^2 = n\rho(1-\rho)$. we do this now:

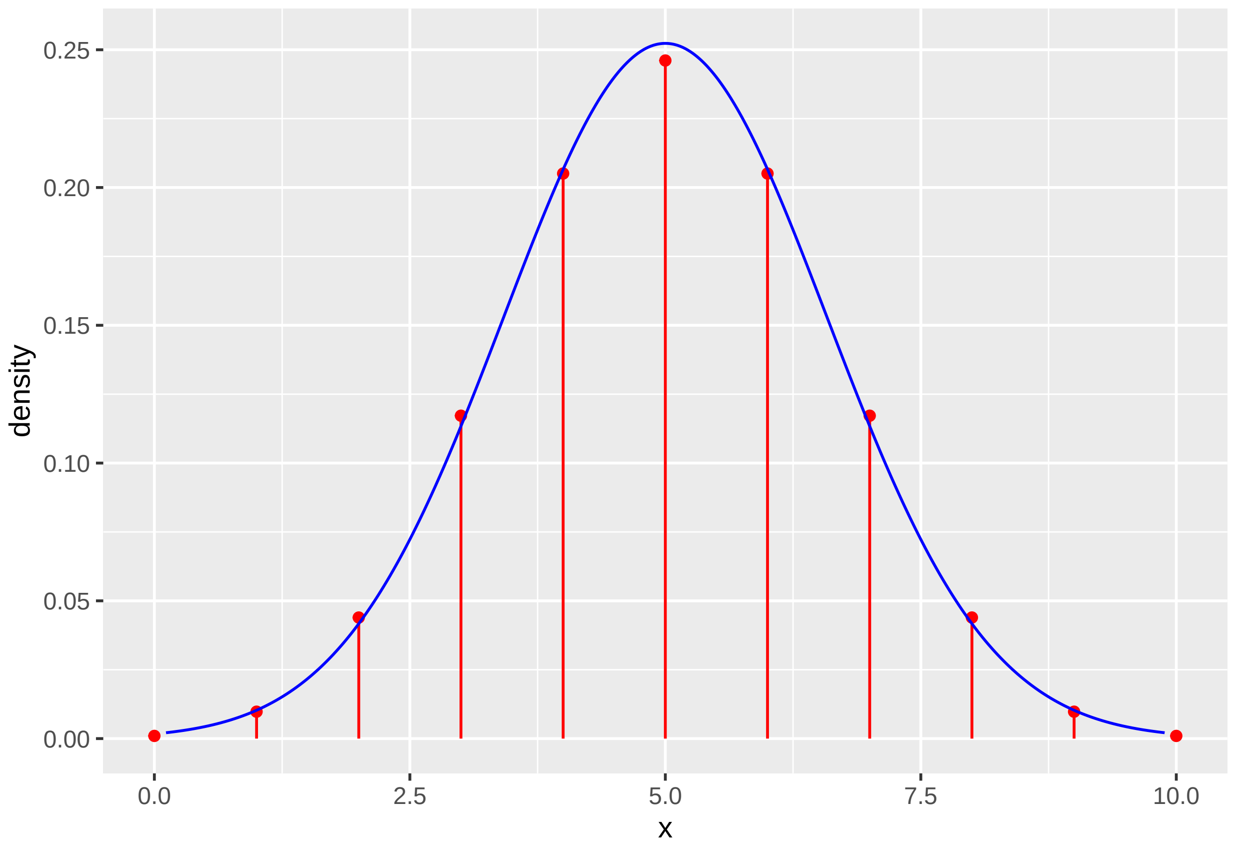
```r
n <- 50
rho <- 0.5
gf_dist("binom",params = list(size=n,prob=rho),
        color="red") %>% gf_dist("norm",params=list(mean=n*rho,sd=sqrt(n*rho*(1-rho))),
                                 color="blue")
```



The red dots are the values of the pmf for a binomial distribution while the blue curve is the values of the pdf for a normal distribution.

Notice what happens if we decrease n:

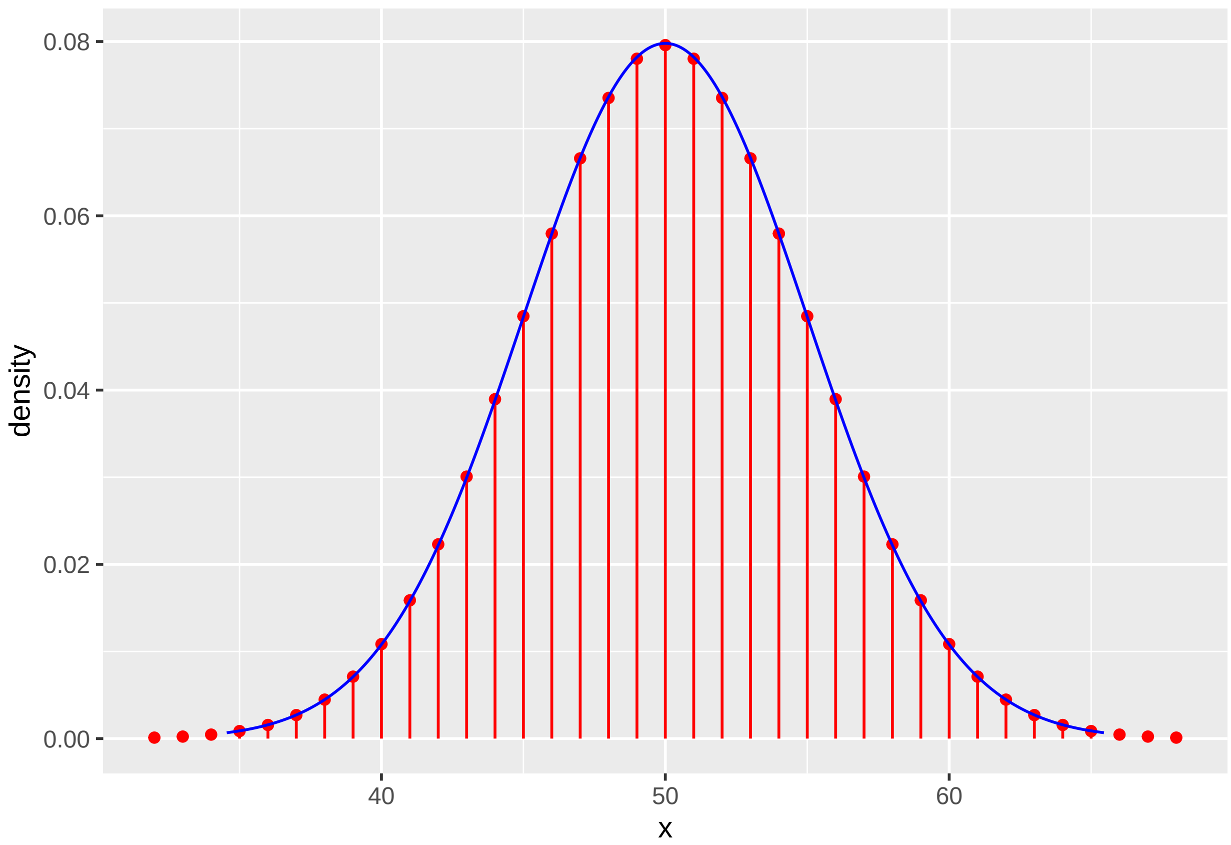
```
n <- 10
rho <- 0.5
gf_dist("binom",params = list(size=n,prob=rho),
        color="red") %>% gf_dist("norm",params=list(mean=n*rho,sd=sqrt(n*rho*(1-rho))),
                                 color="blue")
```



Observe that there is some error in that the dots do not all lie exactly on the blue curve (which is the pdf for a normal distribution). Furthermore, this error is greater for a smaller value of n than for a larger value of n.

Now, let's increase n:

```
n <- 100
rho <- 0.5
gf_dist("binom",params = list(size=n,prob=rho),
        color="red") %>% gf_dist("norm",params=list(mean=n*rho,sd=sqrt(n*rho*(1-rho))),
                             color="blue")
```



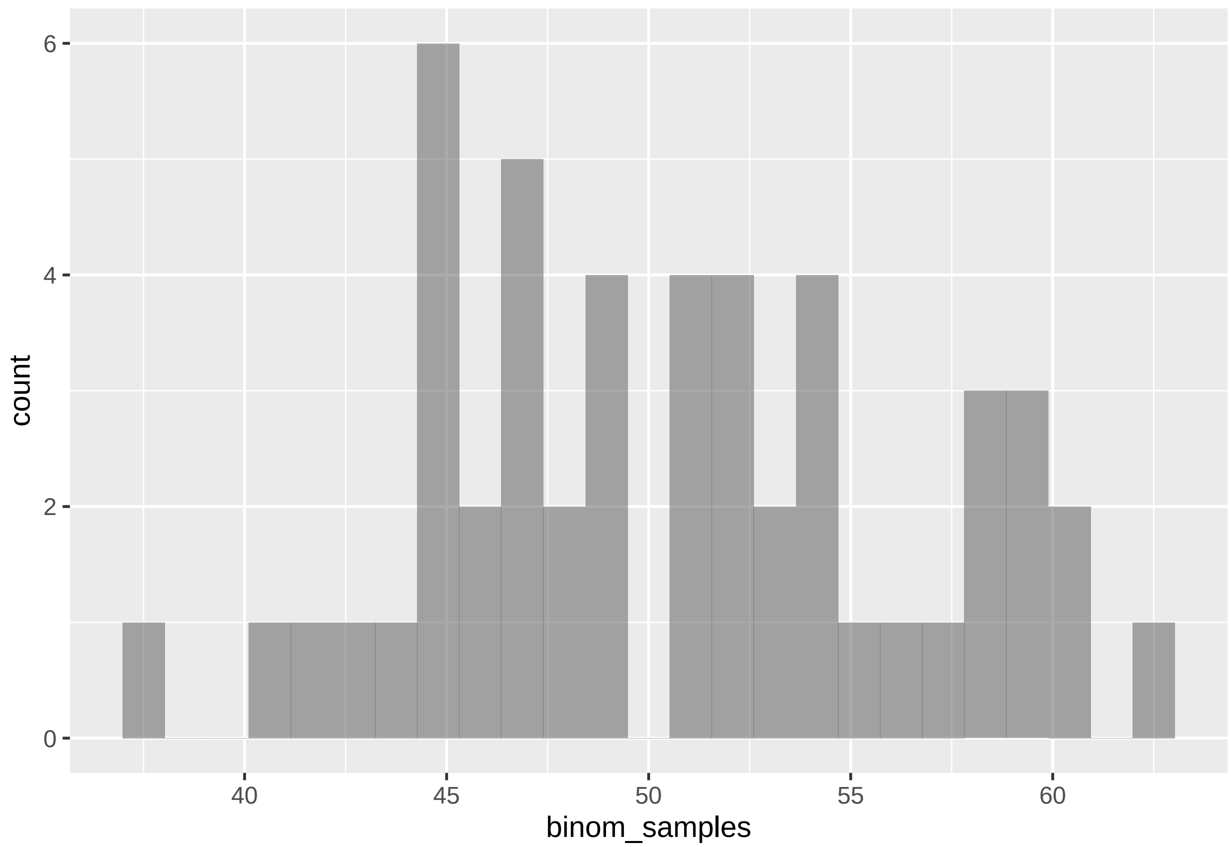
The error is decreased substantially when n is very large.

Now suppose that we take a number of samples from a binomial distribution with large n:

```
binom_samples <- rbinom(50,n,rho)
```

Let's look at a histogram for this sample data:

```
gf_histogram(~binom_samples)
```

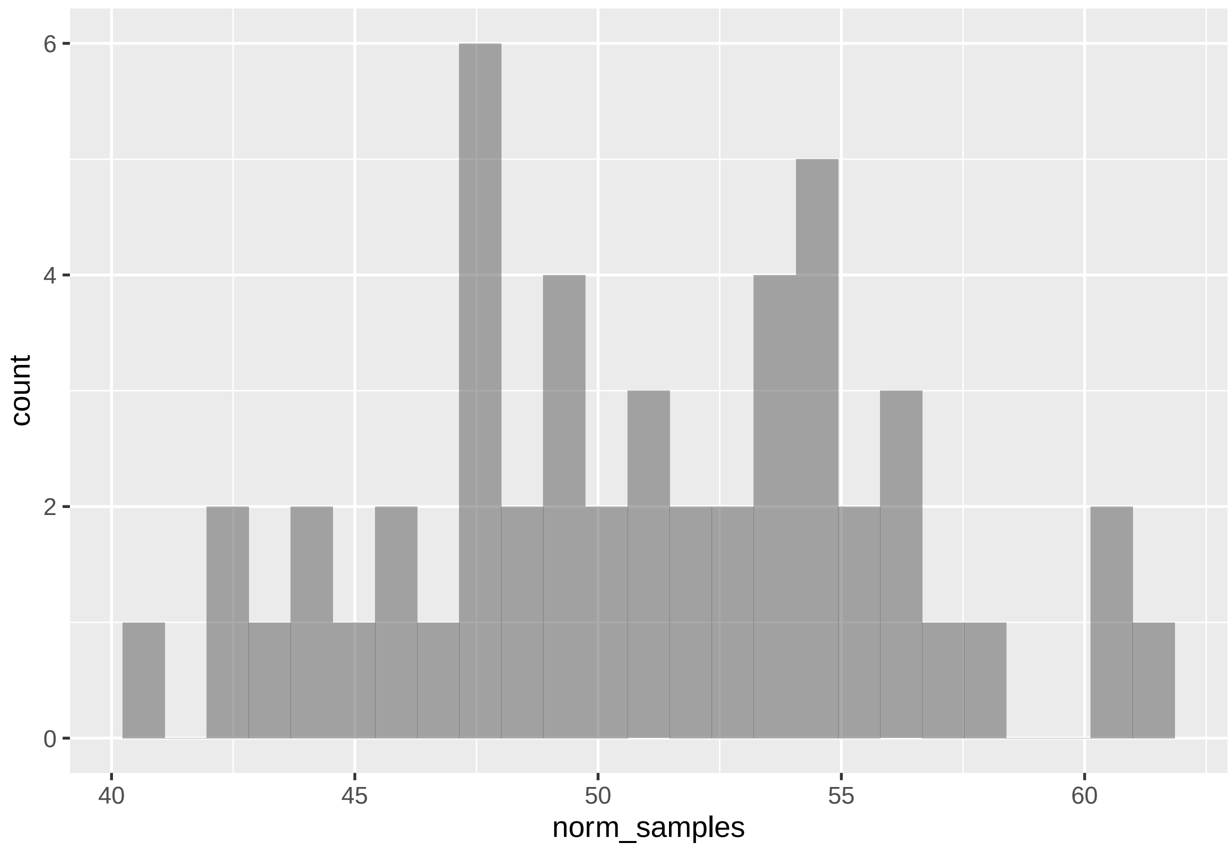


Next, let's take the same number of samples from a normal distribution with $\mu = n\rho$ and $sigma^2 = n\rho(1-\rho)$:

```
norm_samples <- rnorm(50,n*rho,sqrt(n*rho*(1-rho)))
```

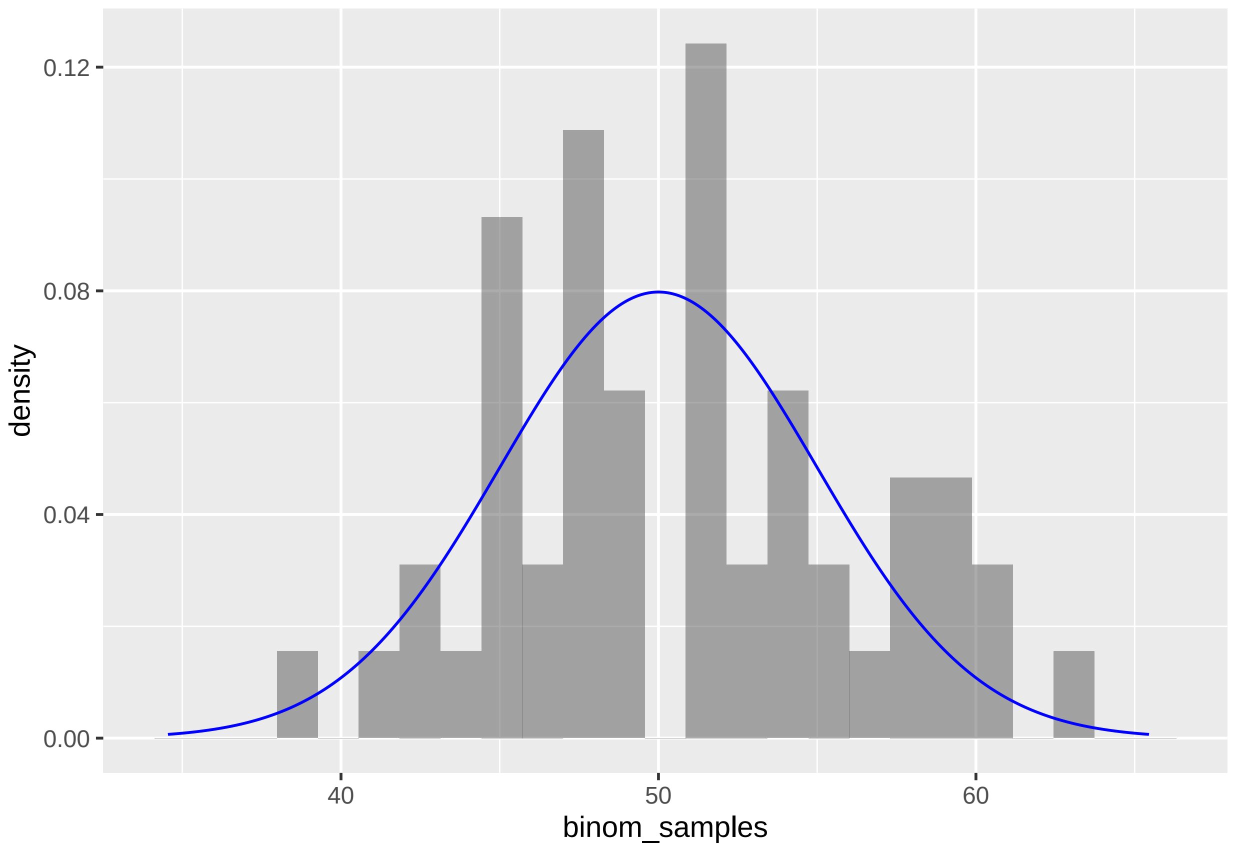
Here is the corresponding histogram:

```
gf_histogram(~norm_samples)
```

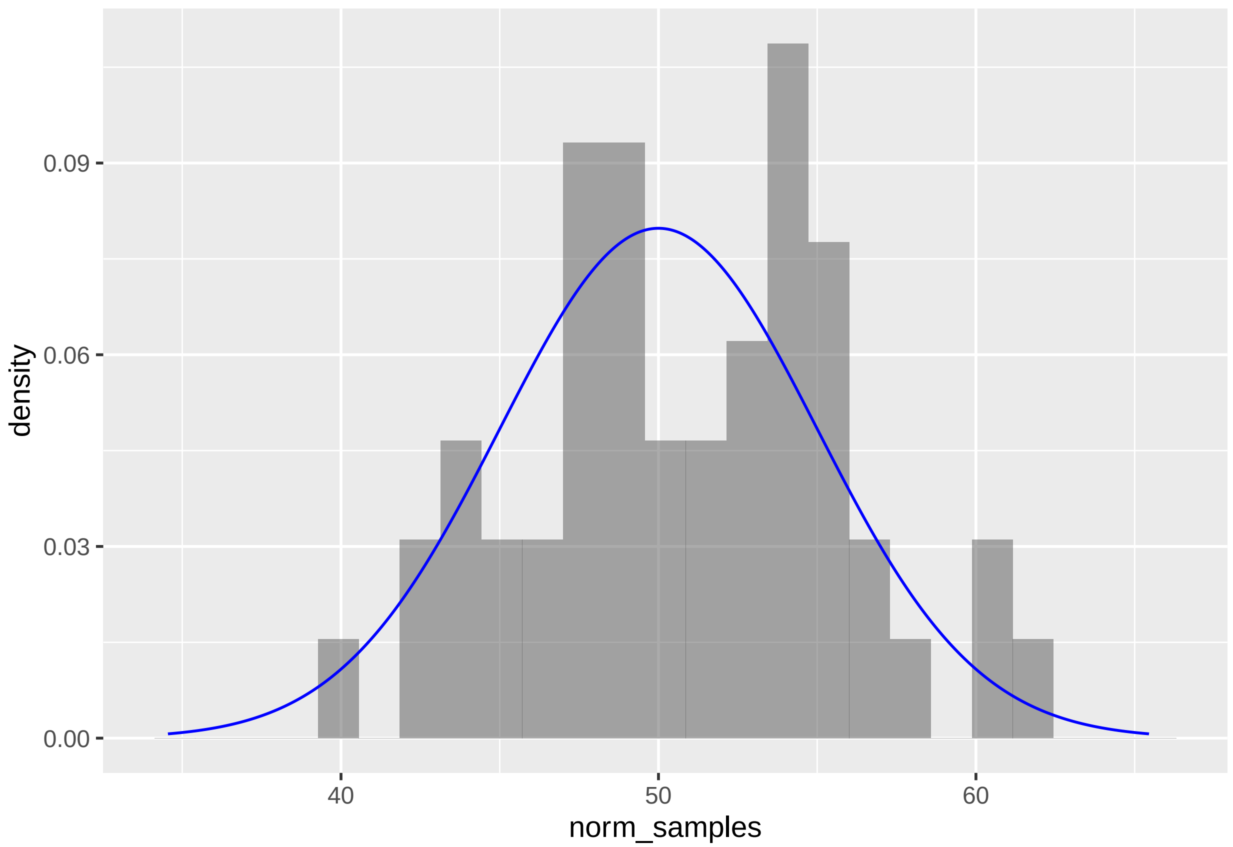


We see very similar historgrams. In fact, let's overlay a normal pdf on top of density plots for both of these data sets:

```
gf_dhistogram(~binom_samples) %>% gf_dist("norm",params=list(mean=n*rho,sd=sqrt(n*rho*(1-rho))),
                                           color="blue")
```



```
gf_dhistogram(~norm_samples) %>% gf_dist("norm",params=list(mean=n*rho,sd=sqrt(n*rho*(1-rho))),
                                   color="blue")
```



The behavior is very similar and gets better as n grows larger.

Sampling Distribution of the Mean

Suppose that we take the mean value of some sample data of size n. For example:

```
n <- 25
data_x <- runif(n,0,10)
(mean_data_x <- mean(data_x))
```

```
## [1] 5.951638
```

The previous command takes a sample of size $n = 25$ from a uniform distribution on the interval $[0, 10]$ and computes the mean of the sample data. Since we are drawing a random sample, each time we rerun this command, we will get a different value. In this code, we are sampling from a **population** (which in this example in known) and then computing the mean of the sample thereby obtaining the sampel mean. Since this value changes with each new sample, the sample mean is a random variable. We can define this formally. Let X be a random variable with some (perhaps unknown) distribution. A sample of size n from X is obtained by creating n independent copies of X which we denote as X_1, X_2, $\ldots$, X_n. Then the sample mean S_n is the random variable defined by

$S_n = \frac{X_1+X_2+\cdots+X_n}{n}$.

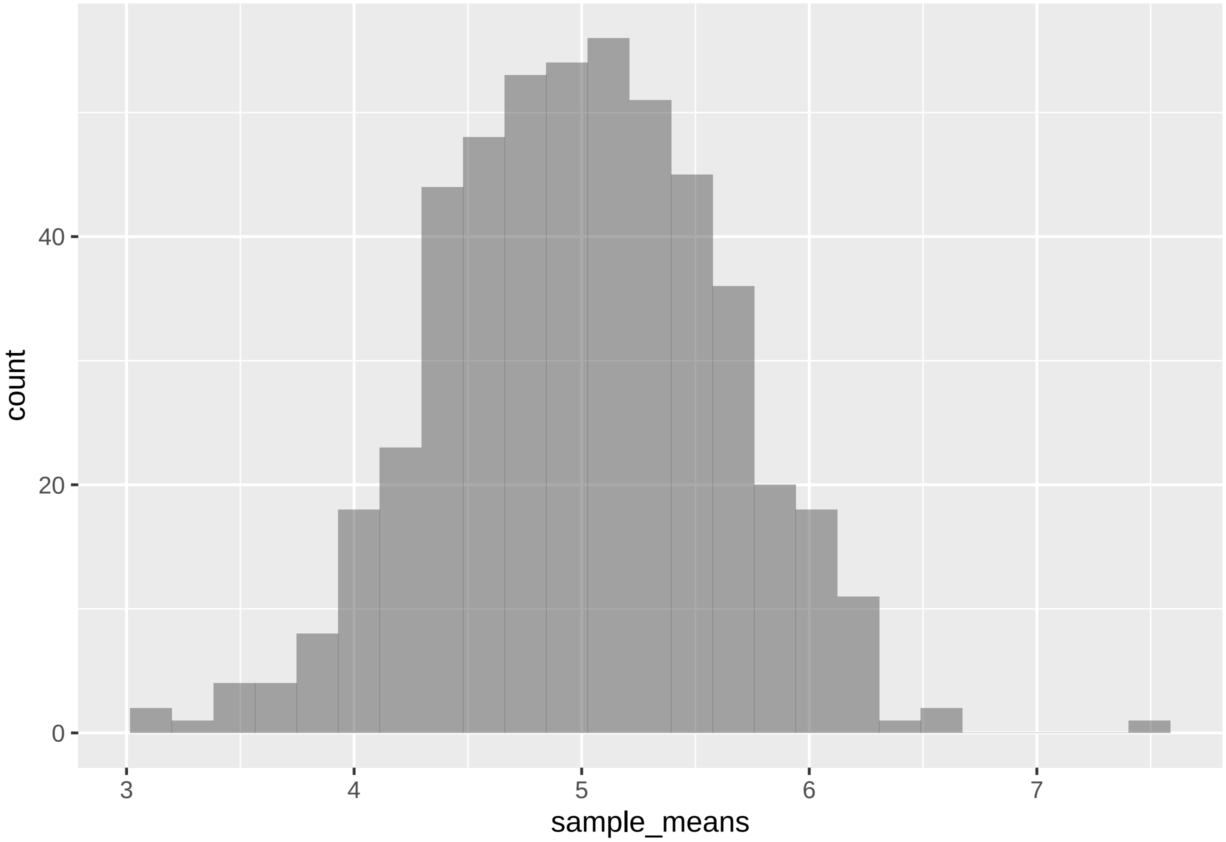
Question: What is the distribution of S_n?

In order to get a sense of the answer to this question, let's perform a simulated experiment. Let's compute the sample mean of a random sample of the same size from the same distribution some large number of times:

```r
N_repeats <- 500
sample_means <- numeric(N_repeats)
for (i in 1:N_repeats){
  sample_means[i] <- mean(runif(n,0,10))
}
```

Specifically, the previous code repeatedly draws 25 values from a uniform distribution on $[0, 10]$ and computes the sample mean for 500 times. Let's examine the histogram of the result:

```r
gf_histogram(~sample_means)
```



It is hard to say for certain but it appears that the distribution of a sample means could be normal with mean value of $\mu = 5$. Observe that if X is a uniform random variable on $[0, 10]$, then

$E[X] = \int_{-\infty}^{\infty} xf(x) = dx = \int_0^{10} \frac{x}{10}\, dx = 5.$

Furthermore, if X_1, X_2, …, X_n is a sequence of independent random samples from the same distribution, then

$E[S_n] = E\left[\frac{X_1+X_2+\cdots+X_n}{n}\right] = \frac{1}{n}E[X_1+X_2+\cdots+X_n] = \frac{1}{n}(E[X_1]+E[X_2]+\cdots+E[X_n]) = \frac{1}{n}nE[X] = E[X]$

Conclusion: If X_1, X_2, …, X_n is a sequence of independent random samples from the same distribution X with expected value $E[X]$, then $E[S_n] = E[X]$.

If we go back to our example of sampling from a uniform ditribution on $[0, 10]$, then by the work we just did, we known that expected value of the corresponding sample mean should be 5.

A more general version of the CLT says the following:

Central Limit Theorem: Let X_1, X_2, …, X_n be a sequence of independent random variables all with the same exact distribution. Let μ be the common expected value for each of these random variables. Then the distribution of the sampel mean S_n is approximately normal with mean μ whenever n (the number of samples) is sufficiently large.

It is possible to make the above statement a good deal more mathematically precise. For our purposes, here is what is most important:

1) Statistics, that is, values such as the sample mean that are computed from a sample from a population are random variables and thus have a distribution themselves. The distribution of a statistic is called the *sampling distribution* of the statistic. Of course as a random variable a statistic will also have a variance and hence a standard deviation. The standard deviation of a statistic is called the *standard error* of the statistic.

2) The statement of the CLT applies to a sample from **any** distribution and thus the sampling distribution of the mean is always approximately normal regardless of the distribution of each individual sample.

When we start our detailed study of statistics (as a field) an important part of all that we will do is centered around trying to determine or estimate the sampling distribution for some particular statistic.