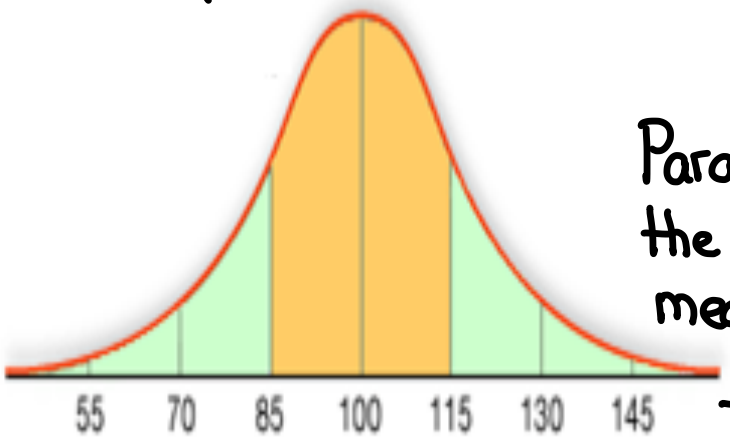
Statistics & Bootstrapping
Part 1

Population: is everyone/everything in a group of interest.

Sample: subset (smaller group) of the population.

Probability Distribution: it's a function that describes the possible values and likelihoods that a specific random variable can take.

Normal Distribution



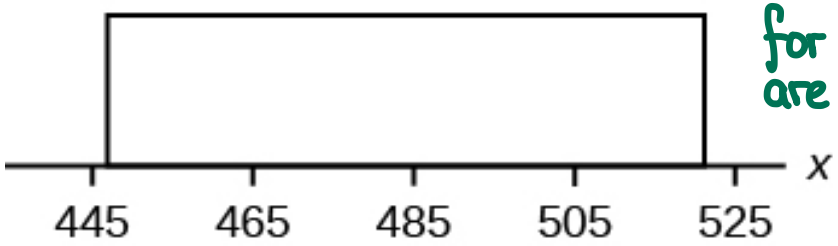
IQ

Parameters for the normal are the mean and std. dev.

Mean = 100

Standard Deviation = 15

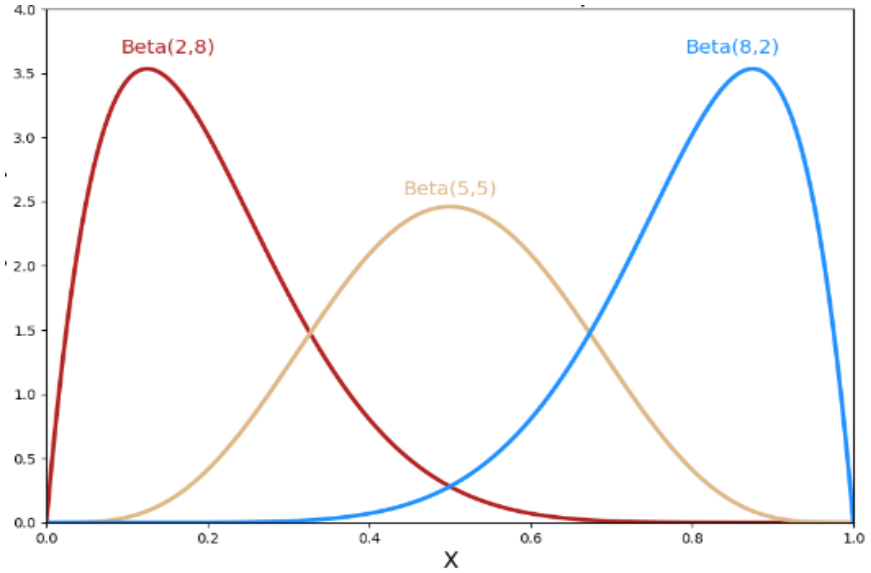
Uniform Distribution



Parameters for the uniform are the min and max.

min = 446

max = 520



Beta Distribution

shape 1 = 2

shape 2 = 8

Parameters for the Beta Distribution are shape 1 and shape 2.

Creating a Population

Normal Distribution: Mean 100, Standard deviation 15.

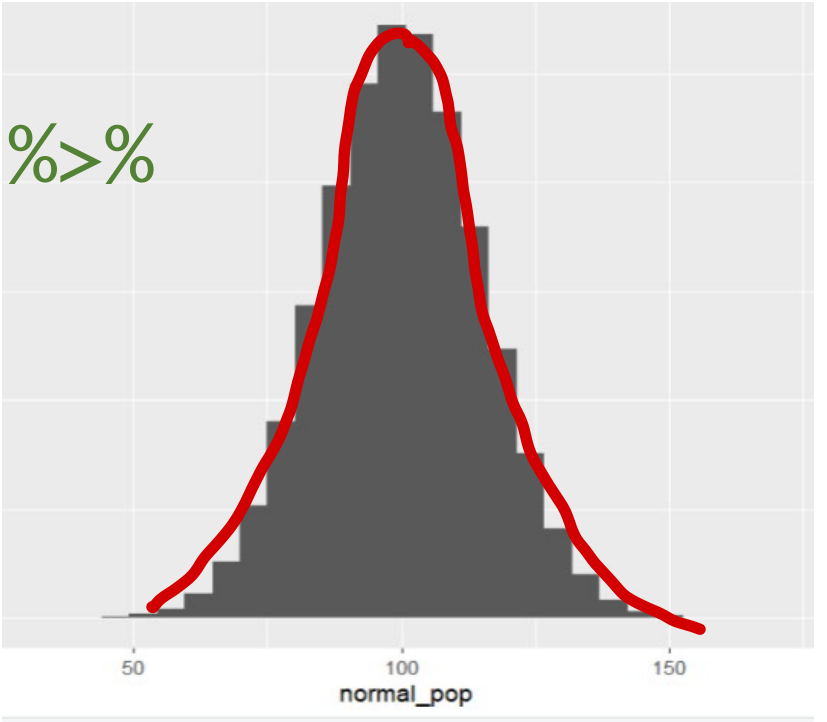
```
N <- 1000000
```

← Population number

```
normal_pop <- rnorm(N, mean = 100, sd = 15)
```

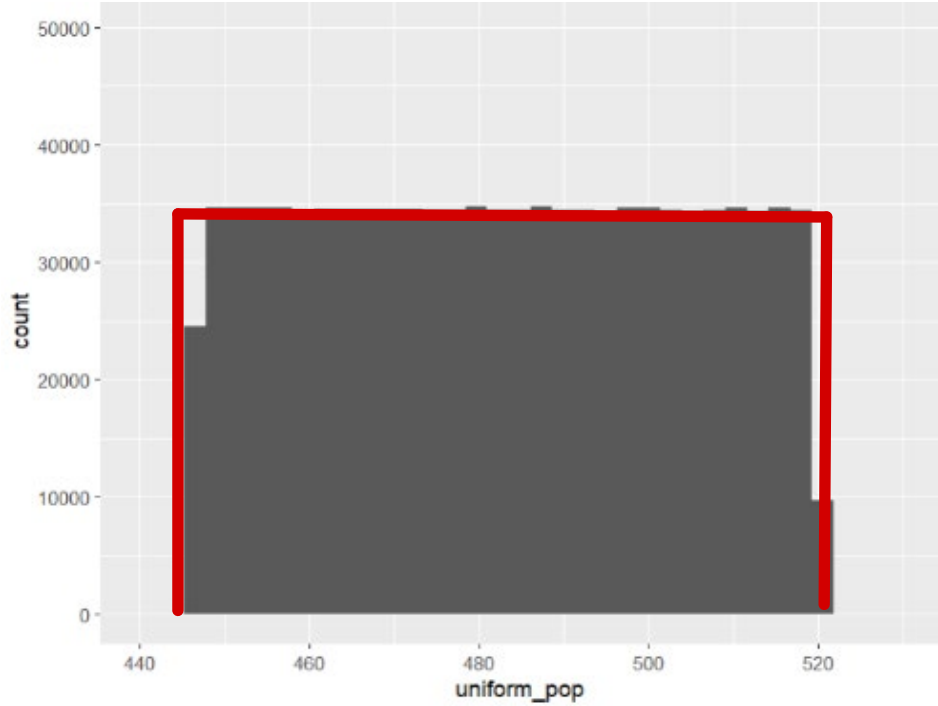
random

```
plot_1 <- data.frame(normal_pop) %>%
  ggplot(aes(normal_pop))+
  geom_histogram()
plot_1
```



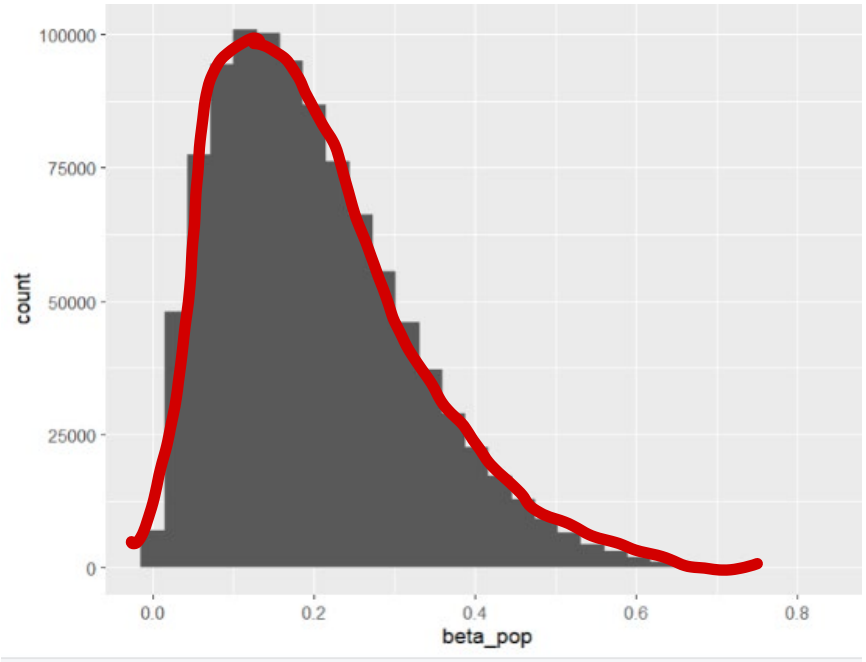
Uniform: min = 446, max = 520

```
N <- 1000000
uniform_pop <- runif(N, min = 446, max = 520)

plot_2 <- data.frame(uniform_pop) %>%
  ggplot(aes(uniform_pop))+
  geom_histogram()
plot_2
```



Beta: shape1 = 2, shape2 = 8.

```
N <- 1000000
beta_pop <- rbeta(N, shape1 = 2, shape2 = 8)

plot_3 <- data.frame(beta_pop) %>%
  ggplot(aes(beta_pop))+
  geom_histogram()
plot_3
```



Creating One Sample

We randomly sample 100 data points from the population of the normal distribution to create our sample then compute the mean.

```
n <- 100
one_sample <- sample(normal_pop, n)
one_sample_mean <- mean(one_sample)
one_sample_mean
```

n <- 100 ← Sample size

sampling from the population (normal) that we created earlier

Output: 99.18 (very close to the population)

Creating a Sampling Distribution of the Mean with n = 100

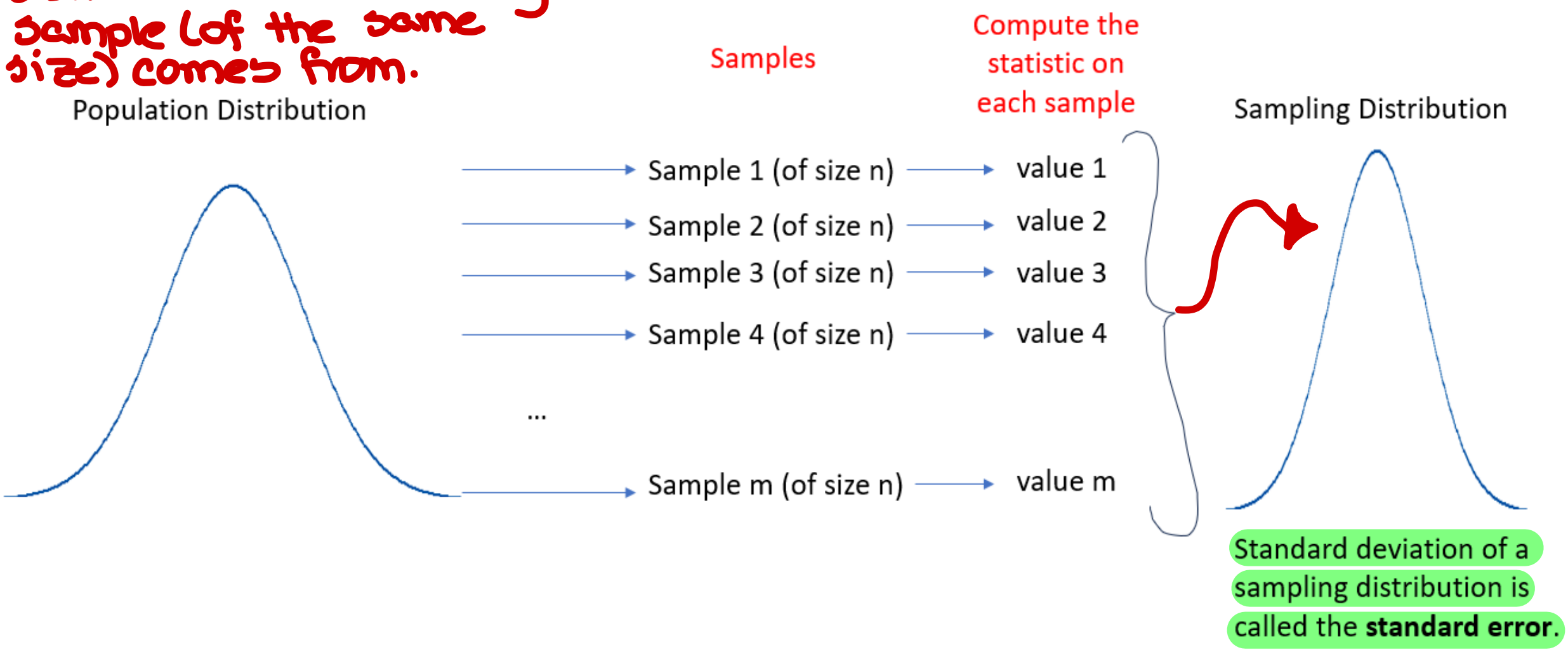
What is a "statistic"?: A statistic is a numerical value or measure that summarizes some aspect of a sample. (i.e., mean, median, sample standard deviation... etc.)

Sampling Distribution: it's a distribution of a sample statistic based on all possible simple random samples of the same size from the same population.

To create a sampling distribution we need a statistic and a sample size n.

Underlying population distribution where every sample (of the same size) comes from.

Sampling Distribution of a Statistic

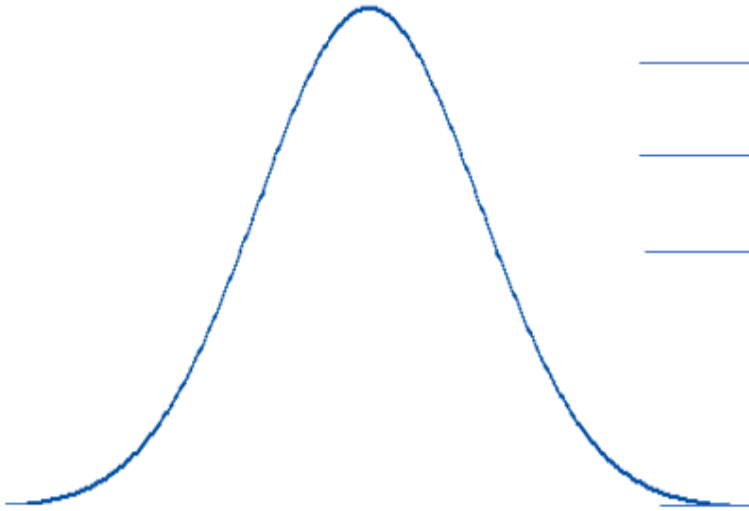


Standard deviation of a sampling distribution is called the **standard error**.

Sampling Distribution of the Mean (n=5)

Every sample is of size n=5

Population Distribution



[4, 5, 6, 7, 8] (of size 5) → 6

[1, 2, 3, 4, 5](of size 5) → 3

[4, 5, 6, 6, 7](of size 5) → 5.6

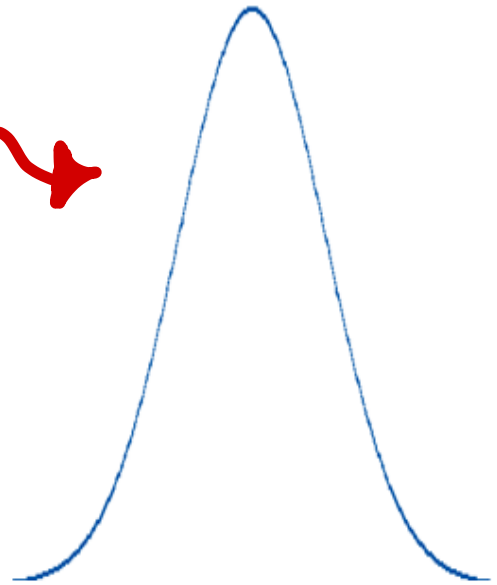
[3, 4, 5, 5, 6](of size 5) → 4.6

...

[1, 4, 8, 5, 6](of size 5) → 4.8

Compute the mean on each sample

Sampling Distribution of the mean (n = 5)



Standard deviation of a sampling distribution is called the **standard error**.

Creating multiple samples and computing the mean of each sample.

```
get_one_sample_mean  <- function(i, population_vector, n) {
  one_sample <- sample(population_vector, n)
  one_sample_mean <- mean(one_sample)
  return(one_sample_mean)
}
```

This function draws one sample of size n, then computes the mean. We combine this with the map function to create multiple samples.

Sampling Distribution of the mean with n = 100.

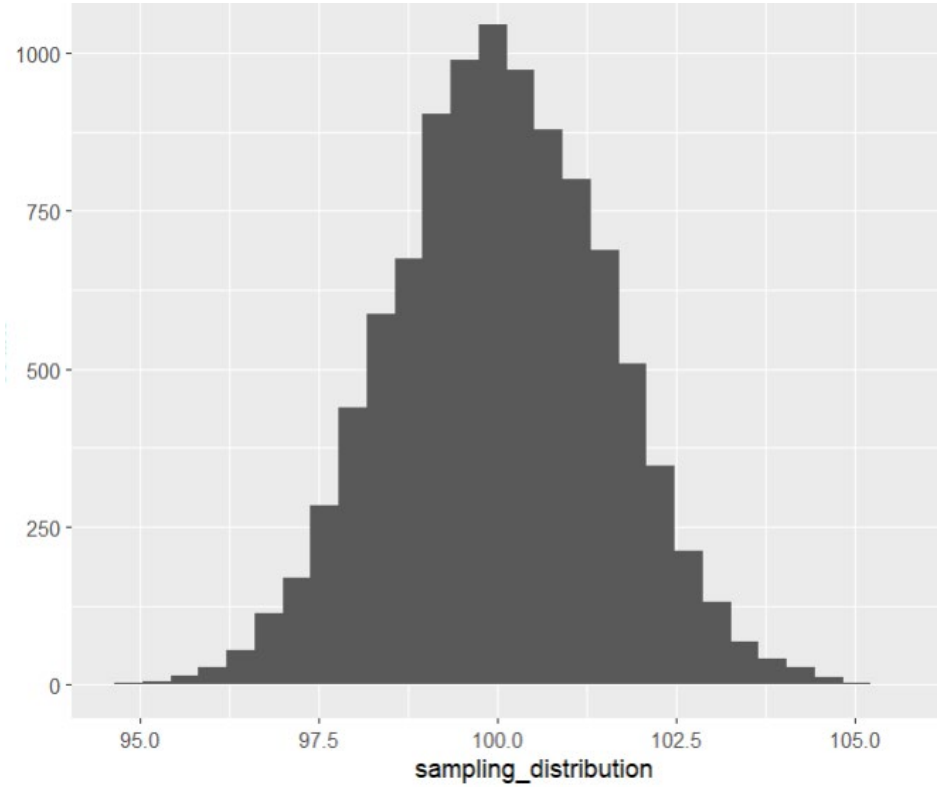
This code creates the sampling distribution

```
sampling_distribution <- map_dbl(1:10000,

get_one_sample_mean, population_vector= normal_pop, n = 100)
```

This code plots the sampling distribution

```
plot_4 <- data.frame(sampling_distribution) %>%
  ggplot(aes(sampling_distribution))+
  geom_histogram()
plot_4
```

This is the sampling distribution of the mean with a sample size of n=100



Standard Error

The standard error is the standard deviation of the sampling distribution.

```
st_error <- sd(sampling_distribution)
st_error
```

Output: 1.48898