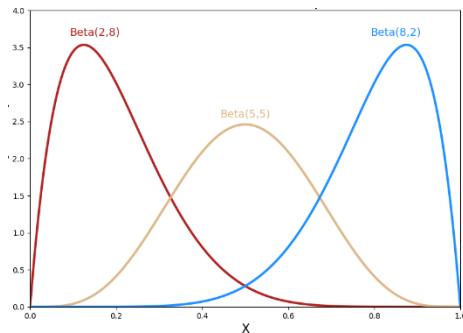
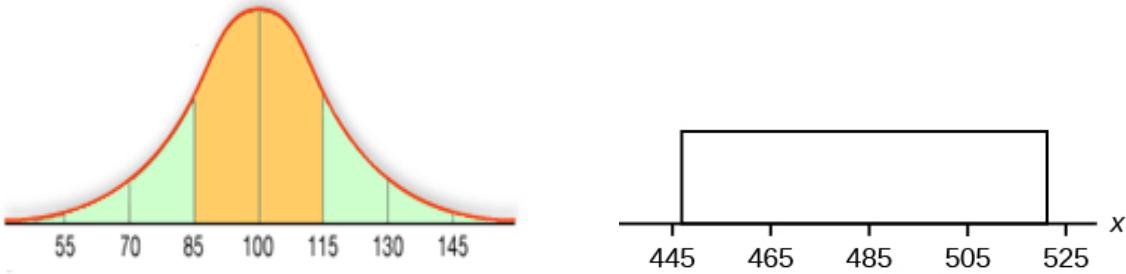


Statistics & Bootstrapping – Part 1

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.

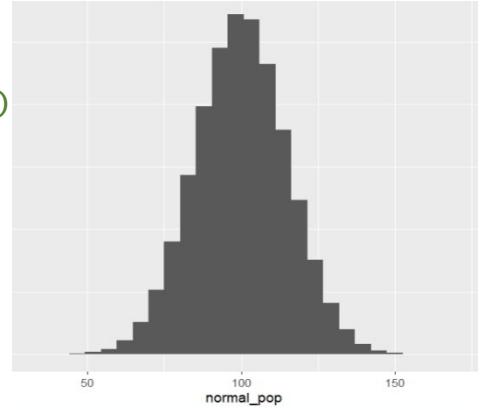


Creating a Population

Normal Distribution: Mean 100, Standard deviation 15.

```
N <- 1000000
normal_pop <- rnorm(N, mean = 100, sd = 15)

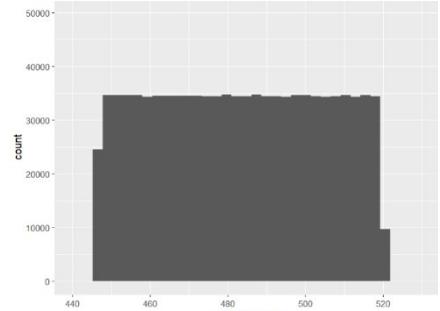
plot_1 <- data.frame(normal_pop) %>%
  ggplot(aes(normal_pop)) +
  geom_histogram()
plot_1
```



Uniform: min = 446, max = 520

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

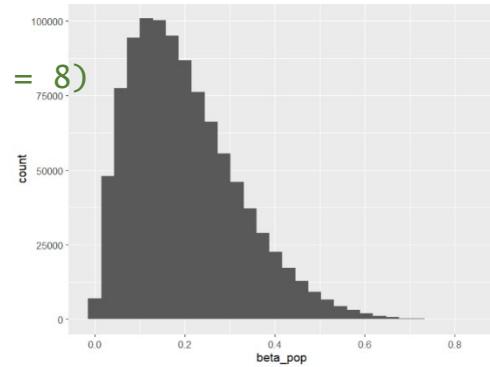
plot_2 <- data.frame(uniform_pop) %>%
  ggplot(aes(uniform_pop)) +
  geom_histogram()
plot_2
```



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
```

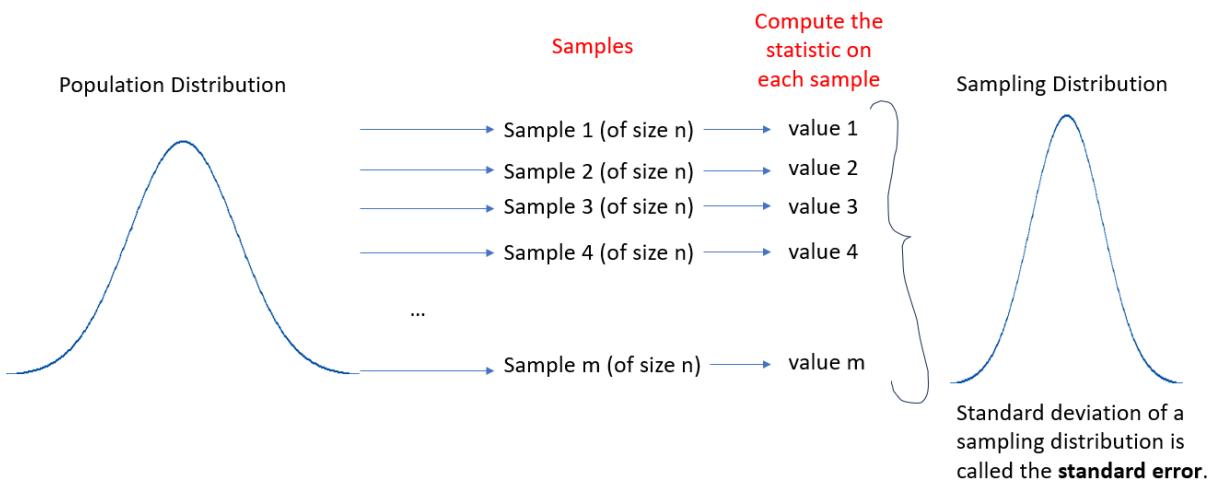
Output:

Creating a Sampling Distribution of the Mean

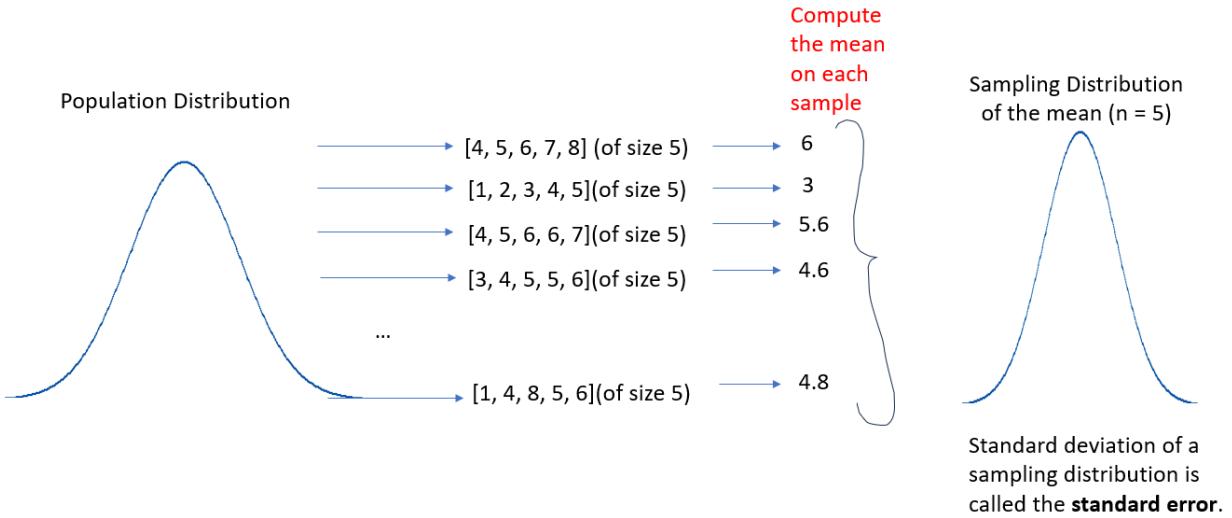
What is a "statistic"?: A statistic is a numerical value or measure that summarizes some aspect of a sample. (i.e., mean, median, sample standard deviation... etc.)

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.

Sampling Distribution of a Statistic



Sampling Distribution of the Mean (n=5)



Creating multiple samples and computing the mean of each sample.

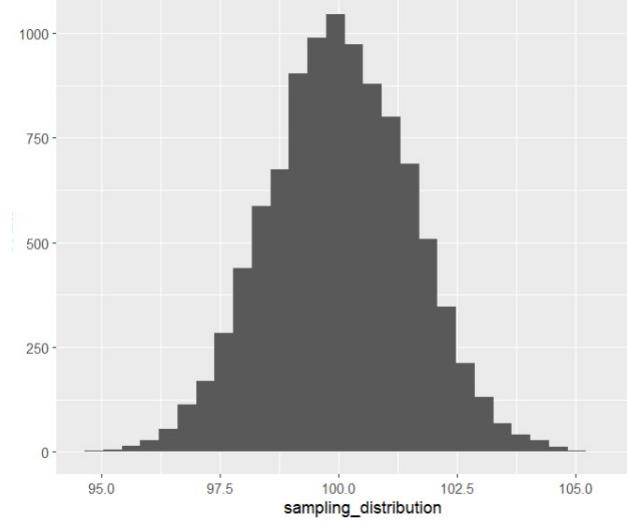
```
get_one_sample_mean <- function(i, population_vector, n) {
  one_sample <- sample(population_vector, n)
  one_sample_mean <- mean(one_sample)
  return(one_sample_mean)
}
```

Sampling Distribution of the mean with n = 100.

```
sampling_distribution <- map_db1(1:10000, get_one_sample_mean,
population_vector= normal_pop, n = 100)
```

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

plot_4



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

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

Output: