Distrbution of Sample Means

John Smith

August 13th, 2014

Let’s find sampling means from a population. If we pull multiple samples from a population, how will the means of the samples be distributed? Will the sampling means follow the distribution of the population?

We start by creating a function to create and plot a skewed population. The function also returns the full population it created.

```
skewed_population_curve <- function() {
    # Create left-skewed population distribution
    a <- rnorm(n = 60000, mean = 600, sd = 50)
    b <- rnorm(n = 30000, mean = 725, sd = 50)
    c <- rnorm(n = 20000, mean = 850, sd = 50)

    pop <- c(a, b, c)
    plot(density(pop), col = "blue", lwd = 2,
        xlim = c(400, 1000))
    abline(v = mean(pop))
    abline(v = median(pop), lty = "dashed")

    return(pop)
}
```

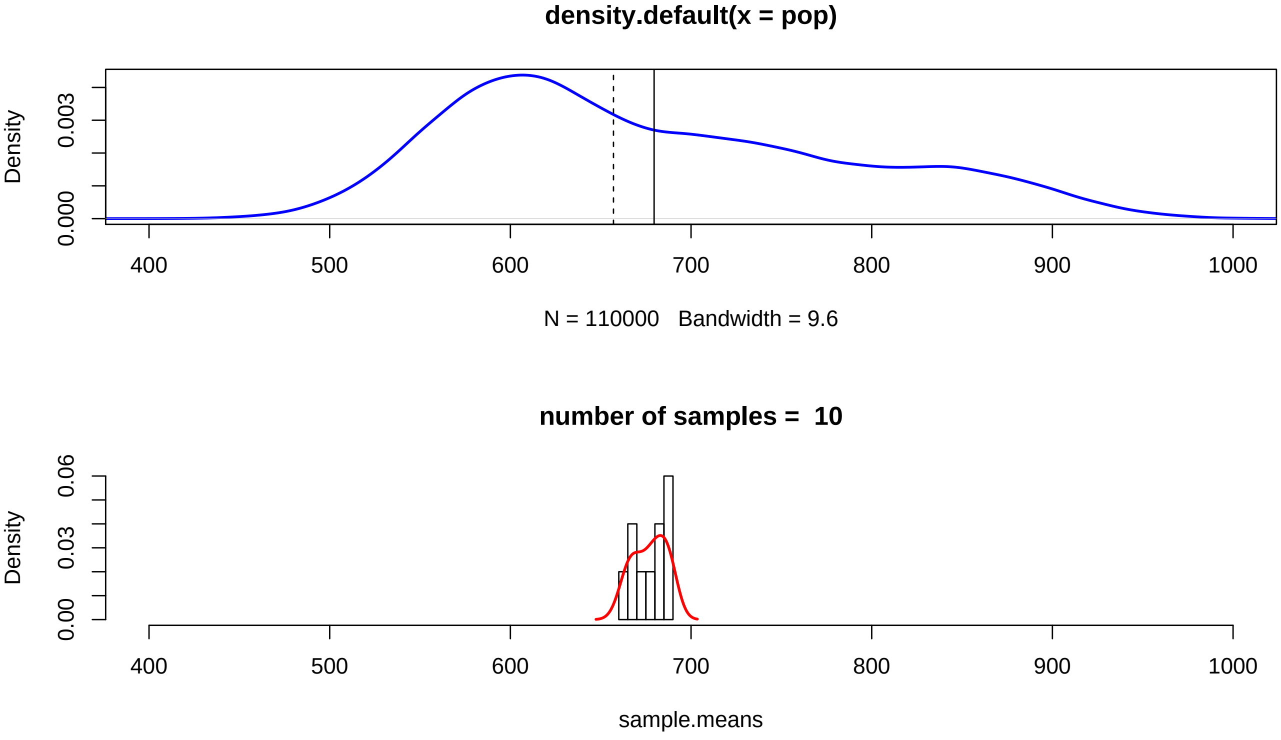
Another function draws multiple samples from a given population. This function plots a histogram of the means of the samples, and return the means as a list

```
sample_histogram <- function(population, num.samples) {
    sample.means <- c(0)
    for (i in 1:num.samples) {
        sample.means[i] <- mean(sample(population,
            100))
    }
    hist(sample.means, prob = TRUE, xlim = c(400,
        1000), main = paste("number of samples = ",
        num.samples))
    lines(density(sample.means), col = "red",
        lwd = 2)
    return(sample.means)
}
```

Now let's compare the distribution of a population to the distribution of sampling means. We can plot those sampling means as a histogram!

We will first create a population (we made a function for that!) and then find the means of 10 samples pulled from that population (we made a function for that too!).

```
par(mfrow = c(2, 1))
population <- skewed_population_curve()
means <- sample_histogram(population, 10)
```



```
par(mfrow = c(1, 1))
```

The 10 means we put in the histogram have their own distribution! This means they have their own 'central tendency' and 'disperson'. What do you think the grand mean of the means is? What do you think the dispersion of the means tells us?

```
means
```

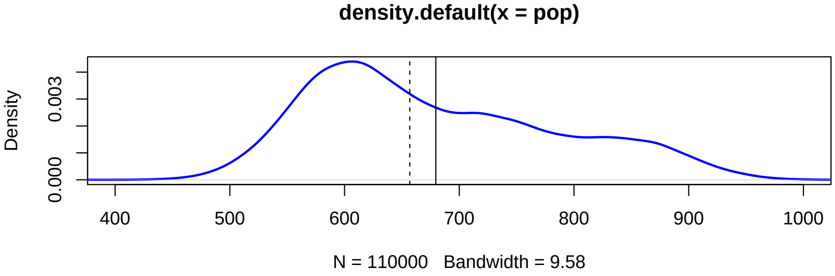
```
##  [1] 667.6 676.9 663.0 686.7 672.7 687.5
##  [7] 680.8 680.5 666.8 687.8
```

Butx can see the means of the samples do not appear to follow a simple distribution. This could be because we did not pull enough samples.

```
set.seed(-1759254325)
```

Let's pull samples from a population one more time.

```
population <- skewed_population_curve()
```



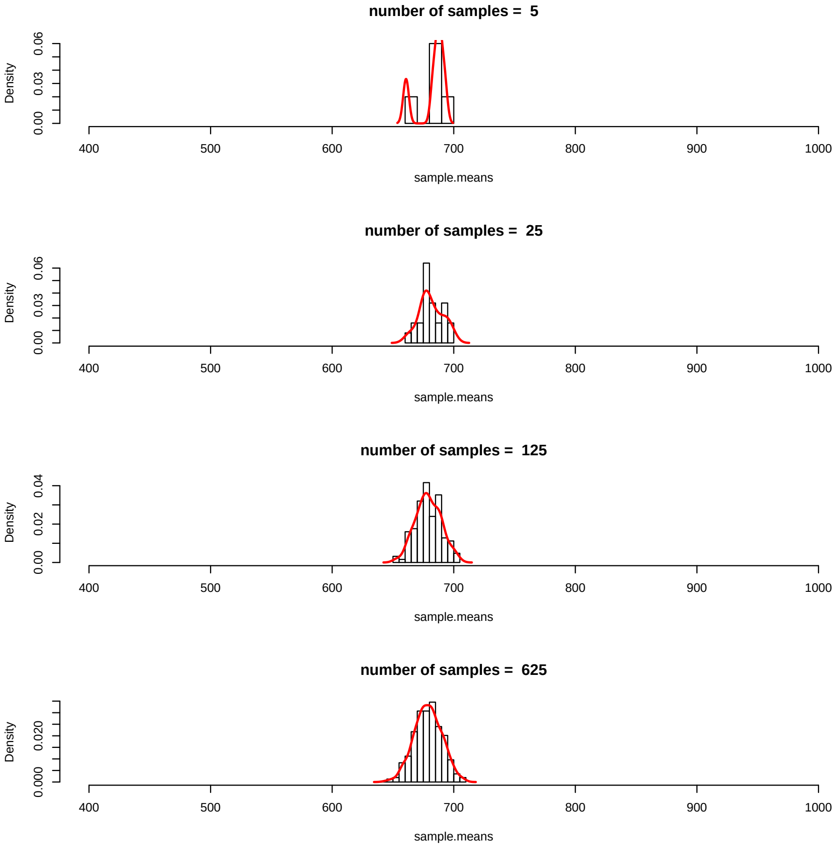
This time, we will pull 5, 25, 125, and 625 samples and plot their means.

We can see that the means of the samples begin to follow a normal distribution, even though our original population was not normally distributed! This is the central limit theorem at work.

We also begin to see why the normal distribution is so important. It means we can say something about the central tendency and dispersion of a statistic (like the mean), without having to know anything about the distribution of the original population.

This makes the central limit theorem one of the most important and useful principles in statistical modeling.

```
sample_size <- 5^seq(1:4)
par(mfrow = c(4, 1))
for (size in sample_size) {
    sample_histogram(population, size)
}
```



```
par(mfrow = c(1, 1))
```