Weekly Tutorial Activity 9

Sean Leggett - BDA201, Winter 2020
March 31, 2020

Option 2

Answer the question below in your own words

What is Central Limit Theorem. Why is it useful in Statistics ? Explain with examples.

Answer

The Central Limit Theorem (CLT) holds that if you sample any given population enough times, the means of the samples will approximate the mean of the population. This means that even if we do not, or cannot, know the mean of the population, we can approximate it with a high degree of confidence by repeating the sample process. Typically, to be meaningful, the sample size must be greater than 30. This becomes a powerful tool in being able to predict characteristics of a population simply by sampling.

One of the most common examples of this is polling for any variety of reasons but often for political data. While a sample size beyond 30, say 40 samples, is helpful, it is not enough in a one-off sample to predict a characteristic of a population. One woudl need to repeat a sample of 40 observations many times (thousands) and average out the findings to use this data effectively. When a high degree of confidence is required, a sample of ~1,000 is acceptable as it yields a degree of confidence of being accurate 97 times out of 100. Substantially, exponentially, larger samples are required to move beyond the 97% confidence level to achieve 98% or 99% confidence. The CLT works in a complementary fashion with the Law of Large Numbers.

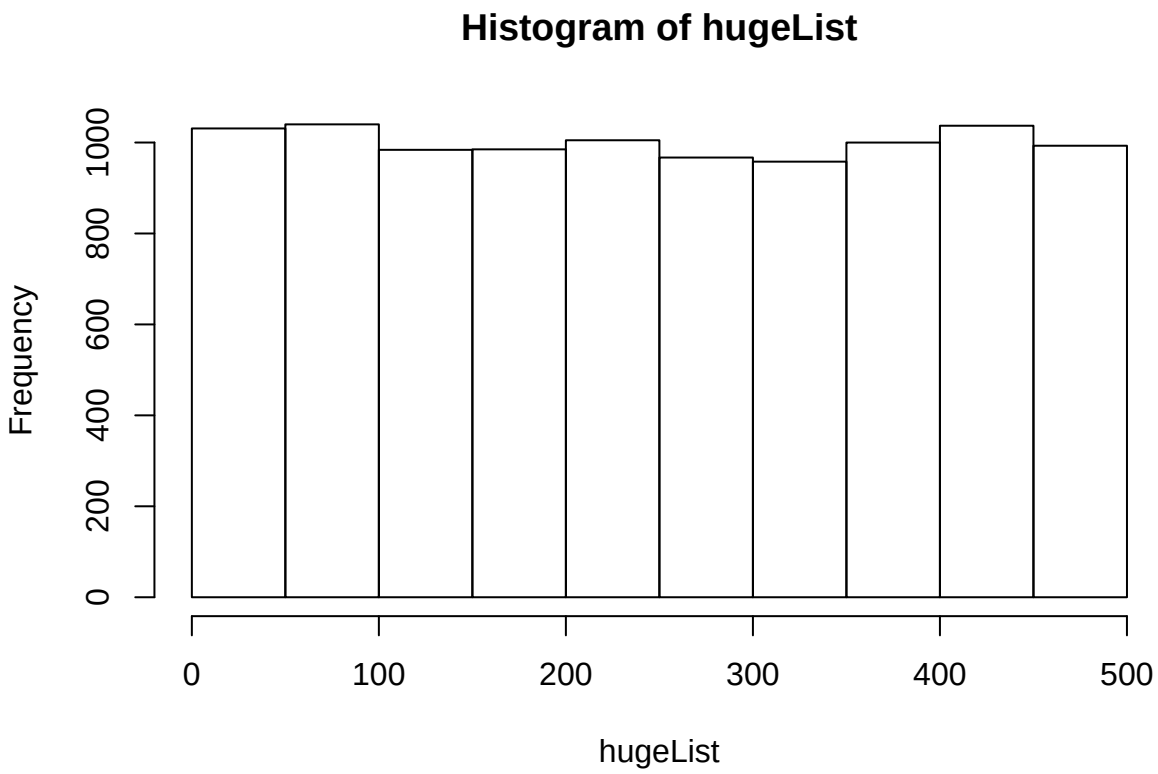
Another interesting aspect of the CLT is that the means of the samples will eventually yield a normal distribution. Even if the original variable is not normally distributed. An interesting fact that we attempt to illustrate below.

Our examples below are comprised of the following:

We built an array with 10,000 randomly chosen numbers from 1 to 500.
We display a histogram of those 10,000 numbers and we are not particularly convinced of their randomness as a clearly uniform distribution emerges, even with replacement used in the code. However, it serves it's purpose as the distribution is clearly not normal.
We took 4 samples from this list. The first sample has 5 observations, the second 10, third 100 and the last 1,000.
In an attempt to display the effect of sample size, four histograms are produced. One for each of the samples. We can see that the fourth (1,000 sample size) approximates the shape of the original population of 10,000. As the sample sizes grow, each histogram more resembles the population.
Finally, we collected the means of our sample and our population and plotted. This is to illustrate a) the means when taken from small and large samples and also, b) the closer the values become as the sample size increases to 100 and 1,000.

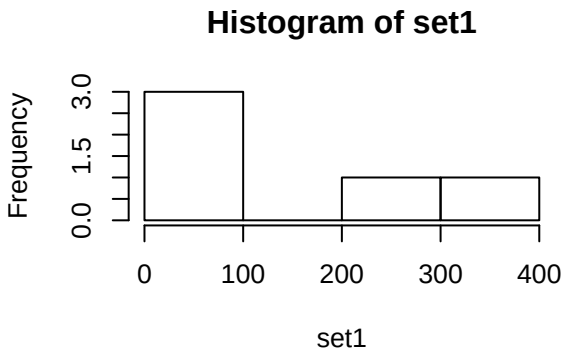
While four samples is by not means a proof, the illustrative value of the effects of the CLT is convincing.

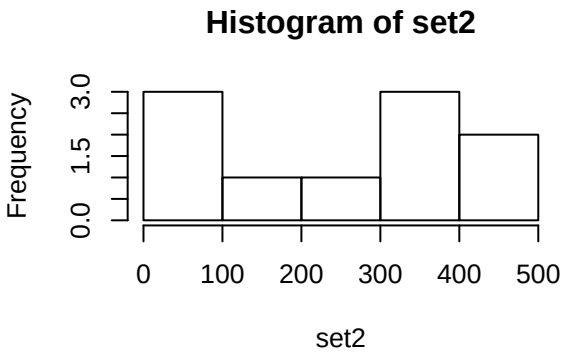
```
## we make a large dataset of integers between 1 and 500. 10,000 in fact.
hugeList <- sample(1:500, 10000, replace = TRUE)
hist(hugeList)
```

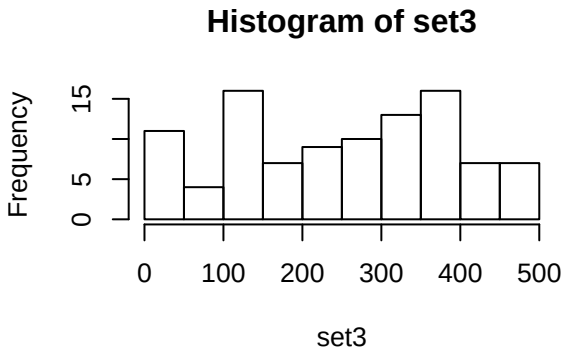


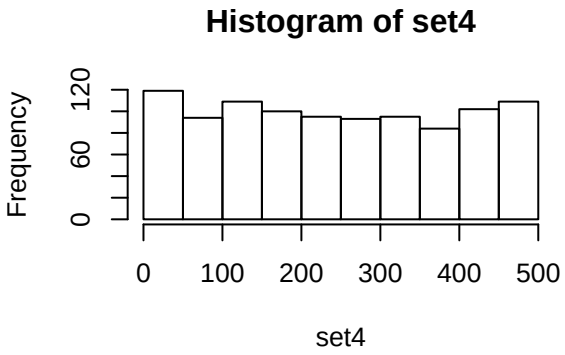

```
## now we take samples of various sizes from hugeList
set1 <- sample(hugeList, 5, replace = TRUE)
set2 <- sample(hugeList, 10, replace = TRUE)
set3 <- sample(hugeList, 100, replace = TRUE)
set4 <- sample(hugeList, 1000, replace = TRUE)

par(mfrow = c(2,2))
hist(set1)
hist(set2)
hist(set3)
hist(set4)
```

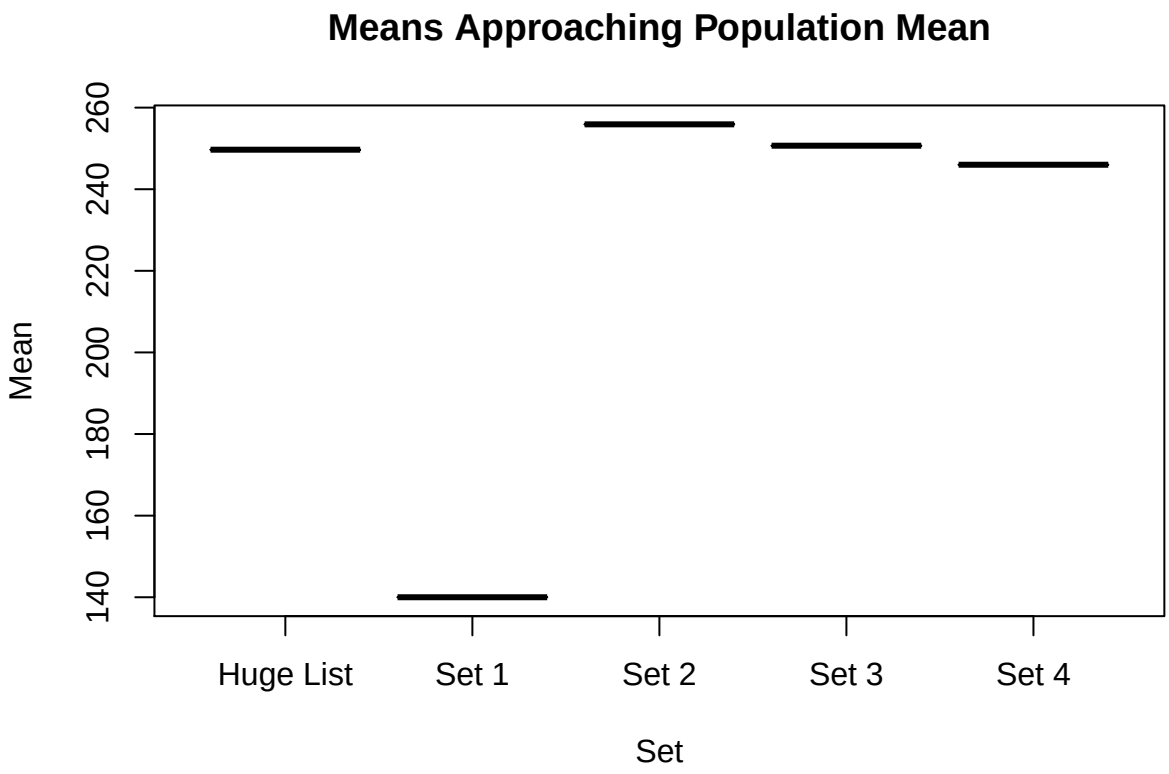









```
meanhuge <- mean(hugeList)
mean1 <- mean(set1)
mean2 <- mean(set2)
mean3 <- mean(set3)
mean4 <- mean(set4)
x <- data.frame("Set" = c("Set 1", "Set 2", "Set 3", "Set 4", "Huge List"),
                "Value" = c(mean1, mean2, mean3, mean4, meanhuge))
plot(x$Set, x$Value,
   main = "Means Approaching Population Mean",
   xlab = "Set",
   ylab = "Mean")
```


Means Approaching Population Mean
Mean
140
160
180
200
220
240
260
Huge List
Set 1
Set 2
Set 3
Set 4
Set