Tutorial: Central Limit Theorem in Stata

We examine BMI at baseline using the Framingham cohort as our reference population. Specifically, we can think of the Framingham population as 'the population of interest' and consider sampling from this population to examine how statistics behave in samples from a population where we know about everyone.

1. Calculate the mean μ standard deviation σ BMI in the Framingham dataset at baseline.

   ```
   . summarize bmi1
   ```

   $\mu = 25.8$ and $\sigma = 4.1$.

2. Take a sample of size 20 from the Framingham dataset. Calculate a sample mean BMI at baseline, $\bar{x}_1$. Then take a second sample from the same population and calculate the sample mean, $\bar{x}_2$. Would you expect $\bar{x}_1$ and $\bar{x}_2$ to be exactly the same? Why or why not?

   ```
   use "fhs.dta", clear
   drop if bmi1 == .
   keep bmi1

   preserve
   sample 20, count
   mean bmi1

   restore
   preserve
   sample 20, count
   mean bmi1
   ```

   We don't expect $\bar{x}_1$ and $\bar{x}_2$ to be exactly the same, because the mean has some stochastic variability.

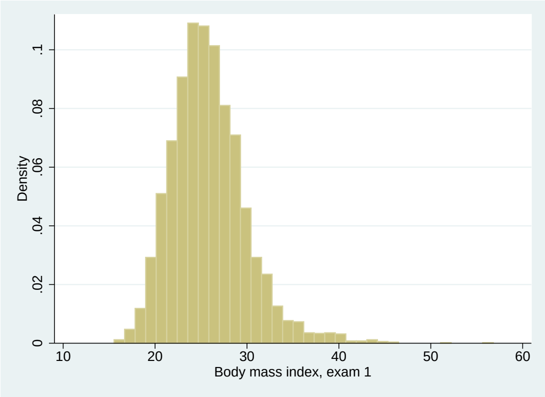
3. Repeat this exercise, but with a sample size of 100. Are $\bar{x}_1$ and $\bar{x}_2$ closer together than those from the samples of size 20? Are $\bar{x}_1$ and $\bar{x}_2$ always going to be closer together using a sample size of 100 versus 20?

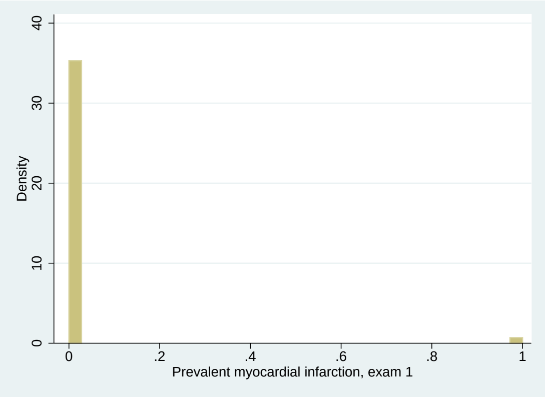
   ```
   restore
   preserve
   sample 100, count
   mean bmi1

   restore
   preserve
   sample 100, count
   mean bmi1
   ```

In my sample, the values of the sample mean are closer with the larger sample. This will usually, but not always, be true.

4. Compare histograms of BMI at baseline and prevalent MI at baseline. Would the central limit theorem apply to the binary indicator prevalent MI at baseline?





Yes, but the more skewed a distribution is, the larger sample size we need to collect before the CLT “kicks in”.