Douglas Brown
CMST Challenge Project
Central Limit Theorem Simulation

USING SIMULATION

TO VALIDATE THE

CENTRAL LIMIT THEOREM

IN STATISTICS

Using Simulation to Validate the Central Limit Theorem in Statistics

Background on the Concept of Sampling Distributions

Understanding the concept of sampling distributions is fundamental to this project. To assist the reader, an explanation is provided here. Within a population, let's say income is the attribute of interest. A sample of size 500 is taken from the population, and the mean of the income is determined. If another sample is taken, a different mean would probably result. Imagine that this sampling process is repeated over and over, so that *every possible sample* of size 500 is collected from the population. If the means for each of these samples were determined, these means would comprise the *sampling distribution of the mean income*. The sampling distribution does not represent the distribution of a single sample, but the distribution of a statistic (such as the mean) for every possible sample of a given size from the population.

Abstract

A major goal of statistics is statistical inference, the ability to make predictions or claims about a population based upon a sample taken from that population. Typical approaches within statistical inference include confidence intervals and significance tests.

There is a large body of mathematics which support the use of statistical inference techniques. The population itself is generally not known (otherwise there would be no need for sampling or statistical inference). However, if the distribution of the population was normal (bell-shaped), it would not be too surprising to find out that the sampling distribution itself is normal. The importance of having a sampling distribution which is

normal cannot be overstated: so much is known mathematically about normal distributions that we can make claims about them.

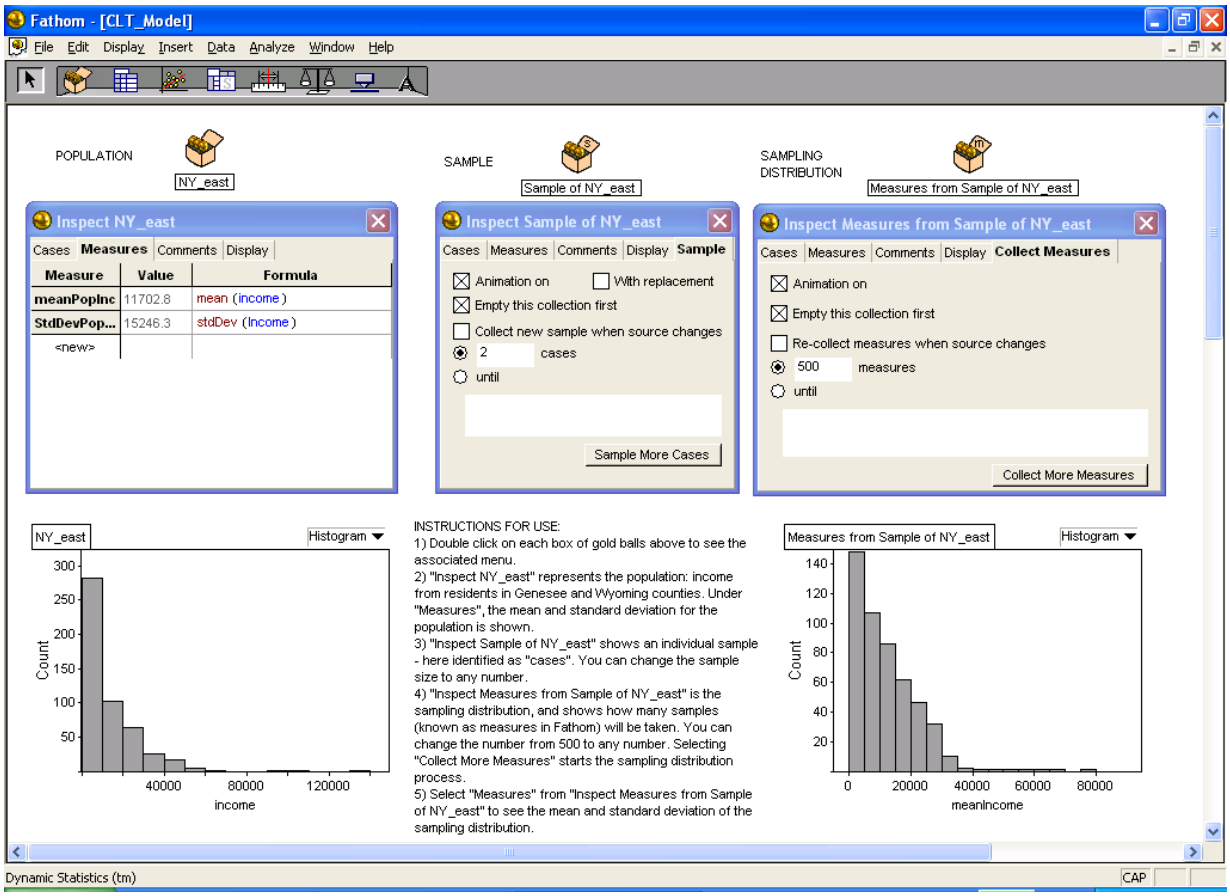
But what happens if the population itself is not normal? Can we still expect the sampling distribution to be approximately normal? Surprisingly, the Central Limit Theorem says that we can, IF the sample size is "large enough". Furthermore, the Central Limit Theorem states that the sampling distribution of the sample will be approximately normal with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size.

This project will use simulation to validate the Central Limit Theorem, using various sample sizes to build their associated sampling distributions, noting how the shape of the sampling distribution changes as sampling sizes increase. The population will be sample census data based on the income for residents in Genesee and Wyoming counties, a population known to be extremely skewed.

Modeling Software

This project used Fathom for the simulation. Fathom is a statistical software package, one of the two major statistical software tools available (Minitab is the other). Fathom was chosen because it is targeted toward education and classroom use, it is more user-friendly for students, and it was recommended as better for classroom use by a Minitab representative that I talked to at the 2004 NCTM annual conference! Fathom had the capabilities to perform the simulation desired, and would also be able to dynamically show the sampling distribution being created as the sampling took place, which I viewed

as a big advantage for deepening student understanding of sampling distributions in general and the Central Limit Theorem in particular.



Decision Making and Planning

I chose this particular topic for a number of reasons. First of all, sampling distributions are central to an understanding of statistical inference. After teaching AP Statistics for three years, however, I have found that students generally don't have an accurate understanding of them. I felt that if students were able to view them being built dynamically as the sampling process occurred, they would gain a greater appreciation of their importance and use. Additionally, the Central Limit Theorem's conclusion is both

significant and counterintuitive, so that a simulation would increase students' conceptual understanding for statistical inference.

With these multiple objectives in mind, I planned what I wanted the model to include: its construction, which inputs would be variable, and what graphics would be necessary to show the dynamic nature. I built the model myself, and although it was my first time using this software, I was able to overcome any hurdles. After getting the major components working properly, I created a lab report where my statistics students would test the model and its functions, using various sample sizes, etc.

Division of Labor

Initially looking for students to build the model, several students expressed an interest. However, this interest apparently evaporated prior to the end of school each day. After several days where there were no students who returned after school, I decided to build the model myself to assure its completion.

Design and Execution of Tasks

Sample census files are included with the Fathom software. Because I wanted students to discover and validate the surprising conclusions of the Central Limit Theorem, I chose income as the attribute from census data for residents of Genesee and Wyoming counties. This choice served a dual purpose: (1) income is an inherently skewed distribution for a population, adding an air of suspense for whether the sampling distributions would take a normal shape, and (2) the fact that the population data came from an area close to Rochester would increase the realism of the simulation.

Having secured the population, the sampling mechanism had to be constructed, and then the dynamic process of building the sampling distributions. For each step, menu

boxes for changing variables had to be added, required statistics needed to be reported, and the appropriate graphics for the population distribution and the sampling distribution needed to be displayed to visually validate the Central Limit Theorem's conclusion.

Problems Encountered

Not having students willing to commit time after school was an initial disappointment. Admittedly, in an urban environment, many of the students work after school. AP Statistics is a rigorous course for many students, and they may have felt overcommitted already. Their interest does not mirror my own, but that is true for most of the students I teach.

I knew the model I wanted to create. However, during its creation there were times when I felt its feasibility was threatened due to my inexperience with the Fathom software, in particular, how to make the sampling process work, and then replicate it 500 times to create the sampling distribution. (Because of classroom time constraints, the student lab exercise took 100 samples rather than 500.) With persistence and scouring of the reference manuals, I was able to overcome these problems, and also get the dynamic viewing of the sampling distribution to work correctly.

Because this was the students' introduction to Fathom, there was some initial confusion on its use. However, with the repetitive nature of this exercise, this was quickly overcome. Additional instructions were added to help guide the students.

Evaluation of Results

Not only did the model operate well, but the dynamic nature of the simulation helped the students gain a deeper understanding of sampling distributions in general and the Central Limit Theorem in particular. Also, the choice of using New York census data

added a sense of realism to the exercise. Using a population as skewed as the one chosen gave the entire exercise a hint of suspense as in: can a population this skewed really produce a normal-shaped sampling distribution, and what sample size is necessary to produce it? The fact that students had control of variables such as sample size and the number of samples to create the sampling distribution stimulated their interest and thinking, providing them with an opportunity to ask "what if" questions and then giving them the tools with which to test them.

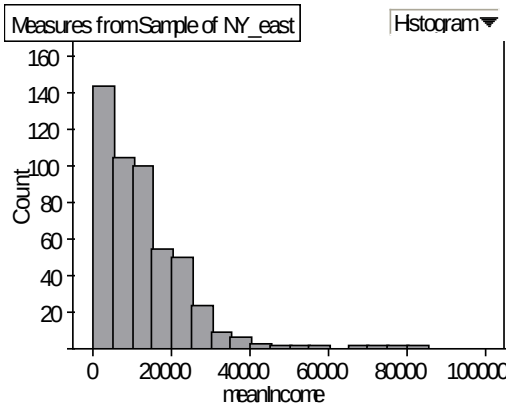
Most significant was the model's dynamic nature: students were able to view the sampling distribution being built as each sample was collected from the population. They were able to see how increasing sample sizes had a direct effect on the shape of the sampling distribution, moving from a skewed distribution similar in shape to the population, to a distribution which became more and more normal in shape as sample size increased.

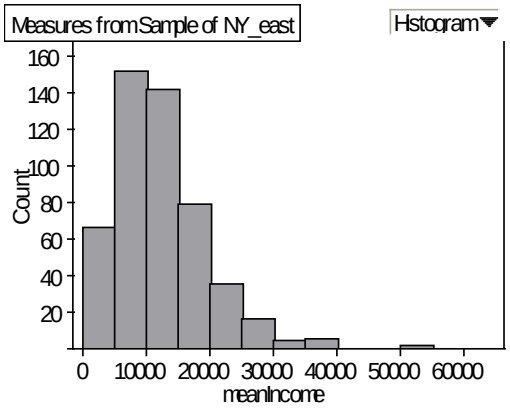
Part of the lab exercise included a section where the students calculated certain measures in accordance with the Central Limit Theorem, and then compared them to the data obtained from the simulation. This helped validate the Central Limit Theorem for the students.

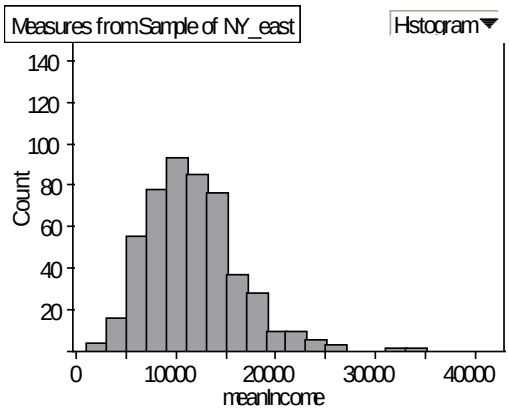
Certain unexpected benefits resulted from the simulation. Although students might be working with the same set of input values, they obtained different sampling distributions. This drove home the point that sampling is a random process, and that each new sample selected will be different from the previous samples.

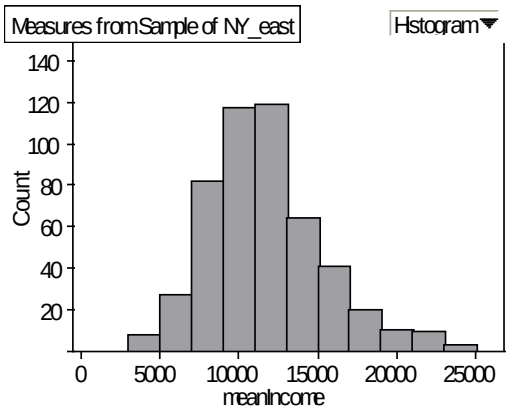
Another unexpected benefit is that increasing sample sizes did not automatically guarantee a sampling distribution whose mean was closer to the population mean than the

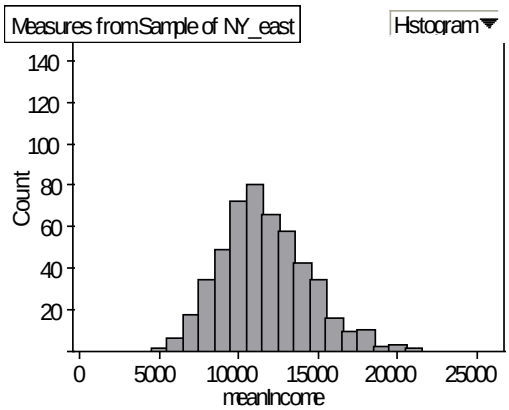
sampling distribution with a smaller sample size. Students became cognizant that their sampling distributions based on taking only 100 samples was not all-inclusive. Once again, it provided the opportunity for students to become more aware of how random the sampling process is by its very nature.











Summary of Experience

The model was an appropriate and significant one in terms of increasing student understanding of sampling distributions and the Central Limit Theorem. Also, there was a feeling of confidence and accomplishment in learning enough about the Fathom software to overcome the initial problems so that the main features of the simulation became operable.

Because this was the students' introduction to the Fathom software, there was some initial confusion on how to change variables or operate the model. I have since included more instructions, and added some text boxes to help clarify the process. Wanting the complete model to be visible on one screen restricts that somewhat.

Features I would like to change or include for future versions include eliminating some of the Fathom menu boxes (with some extraneous information) and replacing them with concise variable boxes such as "sample size" and "number of samples for sampling distribution", and an action button to initiate the simulation.

Overall, student comments regarding the simulation model and the lab exercise were very positive. The fact that the sampling distribution was both visibly graphic and dynamic was a major breakthrough in student understanding. Student engagement was demonstrated by different students testing additional sample sizes or running the simulation with increased samples. Lab questions were answered with a greater use of language, indicating increased critical thinking as a result of the exercise.