Visually Illustrating the Central Limit Theorem

1. What is the central limit theorem?

For practical purposes, the main idea of the *central limit theorem* (CLT) is that the average of a sample of observations drawn from some population can be approximately distributed as a normal distribution if certain conditions are met. In theoretical statistics there are several versions of the central limit theorem depending on how these conditions are specified. These are concerned with the types of assumptions made about the distribution of the parent population (population from which the sample is drawn) and the actual sampling procedure.

One of the simplest versions of the theorem says that if $X_1, X_2, \ldots, X_n$ is a random sample from an infinite population with mean μ and standard deviation σ, then the random variable defined by

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

converges to a standard normal distribution or, equivalently, the sample mean $\bar{X}$ approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as $n \to \infty$. In applications of the central limit theorem to practical problems in statistical inference, however, statisticians are more interested in how closely the approximate distribution of the sample mean follows a normal distribution for finite sample sizes, than the limiting distribution itself. Sufficiently close agreement with a normal distribution allows statisticians to use normal theory for making inferences about population parameters (such as the mean μ) using the sample mean, irrespective of the actual form of the parent population.

It is well known that whatever the parent population is, the *standardized variable* Z_n will have a distribution with a mean 0 and standard deviation 1. Moreover, if the parent population is normal, then Z_n is distributed exactly as a standard normal variable. The central limit theorem states the remarkable result that, even when the parent population is non-normal, the standardized variable Z_n is approximately normal if the sample size is `large enough`.

It is generallly not possible to state conditions under which the approximation given by the central limit theorem works and what sample sizes are needed before the approximation becomes good enough. As a general guideline, statisticians have used the prescription that if the parent distribution is symmetric and relatively short-tailed, then the sample mean reaches approximate normality for smaller samples than if the parent population is skewed or long-tailed.

In this lesson, we will study the behavior of the mean of samples of different sizes drawn from a variety of parent populations. Examining sampling distributions of sample means computed from samples of different sizes drawn from a variety of distributions, allow us to gain some insight into the behavior of the sample mean under those specific conditions as well as examine the validity of the guidelines mentioned above for using the central limit theorem in practice.

2. Objectives

The central limit theorem module (`clt_module`) was designed to enable visual investigation of the following important statistical concepts:

- Under certain conditions, in large samples, the sampling distribution of the sample mean can be approximated by a normal distribution.
- The sample size needed for the approximation to be adequate depends strongly on the shape of the parent distribution. Symmetry (or lack thereof) is particularly important.
- For a symmetric parent distribution, even if very different from the shape of a normal distribution, an adequate approximation can be obtained with small samples (e.g., 10 or 12 for the uniform distribution).
- For symmetric short-tailed parent distributions, the sample mean reaches approximate normality for smaller samples than if the parent population is skewed and long-tailed.
- In some extreme cases (e.g. binomial with $p \approx 1, n = 1$) samples sizes far exceeding the typical guidelines (e.g., 30 or 60) are needed for an adequate approximation.
- For some distributions without first and second moments (e.g., Cauchy), the central limit theorem does not hold.

3. Startup Instructions

On a Vincent workstation

```
% add lisp
% add stat
% sd_module
```

On a PC
On a Macintosh

4. The module interface

When the `clt_module` is started-up, two windows appear on the screen that will be mostly blank except for some control tools which we will use to run the module. The main window (in the upper left) contains controls to choose a distribution, choose sample sizes, and run the sampling simulations. Each of the other windows contains slide-bar controls to adjust the number of bins in the histograms of the sample statistics and to adjust the smoothed estimate of the sampling distributions.

The user begins by choosing a distribution (by holding down `Distributions` button), parameter values, and sample size (using the appropriate slide-bars displayed). A static graph of the density or mass function of the selected distribution appears on the main window. Each time the `New Sample` button is clicked, a new sample is generated and displayed as a dot plot under the density curve. The sample mean is highlighted with the symbol `x`, and is added

to the sample distribution of accumulated sample means displayed as a dynamically updated histogram on the second window. Figure 1 shows `clt_module` windows after 101 simulations from the exponential distribution with mean 3.0.

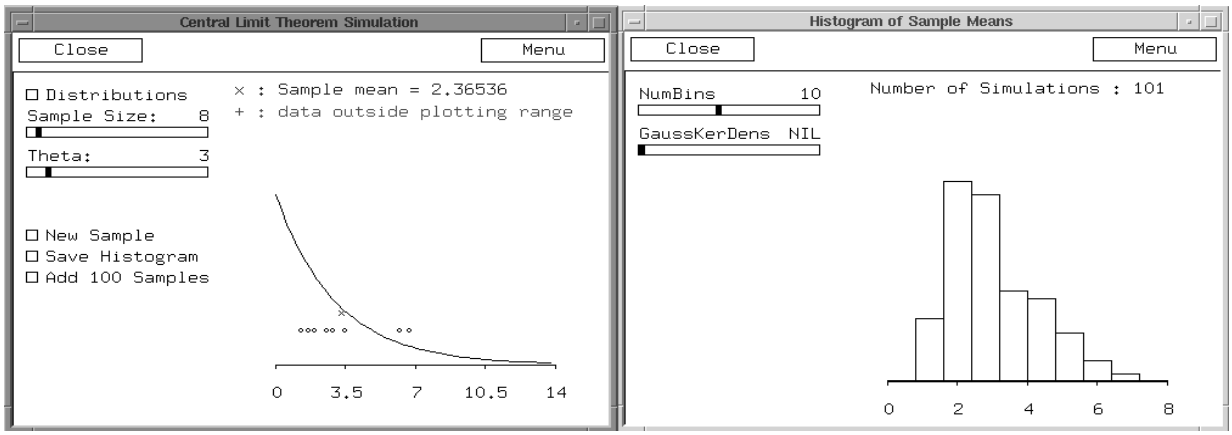

Figure 1: Frame of the Central Limit Theorem Module Windows

If the `New Sample` button is held down, the process of drawing new samples and updating the histogram is repeated continuously. The sampling process could be accelerated using the `Add 100 Samples` button. Using this button causes the histogram to be updated with 100 new sample means immediately. At any point in the process the student can click on the `Save Histogram` button on the histogram window to save a snapshot of the current histogram.

This allows comparisons of sampling distributions obtained under different conditions. Figure 2 shows snapshots of three histogram windows comparing sample means from samples of size from three different parent populations.

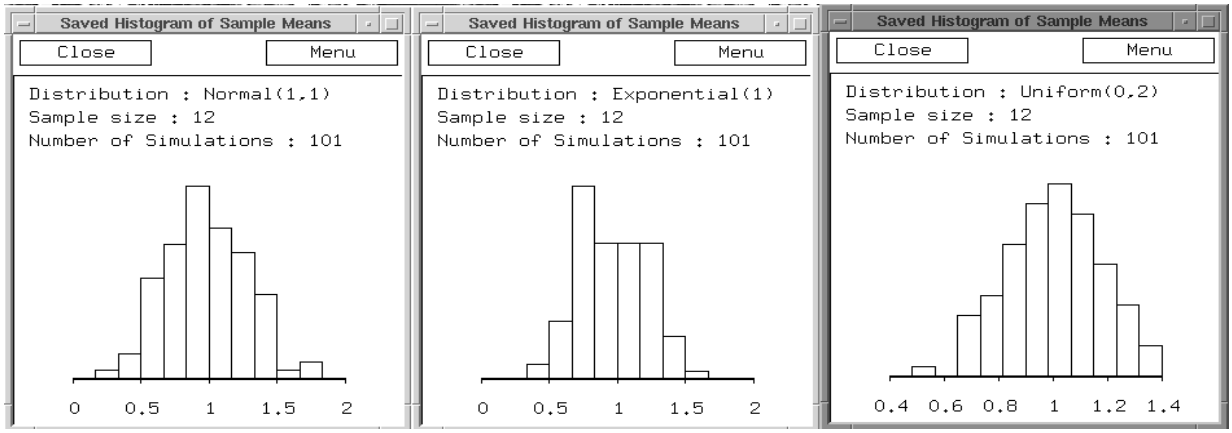

Figure 2: Three Frames of the Central Limit Theorem Software Windows

5. Warm-ups

To gain some familiarity with the `clt_module`, try the following:

- Use the `distribution` button to explore the shapes of the different `parent distributions`.
- For each parent distribution, notice the effect that changing the parameter values [done by moving the appropriate slide bar(s)] has on distribution shape.
- Choose a distribution from the list (e.g. a chi-square with 1 degree of freedom) from which to sample. Choose a sample size (say, 5) using the `Sample Size:` slider. Then click on the `New Sample` button once and examine the sample displayed as a dot plot below the density curve. Does this sample appear to you to be one from the parent distribution you selected ? Also note the sample mean displayed as a blue-colored `x` just above the dot plot. Does this appear to be correct ?
- Repeat the simulation by keeping the `New Sample` button pressed. Examine the appearance of the histogram on the other window and how it is updated as new sample means are generated. As the number of simulated samples increase, the histogram will represent a visual approximation to the sampling distribution of $\bar{X}$.
- Experiment with the number of histogram cells and with the `GaussKerDen` smoothing constant value in the histogram window to see if you can find a most visually appealing smoothed density estimate in each window.
- Use the `Add 100 Samples` button to update the histogram instantly by 100 sample means. Observe the change in the histogram.
- Use the `Save Histogram` button to obtain a copy of the current histogram. Once the copy of the histogram appears as a smaller window move it around on the screen and save it an appropriate place.

6. Exercises

In the following exercises, choose the indicated distribution as your first step. Then, for a selected sample size, use the `Add 100 Samples` button to obtain a histogram for 101 sample means. This will represent a simulated estimate of the sampling distribution of the sample mean. Use the `Save Histogram` to save a copy of this histogram. Repeat this procedure for other distributions or sample sizes specified. Study the histograms in the 3 windows thus obtained to answer the questions. You may want to repeat each experiment more than once to check the consistency of your conclusions.

1. Statistical theory tells us that the means of samples from a normal distribution should also follow a normal distribution. To investigate this result, choose a normal distribution with a mean $\mu = 100$ and $\sigma = 10$ as the parent distribution. Obtain simulations of the sampling distributions of the sample mean for the 3 sample sizes 4, 12, and 30 as described above. Study the histograms in the 3 windows created.

(a) Compare the *shapes* of the histograms in the 3 windows. How do they differ? Explain how your conclusions relate to the statistical theory about sampling from a normal distribution.

(b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

2. A uniform distribution is a choice for a symmetric short-tailed parent distribution. The CLT tells us that the normal distribution could provide a good approximation to the sampling distribution of the mean under certain conditions. Choose a uniform distribution as the parent distribution and obtain the sampling distributions of the sample mean for the 3 sample sizes 4, 12, and 30 as described above. Study the histograms in the 3 windows.

   (a) Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a uniform distribution?

   (b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

3. A chi-square distribution is an example of a skewed distribution whose shape depends on a parameter value. Choose the chi-square with 1 degree of freedom. Obtain the sampling distributions of the sample mean for the 3 sample sizes 4, 12, and 30 as described above. Study the histograms in the 3 windows.

   (a) Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a distribution that is chi-square with 1 degree of freedom?

   (b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

4. Choose the Cauchy distribution with median (location parameter) set to 5. Obtain the sampling distributions of the sample mean for the 3 sample sizes 4, 12, and 30 as described above. Study the histograms in the 3 windows.

   (a) Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a distribution that is chi-square with 1 degree of freedom?

   (b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

5. In this exercise we will study sampling distributions of means of samples of size 16 from each of three different distributions: Uniform between 0 and 2, Exponential with mean 1, and Normal with mean 1 and standard deviation 2. As before, use the `Add 100 Samples` button to obtain histograms for 101 sample means in each case. Study the histograms in the 3 windows. Notice that the means these population distributions are the same.

   (a) Compare the *shapes* of the histograms in the 3 windows. How do they differ? How is the shape of the sampling distributions related to the shape of the parent distribution?

   (b) Compare the spread in the 3 different sampling distributions. How does the choice of population standard deviation affect the spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

7. Solutions to Exercises

1. The results from theory for sampling from a normal distribution can be stated simply and concisely: The mean of a sample from a normal distribution with mean μ and standard deviation σ will follow a normal distribution with mean μ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

   (a) Although contaminated with random "noise," the shapes are all similar and approximately normal, as suggested by statistical theory because the sample mean from a normal distribution also follows a normal distribution.

   (b) The spread in the distribution decreases with increasing sample size. This is as expected from theory where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

2. Note that the uniform distribution has a shape that is quite different from that of a normal distribution, but it is a symmetric distribution.

   (a) For the two larger sample sizes, the sampling distributions appear (again, except for the random "noise") to be shaped like a normal distribution. The approximation improves as the sample size gets larger.

   (b) Again, the spread in the distribution decreases with increasing sample size, and, this is as expected from theory because, even if the underlying distribution is not normal, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

3. (a) The parent distribution is chi-square with one degree of freedom, which is very skewed to the right. The sample means, however, have distributions that are more symmetric. As the sample size increases, the sampling distributions become more symmetric.

   (b) As with the other distributions, the spread in the distribution decreases with increasing sample size, and, this is as expected from theory because, even if the underlying distribution is not normal, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

4. The Cauchy distribution does not have a mean or a variance (i.e., the usual mathematical forms leading to the definition of a distributions mean or variance turn out to be infinite). This will cause some strange behavior in the simulations. In particular, the sample means will often have extremely large deviations from the center of the Cauchy distribution (the distribution is symmetric and does, of course, have a median).

   (a) The histograms are difficult to interpret. The sampling distribution is so spread out that a histogram, with many equally-spaced cells will have most of its cells (all but 3 or 4) with only one observation in it. Most of the observations from a Cauchy will be near to its median, but some of the sample means will be far, far, away.

   (b) See above. The the Cauchy distribution does not meet the conditions for the central limit theorem to hold. Sample means of Cauchy random variables will not, even in very large samples, follow a normal distribution.

5. The object of this experiment is to directly compare shapes and spreads of distributions of means sampled from 3 different distributions independent of the sample size. The 3 selected populations distributions have different shapes but have the same mean (location) so that the spreads of the three sampling distributions should be easier to compare.

   (a) The parent distributions vary from a symmetric, short-tailed distribution to one very skewed to the right. The sample means, however, have distributions that are more symmetric. The sampling distributions are more symmetric for the parent distributions that are symmmetric than skewed; or those that are long-tailed than short-tailed.

   (b) Since the sample size is fixed, the spread in the sampling distribution varies directly with the population standard deviation σ, which are 0.5773, 1.0 and 2.0, respectively . This is as expected from theory where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.