Journal of Statistics Education

Central Limit Theorem: New SOCR Applet and Demonstration Activity

Ivo D. Dinov, Nicolas Christou, and Juana Sanchez
University of California, Los Angeles

Journal of Statistics Education Volume 16, Number 2 (2008), www.amstat.org/publications/jse/v16n2/dinov.html

---

Key Words: Statistics education; Technology-based blended instruction; Applets; Central Limit Theorem; SOCR.

Abstract

Modern approaches for information technology based blended education utilize a variety of novel instructional, computational and network resources. Such attempts employ technology to deliver integrated, dynamically linked, interactive content and multi-faceted learning environments, which may facilitate student comprehension and information retention. In this manuscript, we describe one such innovative effort of using technological tools for improving student motivation and learning of the theory, practice and usability of the Central Limit Theorem (CLT) in probability and statistics courses. Our approach is based on harnessing the computational libraries developed by the Statistics Online Computational Resource (SOCR) to design a new interactive Java applet and a corresponding demonstration activity that illustrate the meaning and the power of the CLT. The CLT applet and activity have clear common goals; to provide graphical representation of the CLT, to improve student intuition, and to empirically validate and establish the limits of the CLT. The SOCR CLT activity consists of four experiments that demonstrate the assumptions, meaning and implications of the CLT and ties these to specific hands-on simulations. We include a number of examples illustrating the theory and applications of the CLT. Both the SOCR CLT applet and activity are freely available online to the community to test, validate and extend (Applet: http://www.socr.ucla.edu/htmls/SOCR_Experiments.html and Activity: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem).

1. Introduction

1.2 General

Contemporary Information Technology (IT) based educational tools are much more than simple collections of static lecture notes, homework assignments, and web-based applets posted on course-specific Internet sites. Over the past five years, a number of technologies have emerged that provide dynamic, linked, and interactive learning content with heterogeneous points-of-access to educational materials (Dinov, 2006c). Examples of such new IT resources include common web-places for course materials (BlackBoard, 2006; MOODLE, 2006), complete online courses(UCLAX, 2006), Wikis (SOCRWiki, 2006), interactive video streams (ClickTV, 2006; IVTWeb, 2006; LetsTalk, 2006), audio-visual classrooms, real-time educational blogs (Brescia & Miller, 2006; PBSBlog, 2006), web-based resources for blended instruction (WikiBooks, 2006), virtual office hours with instructors (UCLAVOH, 2006), collaborative learning environments (SAKAI), test-banks and exam-building tools (TCEXAM) and resources for monitoring and assessment (ARTIST, 2006; WebWork).

This explosion of tools and means of integrating science, education and technology has fueled an unprecedented variety of novel methods for learning and communication. Many recent attempts (Blasi & Alfonso, 2006; Dinov, 2006c; Mishra & Koehler, 2006) have demonstrated the power of this new paradigm of IT-based blended instruction. For instance, in statistics education there are a number of excellent examples where fusing new pedagogical approaches with technological infrastructure allowed instructors and students to improve motivation and enhance the learning process (Forster, 2006; Lunsford, Holmes-Rowell, & Goodson-Espy, 2006; Symanzik & Vukasinovic, 2006). In this manuscript, we build on these and other similar efforts and introduce a general, functional and dynamic central limit theorem (CLT) applet along with a corresponding hands-on activity.

1.2 The CLT

There are various statements of the central limit theorem, but all of them represent weak-convergence results regarding (mostly) the sums of independent identically-distributed (random) variables (Davidson, 1994; Lyapunov, 1906). There are also versions of the CLT for dependent variables (Merlev`ede, Peligrad, & Utev, 2006). Perhaps, the most frequently used and practically important statement of the central limit theorem is in terms of (arithmetic) averages of random variables sampled from a process with well-defined and finite first two moments. In this case, CLT implies that the average will follow approximately a normal distribution, as the sample-size increases (Aberson, Berger, Healy, Kyle, & Romero, 2000). Various generalized versions of the CLT may be phrased in terms of products of random variables (DeGroot, 1970), convolution of densities (Chung et al., 2001), or as the product of characteristic functions. (Formanov, 2002) The CLT is the statistical analogue of the Grand Unified Theory in physics (Hall & Nomura, 2002). As such, the CLT is one of the most celebrated results in the field. The history of the CLT is equally as impressive as its deep scientific value (Tijms, 2004).

Many undergraduate and graduate classes use the following statement of the central limit theorem:

CLT: Let $X_1,\ldots,X_n, \ldots$ be a random sample (independent and identically distributed) from a (native) distribution with well-defined mean μ_X and variance σ_X^2 ($|\mu_X| < \infty$ and $\sigma_X^2 < \infty$). Then as n increases, the sampling distributions of the *sample average*

$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and the *total sum*

$T = \sum_{i=1}^{n} X_i$ approach Normal distributions with corresponding

$\left(\mu_{\overline{X}} = \mu_X,\ \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}\right)$ and $\left(\mu_T = n\mu_X,\ \sigma_T = \sigma_X\sqrt{n}\right)$.

In essence, the CLT implies that the Normal distribution is the center of the universe of all *nice* distributions. This is the reason why we encounter so frequently estimates involving arithmetic-averaging. In a way, the pathway from a *nice* distribution to Normal distribution is paved by sample averages. In other words, the CLT provides a unifying framework for all (*nice*) distributions by facilitating explicit roadmaps from them to Normal distribution. A useful analogue to CLT is the physics Grand Unifying Theory, which similarly attempts to unite the theory behind the fundamental forces in nature.

The ramifications of the CLT go beyond the scope of this interpretation. For example, one may wonder if there are other types of population-parameters or sample-statistics that have similar limiting properties. How large does the sample size have to be to ensure normality of the sample average or total sum? Does the convergence depend on the characteristics of the native distribution (e.g., shape, center, dispersion)? How about weighted averages, non-linear combinations, or more general functions of the random sample? People exposed to the CLT frequently ask many such interesting questions. Some may have known theoretical answers (exact or approximate); other questions may be better addressed empirically by simulations and experiments (Garfield, 2002).

1.3 Other Similar Efforts

Several other groups have developed hands-on activities paired with interactive applets that demonstrate the foundation of the CLT. Among these efforts are the Cal Poly ISCAT collection (Lunsford et al., 2006; Rossman, Chance, & Ballman, 1999), Cal State CLT Applet (Stanton, 2005); Vienna University Learning by Simulation (Lohninger, 2008); Northwestern NetLogo modeling-and-simulation environment (Abrahamson & Wilensky, 2004); Star Library CLT Activity (Andrews, 2005); Rand CLT applet (RANDCLT, 2007), WebStat (West & Ogden, 1998) and many others. A "CLT" keyword search at www.causeweb.org yields a number of instructional plans, activities, aids and resources for technology-enhanced materials for teaching the central limit theorem.

Each of these applets, activities and resources has unique features that make it useful in its specific context. Some affirm ones intuitive beliefs, while others reinforce practice learning or address specific CLT-related questions, typically using one native distribution model. In this manuscript, *native* (distribution or process) refers to the underlying population characteristics.

1.4 CLT Instructional Challenges

We have extensive CLT pedagogical experience based on graduate and undergraduate teaching, interacting with students (and teaching assistants) and evaluating students' performance in various probability and statistics classes. In our endeavors, we have used a variety of classical (e.g., mathematical formulations), hands-on activities (e.g., beads, sums, Quincunx) and technological approaches (e.g., applets, demonstrations). Our prior efforts have identified the following instructional challenges in teaching the concept of the CLT using purely classical and physical hands-on activities. Some of these challenges may be addressed by employing modern IT-based technologies, like interactive applets and computer activities:

- What is a native process (native distribution), a sample, a sample distribution, a parameter estimator, a sample-driven

numerical parameter (point) estimate or a sampling distribution?
- What is the relationship between the inference of the CLT and its applications in the real world?
- How does one improve CLT knowledge retention, which seems to decay over time? Are there paramount characteristics we can demonstrate in the classroom, which may later serve as a foundation for reconstructing the detailed statement of the CLT and improving communication of CLT meaning and results?
- How does one directly involve and challenge students in thinking about CLT (in and out of classroom)?

Traditional CLT teaching techniques (symbolic mathematics and physical demonstrations) are typically restricted in terms of time and space (e.g., shown once in class) and may have the limitations of involving one native process, studying one population parameter and restricting the scope of the inference (e.g., sample-size constraints).

Modern IT-based blended instruction approaches address many of these CLT teaching challenges by utilizing the Internet and the available computational power. For example, a Java CLT applet may be evoked repeatedly under different initial conditions (choosing sample-sizes and number of experiments, native process distributions, parameters of interest, etc.). Such tests may be performed from remote locations (e.g., classroom, library, home), and may provide enhanced interactive features (e.g., fitting Normal model to sampling distribution) demonstrated in different experimental modes (e.g., intense computational vs. visual animated sampling). Such features are especially useful for active, visual and deductive learners. Furthermore, interactive demonstrations are thought to significantly enhance the learning process for some student populations (Demir, 2006; Laakso, Salakoski, Grandell, Qiu, Korhonen, & Malmi, 2005; Masters, Madhyastha, & Shakouri, 2005).

Students in probability and statistics classes are generally expected to master difficult concepts that ultimately lead to understanding the basis of data variation, modeling and analysis. For many students relying on procedural manipulations and recipes is natural, perhaps because of their prior experiences with (deterministic) Newtonian sciences. Various statistics-education researchers have experimented with technology to explore novel exploratory data-analysis techniques that emphasize making sense of data via data manipulation, visualization and simulation. Such investigators refer to statistical literacy as the process of acquiring and utilizing intuition for discovering and interpreting trends, proposing solutions and counterexamples to basic problems in probability, as well as understanding statistical data modeling and analysis (Ben-Zvi & Garfield, 2004).

Because the concepts of distribution, variation, probability, randomness, modeling and estimation are so ubiquitously used and entangled (Wild, 2006), instructors frequently forget that these notions should be defined, explained and demonstrated in (most) undergraduate probability and statistics classes. Various sampling and simulation applets and demonstrations are quite useful for this purpose.

1.5 The SOCR Resource

The UCLA Statistics Online Computational Resource (SOCR) is a national center for statistical education and computing. The SOCR goals are to develop, engineer, test, validate and disseminate new interactive tools and educational materials. Specifically, SOCR designs and implements Java demonstration applets, web-based course materials and interactive aids for IT-based instruction and statistical computing (Dinov, 2006b; Leslie, 2003). Various types of users (e.g., instructors, students and researchers) may find the SOCR resources useful. The SOCR Motto, *"It's Online, Therefore It Exists!",* implies that all of these resources are freely available on the Internet (www.SOCR.ucla.edu).

There are four major components within the SOCR resources: computational libraries, interactive applets, hands-on activities and instructional plans. The SOCR libraries are typically used for statistical computing by external programs (Dinov, 2006a). The interactive SOCR applets are subdivided into Distributions, Experiments, Analyses, Games, Modeler and Charts. The hands-on activities are dynamic Wiki pages (SOCRWiki, 2006) that include a variety of specific instances of demonstrations of the SOCR applets. The SOCR instructional plans include lecture notes, documentation, tutorials and guidelines about statistics education. We have decided to develop and disseminate SOCR educational resources using Wiki pages, as the Wiki infrastructure greatly facilitates building community-based, version-controlled and dynamic materials.

1.6 Goals of the SOCR CLT Activity

The goals of the SOCR CLT activity are as follows:

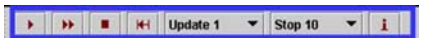
- to provide an intuitive notion of sampling from any process with a well-defined distribution;
- to motivate and facilitate learning of the central limit theorem;
- to empirically validate that sample-averages of random observations (most processes) follow approximately a Normal distribution;
- to empirically demonstrate that the *sample-average* is special and other sample statistics (e.g., median, variance, range, etc.) generally do not have distributions that are normal;
- to illustrate that the expectation of the sample-average equals the population mean (and the sample-average is typically a good measure of centrality for a population/process);

- to show that the variation of the sample average rapidly decreases as the sample size increases $\left(\frac{1}{\sqrt{n}}\right)$;
- to provide empirical evidence for and against the "golden rule" of using sample sizes of 30;
- to reinforce the concepts of native distribution, sample, sample distribution, sampling distribution, parameter estimator and data-driven numerical parameter estimate.

2. Design and Methods

2.1 SOCR CLT Applet

The SOCR CLT applet is designed as a meta-experiment (part of SOCR Experiments, integrating functionality from SOCR Distributions, Charts and Modeler). In this applet, we allow the most flexibility in terms of choosing the native process distribution, population parameters of interest, sample-sizes, number of experiments, animated and computational sampling of data. Figure 1 illustrates the main components of the applet interface. This applet may be accessed directly online by going to the SOCR Experiments (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html) and selecting *Sampling Distribution CLT Experiment* from the drop-down list of experiments (top-left). There are tool-tips included for every widget (active graphical component) in this applet. Applet tool-tips (pop-up information fields describing interface features) are activated by bringing the mouse over a component within the applet window. The most important components of this applet are boxed in blue rectangles on Figure 1 and explained below:

o 

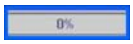
the main action buttons in the CLT experiment;

o 

button for refreshing the summary statistics table after completion of a sampling experiment;

o 

two groups of fields allowing control over the type and sample-size for each of the (two possible) parameters of interest;

o 

progress monitor (for large sampling tasks);

o 

a button to go directly to the SOCR CLT Wiki Activity (see below);

o 

tab-panel selection (Graphical Histogram and Numerical Summaries (default) and Distribution selection);

o 
four distribution graphs (native population, last sample and the sampling distributions of the two parameters of interest chosen by the user).

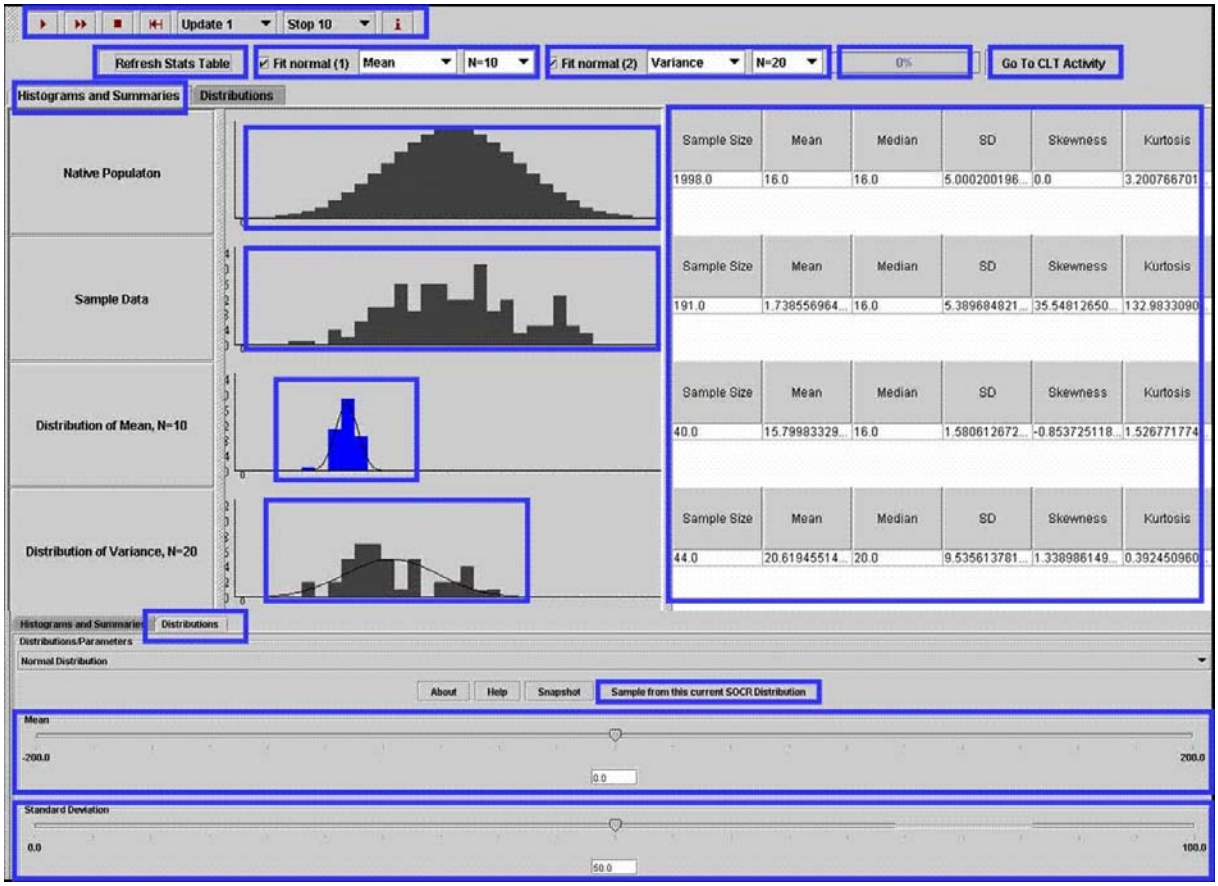


Figure 1: SOCR CLT Applet interface.

In addition to other SOCR libraries, this applet utilizes ideas, designs and functionality of the Rice Virtual Laboratory of Statistics (RVLS) and the University of Alabama Virtual Laboratories in Probability and Statistics (VirtualLabs).

The generality and usability of the SOCR CLT applet comes from the fact that the user has full control over of each of the features listed above and because the applet allows a very large number of possibilities to test and observe the power of the central limit theorem. In particular, all of the CLT activity goals presented in section **I.f**, above, can be addressed using this applet. There is a second panel (called Distributions) in this applet (shown on the bottom of Figure 1), which allows the selection of the native process we wish to sample from. In the CLT Applet, the user may manually generate a distribution or choose one from the 50+ families of SOCR Distributions (including commonly used distributions like the Binomial and Normal, as well as more exotic ones like the Rayleigh and the von Mises, http://wiki.stat.ucla.edu/socr/index.php/About_pages_for_SOCR_Distributions). In the Distribution tab-panel the user may select specific parameters for the chosen distribution, if appropriate. This large space of possibilities makes this applet applicable in a variety of CLT hands-on demonstrations and activities tailored for a specific curriculum. The SOCR CLT activity, discussed below, is just one instance of such specific instructional activity that illustrates the theory, promotes the practice, and emphasizes the usability of the CLT. Using this template, others may develop their own demonstrations of the CLT directly on the SOCR Wiki Resource or on their own web-servers.

2.2 SOCR CLT Activity

The SOCR CLT activity is available online (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem) and accessible from any internet-connected computer with a Java-enabled web browser. The activity includes dynamic links to web resources on the CLT, interactive CLT demonstrations, and relevant SOCR resources.

Following the outline of its goals, the CLT experimental settings are described online in the second paragraph of the SOCR CLT Wiki activity. Each of the four experiments suggested in the activity addresses different feature sets of the applet and demonstrates separate characteristics of the CLT.

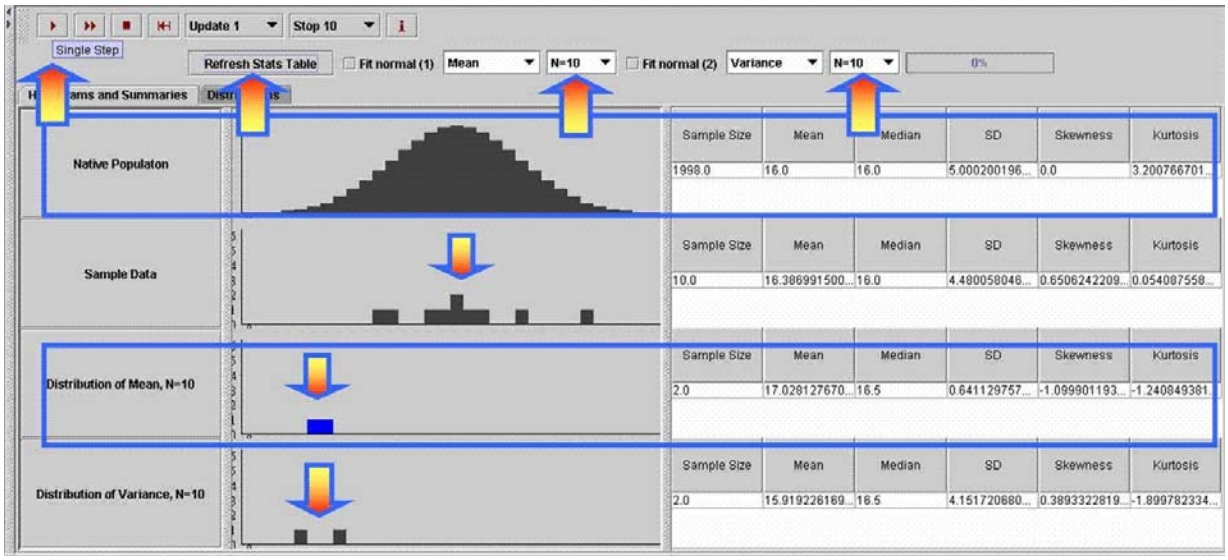


Figure 2: SOCR CLT Activity: Snapshot of the first experiment.

The *first experiment* depicts the effects of small sample-sizes on the sampling distributions of the sample average and the sample variance, **Figure 2**. The arrows in this figure point to the most important features in this experiment (left-to-right and top-to-bottom):

-  button is for taking one sample from the selected distribution;
-  button is for recomputing the sample and sampling distribution summary statistics in the tables to the right;
-  two drop-down lists allow the user to select the sample-sizes for one or two sampling distributions;
-  second graph panel shows the distribution of the last sample;
-  third graph panel represents the sampling distribution of the first parameter (mean, in this case);
-  the bottom graph panel shows the sampling distribution of the second parameter (variance); and
-  the data tables on the right show the summary statistics for all 4 distributions (native, sample, sampling (mean) and sampling (variance)).

The user observes graphically the process of random sampling from a distribution (in this case Normal). At each run of the experiment, the user visually tracks the meaning of the sample statistics calculation and the construction of the appropriate sampling histogram. Using the applet's *Refresh Stats Table* button, one computes the basic summaries for the distributions of the sample statistics (the right panel). Summary statistics are computed for all processes: the native population distribution (row 1), last sample (row 2) and the two sampling distributions, in this case the mean and the variance (rows 3 and 4).

The *second experiment* portion of the SOCR CLT activity explores the numerical characteristics of the sample average and sample-variance distributions. Some of these characteristics computed are the mean, standard deviation, skewness, and kurtosis. **Figure 3**. The user may inspect the goodness-of-fit between the sampling distribution and a Normal distribution with the same first two moments.

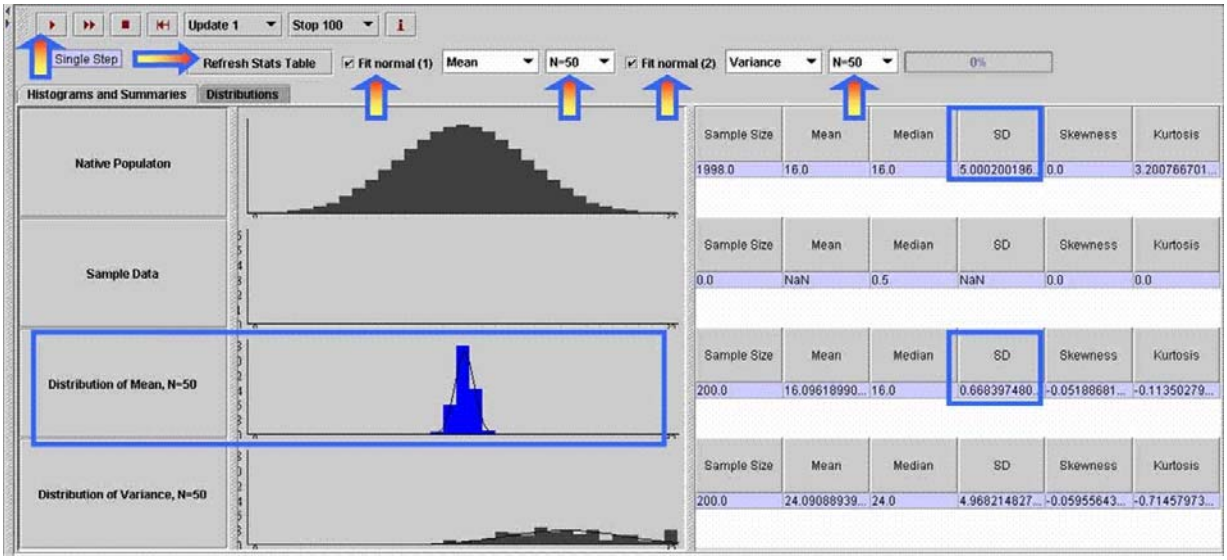


Figure 3: SOCR CLT Activity: Snapshot of the second experiment.

This experiment draws parallels between the sampling distribution statistics (rows 3 and 4) and their analogues from the native population (on the top row). For example, the mean of the multiple sample-averages is about the same as the mean of the native population. On the other hand, the standard deviation of the sample-averages is about $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the original native process.

The *third experiment* involves sampling from an arbitrary SOCR Distribution, Figure 4. This demonstrates that the central limit theorem is valid for numerous families of distributions. This experiment may be used to empirically validate that the sample average is a unique data statistics that has invariant limiting of its sampling distribution. In addition, the convergence of the sampling distribution to a Normal may be validated, relative to the chosen sample-size.

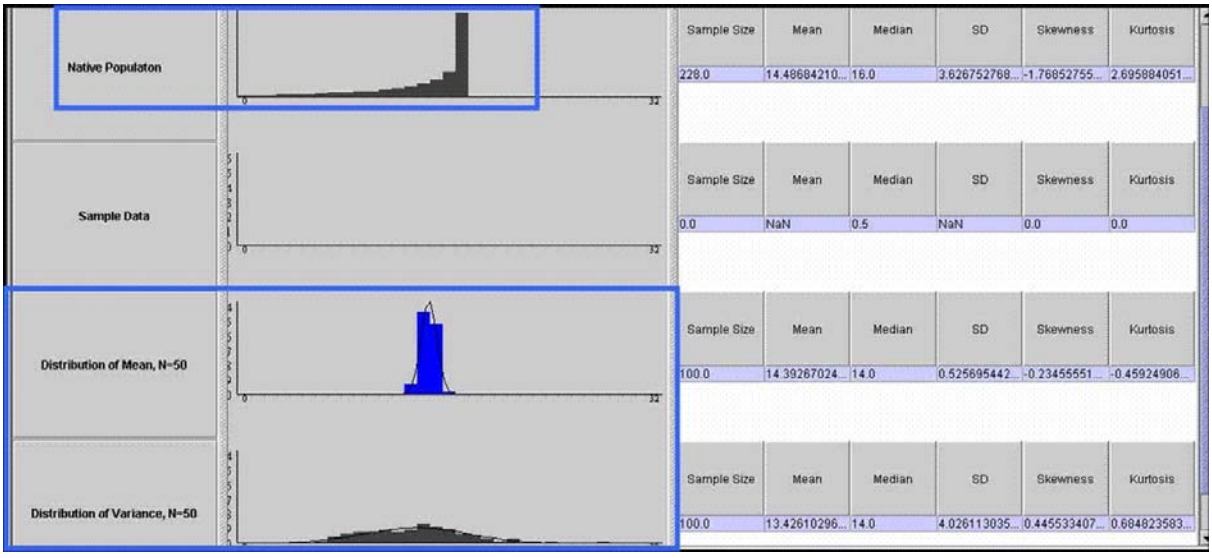


Figure 4: SOCR CLT Activity: Snapshot of the third experiment.

The *fourth experiment* in this activity illustrates how to generate and sample from a new distribution not included in the list of SOCR Distributions, Figure 5. The user may draw the shape of a hypothetical distribution by clicking and dragging the mouse in the top graphing canvas on the applet. This functionality allows exploring the assumptions and convergence guaranteed by the CLT for continuous and discontinuous, symmetric and asymmetric, unimodal and multi-modal, leptokurtic and mesokurtic, and other types of distributions or processes. Also, we explore the differences between two data-driven estimates for the population centrality – the sample mean and the sample median.

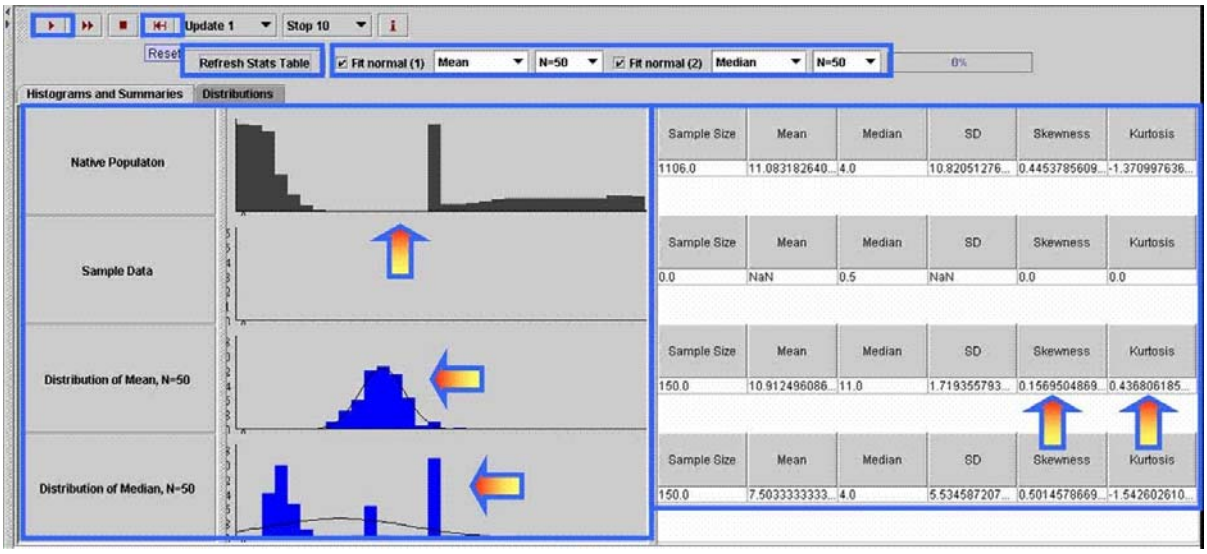


Figure 4: SOCR CLT Activity: Snapshot of the fourth experiment.

3. Discussion and Utilization

3.1 Classroom use of the pair of CLT applet and its corresponding activity.

There are a variety of ossible applications of these resources in a general classroom setting. Instructors should consider fine-tuning these materials to their specific course curricula. In the past, we have experimented in a number of ways with these materials. A step-by-step of illustration of one possible strategy for using these resources in a classroom is described below. We first choose a distribution that we want to start with (typically Normal or constructed by clicking and dragging the mouse in the top graphing panel). Then the instructor begins an interactive discussion with the students, which involves applet usage, CLT properties, examples and counterexamples, student feedback and computer simulations:

Such questions to be discussed are:

- Can you think of a process that may have this distribution?
- Now draw a random sample of size 20 from this distribution.
- What is the mean of **this sample** (look in row 2)? What is the corresponding standard deviation?
- What does the distribution of *this sample* look like? What would change if we increased the sample size to 100 or 1,000? Try it and think of a reason for that phenomenon. (This point helps demonstrate that the distribution of the sample will tend to look like the parent distribution, as the sample-size increases).
- Now let's go back and draw many samples of size 20, again from the same distribution, and compute the mean of each of them. If we draw 100 samples each of size 20 how many sample means do we have? Their distribution is called the *sampling distribution of the sample mean*. Similarly, we can construct the sampling distributions for other parameters (e.g., median, variance, range, etc.)
- What does the distribution of the **sample mean** (row 3) look like? Does it look like the one in the second row? Does it depend on the sample-size?
- Now, compare the **distribution of the last sample** and the **sampling distribution** of the sample mean. What do you conclude? Why?
- Compare also the mean and standard deviations of the distributions in rows 2 and 3. Repeat the exercise with another distribution (selected from the drop-down list of distributions in the second tab-panel in the main window, or drawn one by hand directly in the top row graphing canvas).
- Can we make the same conclusion as before? Do you think that if we try another distribution you will come to the same conclusion? Try it and answer the question empirically. Try it with your most or least favorite distributions.
- Let’s shown the outcome of running the experiment a number of times with the population mean and median as parameters of interest. Notice the sampling distributions of the sample average and the sample median.
- After discussing the situations where the CLT conclusions are valid, the instructor may choose to demonstrate situations where CLT-like results are not valid. These may include small sample sizes, distributions with ill-posed first two moments (e.g., Cauchy) and different sample statistics (e.g., median, range). Such counterexamples also provide valuable information to students.
- Finally, an assessment question may be appropriate (e.g., quiz).

The instructor builds complexity as he or she is satisfied with students’ responses, attentiveness and comprehension. A homework

assignment that reinforces these CLT principles is appropriate in many cases and may improve knowledge retention.

3.2 Real-Life examples of the CLT.

There are many practical examples using CLT to solve real-life problems. Here are some instances, which use CLT to approximate probabilities of interests and can be solved using SOCR resources.

- Suppose a call service center expects to get 20 calls a minute for questions regarding each of 17 different vendors that rely on this call center for handling their calls. What is the probability that in a 1-minute interval they receive less than 300 calls in total?
- Let X_i be the random variable representing the number of calls received about the i^{th} vendor within a minute, then X_i *~Poisson (20)*, as X_i is the number of *arrivals* within a unit interval and the mean arrival count is given to be 20. The distribution of the total number of calls $T = \sum_{i=1}^{17} X_i \sim Poisson(17 \times 20).$ By CLT, $T \sim N\left(\mu_T = 17 \times 20,\ \sigma_T^2 = 17 \times 20\right)$, as an approximation of the exact distribution of the total sum. Using the SOCR Poisson Distribution applet (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html) one can compute exactly the *P(T<300 | T~Poisson(17×20))= 0.014021*. On the other hand, one may use the CLT to compute a Normal approximation probability of the same event,

$$P(T<300 \mid T \sim N\left(\mu_T = 17 \times 20,\ \sigma_T^2 = 17 \times 20\right)) = 0.014896.$$

  The last quantity is obtained using the SOCR Normal Distributions applet, without using continuity correction. Using continuity correction the approximation improves, *P(T<300 |* $T \sim N\left(\mu_T = 17 \times 20,\ \sigma_T^2 = 17 \times 20\right)$ *)= 0.0140309.* Arguably, the CLT-based calculation is less intense and more appealing to students and trainees, compared to computing the exact probability.

- It is believed that the life-time, in hours, of light-bulbs are exponentially distributed, say $Exp\left(\lambda = \frac{1}{2,000}\right)$, mean expected life of 2,000 hours. Recall that the Exponential distribution is also called the *Mean-Time-To-Failure* distribution. You can find more about it from the SOCR Exponential Distributions applet. Suppose a University wants to purchase 100 of these light-bulbs and estimate the average life-span of these light bulbs. What is a CLT-based estimate of the probability that the average life-span exceeds 2,200 hrs?

  Denote $X_i \sim Exp\left(\lambda = \frac{1}{2,000}\right)$ and $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$. Notice that in this case the exact distribution of $\overline{X}$ is (generally) not exponential, even though the density may be computed in closed form (Khuong & Kong, 2006). If we use the CLT, however, we can approximate the probability of interest

$$P\left(\overline{X} > 2,200\right) \cong P\left(\overline{X} > 2,200 \,\middle|\, \overline{X} \sim N\left(\mu_{\overline{X}} = 2,000,\ \sigma_{\overline{X}}^2 = \frac{(2,000)^2}{100}\right)\right),$$

  as we know that the mean and the standard deviation of X_i are $\frac{1}{\lambda} = 2,000$, and the standard deviation of $\overline{X}$ is $\frac{1}{\lambda\sqrt{100}} = 200.$ Therefore, $P(\overline{X} > 2,200) \sim 0.158655$, using the CLT approximation and the SOCR Normal Distributions calculator, Figure 6.

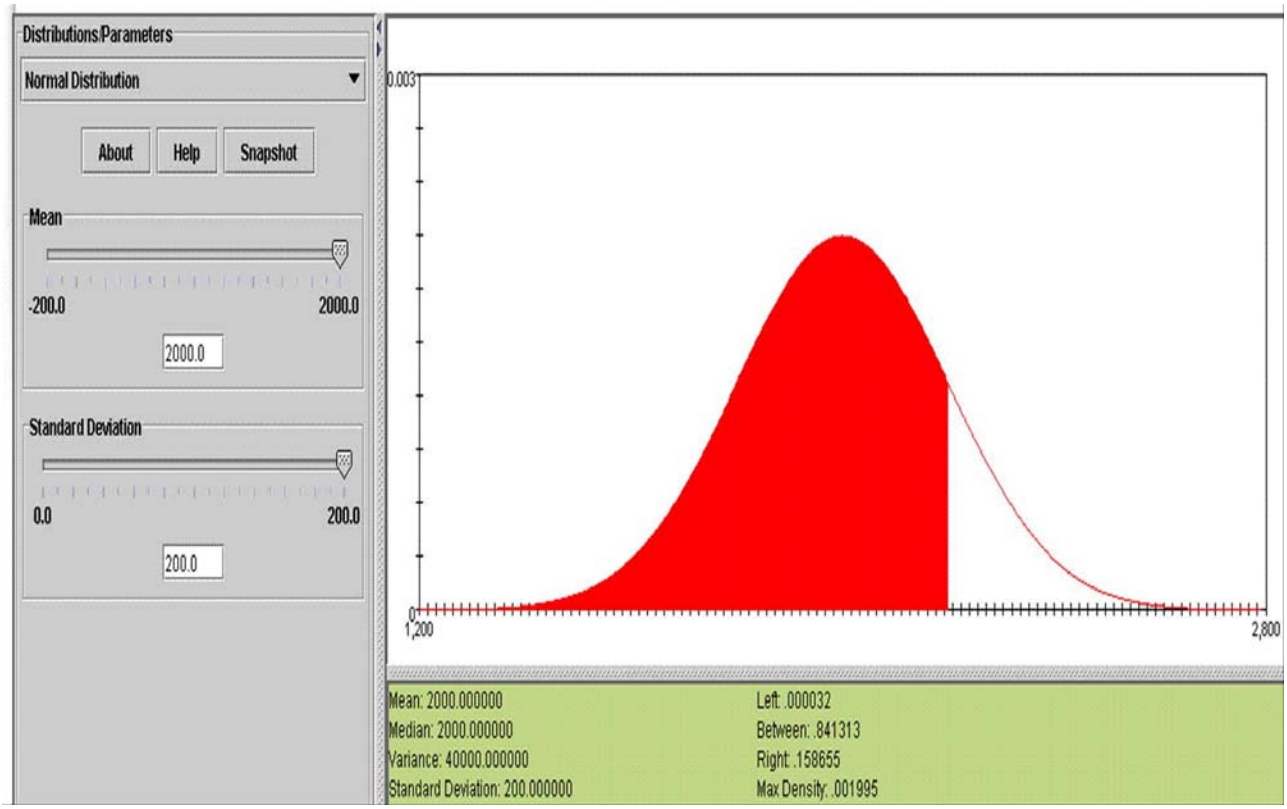


Figure 6: Calculating the approximate probability for the average of the Exponential variables.

- A weekly TV talk show from broadcaster ***U*** invites viewers to call to express their opinions about the program. Many people call, which sometimes results in quite a long wait time until the host replies. The time it takes the host to respond tends to follow an exponential distribution with mean of 50 seconds. A competing TV network ***W*** has another similar talk show and would like to respond to callers faster than broadcaster ***U***. To do that ***W*** executives need to know how long ***U*** takes to respond. So, ***W*** personnel make 25 calls per week to the ***U*** show (for 50 weeks) and measure how long it took the ***U*** host to respond. Then ***W*** executives compute the average length for their weekly samples of size 25. At the end of the year, they plot the distribution of the sample means. What are the center, spread and shape of this distribution? Find out using the SOCR CLT applet. Approximately, what proportion of time is the average weekly wait time for the ***U*** broadcaster exceeding 45 seconds? Figure 7 shows the corresponding sampling simulation. Notice the differences in the summary statistics between the native distribution, the sample distribution and the sampling distribution for the mean. Because of CLT, the answer of this exercise may then be computed approximately using the SOCR Normal Distribution,

$$P(\bar{X} > 45) \cong P\left(\bar{X} > 45 \middle| \bar{X} \sim N\left(\mu_{\bar{X}} = 50,\ \sigma^2_{\bar{X}} = \frac{(50)^2}{25}\right)\right) = 0.691463.$$

This chance may also be computed empirically by counting the number of weekly samples that generate an average wait time over 45 seconds and dividing this number by 50 (the total number of weeks in this survey).

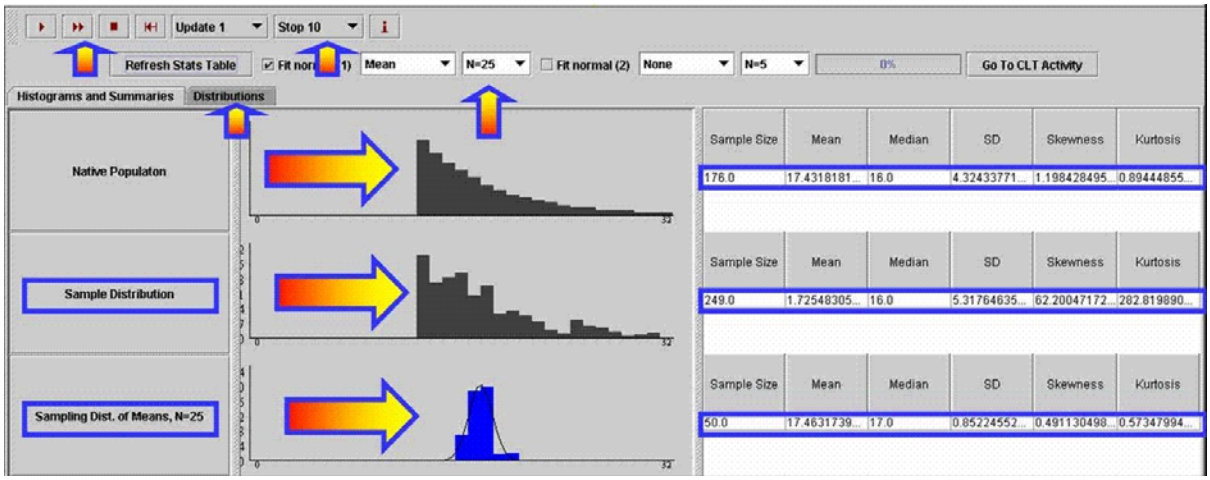


Figure 7: SOCR CLT Applet using Exponential (l=0.02) and sampling 50 samples, each of size 25.

- Suppose a player plays a standard Roulette game (SOCR Roulette Experiment, http://www.socr.ucla.edu/htmls/SOCR_Experiments.html) and bets $1 on a single number. Find the probability the casino will make at least $28 in 100 games. One way to solve this problem is to find the distribution of the casino's payoff first: If ***Y*** is the random variable representing the payoff for the casino in a single game, the probability mass function for ***Y*** is given by $P(Y=1)=\frac{37}{38}$ and $P(Y=-35)=\frac{1}{38}$, as there are 38 numbers in total (0, 00, 1, 2, …, 36). The player may place a bet on any of these numbers, with a player success payoff of $35 (casino loss of $35) and a player loss of $1 (casino win of $1).

  Therefore, the casino expected return of the game is $\mu_Y = E(Y) = \frac{2}{38} = 0.05263$ and the variance of the casino return is $\sigma_Y^2 = V(Y) = 33;\ (\sigma_Y = 5.8)$ and the range of the return is [-35 : 1], for one game.

The exact probability of interest may be computed by using the Binomial Distribution. If the total casino return in 100 games is denoted by $T = \sum_{i=1}^{100} Y_i$, then the expected casino return in 100 games is $5.26 and *P(T>28)=P(X>k)*, where *X~Binomial(p=* $^{37}/_{38}$*=0.97368, n=100)* and *k* is the integer solution of the following dollar amount equation: *k* × *$1* - (100-k) × $35 = $28, k=98. Therefore, $P(T \geq 28)=P(X \geq 98)=0.508326$. The last probability represents the exact solution and is computed using the SOCR Binomial Distribution applet (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html). This exact calculation is numerically intractable for large sample-sizes (*n>200*), even though approximations exist. Again, one could use the CLT to find a very good approximation to this type of probabilities. For example, in the case above (*n=100*), we can estimate the probability of interest by

$$P(T \geq 28) \cong P\left(T \geq 28 \,\middle|\, T \sim N\left(\mu_T = n \times \mu_Y = 100 \times 0.05263,\ \sigma_T^2 = \sigma_Y^2 \times n = (5.8)^2 \times 100\right)\right) = 0.347.$$

Notice that this calculation is sample-size independent, and hence widely applicable, whereas the former exact probability calculation is limited for small *n*. What caused the large discrepancy between the exact and approximate values of the probability of interest: *P(T ≥ 28)= 0.508326* and *P(T ≥ 28) ≈ 0.347?* This is an example where the usual rule of "30 measurements" breaks, because of the skewed underlying Binomial distribution. Such limitations of the CLT even for large sample-sizes have been previously observed and reported for severely skewed distributions (Freedman, Pisani, & Purves, 199). Here one would need much larger sample to get a reasonably good approximation using CLT. For example, if *n=1,000* and we are looking for *P(T>100)*, then *k=975*, the exact probability is *P(T>100)=P(X>975)=0.4493287*, and the CLT approximation is much closer:

$$P(T \geq 100) \cong P\left(T \geq 100 \middle| T \sim N\left(\mu_T = n \times \mu_Y = 1{,}000 \times 0.05263,\ \sigma_T^2 = \sigma_Y^2 \times n = (5.8)^2 \times 1{,}000\right)\right) = 0.3979.$$

Finally, we show how one can use the SOCR CLT applet alone to completely empirically estimate the probability of interest, *P (T>28)*, for *n=100*. Figure 8 demonstrates how one can manually construct the native probability mass function for the random variable ***Y*** (casino payoff of one roulette game). A simple linear transformation is needed to convert the values of ***Y*** to ***W*** $\left(W = \frac{32}{36}(Y + 35)\right)$, so that the range of the ***Y*** variable [-35 : 1] may be mapped to the default range of the CLT Applet native distribution [0 : 32], ***W***. Now, recalling the definitions above (for the *n=100* case) we have that $P(T>28)=P(\overline{Y}>0.28)=P(W>31.36)\approx 0.397614$, see Figure 8. The last equality is obtained by noticing that ***W*** will have approximately Normal (μ = 31.23332, σ^2=0.48811895^2) distribution, with empirical mean and standard deviation obtained from row 3 in the table.

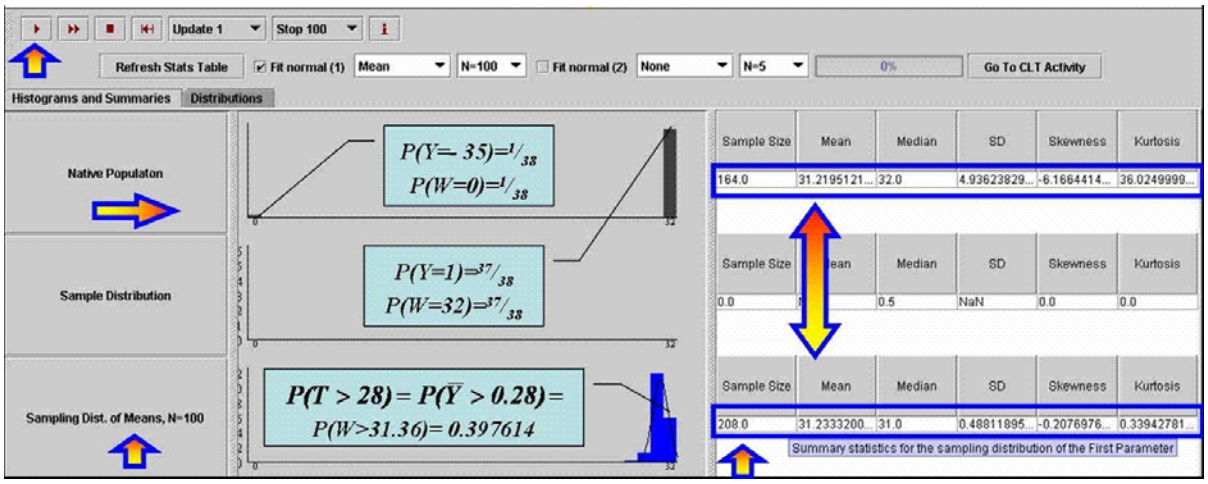


Figure 8: Demonstration of an empirical estimation of the calculation of the roulette probability using the SOCR CLT applet alone (see text).

3.3 What is unique about the SOCR CLT Activity?

A key feature of the SOCR CLT activity is that it demonstrates the fact that, most of the time, the native process distribution is unknown to the user. If one assumes that a curtain hides away the native population, one may only observe samples randomly drawn from the *unknown* process. A good analogy of this is an experiment of observing shadows (samples) of an object (process of interest) projected under different light directions. One may want to answer questions about the object (e.g., its center) in terms of the observed shadow projections (e.g., sample-driven estimates of centrality). Then the CLT allows us to quantify how biased and dispersed are these sample-driven estimates of the object-specific parameter of interest (e.g., its center-of-mass). In the case that the question of interest is the object's center-of-mass, the CLT implies that averaging the x- and y-coordinates of points in the 2D shadow projections yields unbiased estimate of the object center-of-mass, and this estimate becomes tighter as the number of projections increase.

There are two unique characteristics of the SOCR CLT applet and activity. The first one is the large number of families of distributions that are incorporated in the CLT applet. This feature may be used to clearly justify the *quasi*-independence of the CLT result from the distribution of the native process. The second distinct aspect of this activity is the broad spectrum in which the CLT can be demonstrated. An instructor may easily wrap his or her own interpretation of the CLT, tie it to their course curriculum within the SOCR Wiki framework and quickly make this available to their students and the community. For example, one instructor may want to stress the effects of the sample-sizes, whereas another may emphasize the properties of the chosen sample-statistics. And still another may focus just on the random sampling, histogram plotting, counterexamples or sample-size reasoning, all within the same SOCR CLT applet-activity framework.

3.4 Other SOCR Activities

The SOCR resource has been continuously developing other similar hands-on activities. All of these activities are paired with one or more SOCR applets and typically illustrate one possible approach for demonstrating a probability or statistics concept via a SOCR distribution, experiment, analysis, graphing or modeling applet. Most general topics covered in lower and upper division

probability theory or statistical inference courses already have available SOCR Wiki activities (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials). For example, there are two other SOCR CLT activities, which involve the Binomial and Gamma distributions. There are many other instances of SOCR activities covering the distributions, analyses, experiments, graphs and modeler applets. Finally, there are also some SOCR applets and activities, which may be used in more advanced undergraduate and graduate level classes, e.g., law of large numbers, mixture modeling, expectation maximization, Fourier and wavelet signal representation, etc.

Acknowledgements

These new SOCR CLT resources were developed with support from NSF DUE grants 0716055 and 0442992, and from NIH Roadmap for Medical Research, NCBC Grant U54 RR021813. We are also indebted to the anonymous reviewers that contributed significantly to improve the presentation of the manuscript. We also thank Annie Chu, Jenny Cui, Priscilla Chui, Rahul Gidwani, Ariana Anderson and Robert Gould for their improvement ideas and help with validation.

References

Aberson, C. L., Berger, D. E., Healy, M. R., Kyle, D. J., & Romero, V. L. (2000). Evaluation of an Interactive Tutorial for Teaching the Central Limit Theorem. *Teaching of Psychology*, **27,** 289-291.

Abrahamson, D., & Wilensky, U. (2004). ProbLab: A comsssputer-supported unit in probability and statistics, *In M.J. Hoines & A.B. Fuglestad (Eds.), Proceedings of the 28th Annual Meeting of the International Group for the Psychology of Mathematics Education*.

Andrews, D. (2005). Sampling Distributions of the Sample Mean and Sample Proportion, *CAUSE Star Library*.

ARTIST (2006). https://app.gen.umn.edu/artist/.

Ben-Zvi, D., & Garfield, J. (2004). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking.* : Springer.

BlackBoard (2006). http://www.blackboard.com/.

Blasi, L., & Alfonso, B. (2006). Increasing the transfer of simulation technology from R&D into school settings: An approach to evaluation from overarching vision to individual artifact in education. *Simulation Gaming*, **37,** 245 - 267.

Brescia, W., & Miller, M. (2006). What's it worth? The perceived benefits of instructional blogging. *Electronic Journal for the Integration of Technology in Education*, **5,** 44-52.

Chung, S.-Y., Richardson, T., & Urbanke, R. (2001). Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation. *IEEE Transactiond on Information Theory*, **47**.

ClickTV (2006). http://blog.click.tv/.

Davidson, J. (1994). The Central Limit Theorem, *Stochastic Limit Theory* (pp. 345-347): Oxford Scholarship Online Monographs.

DeGroot, M. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

Demir, S. (2006). Interactive Cell Modeling Web-Resource, iCell, as a Simulation-Based Teaching and Learning Tool to Supplement Electrophysiology Education. *Annals of Biomedical Engineering*, **34,** 1077 - 1087.

Dinov, I. (2006a). SOCR: Statistics Online Computational Resource: socr.ucla.edu. *Statistical Computing & Graphics*, **17,** 11-15.

Dinov, I. (2006b). Statistics Online Computational Resource. *Journal of Statistical Software*, **16,** 1-16.

Dinov, I., Sanchez, J., and Christou, N. (2006c). Pedagogical Utilization and Assessment of the Statistic Online Computational Resource in Introductory Probability and Statistics Courses. *Journal of Computers & Education*, in press.

Formanov, S. (2002). The Stein--Tikhomirov Method and a Nonclassical Central Limit Theorem. *Mathematical Notes*, **71,** 550-555.

Forster, P. (2006). Assessing technology-based approaches for teaching and learning mathematics. *International Journal of Mathematical Education in Science and Technology*, **37,** 145 - 164, DOI: 10.1080/00207390500285826.

Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics*: W.W. Norton.

Garfield, J. (2002). The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education*, **10**.

Hall, L. J., & Nomura, Y. (2002). Complete theory of grand unification in five dimensions. *Physical Review D*, **66,** 075004.

IVTWeb (2006). www.ivtweb.com/.

Khuong, H., & Kong, H.-Y. (2006). General Expression for pdf of a Sum of Independent Exponential Random Variables. *IEEE Communications Letters*, **10,** 159-161.

Laakso, M.-J., Salakoski, R., Grandell, L., Qiu, X., Korhonen, A., & Malmi, L. (2005). Multi-perspective study of novice learners adopting the visual algorithm simulation exercise system TRAKLA2. *Informatics in Education*, **4,** 49–68.

Leslie, M. (2003). Statistics Starter Kit. *Science*, **302,** 1635.

LetsTalk (2006). LetsTalk: http://duber.com/LetsTalk.

Lohninger, H. Learning By Simulation.

Lohninger, H. (2008). Learning by Simulations, *Vienna University of Technology, Austria*: http://www.vias.org/simulations/simu_stat.html.

Lunsford, M., Holmes-Rowell, G., & Goodson-Espy, T. (2006). Classroom Research: Assessment of Student Understanding of Sampling Distributions of Means and the Central Limit Theorem in Post-Calculus Probability and Statistics Classes. *Journal of Statistics Education*, **14**.

Lyapunov, A. (1906). About a proposition in the probability theory (Sur une proposition de la théorie des probabilités), *Russian Academy of Sciences*.

Masters, J., Madhyastha, T. M., & Shakouri, A. (2005). Educational Applets for active learning in properties of electronic materials. *Education, IEEE Transactions on*, **48,** 29-36.

Merlev`ede, F., Peligrad, M., & Utev, S. (2006). Recent advances in invariance principles for stationary sequences. *Probability Surveys*, **3,** 1-36.

Mishra, P., & Koehler, M. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, **108,** 1017–1054; doi:10.1111/j.1467-9620.2006.00684.x.

MOODLE (2006). http://moodle.org/.

PBSBlog (2006). http://www.pbs.org/teachersource/learning.now.

RANDCLT (2007). Rand CLT in Action, *http://www.rand.org/statistics/applets/clt.html*.

Rossman, A., Chance, B., & Ballman, K. (1999). A Data-Oriented, Active Learning, Post-Calculus Introduction to Statistical Concepts, Methods, and Theory (SCMT), *funded by the Course, Curriculum, and Laboratory Improvement program of the National Science Foundation, award #DUE-9950476*.

RVLS http://www.ruf.rice.edu/%7Elane/rvls.html.

SAKAI http://www.sakaiproject.org/.

SOCRWiki (2006). SOCR Wiki Resource, *http://wiki.stat.ucla.edu/socr* UCLA.

Stanton, C. (2005). Central Limit Theorem Applet.

Symanzik, J., & Vukasinovic, N. (2006). Teaching an Introductory Statistics Course with CyberStats, an Electronic Textbook. *Journal of Statistics Education*, **14**.

TCEXAM TCExam Testbank: http://sourceforge.net/projects/tcexam/.

Tijms, H. (2004). *Understanding Probability: Chance Rules in Everyday Life*: Cambridge: Cambridge University Press.

UCLAVOH (2006). http://voh.chem.ucla.edu/.

UCLAX (2006). UCLA Extension Online Courses, *UCLA Extension*: http://www.uclaextension.edu.

VirtualLabs http://www.math.uah.edu/stat/.

WebWork http://www.opensymphony.com/webwork.

West, R., & Ogden, R. (1998). Interactive Demonstrations for Statistics Education on the World Wide Web. *Journal of Statistics Education*, **6**.

WikiBooks (2006). http://en.wikibooks.org/wiki/Blended_Learning_in_K-12.

Wild, C. (2006). The Concept of Distribution. *Statistics Education Research Journal*, **5,** 10-25.

Ivo D. Dinov
Department of Statistics and Center for Computational Biology
University of California, Los Angeles
8125 Mathematical Science Building
Los Angeles, CA 90095
dinov@stat.ucla.edu

Nicolas Christou
Department of Statistics
University of California, Los Angeles
8125 Mathematical Science Building
Los Angeles, CA 90095
christou@stat.ucla.edu

Juana Sanchez
UCLA Department of Statistics
University of California, Los Angeles
8125 Mathematical Science Building
Los Angeles, CA 90095
sanchez@stat.ucla.edu
