『学習院大学　経済論集』第50巻　第４号（2014年１月）

# Visualization of the Central Limit Theorem and 95 Percent Confidence Intervals

Yukari Shirota*, Sakurako Suzuki †

**ABSTRUCT**

The paper described our heuristics to teach the central limit theorem and the accuracy of estimates in business math classes. The students are not good at thinking in the abstract and have difficulties in understanding the theorem. We believe that visualization of the theorem and concepts is so helpful. In the paper, our visual teaching materials are shown. In the visual material to explain the confidence intervals, the set of red ink marks are sprayed. The centre of a circle mark is corresponding to a sample average value of one trial. With more trials, the mark color gets deeper and the resulting distribution of the sample averages gets close to a normal distribution. In another visual material for estimates in statistics, we have students count the probability the interval contains the true population mean. We think that these teaching materials by a combination of simulation and its visualization can be powerful teaching materials for the statistics estimate studies.

**Keywords**: statistics, teaching materials, Central Limit Theorem, 95 percent confidence intervals, accuracy estimates.

## 1. INTRODUCTION

The paper offers visual materials to explain 95 percent confidence intervals in the field of statistics. Students of our business math classes have difficulties to understand the concept of the 95 percent confidence intervals in our statistics classes. To understand the concept, they have to firstly understand the central limit theorem. Without the central limit theorem, we cannot explain the confidence intervals. The problem is many students cannot understand the theorem and the concept. While they thought that they had understood that, many students do not rightly understand that. We would like to make more students understand those concepts.

Although the theorem may be difficult for students to study, we think that it becomes easier with better teaching materials. Visualization of the theorem can help students understand those things with ease. The

*） Department of Management, Faculty of Economics Gakushuin University Tokyo, Japan

†） Dr Suzuki has been a part-time lecturer of Gakushuin University in 2013 and will be in 2014.

paper illustrates our heuristics to learn the central limit theorem and the 95 percent confidence intervals. The feature is visualization. In the next section, we offer our materials for the central limit theorem. Then in section three, our materials for the 95 percent confidence intervals will be shown. Finally, our conclusions will be described.

## 2. CENTRAL LIMIT THEOREM

In the section, we shall offer our teaching materials and heuristics to teach the central limit theorem. First, we show you our recommendation description of the central limit theorem [1], which means there are various descriptions/expressions of the theorem and some may not be easy for students to understand, because our target is students in business mathematics classes, not students in a mathematics department. The following is one of the best expressions that we can use in our statistics classes.

The central limit theorem states: Let X1, X2, ⋯ be an infinite sequence of independent random variables with identical distributions (Each X has mean $\mu$ and variance $\sigma^2$.) Then let $\overline{x} = \frac{X_1 + X_2 + \cdots + Xn}{n}$. Then E($\overline{x}$) =$\mu$ and Var($\overline{x}$) =$\sigma^2$/n. The central limit theorem says that, in the limit as n goes to infinity, $\overline{x}$ has a normal distribution.

The sequence of random variables is independent and identically distributed (abbreviated as i.i.d.). Namely, each random variable has the same probability distribution as the others and all are mutually independent. When we teach the central limit theorem, we should use a non-normal population variable, so that a normal distribution should be appeared finally and dramatically. Then, in our teaching materials, we use the non-normal distributed variables as shown in Figure 1. The probability distribution function is not a normal distribution. The population size is 100. The sequence of random variables is independent and identically distributed.

Our visual materials offer the simulation results of the sample average distribution with various sample sizes. On the graphics material, they can change the sample size n, dragging the slider. With a big sample size n, the distribution gets close to a normal distribution. The sampling distribution of means is very nearly normal for the sample size n $\geqq$30 even when the population is non-normal[2].

Some statistics textbooks offer simulation results. Compared to them, our simulation materials' feature is its gradual transformation of the probability distribution function. We would like to show them effects of the fluctuation of the sample size. If the sample size n is one, then the probability distribution function of the sample average is similar to the population distribution. Then with a bigger sample size n, the probability distribution function of the sample average gets close to a normal distribution. Looking at the transformation is interesting.

We have published our graphical teaching materials on our web site[1]. The materials are written in

[1] Yukari Shirota: Graphical Teaching Materials for Statistics, http://www-cc.gakushuin.ac.jp/~20010570/mathABC/.

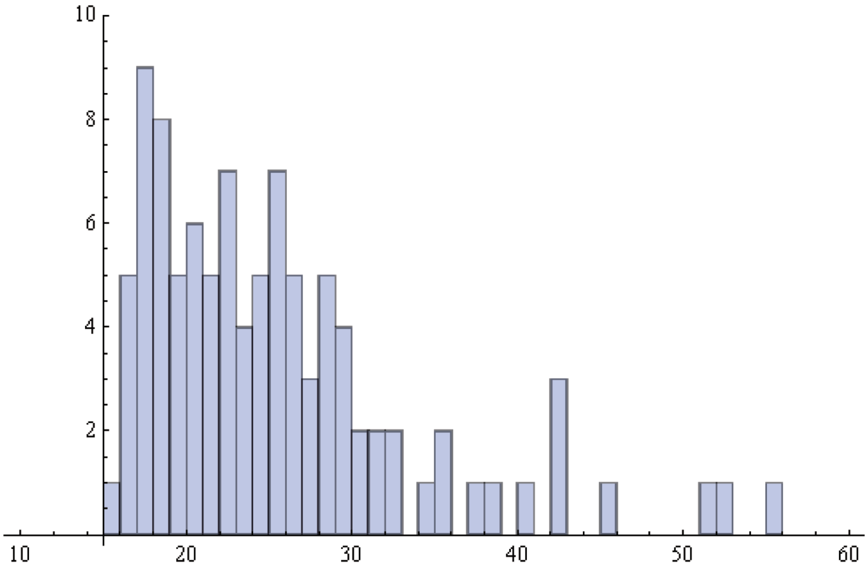
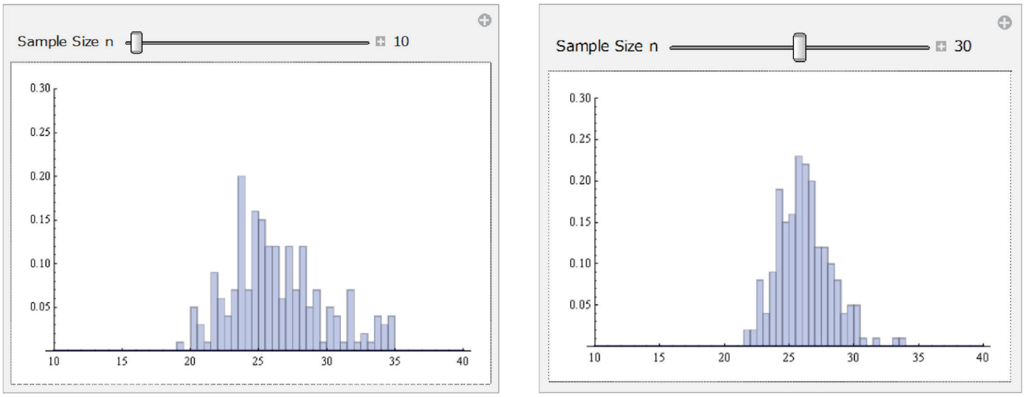


Figure 1: A non-normal population sample. In visualization of the central limit theorem, it is more effective to use a non-normal population distribution like this.



Sample Size n 10

Sample Size n 30

Figure 2: The simulation results of sample averages, $\overline{x}=\frac{X_1+X_2+\cdots+Xn}{n}$. The left one shows a probability distribution functions of the sample size n is 10. The right one shows a probability distribution functions of the sample size n is 30. When we add together many i.i.d. random variables, the resulting average will have a normal distribution. The number of trials is 200 on both trials. The simulation results show that the sampling distribution of means is very nearly normal for the sample size n ≧30 even when the population is non-normal[2].

Wolfram CDF [2] which is a free mathematical software offered by Wolfram, the MATHEMATICA company. You can execute every CDF materials on our web site without installation of the CDF player through the latest web browsers.

We shall show other examples of our teaching materials for the theorem. For other population distributions, consider (1) a uniform distribution, (2) an exponential distribution, and (3) a discrete uniform distribution. Figre 3 illustrates our teaching material with a uniform distribution and an exponential distribution.

The impressive point of these visual teaching materials is that firstly the simulation result does not look like at all like the normal distribution. However, if they increase the sample size then the density function looks more and more like the normal density function. With better simulation and visualizing tools, it is easier to see the detailed transformation to a normal distribution. The visualization can help students understand the central limit theorem. We think that lectures of statistics should present the theory visually for students before teaching the proof of the central limit theorem. We would like to spread the use of the visualization materials in mathematics education much more. Concerning financial mathematics, the author Shirota has already published an e-Book from the O'Reilley JAPAN[3]. Although the book is written in Japanese, the English version of the CDF materials are all available on the web[3] .

## 3. The 95 Percent Confidence Intervals

In the section, we would like to discuss a visual material for a 95 percent confidence interval. We lecturers will have to make students understand the concept of accuracy of estimates. In other words, we have to make our students have the answer to the question “Are the sample estimates likely to be very

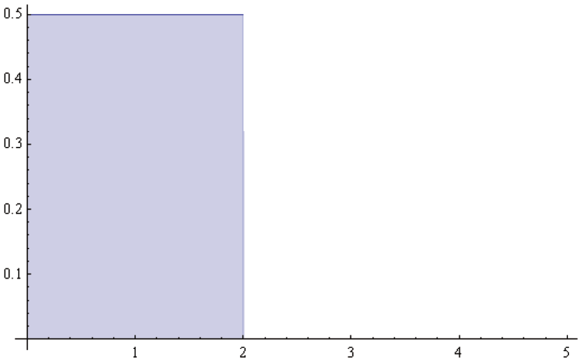
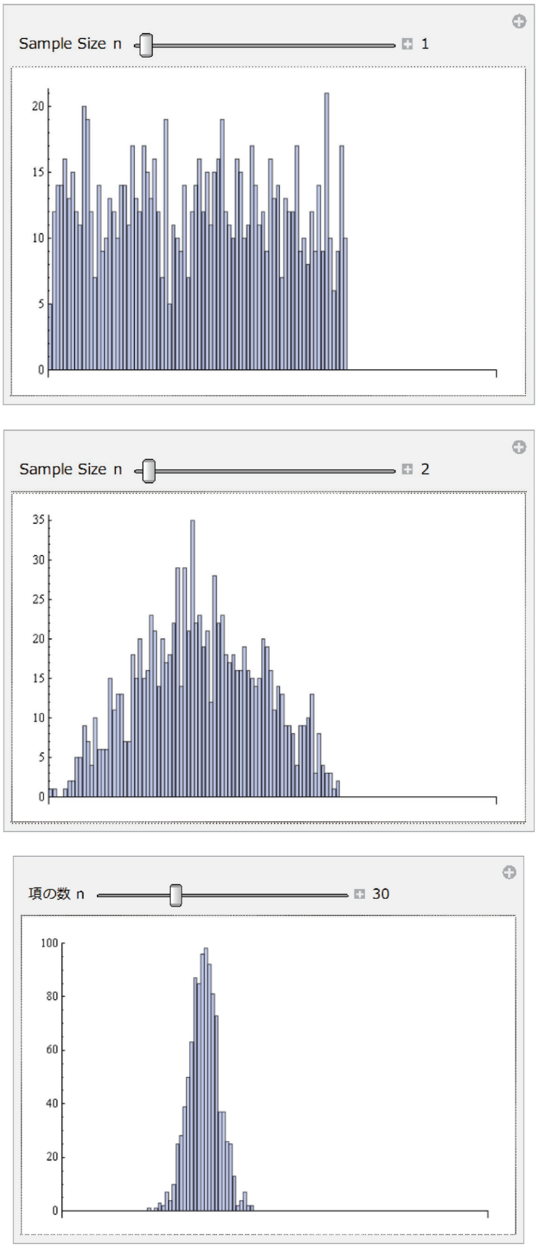


Figure 3a: The random variable has a uniform distribution like this.

[2] Wolfram: Wolfram CDF player (Interactive Computable Document Format) site, http://www.wolfram.com/cdf-player/.

[3] Yukari Shirota: English version of CDF materials in the book titled “感じて理解する数学入門” [3], http://www-cc.gakushuin.ac.jp/~20010570/private/MAXIMA/CDF/



Sample Size n 1

Sample Size n 2

項の数 n 30

Figure 3b: With the uniform distribution as the random variable distribution, the sample average simulation results are shown. The number of trials is 1000 on each simulation. If the sample size n is one, the sample average distribution is similar to the distribution of the random variables itself (See the top figure). If the sample size n is two, the simulation result illustrates a triangle shape (See the middle figure). If the sample size n is greater than 30, the distribution of the sample average gets close to a normal distribution (See the bottom figure).

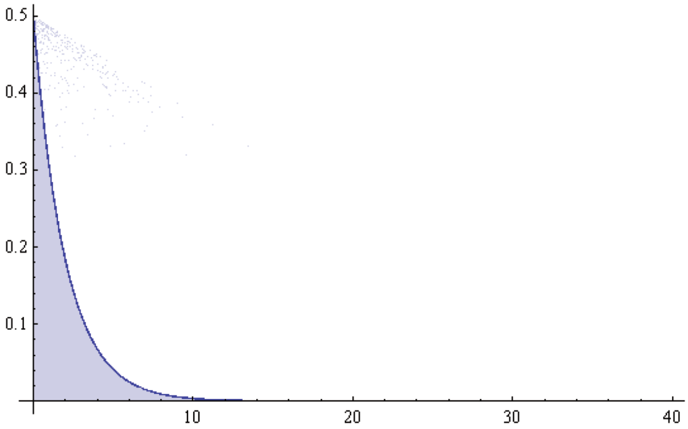
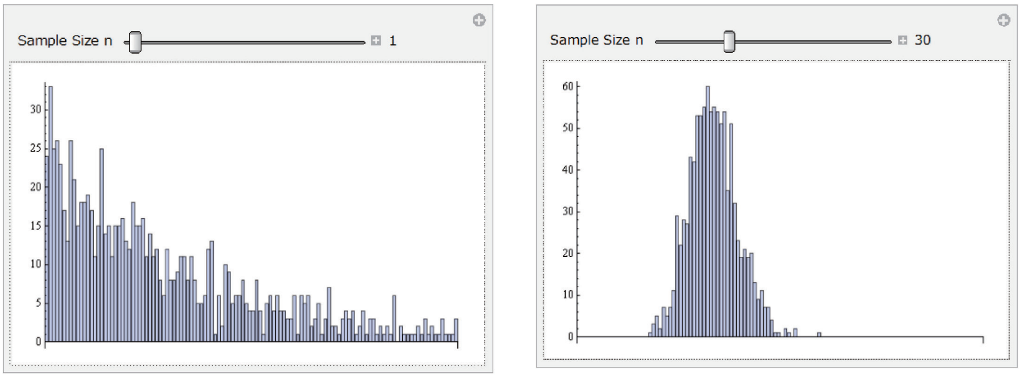


Figure 3c: The random variable has an exponential distribution like this.



Sample Size n 1

Sample Size n 30

Figure 3d: With the exponential distribution as the random variable distribution, the sample average simulation results are shown. The number of trials is 1000 on each simulation. If the sample size n is one, the sample average distribution is similar to the distribution of the random variables itself (See the left figure). If the sample size n is greater than 30, the distribution of the sample average gets close to a normal distribution (See the right figure).

close to the true population mean value?”

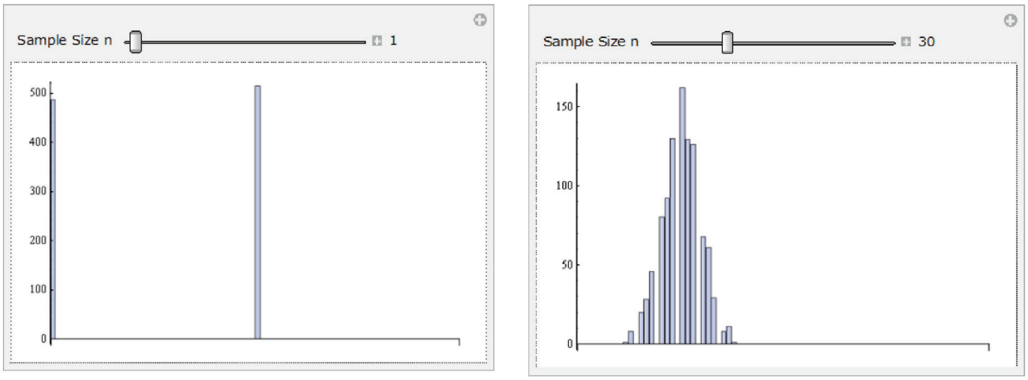
The difficult points for students in business mathematics to understand that may be the followings:

(a) The sample average is itself a random variable.

(b) The true population mean is just estimated stochastically and is unknown for all eternity.

The student may ask the following question after the lecture of the confidence intervals:

*Even if I have tried one estimation and been given a certain confidence interval for the true population mean, the confidence interval may not be very accurate. The gap between the sample average and the true mean may be large. For a closer mean value with more accurate estimates, what should I do?*



Sample Size n 1

Sample Size n 30

Figure 4: With the discrete uniform distribution as the random variable distribution, the sample average simulation results are shown. The number of trials is 1000 on each simulation. If the sample size n is one, the sample average distribution is similar to the distribution of the random variables in which Pr (x =0) =0.5 and Pr (x =1) =0.5 (See the left figure). If the sample size n ≧ 30, the distribution of the sample average gets close to a normal distribution (See the right figure).

If a student asks a question like this, we can say that he/she cannot understand the lesson. Then student is worrying about the instance like the third one in Figure 7 (a). Even if a trial offers such a sample average, including such a case, it is the 95 percent confidence level. We lectures have to make them understand the meaning of the 95 percent stochastically.

The answer to the above mentioned question is: (a) Make the sample size bigger, and (b) make trials repeatedly until the distribution of the sample averages gets close to a normal distribution (Then they can see the true mean value at the centre of the probability distribution function.)

When the population size is so large, in most cases, we can conduct the sampling just only one time. Consider the sampling from a cost standpoint. If the sampling cost is high, we may conduct the sampling just one time. Then, we should make the sample size n to be as large as possible for the narrower intervals. If we make the interval narrower, then we can be more precise about the true population average.

The theory about accuracy of estimates states that there is a 95 percent chance that the true mean value will be between $\overline{\mathrm{X}}_{\mathrm{n}}(\omega)-1.96\frac{\sigma}{\sqrt{\mathrm{n}}}$ and $\overline{\mathrm{X}}_{\mathrm{n}}(\omega)+1.96\frac{\sigma}{\sqrt{\mathrm{n}}}$. There $\overline{\mathrm{X}}_{\mathrm{n}}(\omega)$ is a random variable.

$$\mathrm{P}\left(\left\{\omega:\overline{\mathrm{X}}_{\mathrm{n}}(\omega)-1.96\frac{\sigma}{\sqrt{\mathrm{n}}}<\mu<\overline{\mathrm{X}}_{\mathrm{n}}(\omega)+1.96\frac{\sigma}{\sqrt{\mathrm{n}}}\right\}\right)=0.95$$

where $\omega_1, \omega_2, \omega_3, \cdots, \omega_k$ are a sampling sequence.

Then, after a trial, the single trial offers a fixed interval as follows:

$$\overline{x_1}-1.96\frac{\sigma}{\sqrt{\mathrm{n}}}<\mu<\overline{x_1}+1.96\frac{\sigma}{\sqrt{\mathrm{n}}}$$

where for $\omega_k$, $\overline{\mathrm{X}_n}(\omega_k)=\overline{x_k}$. The interval is not a random variable. The interval value is just a fixed value. The fixed interval may include the true mean value or not. The theory states that after many trials, the chance that an interval includes the true mean value is 95 percent.

In a realistic way, when we would like to be more precise about the true population average, we will conduct trials repeatedly. This is a cost problem. As funds permits, you should conduct many trials. Then the sample average distribution gets close to a normal distribution.

Then we shall describe our heuristics concerning the confidence intervals. To make the students understand the confidence intervals, we use the following analogy:

*Imagine the Cheshire Cat which is a famous character in Alice's Adventures in Wonderland. The cat can sometimes disappear so he can be sometimes invisible. Suppose that the invisible cat squats down on the tree branch just in from of the big white wall. You would like to find the location of the cat. Then, you thought of a good idea: Using an ink spray gun, I can mark the location. The sprayed ink mark shape is a circle of radius r. Suppose that the radius r is greater than the cat size. In addition, suppose that your gun has temperature sensors and that even if the target is invisible, when you spray the ink with the gun, there is a 95 percent chance that the ink mark contains the true location value every time. Although the cat is invisible, the sprayed ink marks gradually become deeper in color. After many sprays, you can see the unknown as a point where the color is the deepest.*

In the analogy, the corresponding relationship is as follows:

- The cat location ⟶ the true population mean value
- The ink marked circle centre ⟶ the sample average of one trial
- The ink marked circle radius r ⟶ a half of a 95 percent confidence interval

The illustration to explain the concept that we use in our classes is shown in Figure 5. The corresponding visual teaching material is shown in Figure 6.

In our lectures, after offering the illustration to make them have an image, and then, the visual tool shown in Figure 7 is offered. We have students count the probability the interval contains the true population mean, using the tool. By simulations, the 95 percent confidence intervals are illustrated on the graphics. The black dot on the interval line corresponds to the sample average. In this case, we set the confidence level to be 0.95, so the probability the interval contains the true mean gets close to 0.95 if the sample size n $\geqq$30.

## 4. CONCLUSIONS

The paper described our heuristics to teach the central limit theorem and the accuracy of estimates in

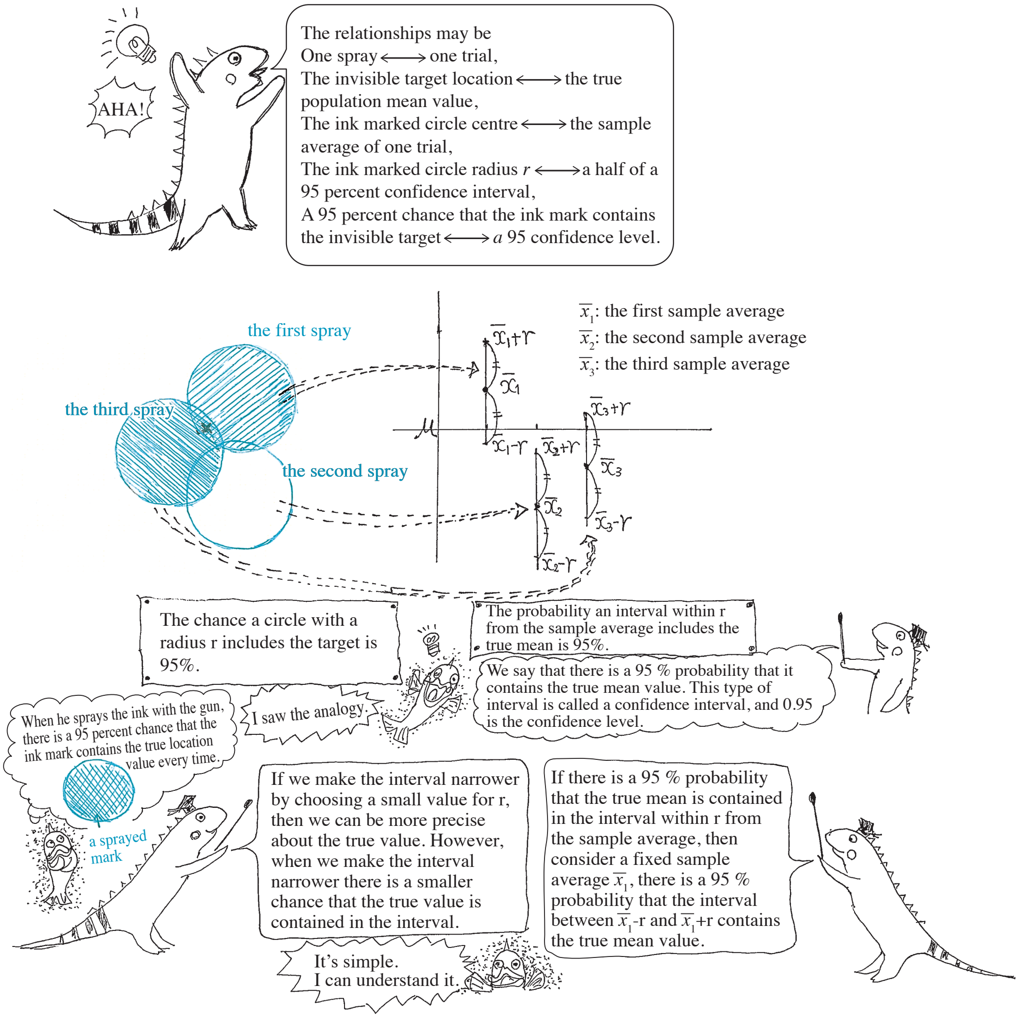
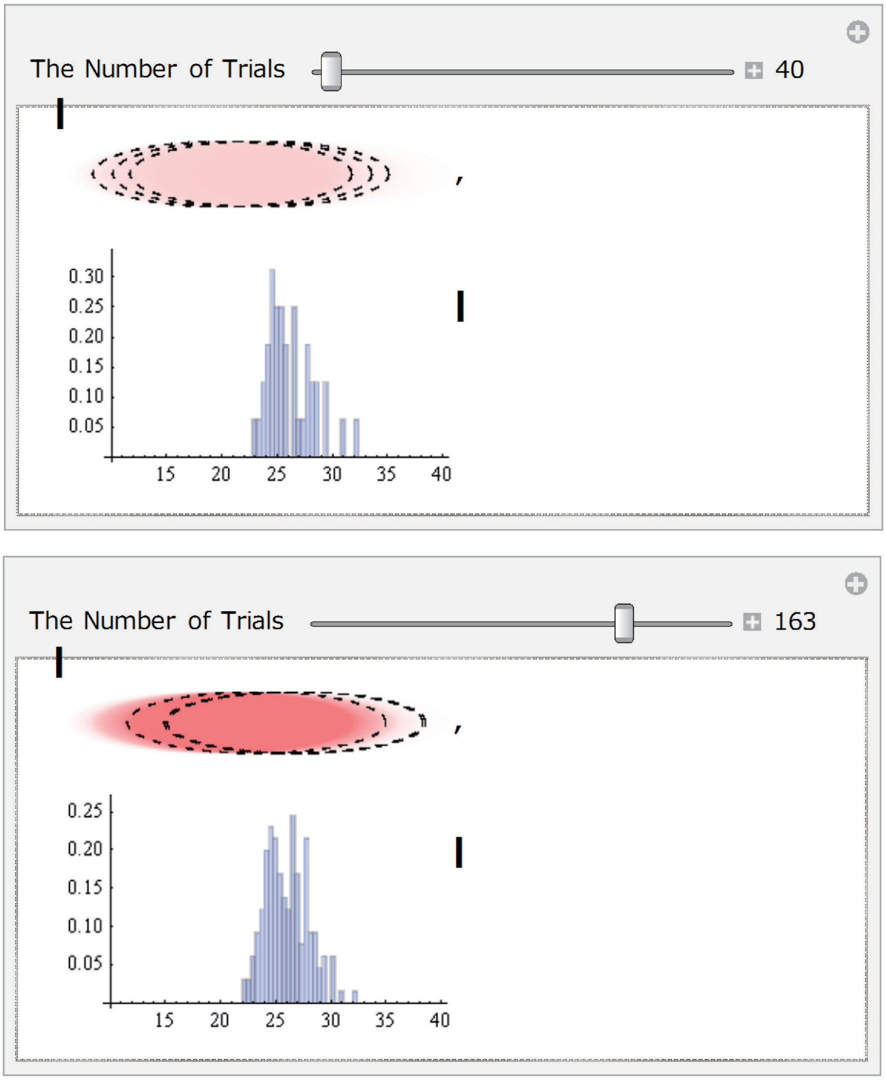

AHA!
The relationships may be
One spray ⟷ one trial,
The invisible target location ⟷ the true population mean value,
The ink marked circle centre ⟷ the sample average of one trial,
The ink marked circle radius $r$ ⟷ a half of a 95 percent confidence interval,
A 95 percent chance that the ink mark contains the invisible target ⟷ $a$ 95 confidence level.
the first spray
the third spray
the second spray
$\bar{x}_1$: the first sample average
$\bar{x}_2$: the second sample average
$\bar{x}_3$: the third sample average
$\bar{x}_1+r$
$\bar{x}_1$
$\bar{x}_1-r$
$\mu$
$\bar{x}_2+r$
$\bar{x}_2$
$\bar{x}_2-r$
$\bar{x}_3+r$
$\bar{x}_3$
$\bar{x}_3-r$
The chance a circle with a radius r includes the target is 95%.
The probability an interval within r from the sample average includes the true mean is 95%.
I saw the analogy.
We say that there is a 95 % probability that it contains the true mean value. This type of interval is called a confidence interval, and 0.95 is the confidence level.
When he sprays the ink with the gun, there is a 95 percent chance that the ink mark contains the true location value every time.
a sprayed mark
If we make the interval narrower by choosing a small value for r, then we can be more precise about the true value. However, when we make the interval narrower there is a smaller chance that the true value is contained in the interval.
If there is a 95 % probability that the true mean is contained in the interval within r from the sample average, then consider a fixed sample average $\bar{x}_1$, there is a 95 % probability that the interval between $\bar{x}_1$-r and $\bar{x}_1$+r contains the true mean value.
It's simple.
I can understand it.

Figure 5: The mathematical illustration that we use to explain the confidence intervals in our lectures.

business math classes. The students are not good at thinking in the abstract and have difficulties in understanding the theorem. We believe that visualization of the theorem is so helpful. Generally speaking, to make them understand the probability, real experiments such as tossing dices and counting the bars are important. Such repetitions make them understand the concept deeply. In addition, simulation experiments are useful because that can conduct a lot of trials in a second. We presented visual teaching materials of the central limit theorem and the 95 percent confidence intervals. We use these in our classes.

In the graphics, the sample average distribution gets close a normal distribution even if the population distribution is non-normal. Such step-by-step increase of the sample size n makes the students interested

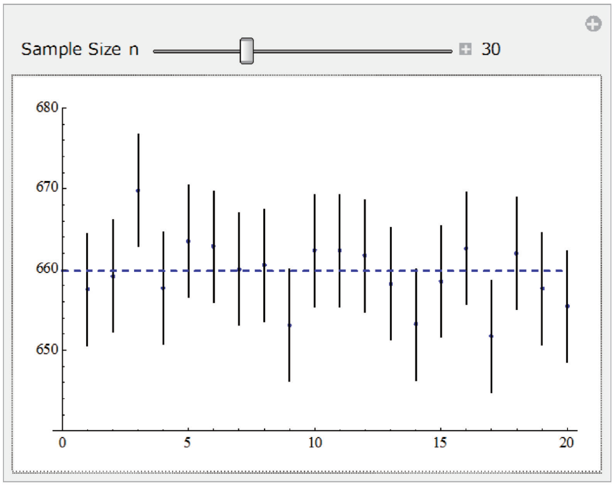


The Number of Trials 40

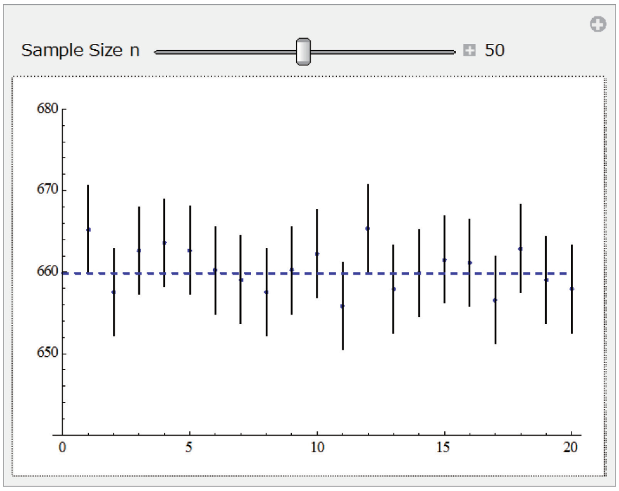
The Number of Trials 163

Figure 6: A graphics material to explain the confidence intervals. The set of red circles correspond to the set of the ink marks sprayed. The centre of a circle is corresponding to a sample average value of one trial. With more trials, the mark color gets deeper. The upper one is a 40 time trial one. The bottom one is a 163 time trial one.

With more trials, the resulting distribution of the sample averages gets close to a normal distribution. The dot lined circles are the last three sprayed marks. Sometimes, a circle centre is largely shifted from other circles.



(a) The sample size is 30. The probability the interval contains the true mean is 18/20=0.9.



(b) The sample size is 50. The probability the interval contains the true mean is 19/20=0.95. By using the tool, counting the number of the intervals that contains the true mean, the students get to understand the concept of the accuracy of estimates.

Figure 7: The visual material to count the probability the interval contains the true population mean. Although the true mean in the case is 660, firstly the true value line is hidden so that students would estimate that.

in the normal distribution and accuracy estimates.

We think that the simulation and its visualization is a very helpful material in statistics education. In old times, we could not use computer graphics. Then all the graphs were handwritten. However, now graphics can be easily drawn by mathematics software. In addition, the graphics can be operated interactively. We think that the interactive graphics should be used as teaching materials more and more. We would like to continue to develop visual teaching materials in the future.

## References

[1] D. Downing and J. Clark, *Statistics The Easy Way*: Barron's, 1989.

[2] M. R. Spiegel and L. J. Stephens, "Theory and Problems of STATISTICS (Third Edition)," in *Scgaum's Outline Series*, ed: McGRAW-Hill, 1988.

[3] Y. Shirota, T. Hashimoto, and S. Iitaka, *感じて理解する数学入門 "Introduction to Financial Mathematics" (e-Book)*: O'Reilley JAPAN, 2012.