Process Book:

Visualizing the Central Limit Theorem

CS 6630-001 Fall 2020

Vai Suliafu, Li Zhang

1. Basic Information

Project Title: Visualizing the Central Limit Theorem

Group Members:

- Vai Suliafu, Email: u0742607@utah.edu, UID: u0742607
- Name: Li Zhang, Email: u1327890@utah.edu, UID: u1327890

GitHub Link:

Repo: https://github.com/VaiSuliafu/dataviscourse-VisualizingTheCentralLimitTheorem

Website: https://vaisuliafu.github.io/dataviscourse-VisualizingTheCentralLimitTheorem/

2. Overview and Motivation

We both share an interest in statistics from our educational background. We were both looking for a project that was related to scientific or mathematical theory so that we could learn the ins-and-outs of the theory while at the same time developing our visualization skills. Statistics presented a unique opportunity to achieve these two goals by way of statistical simulations.

In particular, we were intrigued by the idea of showing the fundamental traits of the Central Limit Theorem. Being one of the most important theorem in probability theory, central limit theorem states that, in many situations, when independent random variables are added, the sample mean tends toward a normal distribution even if the original variables themselves are not normally distributed. Central limit theorem can be, as many people realized, not entirely intuitive and hard to grasp. An interactive visualization of central limit theorem, where viewer can tune the parameters for sampling and distribution to directly see how central limit theorem works, can help establish a first-hand, intuitive understanding of the essence and fundamentals of central limit theorem.

3. Related Work

Our inspiration was to create an educational visualization combined with user interaction. Our design was inspired by several other visualization projects

focusing on statistical and probability theories. Among them, we are most inspired by the See Theory visualization project developed by Brown University, where they created interesting visualizations allowing users to manually tune various parameters, and visualization then responds to the user input and update the view to provide straightforward, easy-to-understand charts that shows various aspects of statistical properties. For central limit theorem, viewer can control both distribution curve as well as sampling draws and size, and visualize how central limit theorem works with different parameters.

Link: https://seeing-theory.brown.edu/probability-distributions/index.html

4. Questions and Objectives

The primary questions we wish to answer through our visualization are:

- What is central limit theorem?
- How does central limit theorem work?
- What types of distributions converge under the central limit theorem?
- How does the sample size and number of draws affect convergence?
- At what point is a sampling distribution of means indistinguishable from the normal distribution?
- What parameters can be used to quantify the goodness of fit for sampling distribution vs theoretical normal distribution?

For this visualization, our primary goal is to communicate an intuitive and visual understanding of the central limit theorem so that it can be understood by audiences without statistical education. From this visualization, we wish to help audiences establishing a deep understanding of the Central Limit Theorem, one of the most important theorem in probability theory. Therefore our main objectives for creating this visualization include:

- Developing a much deeper, more intuitive understanding of the Central Limit Theorem
- Review our knowledge of probability distributions as well as how to generate kernel density plots for these various distributions
- Develop our skillset with interactive sliders and transitions
- Develop our skillset with linked views
- Develop our skillset with transitions
- Strengthen our intuitions for convergence with different distributions and distribution parameters

5. Data and Data Processing

We will not use existing datasets, but instead generate simulated data points according to user inputs.

5.1 Data to Generate Distribution Curve

Some fixed number of simulated data points will be pre-loaded to generate an initial view, but the data being used for updating visualization will be generated on the fly responding to user input. Specifically, the data points being generated will represent the coordinates of a path element showing the distribution curve, and we allow user to manipulate distribution parameters (α and β) through slider bars, thereby updating the data points representing the curve and updating the visualization.

5.2 Simulated Random Data Generation for Sampling

We also allow user to manipulate sample size and number of draws via an interactive sampling menu. Randomized new data points will be generated based on user inputs. These data points may exist momentarily with an associated SVG element, but the data will be absolved into existing and fixed scalar counts. These scalar counts will be represented with growing SVG elements.

Essentially our data will be defined number of points whose attributes are manipulated by the user. We will also have temporary data elements but will be removed after transitions. Sample means and other statistical parameters needed for visualization and story-telling will be updated based on the data generated.

We do not anticipate any further data processing needs because our data will be generated by simulated random sampling. For same reason, exploratory data analysis is not needed for this project.

6. Design Evolution

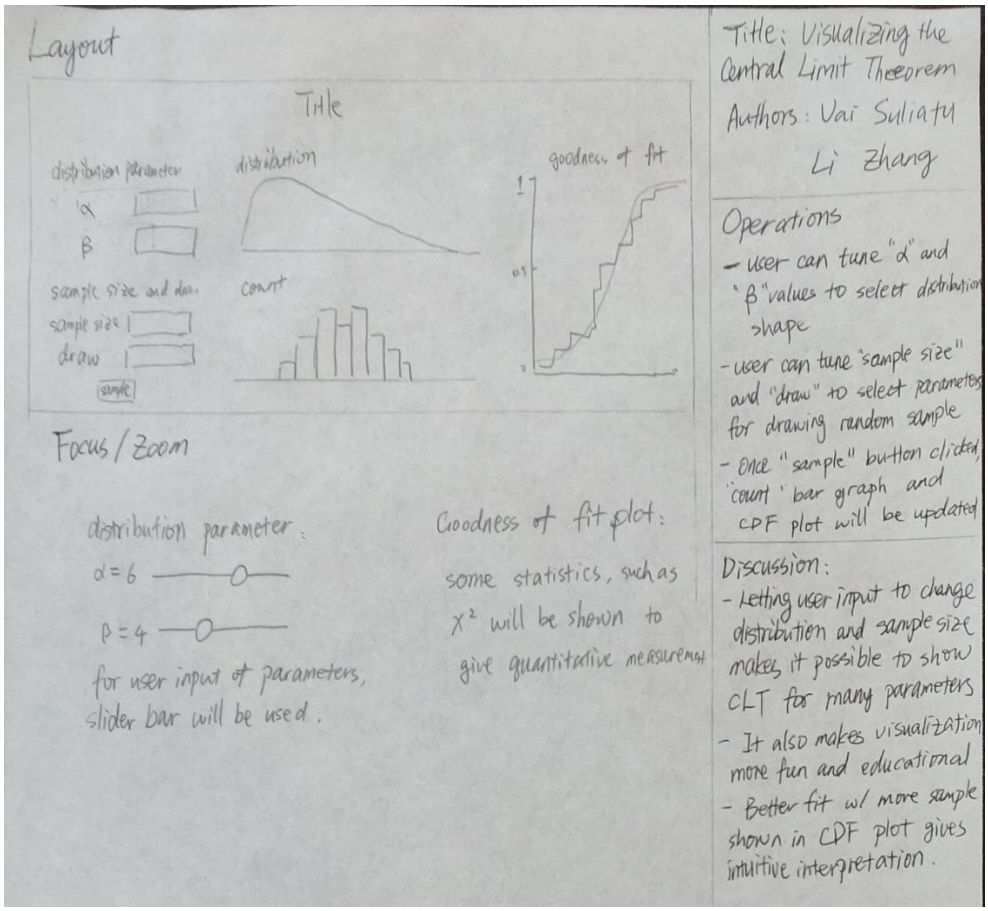
6.1 Change in Data

During peer feedback session, one of the major concern we received from the other group was on computational speed for generating random data. If the

viewer decide to draw a very large sample size, then the calculation can take considerable amount of time, and the visualization will be rendered very slowly. To address this potential issue, we decide to limit the sample size viewer can choose to be within the range of 50 – 100.

6.2 Change in Visualization Design

Our initial design presented in project proposal is shown below. It is mainly composed of 2 major views: On the left, there's a distribution curve where samples will be drew from, and a bar chart to show distribution of means from sample draw. On the right, we propose to use an ECDF curve to measure goodness of fit to compare the bar chart on the left vs the theoretical normal distribution.



During peer evaluation, an issue was raised that, though generating the visualization is non-trivial task, the visualization itself might seem to be somewhat too simple. To address this concern, we decide to add some story-telling feature

using both visualization channels and text labels. We also want to add a qq-plot alongside the ECDF plot as a second measurement for goodness of fit. ECDF plot is more intuitive and easier to understand, but the actual-vs-theory comparison is not easily quantifiable by human eyes. On contrary, qq-plot might require some statistical training to understand, but the theory distribution is shown as a straight, diagonal line, making the comparison between theoretical and actual distribution more profound.

6.2 Final Design

Our final design is a much more simplified version of what we initially proposed. We found ourselves severely limited by outside factors over the course of the project, and we had to concede many of the features we hoped to accomplish. Still, we felt we still accomplished something with the main views included. Namely, we kept the slider panel, the user configurable population distribution, a key goodness of fit plot which was an improvement over previous designs, and the sample histogram.

We believe that the main themes we wanted to communicate are still communicated, although not as obviously as we hoped.

7. Implementation

The visualization is divided in 2 major views, as shown below:
1) On the left is to visualize central limit theorem: there are slider bars for user interaction, juxtaposed with 2 graphs: an area chart to show the original distribution curve to draw sample from, and a bar chart to show the distribution of sample means.
2) On the right is to visualize goodness of fit: an ECDF plot and a QQ-plot.

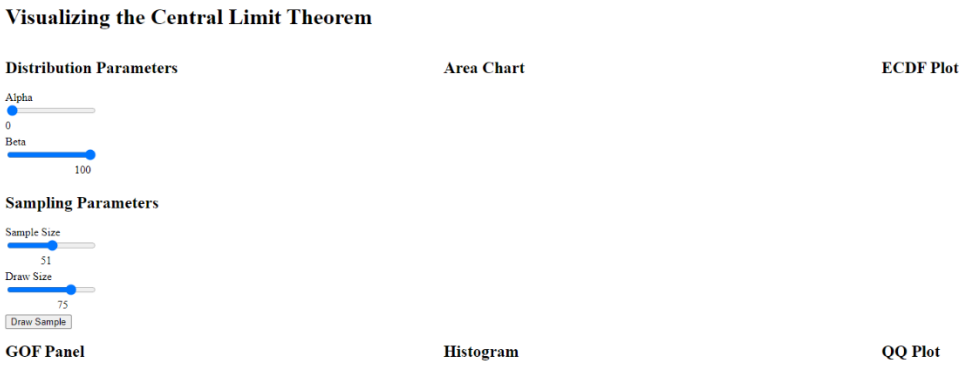
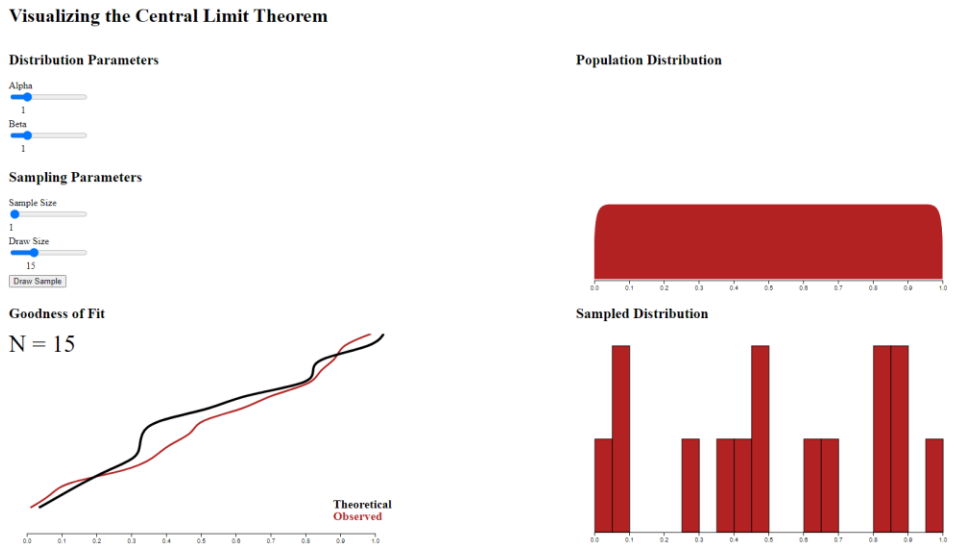


Fig 1. Overview of the complete visualization

** Ultimately do to unforeseen circumstances we had to drop the above view and opt for a simplified view with reduced features, as shown below:



7.1 Distribution Curve

An area chart is used to display the distribution where the random sample will be drawn from. The distribution curve is a beta-distribution, with its parameters α and β can be interactively changed by user using the slider bars to the left (Fig 2a). We use jStat package to calculate distribution function from α and β:

```
jStat.beta.pdf(x, alpha, beta)
```

Initial distribution curve was set at alpha = beta = 1, as shown in Fig 2b. Once the user change the alpha and beta value using slider bars, the distribution curve updates accordingly to display the corresponding beta-distribution. As example, Fig 2c-e show several distribution curves from different alpha and beta values.

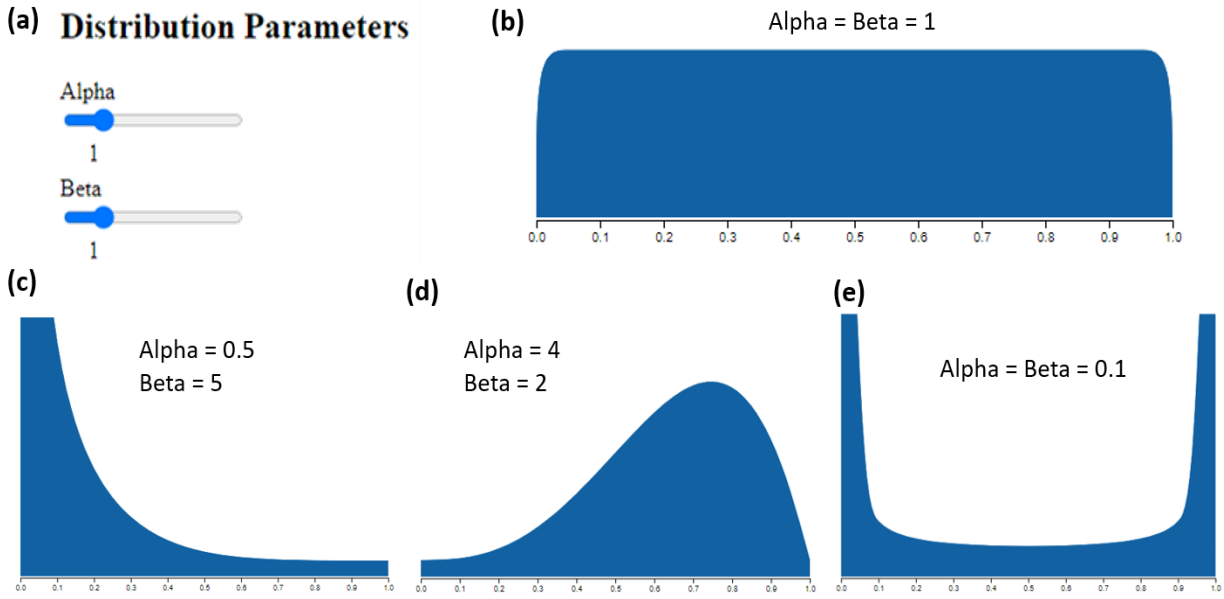


Fig 2. Distribution curve to draw sample from. (a) slider bars for user-input alpha, beta values; (b) default distribution curve when page is rendered; (c)(d)(e) distribution curves with different (alpha, beta) values.

7.2 Central Limit Theorem

To visualize central limit theorem, we allow user to control the sample size and number of draws using slider bars, as shown in Fig 3a. The sample mean distribution is displayed using a bar chart ranging from 0 to 1 with 20 bins. The default bar chart is displayed with evenly distributed bars.

Sample data will be generated based on beta-distribution defined by the user. The data are randomly generated using the function:

```
jStat.beta.sample(x, alpha, beta)
```

Once clicking the “Draw Sample” button, m sets of random data of size n will be generated and mean for each set is calculated (m: draw size, n: sample size). Fig 3b is an example of data generated from a sample draw with sample size = 10, draw size = 20. Distribution of mean is plotted into bar chart as shown in Fig 3c. If

“Draw Sample” button is clicked multiple times, the bar chart will update by accumulating all results.

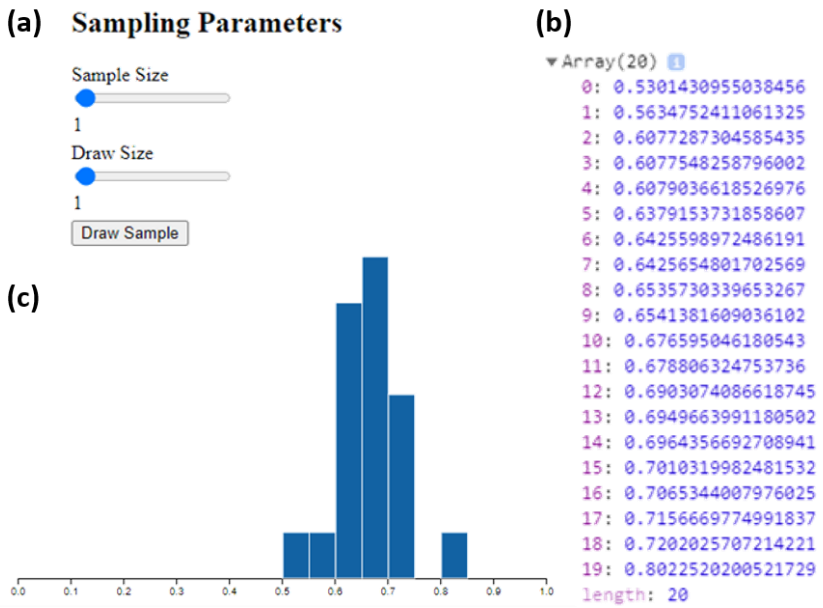


Fig 3. Bar chart visualization of central limit theorem. (a) slider bars for user-input sample size, draw size, and button to initialize random sample draw; (b) example of distribution of means in a sample draw; (c) distribution of means visualization with bar chart.

7.3 Goodness of Fit Panel

We ultimately had to drop this feature given unforeseen circumstances. We were calculating the chi-squared test statistic between the observed sample distribution and the normal distribution. Since these are real valued functions and the chi-squared test requires categorical groupings, we had to implement a binning procedure to force the real values into one of 20 bins. We then used functions from the jStat package to calculate the mean and variance of our population distribution given the current parameters alpha and beta. These mean and variance calcs then went into expected values from the normal distribution for each bin. Ultimately, we realized that the Chi-squared test procedure was not numerically stable when the number of observations in each bin was less than 5, and we did not have sufficient time to make the necessary modifications to the code.

Results of removed chi-square calculation, along with other statistics, are shown on the left in GOF panel beneath the slider bars, as shown in Fig 5.

GOF Panel

 # Draws: 1850

Chi-Squared Test Statistic: 2249.239481652558

Degrees of Freedom: 1849

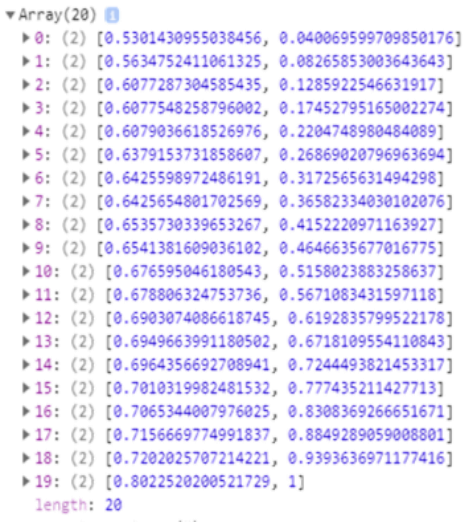
P - Value: 3.419603489263068e-10

Fig 5. Goodness of fit panel

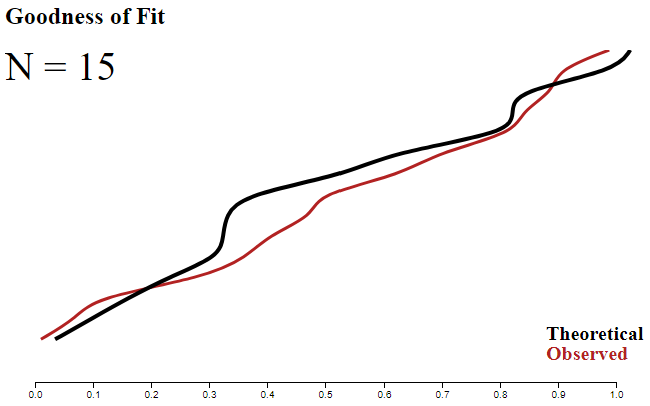
7.4 ECDF plot

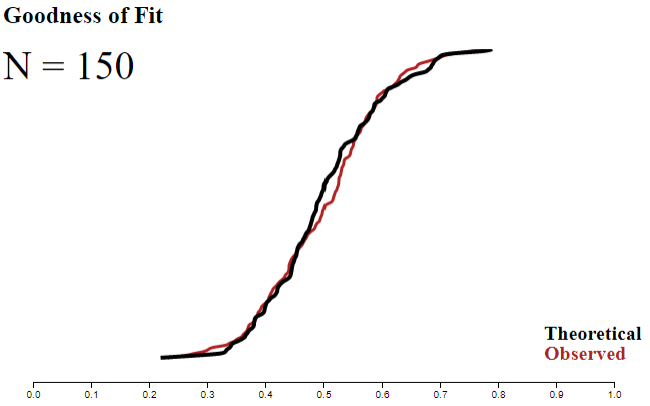
ECDF chart is used to visualize the difference between the actual sample drawn by the user, and the theoretical normal distribution. The mean distribution data from sampling is converted to accumulated frequency, as shown in Fig 4. The accumulated frequency is plotted into ECDF plot as line plot, and overlaid with theoretical cdf plot of normal distribution with mean and variance calculated from beta-distribution:

```
jStat.normal.cdf(x, mean, std)
```



Essentially we calculate the mean and variance from the samples. We then use these calculations as parameter estimations and calculate a normally distributed sample. Then we plot the normally distributed sample with the actual sample and show that the plots converge because the sample is becoming normally distributed:





7.5 QQ plot

We had to drop this feature given unforeseen circumstances.

8. Evaluation

8.1 What We Learnt from Visualization

The visualization clearly shows how the Central Limit Theorem works: regardless of the beta-distribution shape, distribution of means from random sampling follows normal distribution with enough samples being drawn. The mean and variance of the resulting normal distribution is defined by the beta-distribution. We also learned just how important some seemingly subtle assumptions are for chi-squared test. Specifically the fact that each bin needs to have at least 5 observations for numerical stability. This small nuance result in the complete removal of an otherwise complete feature.

8.2 Possible Improvements

Central limit theorem is a highly conceptual and theoretical notion sometimes difficult to be grasped by new learners in statistics. Visualization provides a powerful tool to aid understanding of this concept using visual channels. In our current implementation of visualization, the “mean of random sampling” is not yet clearly conveyed. Including transitions or animations to clearly show how random samples converged to means will greatly improve the visualization.

The summary statistics of our visualization gives a quantitative measurement of goodness of fit. It will be more informative if it can be presented in a more story-telling style, such as adding some visual elements to highlight the point where enough samples were drawn that the distribution becomes statistically indistinguishable from a theoretical normal distribution.