# An Ultimate Guide to Statistical Distributions: An Interactive Visualization Tool for Statistics Students

Bohan Song, Dongyang Wang, Peihong Zhang, Wenjin Zhang

*Department of Statistics, University of Washington*

*{bs801, dwang30, phzhang, wjzh} @uw.edu*

## ABSTRACT

Our goal is to develop a visualization tool that synthesizes information on statistical distributions, and provide a concise way to learn them effectively. We have used Python, primarily with *Altair*,[1] *pandas*, and *scipy* packages, to construct our visualization tool. The visualization contains three components, namely the distribution plots, a network graph, and text information. These elements work together to constitute the visualization tool, which explains the details about commonly used statistical distributions and elucidates the connections between them. As a result, the visualization tool will help educate statistics students, especially MS Statistics students who will soon take the theory exam.

***Keywords*:** *Data Visualization, Education, Interactive Visualization, Statististics, Statistical Distribution*

## INTRODUCTION

### *Motivation*

Familiarity and understanding with different distributions are essential for students in statistics. Although there are dispersed resources online regarding this, there is no summative tool for comparison across different distributions and very few interactive visualization tools to clarify their relationships. Inspired by Leemis's explanation of the relationships between univariate distributions,[2] we are motivated to create a visualization tool.

### *Goal*

Among other things, the probability density/mass functions and the cumulative distribution functions are the building blocks of statistical distributions. A few other important topics include the moment generating function, support, mean and variance, etc. Therefore, our goal is to create a visualization tool to help statistics students understand how each of the commonly used distributions can be visualized and how they relate to each other. Ideally, this tool can also be used to prepare for the MS theory exam for statistics students.

## DATASET

Our dataset has been generated with the *scipy* library in Python. From our prior knowledge and online resources about classic forms of a distribution, we adjusted the parameters to display

| Distribution | type | # of parameters | P1 | P2 | # of transformations |
|---|---|---|---|---|---|
| Normal | continuous | 2 | $\mu$ | $\sigma^2$ | 3 |
| Exponential | continuous | 1 | $\lambda$ | Null | 3 |
| Gamma | continuous | 2 | $\alpha$ | $\beta$ | 3 |
| Beta | continuous | 2 | $\alpha$ | $\beta$ | 2 |
| Uniform | continuous | 2 | a | b | 1 |
| Chi-square | continuous | 1 | n | Null | 3 |
| F | continuous | 2 | dfn | dfd | 1 |
| Binomial | discrete | 2 | n | p | 1 |
| Poisson | discrete | 1 | $\lambda$ | Null | 2 |
| Geometric | discrete | 1 | p | Null | 2 |

Table 1: A list of distributions included

[1] Zsailer. *nx_altair*. GitHub. https://github.com/Zsailer/nx_altair

[2] Leemis, Lawrence. *Introduction to" Probability"*. Lawrence M. Leemis, 2018.

various shapes of PDF and distinct curvature of CDF. Our data is of size 4693 by 9, 4693 data points and 9 attributions for each data point.

The above table includes a list of distributions we included in the visualization tool, their type, number of parameters, what the parameters are, and number of transformations they have with other distributions.Note 10-11 points are generated for discrete distribution and 200-500 points are generated for continuous distribution..

## METHODS

We have used Python Altair for most of our visualization designs. We also made use of GitHub for data storage and hyperlink connection, as well as GitLab as part of the publication process.

While deploying the visualization tool, we mainly had three elements: the basic statistical distribution plots, namely for the PDF/PMF and CDF; the network graph; and the information part for each particular distribution.

### *Visualization Elements*

For the network graph, we used Altair network nodes chart. We first generated the "source to target" data frame indicating the relationships between distribution nodes. Then we got the network chart by allocating the position of nodes and combining the nodes chart and edges chart. Each node represents a certain type of distribution with corresponding legend color. The size of the node is proportional to its connecting edges.

Based on the simulated data we built each distribution with different parameters, we sketched out their PDF/PMF and CDF plots separately. We implement color encoding to group distributions of the same type and dashed lines to differentiate discrete distribution from continuous distribution.

The detailed information of distributions is an image mark scatter plot with each image located in (0, 0), thus the images of detailed information will be displayed alternatively with selection of different distributions. The images are written in Latex and uploaded to Github.

The distribution information, namely the details and properties section, has been drafted in Overleaf, a Latex tool. The information has been obtained from various sources, including textbooks, websites, and Wikipedia. Later they were compiled and generated as images and published on GitHub, and were loaded into our visualization tool through urls.

### *Interaction*

Audiences are able to make selections of distribution of their interest among the network graph, PDF/PMF and CDF charts according to the legend colors and tooltips. The unselected elements display opacity effects and none elements are selected in the original status of the graphs. By clicking nodes on the network graph, the selected distribution will be highlighted and corresponding PDF/PMFs, CDFs and detailed information on other graphs are displayed simultaneously. Selections made on PDF/PMFs or CDFs will lead to the same results as those made on the network graph.

After selecting a certain distribution by clicking on the node in the network or clicking on the lines in the distribution plots, several lines marked with selected distribution will be highlighted while others stay with low opacity. Then the viewer can choose the parameters for the selected distribution utilizing the drop down menu. The selected distribution's plot with the selected parameters will become bold, making it distinguishable from the rest.

### *Connection*

We merged simulated data and network data and got four combined charts. Then we added selection and filtering effects to each chart and concatenated them to interactive ones.

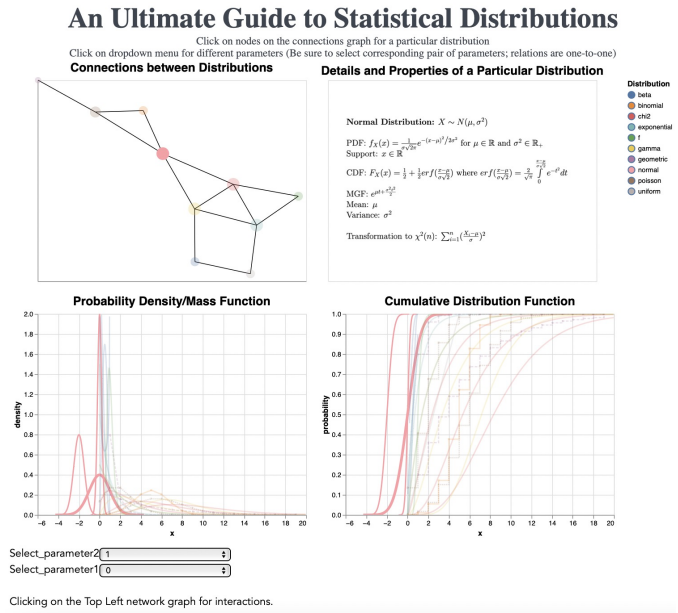


FIGURE 1: the entire visualization after concatenation.

## RESULTS

After deploying each of the above plots and connecting them together, the resulting visualization tool is ready for use. When the user is interested in a statistical distribution, s/he can select by clicking either on the nodes of the network graph, or on the line plot of a particular distribution (either the PMF/PDF or the CDF). In this way, the user will be able to observe how each distribution compares and connects with other distributions, as well as how it will distribute on its own support.

The goals of this tool have therefore been achieved, as three layers of knowledge have been synthesized and compiled into this one single visualization tool. First, the interaction within distribution plots enables users to make comparisons inside the same distribution with different parameters. This helps them understand visually how each individual distribution will change when parameters are adjusted. Second, the text information provides details about each distribution. If the user wants to use a PDF or find the variance, no calculations are needed. A single click on the node on the network graph or the line plots will suffice for basic mathematical knowledge of each distribution. Third, the network graph provides a clear illustration of the relationships between distributions of interest. If the user, after sampling with one distribution, intends to sample from a different distribution, one simple option is to apply the transformation as detailed in our tool, and no further generation is needed. With these being said, the user can benefit from our tool in an effective manner.

## FUTURE WORK

Based on feedback from undergraduate statistics students and MS statistics students, this visualization tool serves them well both in terms of expressiveness and effectiveness. They think that the tool overall gives a good snapshot of the commonly used distributions, and students will benefit from using this tool because of its convenience. They do also provide some valuable insights on the disadvantages of our tool, for the majority of which we found solutions. Admittedly, there remains room for future improvement.

For one, the inclusion of distributions has been selected in a way to be representative of common distributions that statistics students might encounter on a daily basis. It is far from being complete. For example, the fact that the Cauchy distribution has no expected value might be noteworthy.[3] Therefore, as students go into more advanced topics in statistics, more distributions can be included in our visualization tool to help them understand specific distributions based on their particular needs.

For the other, an upgrade to the existing visualization might take place by utilizing a different programming language or data visualization platform. D3 and Vega-Lite might improve the user experience for interactions, although the data generation process might be more onerous. Several limitations of Python Altair are as

[3] Chen, Yen-chi. *Stat 512: Statistical Inference*. University of Washington. Fall 2021.

follows: the inability to include node or edge labels on the network graph, the inability to include hierarchical dropdown menus for interactive selection, and the restrictions regarding text interactions.

## REFERENCES


Chen, Yen-chi. *Stat 512: Statistical Inference*. University of Washington. Fall 2021.

Leemis, Lawrence. *Introduction to" Probability"*. Lawrence M. Leemis, 2018.

Zsailer. *nx_altair*. GitHub. https://github.com/Zsailer/nx_altair


## TEAM CONTRIBUTIONS

We have clearly divided our tasks such that everyone can benefit the most from this project and make the best use of his knowledge.

We maintained meetings at a frequency of twice a week, including brainstorming, updating on what we did, and investigating the problems we had encountered. To be specific, we played different roles in the team as described below.

Bohan worked closely with Wenjin to generate the data set for the distributions characterized by parameters. He built the prototype of the bottom two interactive distribution plots where each plot is layered with a scatterplot layer and a line chart layer for discrete and continuous distribution respectively. He was able to implement a dropdown menu for parameter selection onto the distribution plots. He helped Peihong build interaction between the network graph and the distribution plot.

Dongyang was responsible for collecting necessary information for successfully starting this project, such as researching project ideas, gathering references and coding resources. He is also responsible for providing a guideline for data generation, as well as starting part of the Latex and this writeup. Besides, he was in charge of communications with Eunice Jun, who provided advice on this project.

Peihong was responsible for the network graph. He generated the relationships data between nodes and relative data used in the network chart and image mark chart, and merged them with simulated distribution data. He provided the idea of using the image mark scatter plot with each detailed information of the distribution image located in the same location. He also tackled some interactive problems within and between charts including setting the initial selection and opacity status.

Wenjin mainly shared the responsibility of drawing distribution and density plots with Bohan. He solved the restriction that data should be less than 5000 rows by generating and storing data on GitHub and input the dataset through url. He also contributed to coming up with the plot with the interaction that the line nearest to the mouse can become bold. He cooperated with Bohan in polishing the distribution and density plots.