UCLA
UCLA Electronic Theses and Dissertations

Title
Distributome: An Interactive Web-based Resource for Probability Distributions

Permalink
https://escholarship.org/uc/item/7757h2dn

Author
Chang, Bo

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California

Los Angeles

Distributome: An Interactive Web-based Resource for Probability Distributions

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Bo Chang

2012

© Copyright by
Bo Chang and SOCR Resource
CC-By Licensed
2012

Abstract of the Thesis

Distributome: An Interactive Web-based Resource for Probability Distributions

by

Bo Chang

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Alan Yuille, Co-chair

Professor Nicolas Christou, Co-chair

The advent of technology has facilitated computer based techniques for statistical analysis of data. In addition, the web-based graphical user interface has enabled the development of highly interactive programs. The Distributome project (`http://www.distributome.org`) is an open-source, open content-development project for exploring, discovering, navigating, learning, and computational utilization of diverse probability distributions. It provides several interesting functions about probability distributions.

In this thesis we present the motivation of Distributome and the detailed design of different parts of the navigator and various tools. Also, an example is provided to show the application of Distributome in statistics education of hypergeometric distribution. Finally, we discuss the future directions of Distributome project, including expanding current tools and designing new features, and adopting the latest web development technology.

The thesis of Bo Chang is approved.

Ivo D. Dinov

Frederic Paik Schoenberg

Nicolas Christou, Committee Co-chair

Alan Yuille, Committee Co-chair

University of California, Los Angeles

2012

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Probability Distribution

In probability and statistics, a probability distribution assigns a probability to each of the possible outcomes of a random experiment. For example, when throwing a die, each of the six values 1 to 6 has the probability 1/6. In contrast, when a random variable takes values from a continuum, probabilities are nonzero only if they refer to finite intervals: in quality control one might demand that the probability of a "500 g" package containing between 490 g and 510 g should be no less than 98%. [Wik12]

The concept of the probability distribution is widely used in many areas. It is fundamental in probability theory and statistics. What's more, in physics many processes are described probabilistically, from the kinetic properties of gases to the quantum mechanical description of fundamental particles. For these and many other reasons, simple numbers are often inadequate for describing a quantity, while probability distributions are often more appropriate.

Because of the complexity of the world in nature, there exist numerous probability distributions to describe the real-world phenomena. Many of them are so important in theory or applications that they have been given specific names.

Basically, there are two types of probability distributions: discrete and continuous. A discrete probability distribution shall be understood as a probability distribution characterized by a probability mass function. For a discrete random

variable X,

$$\sum_u \Pr(X = u) = 1$$

as u runs through the set of all possible values of X. Among the most well-known discrete probability distributions that are used for statistical modeling are the Bernoulli distribution, the Poisson distribution, the binomial distribution, and the geometric distribution. In contrast, a continuous probability distribution is a probability distribution that has a probability density function. If X is a continuous random variable, then it has a probability density function $f(x)$, and therefore its probability of falling into a given interval $[a, b] \in \mathbb{R}$ is given by the integral

$$P(X \in [a, b]) = \int_a^b f(x)dx.$$

The commonly used continuous distributions includes normal distribution, exponential distribution, and uniform distribution. [KBJ00] [JKK05]

1.2 Relationships among Probability Distributions

Among these various distributions, there are mainly three types of relationships: special cases and transformations relationships, asymptotic relationships, and Bayesian relationships. [LM08] Below are some examples about the relationships.

Special cases and transformations relationships

Let X_1 and X_2 independently follow standard normal distributions. Then X_1/X_2 follows standard Cauchy distribution.

Asymptotic relationships

Let X_n follow a t distribution with n degrees of freedom. Then X_n follows an asymptotic normal distribution as n goes to infinity.

Bayesian relationships

Let X follow a Poisson distribution. If the prior distribution for the rate

parameter λ is a Gamma distribution, then X follows negative binomial distribution or Gamma-Poisson (mixture) distribution.

Another important distribution about asymptotic relationship worth mentioning is the normal distribution. In probability theory, the central limit theorem (CLT) states that, given certain conditions, the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. [DCS08]

Formally, suppose $\{X_i\}$ is a sequence of i.i.d. (independent and identically distributed) random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then

$$\sqrt{n}((\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu) \xrightarrow{d} N(0,\ \sigma^2),$$

where $\xrightarrow{d}$ denotes "converge in distribution". A sequence $X_1,\ X_2, \ldots$ of random variables is said to converge in distribution to a random variable X if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous function. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively.

Therefore, regardless of the distribution of X_i, the sum of X_i is asymptotic normal distributed. A practical consequence of the central limit theorem is that certain other distributions can be approximated by the normal distribution.

1.3 Application of Probability Distributions

Probability distribution is the fundamental concept in statistics, especially in statistical inference. Some common forms of statistical proposition includes parameter estimation, confidence interval, and hypothesis testing, all of which are based on the assumption that the experimental data follow an unknown distribution.

1.4 Motivation of Distributome

As mentioned above, probability distributions are mathematically rich objects with various applications in solving real-world problems as well as academic research. What's more, there are also quite complicated relationships among them. Therefore, a powerful tool is necessary in analyzing probability distributions. The Distributome project was initiated in 2008 by the UCLA Statistics Online Computational Resource, the UAH Virtual Laboratories in Probability and Statistics, and the OSU Mathematical Biosciences Institute. It is an open-source, open content-development project for exploring, discovering, navigating, learning, and computational utilization of diverse probability distributions. It provides several interesting functions about probability distributions.

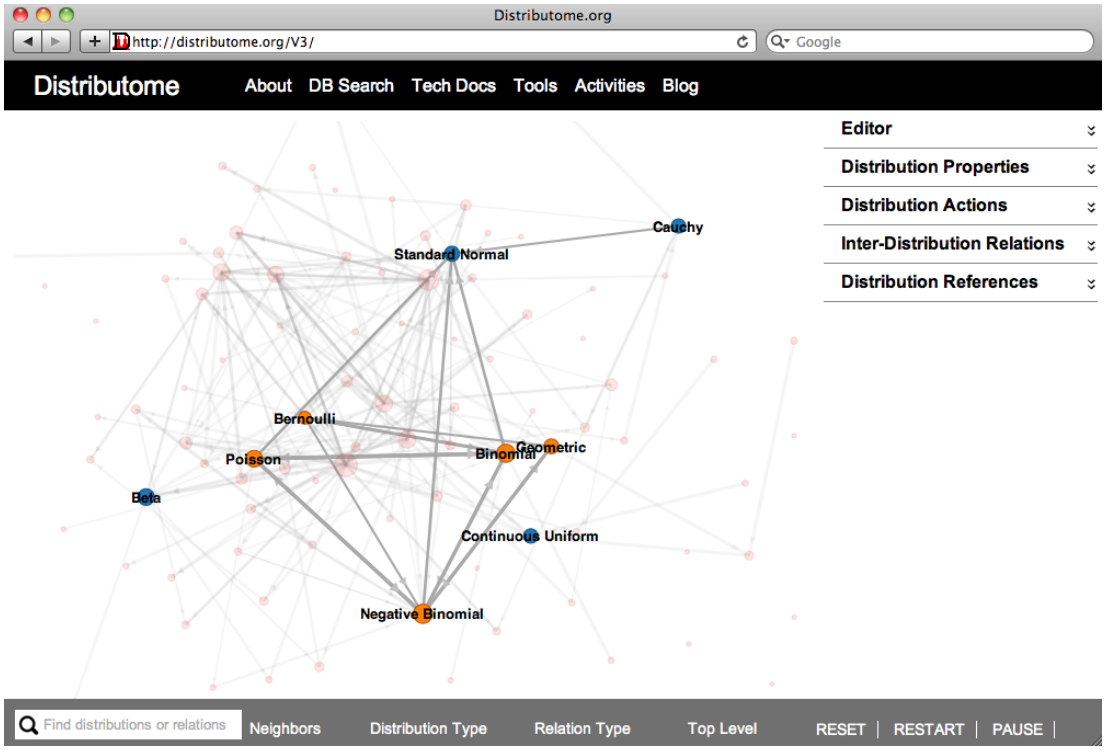


Figure 1.1: Distributome Navigator

A user can visually traverse the space of all well-defined distributions. By clicking each distribution node, the basic distribution properties are shown, including

both qualitative and quantitative information, such as density function and critical and probability values. Also, the references and additional distribution resources are included for further reading. The relationships between different distributions are denoted as edges between nodes. Similarly, if one clicks an edge, the basic information about the relationship will be shown. What's more, a search engine is included in the navigator so that one can easily search for a certain distribution by keyword, property and type. Finally, Distributome project encourages the collaboration of the community. Users can easily revise, add and edit the properties, interrelations and meta-data for various distributions, in order to make all the information proofread and up to date.

The main interface of the Distributome navigator is shown in Figure 1.1. The left panel is the network representation of probability distributions and the relationships among them. In the right sidebar panel, there are several sections, including Distribution Properties, Distribution Actions, Inter-Distribution Relations, and Distribution References. On the top, there is an probability distribution search engine, with which users can search for distributions or relations. Also, users can filter the distributions and relations according to distribution type, relation type, and display ontology. The design of each part mentioned above will be discussed in detail in Chapter 2.

CHAPTER 2

Design of Distributome

In this chapter, we describe the design of different parts of Distributome. The main interface of the latest version Distributome is shown in Figure 1.1. In Section 2.1, we are going to introduce the graphic representation of distribution relationship, in the left panel. In the right panel of the page, there are different information about the distributions and relations, which will be discussed in detail in Section 2.2 to Section 2.5. Section 2.6 talks about the bottom panel about search and filtering.

2.1 Graphic Representation

The main part of the Distributome is the distribution network in the left panel. In order to exhibit the complex network relationships among probability distributions, an elegant data visualization is necessary.

Protovis is a JavaScript library to display given digital data into graphic, dynamic forms. It's an important tool in W3C compliant computing, using largely available SVG, JavaScript, and CSS languages for data visualization. Graph visualizations often seek to reveal relationship patterns between entities and groups in the underlying dataset. [BH09]

The visualization model we used is called Force-Directed Layouts. (Figure 2.1) It is an intuitive approach to network layout modeling the graph as a physical system: nodes are charged particles that repel each other, and links are dampened

Figure 2.1: Force-Directed Layouts of Protovis

springs that pull related nodes together. Such a force-directed layout is a good starting point for understanding the structure of a general undirected graph.

In our case, the probability distributions and their relationships are represented as Figure 2.2. Each node denotes a probability distribution and an edge denotes the relationship between the neighbor distributions. Within these distributions, there is a hierarchical structure. The 3-level hierarchy is stored in a preferences file "Distributome.xml.pref".

Top Hierarchy containing only about 12 commonly-used distributions: Bernoulli, Binomial, Geometric, Hypergeometric, Poisson, Negative-Binomial, Uniform, Cauchy, Pareto, Students T, Standard Normal, Chi-square, F-distribution, and Exponential.

Middle Hierarchy consisting of about 30 common distributions, apart from Top Hierarchy: Benford, Multinomial, Laplace, Beta, Gamma, Log-Normal, Rice, Logistic, Rademacher, Arcsine, Beta-binomial, Log-Gamma, Erlang, Wald, and Levy.

All Hierarchy is an implicit category containing ALL distributions.

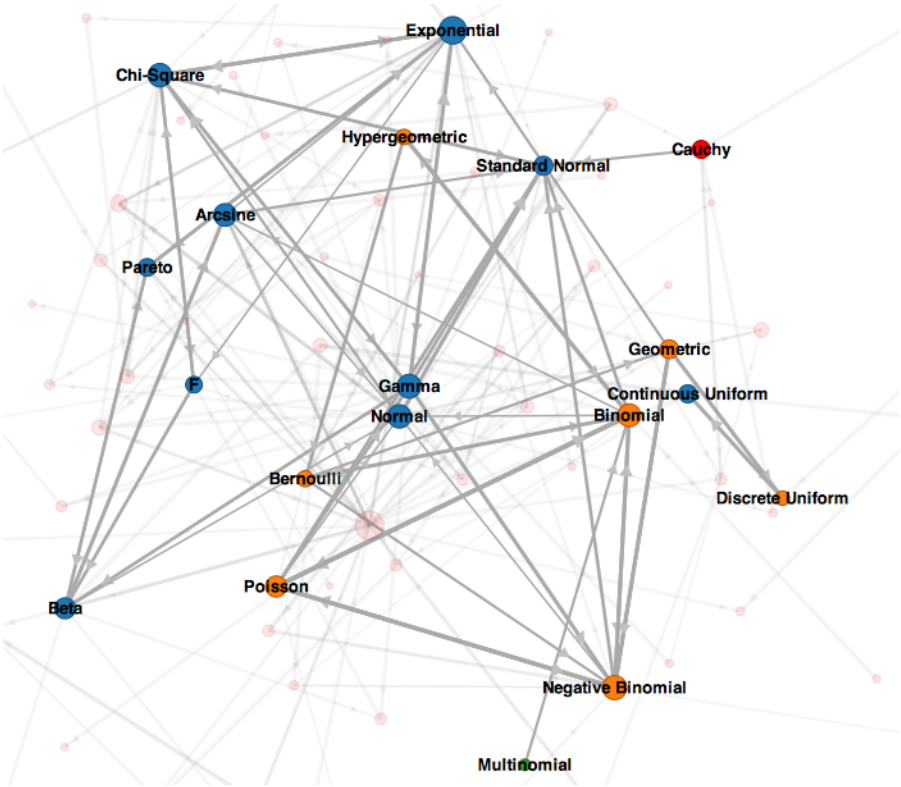


Figure 2.2: Network Representation of Probability Distributions

These hierarchies are used to simplify (or increase the complexity of) the graph of distributions rendered by the Navigator. Typically top hierarchy view is appropriate for general audiences (e.g., undergraduate students), whereas the lower-level hierarchies provide additional complexity appropriate for more advanced users (e.g., graduate or professional students). Users can view the hierarchy with the filter, which will be discussed in Section 2.6.

It is easy to move a node by pulling it with mouse. For example, if we want to see the information for a certain distribution, but the nodes are too crowded, we can simply pull out the node that interests us to make the network clear. Also, the whole canvas can be zoomed in and out with mouse scroll.

Once a node or an edge is selected, it is highlighted with red color and corresponding information is shown in the sidebar panel, including Distribution Properties, Distribution Actions, Inter-Distribution Relations, and Distribution References. We are going to discuss the distribution properties and references in the sidebar in Section 2.2 and 2.3 respectively.

2.2 XML Database

In the sidebar panel on the right, there are two sections: Distribution Properties and Inter-Distribution Relations. They are about the basic information about the distributions and relationships.

As mentioned in Chapter 1, there are numerous probability distributions and the relationships among them are complicated. Therefore, a well-organized database is needed to store the information. Extensible Markup Language (XML) is a markup language created to structure, store, and transport data by defining a set of rules for encoding documents in a format that is both human-readable and machine-readable. The XML specification defines an XML document as a text that is well-formed. [BPS97]

```
<distribution id="exponential">
    <name>exponential distribution</name>
    <name>negative exponential distribution</name>
    <type>continuous</type>
    <model>The exponential distribution models the time between random points in the Poisson model.</model>
    <parameter>\(r \in (0, \infty)\), rate</parameter>
    <support>\([0, \infty)\)</support>
    <pdf>\(f(x) = r e^{-r x}, \; x \in [0, \infty)\)</pdf>
    <mode>\(0\)</mode>
    <cdf>\(F(x) = 1 - e^{-r x}, \; x \in [0, \infty)\)</cdf>
    <qf>\(Q(p) = \frac{- \ln(1 - p)}{r}, \; p \in [0, 1)\)</qf>
    <mgf>\(\left(1 - \frac{t}{\lambda}\right)^{-1}\,\)</mgf>
    <cf>\(\frac{r}{r - i t}, \; t \in (-\infty, \infty)\)</cf>
    <mean>\(\frac{1}{r}\)</mean>
    <variance>\(\frac{1}{r^2}\)</variance>
    <skew>\(2\)</skew>
    <kurt>\(6\)</kurt>
    <entropy>\(1 - \ln(r)\)</entropy>
    <median>\(\frac{\ln(2)}{r}\)</median>
    <q1>\(\frac{\ln(4) - \ln(3)}{r}\)</q1>
    <q3>\(\frac{\ln(3)}{r}\)</q3>
    <family>exponential </family>
    <family>scale</family>
    <history>The exponential distribution was named by Karl Pearson in 1895.</history>
    <cite>siegrist2007exponential</cite>
</distribution>
```

Figure 2.3: Distribution Data in XML

In Distributome, the distribution properties and distribution relations are stored in the file *Distributome.xml*. Part of the XML database is shown in Figure 2.3, which includes the basic properties of exponential distribution, such as density function, moment generating function, and some common statistics, each as a tag. This data representation structure makes it clear and readable. Also, it is easy to display the XML data in the navigator.

When displaying the distribution properties and relations, there are always abundant mathematical equations, such as the probability density function and other statistics about the distribution. Therefore, a systematical method to display the mathematical equations is need to improve the user experience. In Distributome, MathJax is adopted to deal with this issue. MathJax is a cross-browser JavaScript library that displays mathematical equations in web browsers, using LaTeX math and MathML markup. MathJax is released as open-source software under the Apache license. It downloads with web page content, scans the page content for equation markup, and typesets the math. Thus, MathJax requires no installation of software or extra fonts on the reader's system. [Cer12] In Figure 2.4

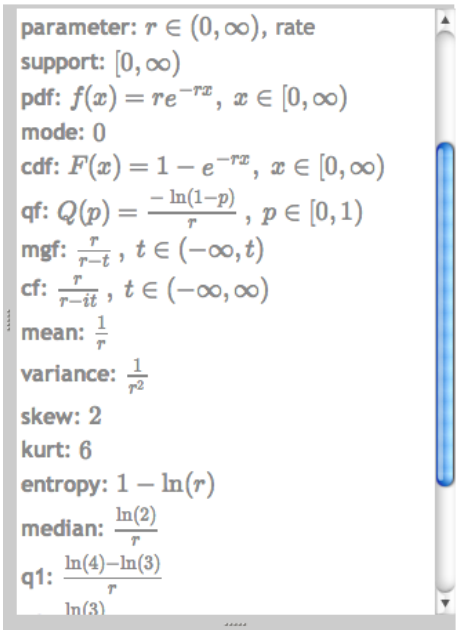


Figure 2.4: Distribution Properties Display with MathJax

we can see that the distribution properties information carried by the XML file shown in Figure 2.3 are displayed clearly as normal mathematical equations. Also, Figure 2.5 shows the inter-distribution relations between Poisson distribution and binomial distribution.

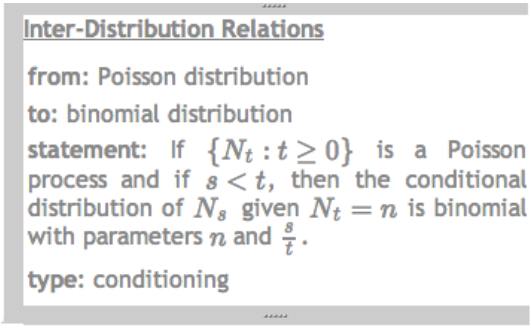


Figure 2.5: Inter-Distribution Relations Display with MathJax

2.3 BiBTeX Format

Also in the sidebar panel, there is a section called Distribution References. It is important to properly and appropriately cite references in scientific research project like Distributome in order to acknowledge your sources and give credit where credit is due. For Distributome, each distribution property or inter-distribution relation displayed has clear reference sources for further reading. In the XML database *Distributome.xml*, there is at least one tag named "cite", indicating the source index. For example, the cite tag of exponential distribution is "siegrist2007exponential", which is an index stored in the reference database.

In order to make the reference database less dependent on the XML database and expandable, we store all the reference data in a Bibtex file. BibTeX is a reference management software for formatting bibliography. The BibTeX tool is typically used together with the LaTeX document preparation system. Within the typesetting system, its name is styled as BibTeX. BibTeX makes it easy to cite sources in a consistent manner, by separating bibliographic information from the presentation of this information, similarly to the separation of content and style supported by LaTeX itself. [Pat88] BibTeX has been widely in use since its introduction. People can easily cite a reference with Bibtex.

```
@article{siegrist2007exponential,
  title={Exponential and gamma distributions on positive semigroups, with applications to Dirichlet distributions},
  author={Siegrist, K.},
  journal={Bernoulli},
  volume={13},
  number={2},
  pages={330--345},
  year={2007},
  publisher={Bernoulli Society for Mathematical Statistics and Probability}
}
```

Figure 2.6: Reference Database in Bibtex

A sample reference entry is shown in Figure 2.6. It specifies the detailed information about the reference, including title, author and so on. In the Bibtex file, each entry is corresponding to a unique index. For example, the index of the

sample entry is "siegrist2007exponential", therefore, this entry will be displayed in the section Distribution References, when the exponential distribution is selected. Figure 2.7 shows the displayed reference information in the sidebar panel. If there are multiple citation, all of them will be displayed.

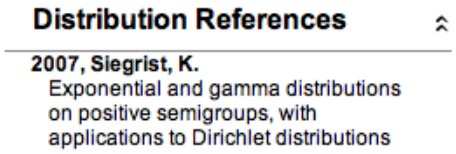


Figure 2.7: Reference Displayed in Distribution References

2.4 Editor

Distributome is a project that is open to public. It encourages users to add new information to enrich its database. In the latest version Distributome V3, there is a new section "Editor" in the sidebar. (Figure 2.8) As mentioned before, all the distribution and inter-distribution relationship properties are stored in a XML file and the reference data in a Bibtex file. With the Editor, users can expand the databases easily.

Figure 2.8: Editor Section

In Editor section, there are three buttons: "Create a new distribution", "Create a new Relation", and "Create a new Bibtex". After clicking a button, there will be a form to enter attributes of the newly created distribution of relationship,

such as distribution name, pdf, mean, variance, etc. For example, after clicking "Create a new distribution", the form in Figure 2.9 shows up. More attributes can be added by clicking "Add Attribute" button.

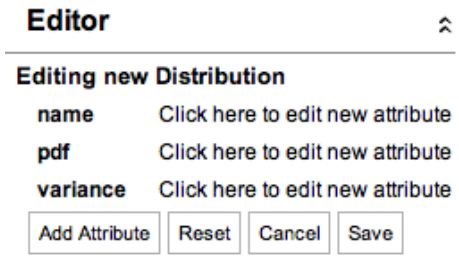


Figure 2.9: Editing new distribution

After editing, users can click "Save" button. After that, in a new browser tab, as shown in Figure 2.10, all the information entered is encoded into an XML file shown in the text field. User should enter his/her name and email address to submit the new XML to Distributome Project. After review, the submitted data will be appended to the current database.

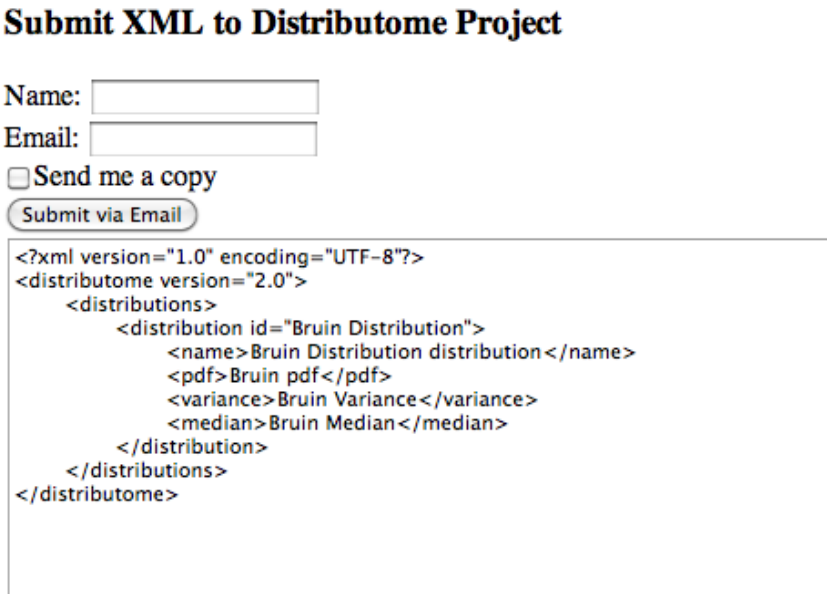


Figure 2.10: Submitting XML

2.5 Distribution Actions

In the sidebar panel, there is a section called Distribution Actions, in which there are three kind of distribution tools: Calculator, Experiment and Simulation. When a distribution node is selected, users can click on the links to the tools. (Figure 2.11) All the links are dynamic and specific for the user-selected distribution.

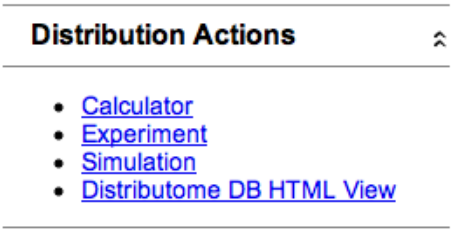


Figure 2.11: Distribution Actions

2.5.1 Calculators

For each distribution, the calculators show the probability density function or cumulative distribution function; and also calculate the probability $P(X \leq x)$, where X is a random variable with following the distribution and x is a value specified by user. Also it can do the reverse process, calculates x by providing $P(X \leq x)$.

For example, Figure 2.12 is a calculator for normal distribution. User first specifies the parameters of normal distribution, μ and σ^2, which are 0 and 1 respectively by default. There are two options in the drop-down list: PDF and CDF. Finally, user can enter the cutoff value x to get p or vice versa. In this example, we are calculating $P(X \leq 1)$; then we enter $x = 1$ and it shows the result that $p = 0.841$.

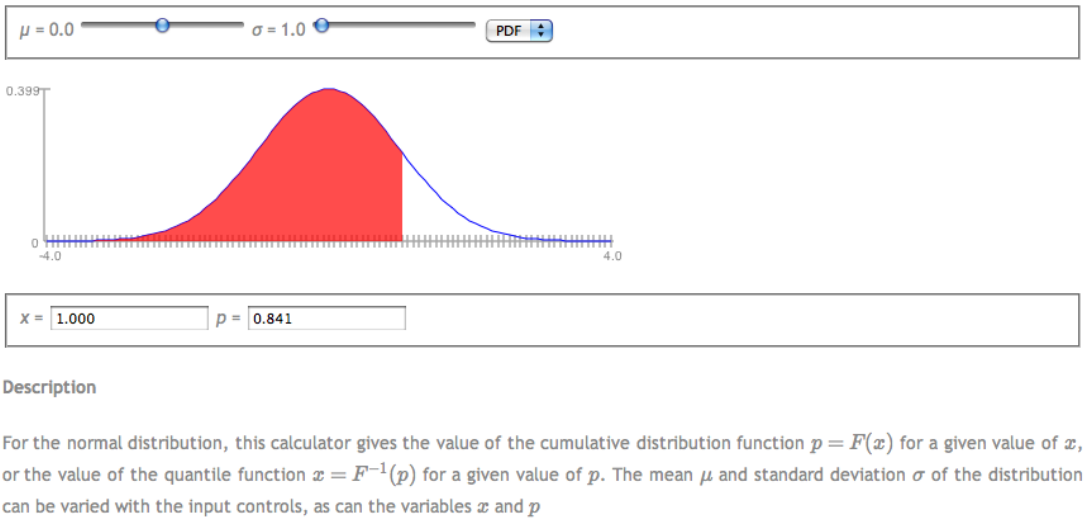


Figure 2.12: Calculators

2.5.2 Experiments

This tool shows an experiment with regard to a certain distribution graphically.

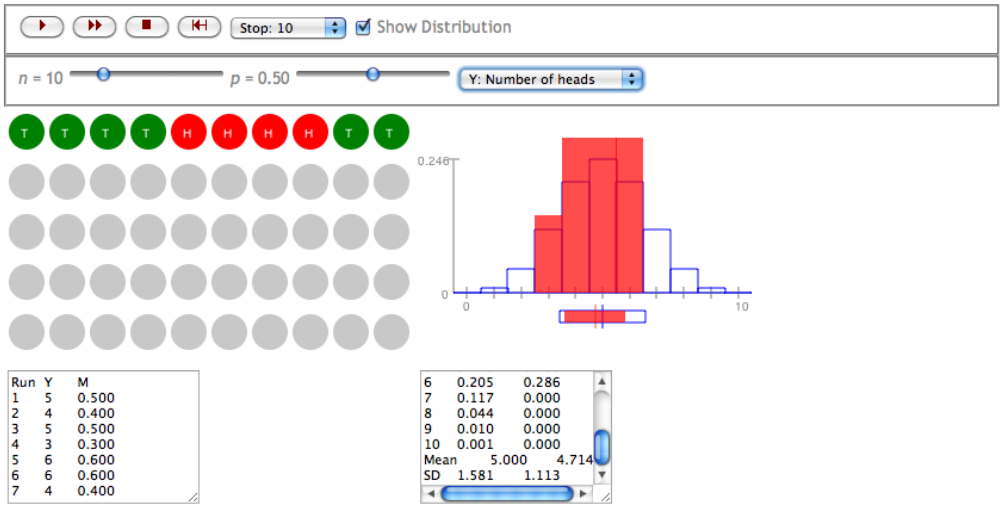


Figure 2.13: Experiments – Binomial Distribution

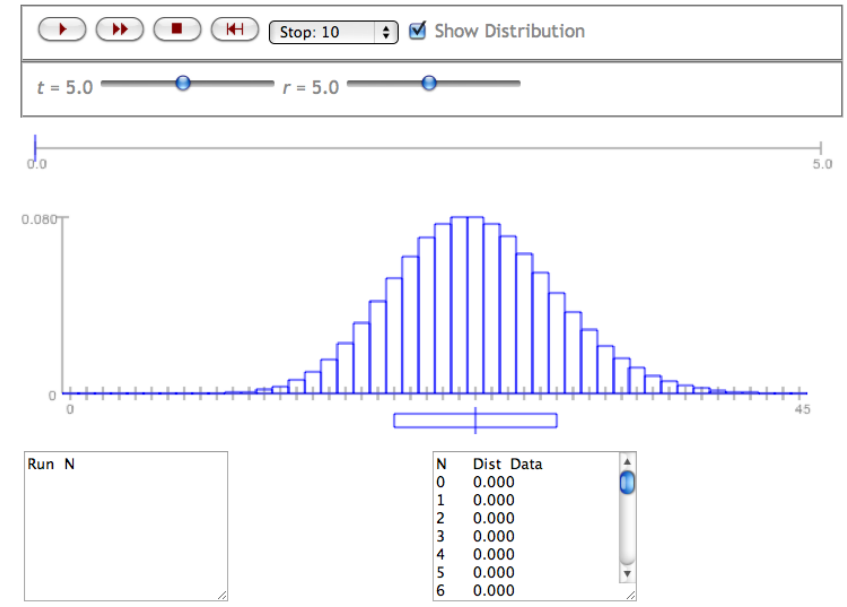
Figure 2.13 is an experiment tool for Binomial distribution. The random experiment consists of tossing n coins, each with probability of heads p. Similar with calculator tool, the parameters n and p can be varied with scroll bars. Random

variable Y gives the number of heads and has the binomial distribution with parameters n and p. The probability density function and moments of Y are shown in blue in the distribution graph blue and are recorded in the distribution table. On each run, the empirical density function and moments of Y are shown in red in the distribution graph and are recorded in the distribution table.

Another example is Poisson distribution shown in Figure 2.14, where the two figures show the tool before and after clicking the start button respectively. The experiment is to run a Poisson process until time t. The arrivals are shown as red dots on a timeline. The random variable of interest is the number of arrivals N; this variable has a Poisson distribution. The value of N is recorded on each update in the data table. The probability density function and moments of N are shown in blue in the distribution graph and are recorded in the distribution table. On each update, the empirical density function and moments are shown in red in the distribution graph and recorded in the distribution table. The parameters of the experiment are the rate of the process r and the time t, which can be varied with the input controls.

2.5.3 Simulations

The tool of simulations shows the process of sampling from certain distribution. For example, Figure 2.15 is the continuous uniform distribution simulation, where the histogram of the samples is shown in red and the PDF of the hypergeometric distribution in blue. It simulates a value of a random variable with a continuous uniform distribution on the interval $[a, a + w]$. The value is recorded on each update. The left endpoint a and the width h can be varied with the input controls.



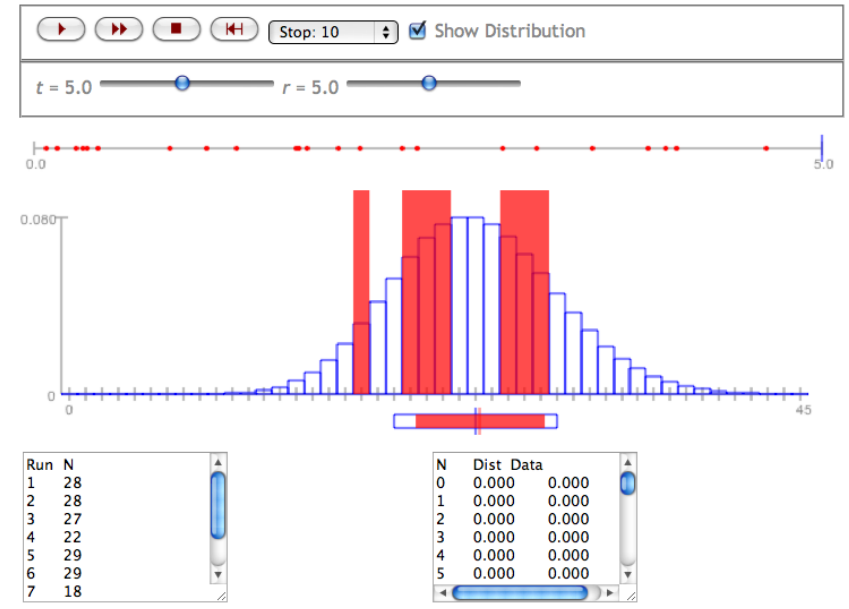
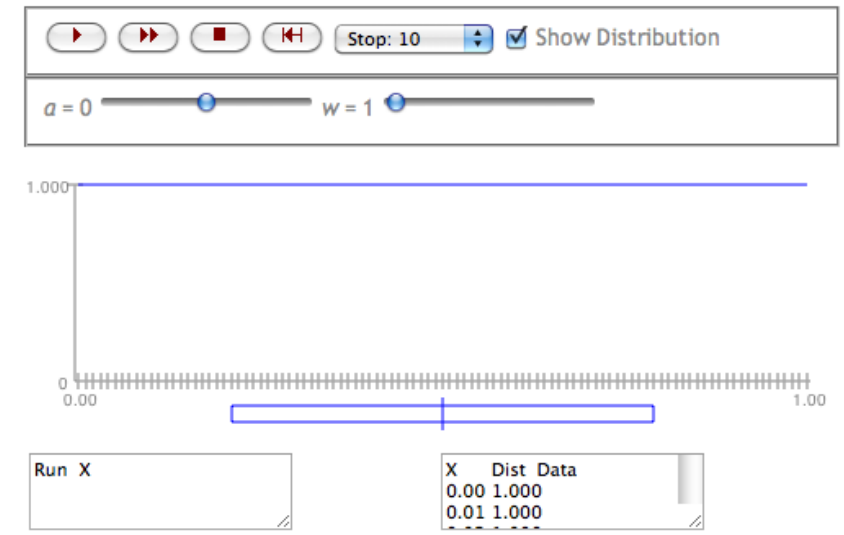


Figure 2.14: Experiments – Poisson Distribution



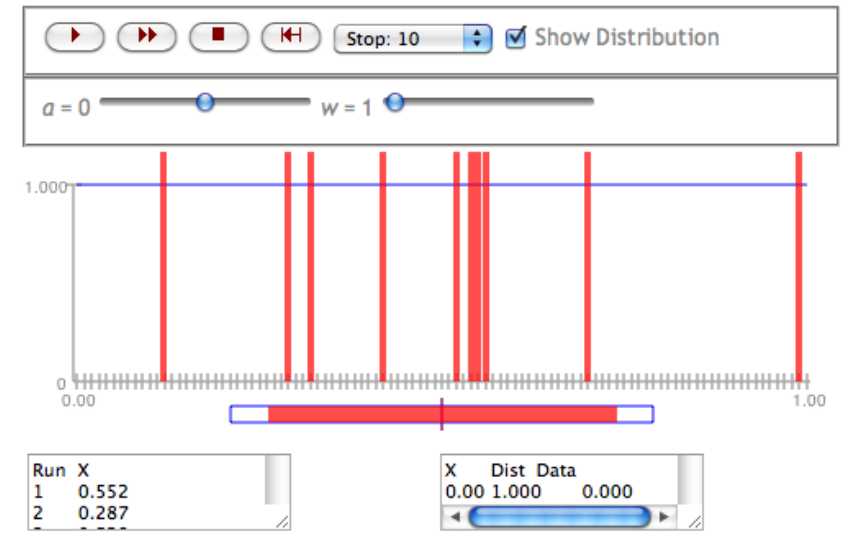


Figure 2.15: Simulations

2.6 Keyword Search and Filtering

As is mentioned, there are numerous probability distributions and inter-distribution relations stored in our database. Therefore, a powerful search engine and different filters are necessary to assist users locate certain distributions. In the bottom of Distributome navigator, there is a banner to search and filter distributions and relations.

2.6.1 Search

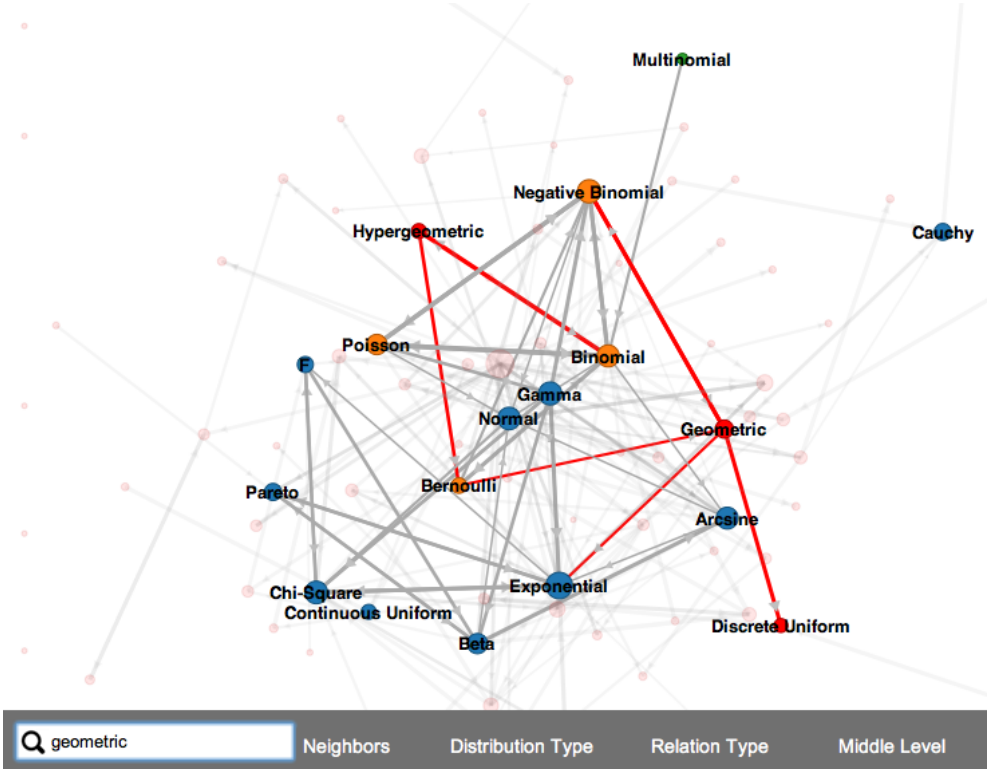


Figure 2.16: Search Engine

For each probability distribution included in the Distributome database, there is a unique URL page. The syntax of these distribution specific pages is "http://www.distributome.org/js/DistributomeDBSearch.xml.php?s=geometric", for geometric distribution as an example. Within this page is the distribution properties

and reference information. What' more, users can use AND, OR and NOT logic operators for complex Boolean search, for example Normal AND Cauchy NOT nonlinear.

Also, users can use the search engine in the bottom of the navigator to get a visual search result. Figure 2.16 shows the search result of keyword "geometric". All distributions related to "geometric" are labeled red, including geometric distribution and hypergeometric distribution, and also the relations corresponding to them.

Apart from search engine, users can use filters in the bottom of page. As is shown in Figure 2.16, there are four filters available: Hierarchical Relation, Distribution Type, Relation Type, Hierarchical Level.

2.6.2 Hierarchical Relation

There are four items in the drop-down list: Neighbors, Parents, Children, Parents & Children. As is mentioned in Section 2.1, there are three hierarchical levels, top hierarchy, middle hierarchy, and all hierarchy. For example, in Figure 2.17, the node Negative Binomial is selected, if we use the filter to select its parents, all the parent nodes and their corresponding edges are highlighted in green.

2.6.3 Distribution Type

Similarly, in the drop-down list, users can select from the items in Table 2.1 to highlight certain distributions.

2.6.4 Relation Type

Also, in the drop-down list, users can select from the items in Table 2.2 to highlight certain relations.

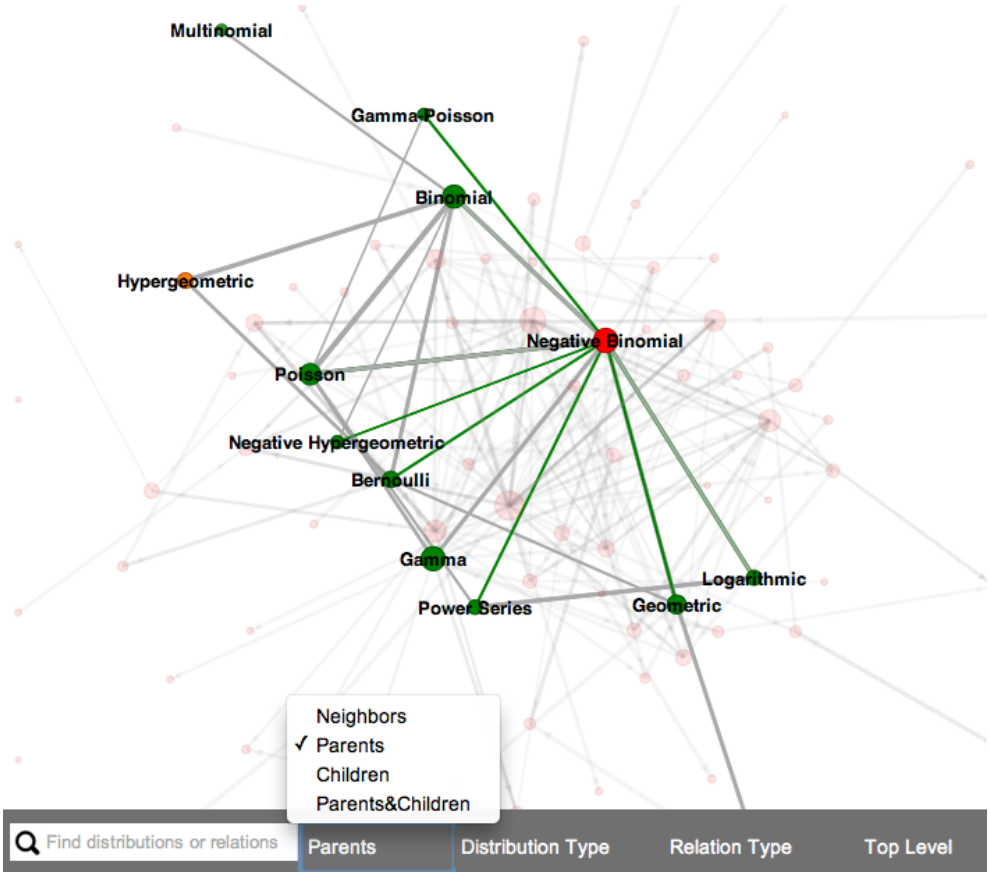


Figure 2.17: Hierarchical Relation

No.	Distribution Types
0	No Type Given
1	Convolution (Conv)
2	Memoryless (Mless)
3	Inverse (Inv)
4	Linear Combination (LinComb)
5	Minimum (min)
6	Maximum (max)
7	Product (Prod)
8	Conditional ResIDual (CondRes)
9	Scaling (Scale)
10	Simulate (Sim)
11	Variate Generation(VGen)

Table 2.1: Distribution Types (Full-Name and Abbreviation)

No.	Distribution Types
0	No Type Given
1	Special Case (SC)
2	Transform (T)
3	Limiting (Lim)
4	Bayesian (Bayes)

Table 2.2: Relation Types (Full-Name and Abbreviation)

2.6.5 Hierarchical Level

Users can use the filter to display distributions in top hierarchy, middle hierarchy, or all hierarchy. For example, in Figure 2.18, "Display All" is selected, compared with Figure 2.16 and Figure 2.17, showing the middle level distributions only.

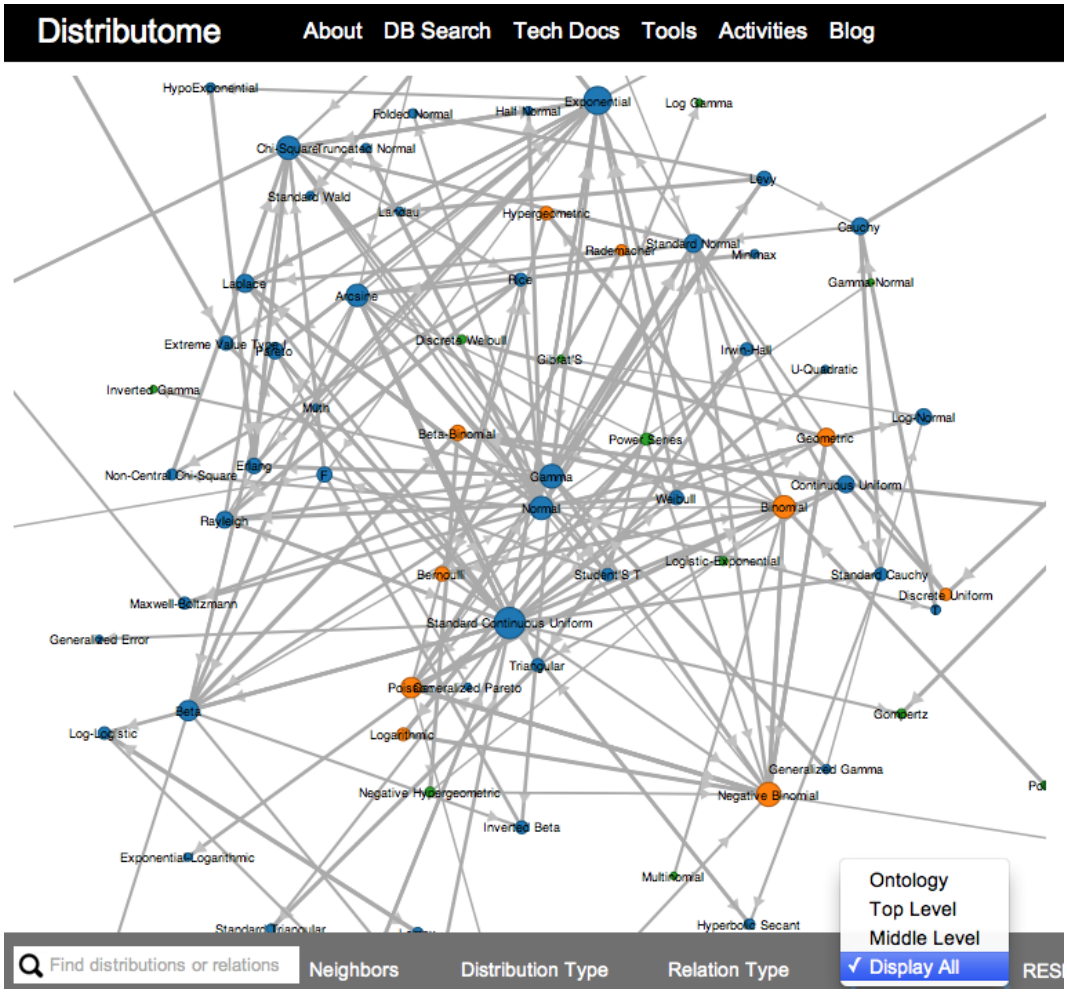


Figure 2.18: Hierarchical Levels

CHAPTER 3

An Example of Distributome Activities

Because distribution is a fundamental concept in Probability and Statistics, one of the main purpose of Distributome project is for education. Below we will give an example of teaching student about hypergeometric distribution.

3.1 Hypergeometric Distribution

In statistics, the hypergeometric distribution is a discrete probability distribution that describes the probability of k successes in n draws from a finite population of size m containing r successes without replacement. [Cha51] According to its definition, the probability mass function is defined by

$$P(X = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}.$$

3.2 Hypergeometric Calculator

Here is an example of hypergeometric distribution.Suppose we randomly select 20 cards without replacement from an ordinary deck of playing cards. What is the probability of getting at most 9 red cards (i.e., hearts or diamonds)?

This is a hypergeometric experiment in which we know the following:

- $m = 52$; since there are 52 cards in a deck.
- $r = 26$; since there are 26 red cards in a deck.

- $n = 20$; since we randomly select 20 cards from the deck.
- $x = 9$; since 9 of the cards we select are red.

Then

$$P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}},$$

which is a little complicated to compute manually.

With distribution calculator we introduced in Section 2.5, we can easily compute the probability $P(X \leq x)$. See Figure 3.1. First, input the parameters: $m = 52$, $r = 26$ and $n = 20$; select PDF for probability density function. Then enter $x = 9$ and press return key; we can see the calculated probability $P(X \leq 9) = 0.388$. It means that if we draw 20 cards from a deck, then we have a probability of 0.388 to get at most 9 red cards.

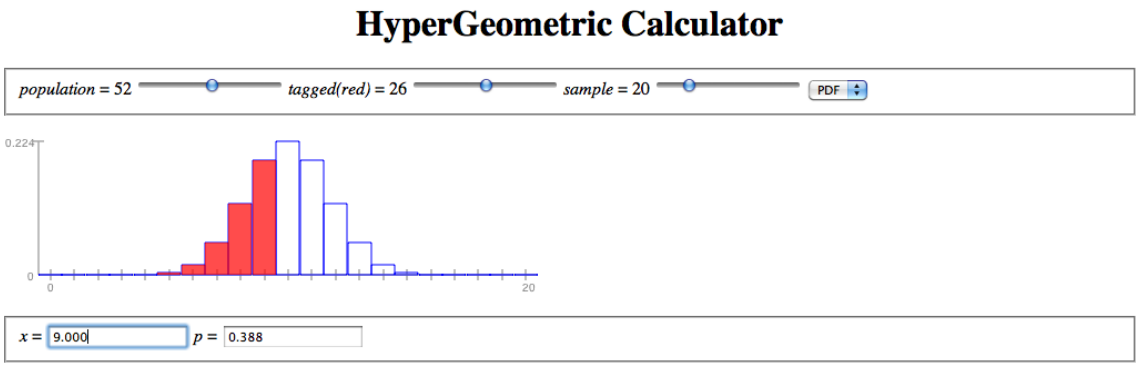


Figure 3.1: Hypergeometric Calculator

3.3 Hypergeometric Experiment

Apart from the calculator, the hypergeometric experiment tool can also help students get an even better understanding of the distribution.

Figure 3.2 shows the Hypergeometric Experiment tool. Similar with the calculator tool, we first specify the parameters, and click "play" button. Then 20

samples are drawn, where red balls denote success while green balls failure. We can even see the unique number on the ball, ranging from 1 to 52. Compared to our example, each ball is a card, red ball corresponding to red cards and green balls for black cards. In the sample shown in Figure 3.2, by counting the number of red balls, we can see $X = 13$.

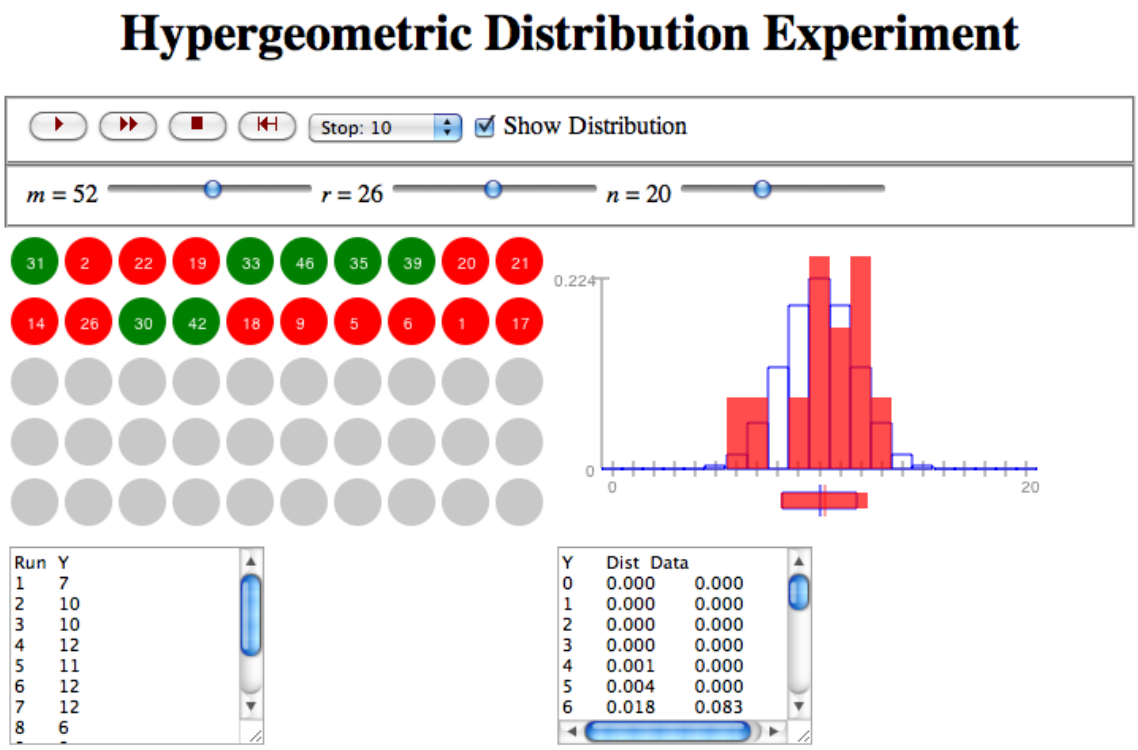


Figure 3.2: Hypergeometric Experiment

What's more, user can specify the number of samples, by selection from the "stop" drop-down list. For example, we selection stop=10 in Figure 3.2. Then there shows the sample values of the random variable on the left and also the histogram on the right.

3.4 Law of Large Numbers

With the distribution experiment, we can also demonstrate the Law of Large Numbers. In the experiment tool, there is a drop-down list to control the number of sampling iterations. Also the histogram as well as the probability density function of the random variable are shown.

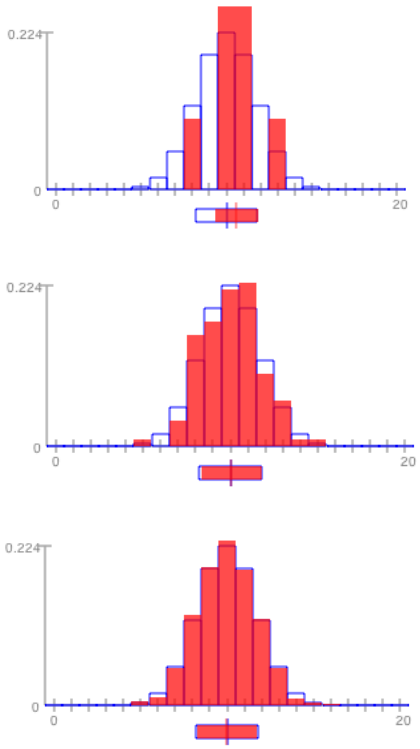


Figure 3.3: Hypergeometric Histogram, with 10, 100 and 1000 Samples.

Figure 3.3 shows the histogram of the samples in red and the PDF of the hypergeometric distribution in blue, with different numbers of iterations from 10 to 1000. It seems like that the histogram is converging to the density function as the number of samples goes to infinity. Actually it can be shown by the Law of Large Numbers.

The law of large numbers is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of

the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed. That is the sample average

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

converges to the expected value

$$\overline{X}_n \to \mu \qquad \text{for} \qquad n \to \infty,$$

where $X_1, X_2, \ldots$ is an infinite sequence of i.i.d. random variables with expected value $E(X_1) = E(X_2) = \ldots = \mu$. [DCG09]

In our case, let Y_i denotes the ith sample of a hypergeometric distribution. For each k, define

$$X_i = \mathbf{1}_{\{Y_i = k\}}.$$

Then

$$\mu = E(X_i) = E(\mathbf{1}_{\{Y_i = k\}}) = P(Y_i = k),$$

which is the PDF of the hypergeometric distribution at point k;

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}(\mathbf{1}_{\{Y_1 = k\}} + \cdots + \mathbf{1}_{\{Y_n = k\}}) = \frac{\#\{Y_i = k\}}{n},$$

which is the histogram at point k. Finally, according to the law of large numbers, $\overline{X}_n \to \mu$, that is to say, the histogram converges to the probability density function pointwise.

CHAPTER 4

Future Work

Distributome is adopting an iterative and incremental development method. Currently it is working well for the main functions mentioned in the thesis. However, there are still some defects to fix. What's more, there are a number of more advanced features going to be designed and implemented in the near future.

Currently, users could submit additional distribution and relation properties via the Distributome editor. The Editor is still being developed and will allow crowd-based contributions in editing, expanding, correcting and updating the Distributome Database in the future.

For Protovis.js, the Protovis team is now developing a new visualization library, D3.js, with improved support for animation and interaction. D3 builds on many of the concepts in Protovis. However, there are plenty of important differences, too. While Protovis excels at concise, declarative representations of static scenes, D3 focuses on efficient transformations: scene changes. This makes animation, interaction, complex and dynamic visualizations much easier to implement in D3. Also, by adopting the browsers native representation (HTML & SVG), D3 better integrates with other web technologies, such as CSS3 and developer tools. [BOH11]

Distribution tools are very important for users to understand certain distributions, as shown in Section 2.5. However, we are still expanding the tools to support more distribution. Table 4.1 shows the current status of Calculator, Experiment and Simulator for all the distributions in top and middle levels. Many

of the missing tools are under development.

Hierarchy	Distribution	Calculator	Experiment	Simulator
Top Level	Bernoulli	OK	missing	missing
	Binomial	OK	OK	missing
	Discrete Uniform	OK	missing	OK
	Geometric	OK	missing	missing
	Hypergeometric	OK	OK	missing
	Multinomial	missing	missing	missing
	Negative Binomial	OK	OK	missing
	Poisson	OK	OK	OK
Middle Level	Beta	OK	missing	OK
	Cauchy	OK	missing	OK
	Chi-Square	OK	missing	OK
	Continuous Uniform	OK	missing	OK
	Exponential	OK	missing	missing
	F	OK	missing	OK
	Gamma	OK	OK	OK
	Normal	OK	missing	OK
	Pareto	OK	missing	OK

Table 4.1: Tools Status

Apart from the functions mentioned above, other features, improvements and customizations are also ongoing.

References

[BH09] M. Bostock and J. Heer. "Protovis: A Graphical Toolkit for Visualization." *Visualization and Computer Graphics, IEEE Transactions on*, **15**(6):1121 –1128, nov.-dec. 2009.

[BOH11] M. Bostock, V. Ogievetsky, and J. Heer. "D3; Data-Driven Documents." *Visualization and Computer Graphics, IEEE Transactions on*, **17**(12):2301 –2309, dec. 2011.

[BPS97] T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, and F. Yergeau. "Extensible markup language (XML)." *World Wide Web Journal*, **2**(4):27–66, 1997.

[Cer12] D. Cervone. "MathJax: A Platform for Mathematics on the Web." *Notices of the AMS*, **59**(2):312–316, 2012.

[Cha51] D.G. Chapman. *Some properties of the hypergeometric distribution with applications to zoological sample censuses*, volume 1. University of California Press, 1951.

[DCG09] I.D. Dinov, N. Christou, and R. Gould. "Law of Large Numbers: the Theory, Applications and Technology-based Education." *Journal of Statistics Education*, **17**(1):n1, 2009.

[DCS08] I.D. Dinov, N. Christou, and J. Sanchez. "Central Limit Theorem. New SOCR Applet and Demonstration Activity." 2008.

[JKK05] N.L. Johnson, A.W. Kemp, and S. Kotz. *Univariate discrete distributions*, volume 444. Wiley-Interscience, 2005.

[KBJ00] S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous Multivariate Distributions, Models and Applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, 2000.

[LM08] Lawrence M Leemis and Jacquelyn T McQueston. "Univariate Distribution Relationships." *The American Statistician*, **62**(1):45–53, 2008.

[Pat88] O. Patashnik. "Designing BIBTEX styles.", 1988.

[Wik12] Wikipedia. "Probability distribution — Wikipedia, The Free Encyclopedia.", 2012. [Online; accessed 22-July-2012].