Animating Statistics: A New Kind of Applet for Exploring Probability Distributions

David Kahle
Baylor University

Journal of Statistics Education Volume 22, Number 2 (2014),
www.amstat.org/publications/jse/v22n2/kahle.pdf

Key Words: Probability distributions; Wolfram CDF technology; Interactive learning environments; Monte-Carlo simulation

Abstract

In this article, I introduce a novel applet ("module") for exploring probability distributions, their samples, and various related statistical concepts. The module is primarily designed to be used by the instructor in the introductory course, but it can be used far beyond it as well. It is a free, cross-platform, stand-alone interactive application based on Wolfram Research's novel computable document format (CDF) technology. It features over thirty common discrete and continuous distributions and can be used to illustrate concepts such as random samples, population and sample means and medians, histograms, kernel density estimators, boxplots, and cumulative distribution, survival, and hazard functions all while dynamically linking samples and estimators to adjustable distribution parameters in real-time. Additionally, the module includes real-world datasets to aid in communicating the concept of fitting a distribution to data. It is hoped that the module will be helpful to instructors at both the high school and college levels for the conceptual understanding of distributions. A simplified version geared specifically toward out-of-class student learning in the introductory course is also made available for students' use. Both are accessible from http://www.baylor.edu/statistics/disttool.

1. Background and Introduction

In my experience, one of the difficulties encountered in the introductory statistics course is establishing the relationship between the theoretical objects of statistics—probability distributions—and the raw objects of applied statistics—the data. In this article I introduce a freely available computer program ("module") that I hope will help instructors overcome this hurdle.

The introductory statistics course is nearly a century old and has had a long and varied past (Aliaga et al. 2005). Over the past few decades, interest in the course has exploded, both in course enrollment and pedagogy. As a consequence, statistics education now has several publications, meetings, and even graduate programs. In addition to these, two focus groups have produced reports giving recommendations for how the introductory statistics course should be taught. The first, commonly called the "Cobb Report" after its initiator, George Cobb, is a synthesized, narrated series of quotes from various members of the statistics education community published as a note in the Mathematical Association of America's Notes and Reports Series. It forwards three primary recommendations (Cobb 1992):

- *Emphasize statistical thinking.* This is open-ended but includes understanding the need for data in problem-solving, the difficulty in obtaining it, and understanding variability through distributions, models and fitting.

- *Focus on data instead of theory and/or recipes.* "Almost any course in statistics can be improved by more emphasis on data and concepts, at the expense of less theory and fewer recipes. (To the maximum extent feasible, calculations and graphics should be automated.)" (David Moore as quoted in Cobb 1992, p. 7) The notion of eliminating calculations where they have no redeeming pedagogical value has recently been resounded by Conrad Wolfram, an advocate for the technology that drives the proposed module (Wolfram 2010).

- *Teach through active learning*. While active learning is a disputed term, the basic principle here is to have students actively engaged in meaningful learning activities where analytical and/or synthetic thought is required; it can include collaborative work. This has been found to have broad but varied support in areas such as student experience and retention (Prince 2004).

The second report, called "GAISE" for Guidelines for Assessment and Instruction in Statistics Education and funded and published by the American Statistical Association, builds on the Cobb Report with six recommendations:

- *Emphasize statistical literacy and develop statistical thinking.* In addition to the Cobb Report, GAISE stresses the need for students to be able to read scholarly reports evidenced with statistical methods.

- *Use real data*. Resounding the Cobb Report, GAISE encourages using real datasets that resonate with students. (e.g., perhaps examples with data from Facebook are preferable to data collected on the anatomy of a mundane species of dragonfly.)

- *Stress conceptual understanding, rather than mere knowledge of procedures.* Recipes are thought to be of far lesser value than a more meaningful understanding of purpose and reasoning strategy.

- *Foster active learning in the classroom.* Essentially a reiteration of the Cobb Report's third recommendation.

- *Use technology for developing conceptual understanding and analyzing data*. At the time of the Cobb Report (early 1990's), software tools for statistics education were in their infancy and mostly localized; so its lack of emphasis there is not surprising (Rossman and Garfield 2011). However, by the time of GAISE (2000's), laptop computers had become mainstream among students. The recommendation here applies not only to simplifying common computations and data analysis but also as a means of reinforcing and exploring conceptual understanding, especially via simulation.

- *Use assessments to improve and evaluate student learning*. This is a novel and substantive suggestion of the GAISE report, including suggesting techniques outside of the traditional homework/testing framework as well as varying the content and frequency of the traditional assessments. Research in this area in a general context is rapidly evolving for both traditional lecture-hall settings (see, e.g., Szpunar, McDermott, and Roediger III 2008) and online settings (Szpunar, Khan, and Schacter 2013). It is also being pursued for statistics education specifically (delMas, Garfield, Ooms, and Chance 2007).

Anecdotally, at least, the two advisories seem to have had widespread readership, and references to the report can be found in many introductory texts (Peck, Olsen, and Devore 2011 and Weiss 2012, to list a few). Meaningful adoption of these advisories among statistics educators, however, appears less certain (Rossman and Garfield 2011).

The module presented in this work is a direct answer to the call put out in the fifth recommendation of the GAISE report to use technology—when helpful—to develop conceptual understanding. In the process, it indirectly enables several other elements of both reports.

Specifically, in this article I introduce a new interactive module rooted in next generation technology and geared towards interactively introducing univariate probability distributions. The software is cross platform, freely available online and executed locally with a free player. In other words, once the module (and player) is downloaded, an internet connection is not required. Broadly speaking, the purpose of the module is to introduce students to the concept of univariate probability distributions, random samples from those distributions, and (to a lesser extent) fitting distributions to real data. Specific topics included are probability density functions (PDFs/PMFs), cumulative distribution functions (CDFs) and survival functions, hazard functions, parameters, random samples and sample size as well as boxplots and distribution estimates based on histograms and kernel density estimators. Over thirty common parametric families are available, including the continuous families of the normal, gamma (exponential, chi-squared), t, F, and beta (uniform) and the discrete families of the binomial (Bernoulli), Poisson, hypergeometric, negative binomial (geometric), and discrete uniform.

The implementation is highly interactive, allowing for dynamic manipulation of parameters, hover-over capabilities, and random number generation. Moreover, many of the interactive objects are dynamically linked for a more natural user experience. For example, when viewing ten samples from a standard normal distribution, if the user slides the standard deviation parameter from one to two (in a continuous motion), the plotted points and any statistics/graphs based on them also change in real-time to reflect the samples/statistics from the "new" distribution N(0, 2). The result is a streamlined, fluid interface that is intended to minimize user frustration and learning time while maintaining a wide range of statistical content. For this example and for the rest of this article, the reader is encouraged to download the module and follow along with the examples in the text.

1.1 Previous work

There are several very useful interactive applets that currently exist online to illustrate statistics concepts, but none operates with qui te the same flexibility, speed, and overall user-experience as the proposed tool, which can be attributed to the novel technologies used.

The Rice Virtual Lab in Statistics (RVLS) contains a collection of predominantly Java-based applets along with an online textbook, but it does not have an applet for exploring probability distributions (Lane 2008). The University of Alabama in Huntsville offers a similarly nice online resource known as the Virtual Laboratories in Probability and Statistics (VLPS), which contains a "Special Distribution Simulator" and a "Special Distribution Calculator" that are somewhat similar to but not quite as polished as the current tool (Siegrist 2006). A third similar, but significantly more substantial, resource is the Statistics Online Computational Resource (SOCR), maintained by developers at UCLA (Dinov 2006). Each of these resources focuses on recreating an activity-based laboratory setting where students can generate data and then analyze it in various ways. Another popular such online laboratory is the Rossman-Chance Collection (Rossman et al. 2011). These are each in contrast to the module proposed here: while they offer primarily activity-based self-learning tools, the module here is primarily designed for in-class instruction by a trained instructor. Moreover, while the module is merely a single application for instructors to use to illustrate concepts, it effectively combines and enhances the functionality of several of the applets offered by the previous resources.

In 2008, Kyle Siegrist (creator of the VLPS) and Ivo Dinov (creator of the SOCR) teamed up with Dennis Pearl to create the Distributome Project, "an open-source, open content-development project for exploring, discovering, navigating, learning, and computational utilization of diverse probability distributions" (Dinov, Pearl, and Siegrist 2008). The primary web technologies used for the Distributome are HTML5 and JavaScript, which are both very powerful state-of-the-art tools for web development. While the vision is excellent—it even allows for user-contributed distributions—Distributome's progress is still actively being developed, and the advanced web technologies it uses require a considerable amount of time to create. Consequently, many elements of the site still use legacy materials in Java from VLPS and SOCR. For example, its distribution calculator, experiment, and simulation applets all appear to be identical to the VLPS tools. As a related side note, Distributome also includes an elegant (force-directed) interactive network showing the *relationships* between distributions, another nice resource for which is the interactive Univariate Distribution Relationships chart maintained by The College of William and Mary (Leemis, Luckett, Powell, and Vermeer 2012).

2. Wolfram CDF Technology

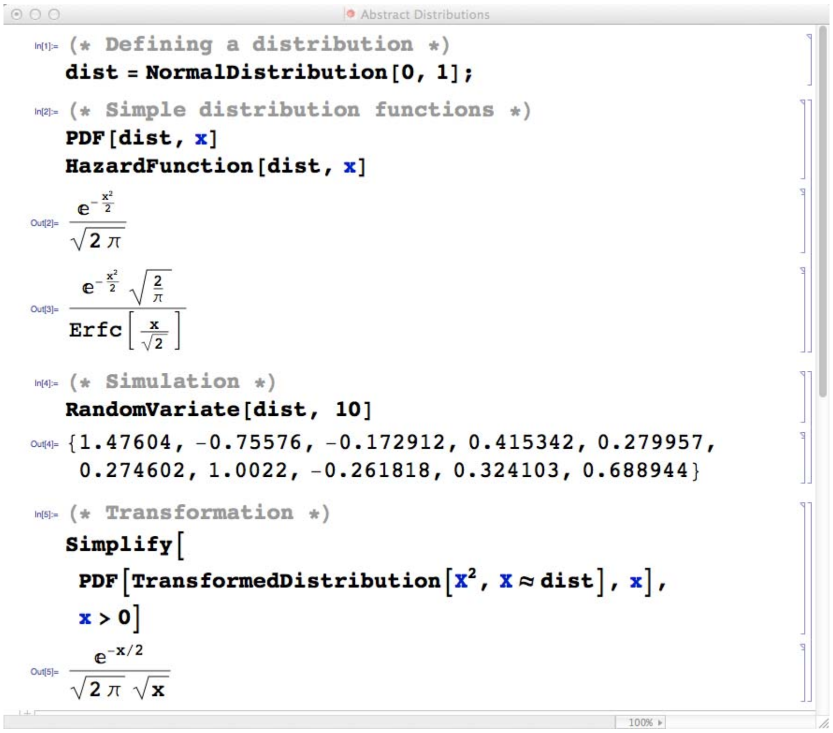
In Summer 2011, Wolfram Research Inc. (WRI), the developers of the blockbuster computer algebra system Mathematica and online computational knowledge engine Wolfram|Alpha, released a new technology it called the computable document format (CDF). The innovation comes in the form of (1) a new kind of file format with extension .cdf and (2) a viewer called the CDF Player. The tool presented in this work is a program—for lack of a better descriptive—a single CDF file made with a Mathematica script. In the statistics literature, similar programs have often been called applets (after Java applets, "little applications," and the corresponding HTML applet tag) or widgets. I use the term module to make the distinction in technologies and to lay the foundational language for a more substantial future work comprised of several modules.

In some ways, CDFs are analogous to Adobe System's ubiquitous portable document format (PDF), and the CDF Player analogous to Adobe Reader. Following the analogy a bit further, Wolfram's flagship mathematical software Mathematica plays a role similar to Adobe Acrobat. CDFs are written and deployed using Mathematica and can be viewed with either Mathematica or the CDF Player, just as PDFs are often authored and edited with Acrobat and viewed with either Acrobat or Reader. Acrobat and Mathematica are both purchased applications with various levels of licensing, and Reader and the CDF Player are free downloads which allow for the viewing of their associated file types by non-paying customers.

While the analogy with Adobe products is useful for understanding the suite of software, there are a number of distinctive and revolutionary features of Wolfram's suite. First and foremost, CDFs are interactive documents, whereas PDFs are by and large static documents. In particular, CDFs can draw from the vast interactive resources and real-time computing ability of Mathematica. Interactive plots, diagrams, and general viewing content can be manipulated with sliders, buttons, drag down menus, mouse-over capabilities and other event handling, and input fields. (Field entries are limited in the free versions.) Each of these interactive capabilities and dozens of others come together to provide the seamless, natural user experience that has characterized the day-to-day mainstream technologies used by today's students. Additional distinguishing features of the CDFs include access to optimization routines and other sophisticated algorithms (indeed, much of the full power of Mathematica), internet connectivity and data acquisition, and the ability to be seamlessly integrated into webpages using a simple JavaScript tag, also made by WRI.

The ability of the CDFs to draw from the tremendous resources of Mathematica is an incredible boon to the development and practical utility of the module. For the past few versions of Mathematica (7, 8, and 9, the current version), WRI has been boosting Mathematica's statistical capabilities. For example, it now boasts the largest assortment of probability distributions of any computer software (Wolfram Research Inc. 2010). The distributions have an easy, systematic referencing system that is also helpful. Most importantly for distributions, however, Mathematica has the ability to handle probability distributions as abstract objects distinct from any particular representation (e.g., its pdf). It then has several built-in functions to manipulate the abstract distributions. For example, simple functions exist which compute special functions associated with the distributions (pdfs, cdfs, mgfs, hazard functions, etc.) that are then also symbolically represented and can be evaluated on demand. Other functions exist to transform distributions, compute summaries (e.g. expectations and quantiles), and sample from distributions. An example of this is shown Figure 1. This is in stark contrast to the preeminent statistical computing environment R, which has different functions that numerically evaluate—not symbolically represent—special distribution functions and simulate from those distributions (and even then not to the extent of Mathematica's capabilities). For example, in R, one uses `dnorm(0)` to give the value of the pdf of the standard normal distribution at 0, `pnorm(0)` to give the value of the cdf at 0, and `rnorm(10)` to generate 10 samples from the distribution, but there is no `mgfnorm` function for computing values of the moment generating function or `hnorm` for computing the values of the hazard function. Thus, Mathematica provides a systematic and concise framework for reliable statistical computing not currently offered by other applications. (Of course, Mathematica's implementation has limitations that don't exist in R. For example, it handles data in a way that is counter-intuitive for many R users.) Finally, to add to these distribution-related capabilities Mathematica has built-in functionality for several statistics concepts: histograms, maximum likelihood estimators, and kernel density estimation to name a few. Each of these aspects combines to make a tremendously powerful framework for the creation of the module discussed in this work.

Figure 1. Mathematica can manipulate abstract distributions.



```
In[1]:= (* Defining a distribution *)
        dist = NormalDistribution[0, 1];

In[2]:= (* Simple distribution functions *)
        PDF[dist, x]
        HazardFunction[dist, x]
```

Out[2]= $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$

Out[3]= $\frac{e^{-\frac{x^2}{2}}\sqrt{\frac{2}{\pi}}}{\text{Erfc}\left[\frac{x}{\sqrt{2}}\right]}$

```
In[4]:= (* Simulation *)
        RandomVariate[dist, 10]

Out[4]= {1.47604, -0.75576, -0.172912, 0.415342, 0.279957,
         0.274602, 1.0022, -0.261818, 0.324103, 0.688944}

In[5]:= (* Transformation *)
        Simplify[
         PDF[TransformedDistribution[X^2, X ≈ dist], x],
         x > 0]
```

Out[5]= $\frac{e^{-x/2}}{\sqrt{2\pi}\sqrt{x}}$

The practical demands of computers in today's learning environments require a bit more than a powerful development framework; they require flexible and diverse dissemination outlets. To match the problem of different students having different machines, CDFs and the CDF Viewer are portable across most current operating systems: Windows, Mac, and several Unix-based operating systems. Additionally, although the CDF technology is currently limited to traditional computer input (mouse, keyboard, etc.), WRI has committed to developing the technology for multi-touch devices such as iPhones and iPads, which is certain to increase module functionality and interest among today's students (Wolfram Blog Team 2012).

In brief, WRI's new CDF technology provides an exceptional platform for making interactive statistics modules that contributes directly to the current modules flexibility, stability, and user-friendliness. Of course, creating your own CDFs does require (1) a Mathematica license and (2) a modest investment of time to learn the basics of the Mathematica language, but there are several resources that can help, including an active mailing list and stackexchange.com listing. Mangano (2010) is another excellent resource. If additional features are needed, for example if you want to create a module that can load user-input data, more expensive licenses are required from Wolfram. The reader is referred to http://www.wolfram.com/cdf/ to answer such queries.

It has not escaped the author's attention that the module in the current article is only scratching the surface of the CDF's potential in statistics education. An ongoing work called the BaylorISMs Project is actively being developed which brings together the CDF with several other current and next-generation computer and web technologies for a single, systematic online statistics resource. A summary exposition of the workings of this research is the focus of a future article. The current article discusses a single deployed CDF, which can be viewed by the free CDF Player. It is freely distributed under a CC BY-SA 3.0 license (the same kind of license as Wikipedia) and can be downloaded from http://www.baylor.edu/statistics/disttool (Creative Commons).

3. Capabilities

Figure 2 illustrates what the user sees upon opening the module: a heavy mix of buttons, sliders, fields, dropdown menus, and a large display window showing the pdf of the standard normal distribution. Since the module's functionality is immense and detailed, rather than go into the minutiae of the inner workings of the code the purpose of this section is provide an overview of the basic capabilities of the module, which can be divided into (1) distribution settings, (2) simulated data settings, and (3) real data settings. This section introduces the capabilities of modules by explaining them by type.

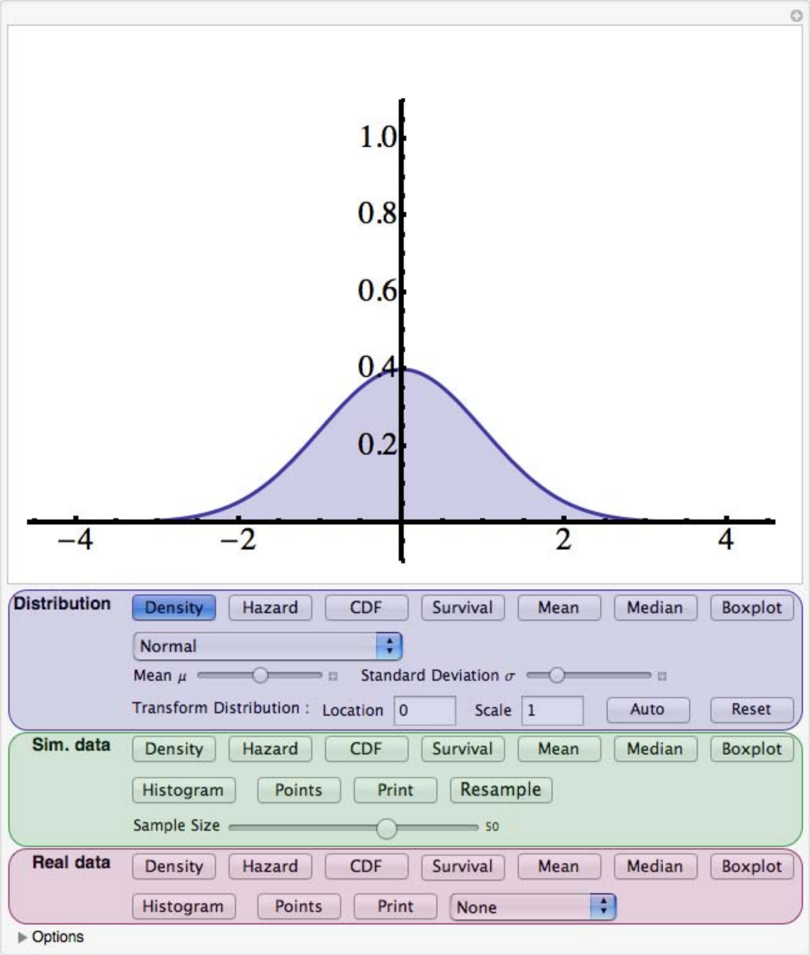
3.1 Distributions and equivalents

The primary purpose of the module is to give the instructor a flexible tool for presenting (univariate) probability distributions. Thus, it is natural that the first controls the user sees when opening the module manipulate which distribution is being referred to, and how that distribution should be presented.

In order to provide a uniform layout of the options for distribution, simulated data, and real data, "presentation" buttons for displaying the density, hazard, cdf, survival, mean, median, and boxplot of the distribution are listed first. Clicking any one of these will activate the button and illustrate the associated function or point in the display window. By default, the Density button is depressed, and the standard normal density is displayed.

Below the presentation buttons is where the distribution family and parameters are set. The module contains the 37 separate distribution families in Table 1. Both discrete and continuous distributions are represented, but more continuous distributions (26) are represented than discrete distributions (11). These families cover most of the distributions taught in a general statistics curriculum at any level.

Figure 2. Basic breakdown of the module controls upon launch. The top section contains the controls for the abstract distribution; the middle the controls for the simulated data; the bottom the controls for the real datasets.



Normal

Transform Distribution : Location 0 Scale 1

Sample Size 50

None

All the distributions are parametric families. Once a family is selected, the parameters indexing that family appear as a labeled series of sliders. When the parameters are changed, the distribution displayed is dynamically (continuously) updated to reflect the new distribution. The sliders allow for a pre-specified range of the parameters that index the distribution. For instance, when the binomial family is selected, two sliders pop up: one for the number of trials parameter N, and one for the probability of success parameter π. Since in theory π is any number between 0 and 1, the practical range of the π variable is a mesh on 0–1; however, since N can be any positive integer, a reasonable range had to be selected, N = 1 to N = 50. Similar decisions were made for every family; the decision was generally made in an attempt to best portray the variability of the distributions within that family or some special property of it (e.g., the binomial distribution becomes bell-shaped for large N for any fixed π). Viewing windows were also taken into account; see Section 4 on options.

Table 1. Distributions represented in the module.

Discrete distributions	Bernoulli, beta-binomial, beta-negative binomial, binomial, discrete uniform, geometric, hypergeometric, log-series, negative binomial, Poisson, and Zipf
Continuous distributions	Beta, Cauchy, chi, chi-square, exponential, extreme value, F, Gamma, Gumbel, half-normal, inverse chi-square, inverse gamma, inverse Gaussian, Laplace, Levy, logistic, lognormal, Maxwell, normal, Pareto, Rayleigh, t, triangular, uniform, Weibull, normal mixture

To end this subsection, a few additional notes are helpful:

A Boxplot button is available under the distribution section. This displays the boxplot constructed using the quantiles of the theoretical distribution, which is not done very often. In some cases it may produce seemingly strange results, particularly with discrete distributions. This is because the percentiles are computed using the quantile function (inverse cdf) as defined in the formal probabilistic sense, so the result may be slightly different than expected (for details see Resnick 2005, p. 179).

Although the standard ranges of the parameters are pre-specified, they can be manually set to any value and even automatically played like a video by clicking the small plus sign to the right of the sliders. This is particularly useful for illustrating specific distributions.

Most elements plotted in the display window are accompanied by mouse-over tooltips. For example, in discrete distributions when the pmf is displayed, hovering the mouse over one of the plotted points displays the value of that point—the probability of observing that value. For continuous random variables, the specific pdf is displayed. The same is true of the non-density distribution related functions (e.g. the cdf) and the distributional summaries (e.g. the mean).

All of the visuals related to the specified distribution are colored a deep blue. This is to distinguish between those of simulated data (always colored in green) and those of real data (colored in dark red). These are changeable; see Section 4 on options.

To reset the module, a circled plus button is available in the top-right of the module with an “Initial Settings” option.

The Transformed Distribution row of the distribution section computes a simple location-scale transformation of the (base) distribution. This is particularly useful for use with real data and so will be discussed more in that section.

The multimodal distribution is a mixture of normal distributions. The mixture weights π_k have sliders that affect the ultimate distribution on a relative scale.

3.2 Distribution samples (simulated data)

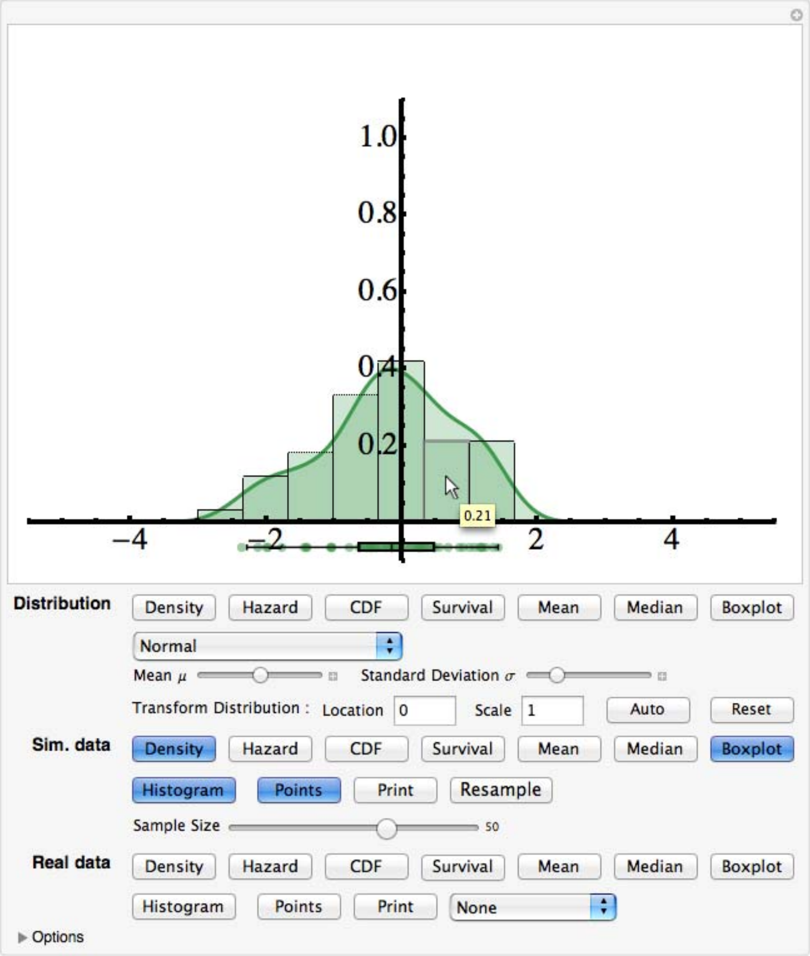
The module itself contains a random number generator that can simulate random samples from any of the distributions listed. This is an incredibly useful and powerful capability that allows instructors to interactively illustrate the relationship between a distribution, its parameters, and its samples. An example illustration is Figure 3.

The best place to begin with the simulated data section of the module is by clicking the Points button. When the Points button is depressed, the individual samples of the distribution are plotted slightly below the *x*-axis for easy viewing. They are green and semi-transparent to reduce overplotting and can be jittered through options. The user can see the individual sample numbers by clicking the Print button.

The sample size is defaulted to n = 50. This can be changed with the slider directly below the buttons and, as it is varied, points are added and subtracted from the same simulated sample. This is helpful as it aids in illustrating the effects of the law of large numbers—with any of the distribution and simulated data summary buttons depressed, as the sample size gets larger the sample estimates approach the population values. (All of the estimators in the module are consistent.) As an example of how this can be used, when the boxplot of the simulated sample is compared to that of the distribution, the two are seen to converge as the sample size increases. The sample size is the one place in the module where flexibility is somewhat limited; for technical reasons the user is only allowed the pre-specified sample sizes. Nevertheless, sample sizes up to 500 are allowed.

Perhaps the most useful capability of the simulated data section for students is the histogram option. The *y*-axis of the histogram is determined on a density scale as opposed to a count (frequency) or probability (relative frequency) scale and consequently is a proper nonparametric density estimator (Scott 1992). However, more relevantly, when both the true distribution density is displayed with the histogram, the practical purpose of the histogram is obvious at a glance, and so are concepts such as its variability through resampling (with the Resample button) and histogram bin width (under Options > Histogram and density options > Simulated data). Moreover, here as before the concept of the law of large numbers can be seen—as the sample size increases, the bins heights approach the density-scaled probabilities of landing in the individual bins.

Figure 3. The histogram, kernel density estimate (KDE), and boxplot of a simulated dataset. Note the mouse-over abilities in the histogram and boxplot images and the single kernel visible in the KDE for the solitary point near -3.



Normal

Mean μ

Standard Deviation σ

Transform Distribution : Location 0 Scale 1

Sample Size 50

None

The first row of the simulated data controls contains the same presentation bar as for the distribution controls. Here the sample density, CDF, survival and hazard functions are those of the distribution given by the kernel density estimate, the mean and median are those of the sample, and the boxplot is the ordinary boxplot of the sample; moreover, mouse-over capabilities abound.

All of the elements of the simulated data section are dynamically linked to those in the distribution section, which creates for a very fluid feel. As the parameters are varied, the data points are continuously transformed to samples from the new distribution. How? The samples themselves are simply continuous functions of uniform 0—1 samples and the parameters of the distribution (excluding rounding in discrete cases); thus as the parameters are varied, the points are smoothly moved to being proper samples from the newly set distribution. The sample estimates are also updated dynamically in real-time as the samples themselves are, making for a natural user experience that highlights conceptual understanding.

3.3 Distribution samples (simulated data)

A repeated advisory in both the Cobb report and GAISE is an emphasis on real world data (Cobb 1992, Aliaga et al. 2005). The module tries to help instructors illustrate aspects of real world data and fitting distributions by including five real world datasets:
The variable petal length from R A Fisher's famous iris dataset, n = 150 (Fisher 1936).
Eruption durations from the Old Faithful dataset, n = 272 (Azzalini and Bowman 1990).
Michelson's 1879 speed of light measurements, n = 100 (Stigler 1977).
Tip amounts by table collected by a waiter, n = 244 (Berenson, Krehbiel, and Levine 2006).
A collection of test scores from an intro course I taught at Baylor, n = 89.
The control panel for the real datasets is identical to that for the simulated datasets and operates identically but independently. Thus, you can view the boxplots, means, medians, histograms, and density estimates as before, and compare those to theoretical distributions.

This is where the location-scale controls come in for the theoretical distributions. As mentioned in the discussion on distribution controls, the sliders controlling their parameters only allow for certain ranges. But the datasets are typically very different from these ranges. Michelson's speed of light data provides a good example. When looking at the histogram of the data with the default 30 bins, the data look quite bell-shaped, indicating a normal distribution would be a good approximation. The sample mean is 299852 with standard deviation 79; however, the normal μ parameter slider only ranges from -3 to 3 and the standard deviation slider from .01 to 5—well outside any reasonable viewing window for the data. The location/scale transformation fixes this problem by allowing the user the ability of transforming the base distribution. While the transformation can be put in manually into a field, an Auto button is available which runs a simple heuristic optimization routine to find a reasonable base location/scale value. Once at the right location/scale, the controls for μ and σ then work exactly as you would like them to—they manipulate the distribution on the new scale. The result is the ability to visually "fit" the distribution to (say) the histogram of the data. An example is provided in Figure 4, where Fisher's iris petal length data are fit with a normal-mixture distribution ("multimodal" in the module) by first visually comparing the histogram to the normal-mixture density and then checked by visually comparing the data simulated by that normal-mixture with the raw data.

Each of the datasets is interesting in its own way. These are discussed further in the next section.

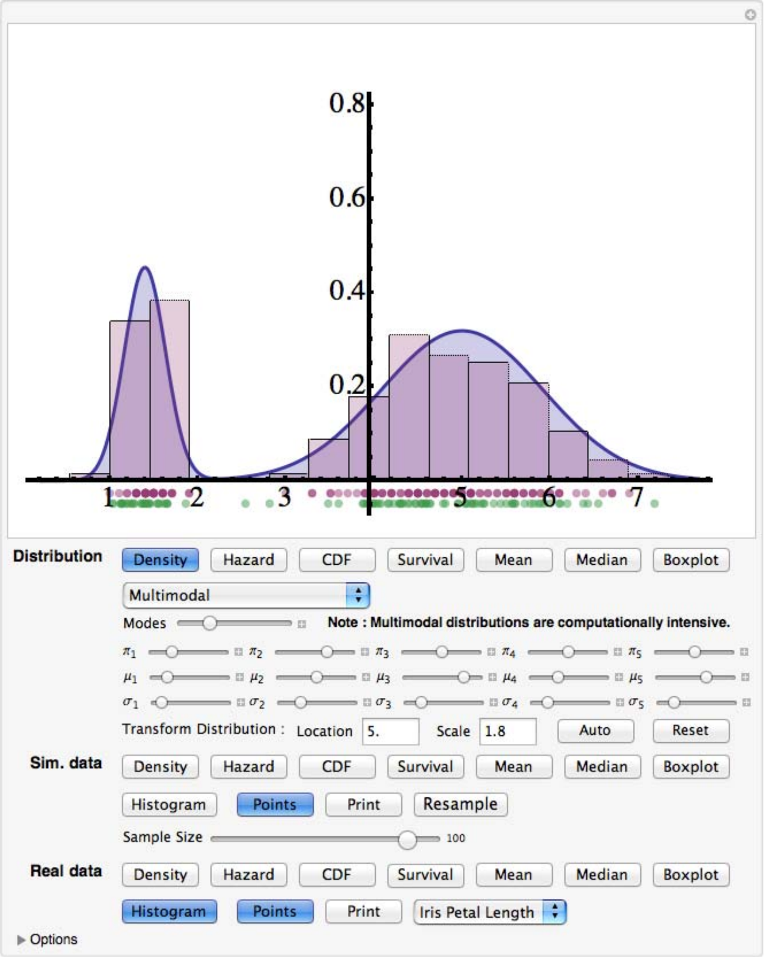
4. Teaching tips

There are several statistical concepts that can be explored with the module. In this section, I highlight some that I find particularly helpful. I begin, however, by explaining some of the options available in the module.

4.1 Options

The module comes equipped with several built-in options accessible via the drop-down menu in the bottom left of the module. When clicked, the user is presented with a "Jitter points" button, which simply adds noise to the plotted points to further reduce overplotting, and four more drop-down menus. A cropped screenshot of the options menus can be seen in Figure 5.

Figure 4: Fisher's iris petal length data visually fit by a mixture of two normal distributions after an auto location/scale transformation.



Multimodal

Note : Multimodal distributions are computationally intensive.

Transform Distribution : Location 5. Scale 1.8

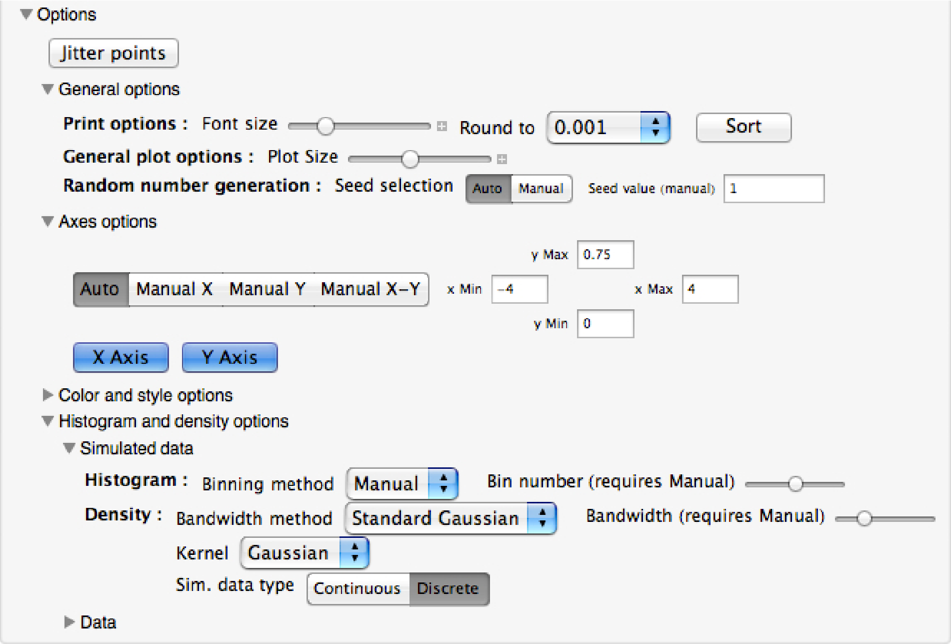
Sample Size 100

Iris Petal Length

General options. In this menu, the user can select the font size for when the simulated or real data are printed, along with rounding options and sorting. Additionally, the user can manipulate the plot size (useful when using projectors) and control random number generation for reproducible graphics.

Axes options. While the default axes specifications are sufficient in most cases, there are still countless illustrations which require a changing of the default specifications—removing the *y*-axis for better visibility, taking the *x*-axis out further or bringing it in, etc. All of these and more can be set under the Axes options menu.

Figure 5. Module options from color to axes to histogram bin width are available by clicking the drop-down triangle next to Options.



Print options : Font size; Round to 0.001

General plot options : Plot Size

Random number generation : Seed selection [✓ Auto] Manual; Seed value (manual) 1

[✓ Auto] Manual X Manual Y Manual X–Y; y Max 0.75; x Min –4; x Max 4; y Min 0

Histogram : Binning method Manual; Bin number (requires Manual)

Density : Bandwidth method Standard Gaussian; Bandwidth (requires Manual)

Kernel Gaussian

Sim. data type Continuous [✓ Discrete]

Color and style options. By default, every function related to one of the schemes (distribution / simulated data / real data) has the same color (blue / green / red). They are fairly thick solid lines, and contain a color fill below them of a certain transparency. Each of these options can be easily changed so that the user is not limited by default color and style schemes.

Histogram and density options. Although the histograms and density estimators can be plotted in a single click, much more goes into creating them. Selections of the binning method for histograms (manual or automated, e.g. Sturges' rule) or the kernel type/bandwidth for the density must be made. These are largely defaulted to reasonable choices that can be manipulated by the user.

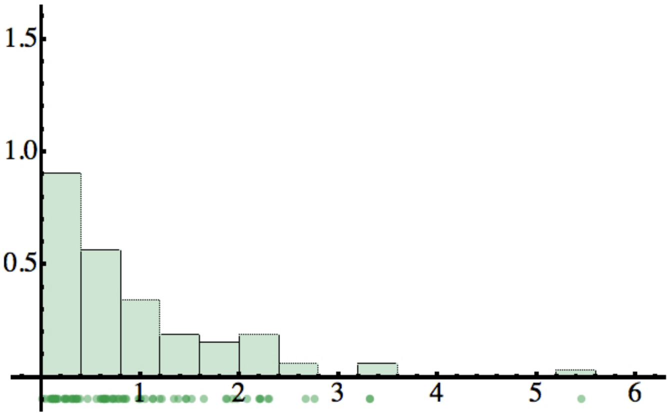
All simulated and real data are, by default, considered continuous, and their density is estimated using kernel density estimation. Once the density estimate is known, the CDF/survival/hazard functions key off of the density estimate. As an alternative, for both the simulated and real data the user is allowed to treat the data as discrete. In this case, the empirical CDF $F_n(x)$ is used and the other functional forms of the distribution key off of it (the density in that case is just the relative frequencies of each occurrence).

4.2 Highlights

There are several individual topics that I have found the modules particularly effective at illustrating:

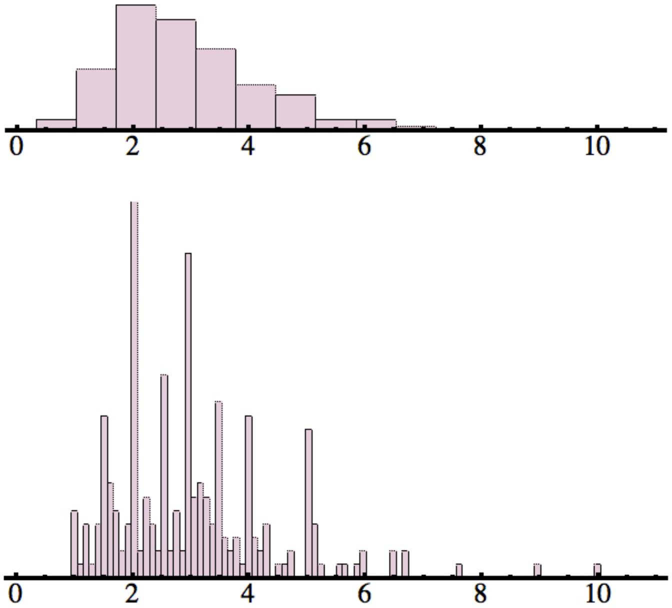
Example making. One place where I have found the module useful is the creation of examples for supplement handouts and tests. For example, test figures assessing student understanding of distribution shape through histograms can be very easily generated by selecting the appropriate distribution (normal, beta, exponential, etc.) and clicking the Histogram button under simulated data. The figures can be easily extracted using third party software such as Jing (free). An example is Figure 6. Example questions include: Name the shape of this distribution; Given the histogram below, would you expect the population to be skewed and if so, how?; Given the histogram below, would you expect the mean to be greater than, smaller than, or equal to the median?

Figure 6. Making example/test figures with screenshots.



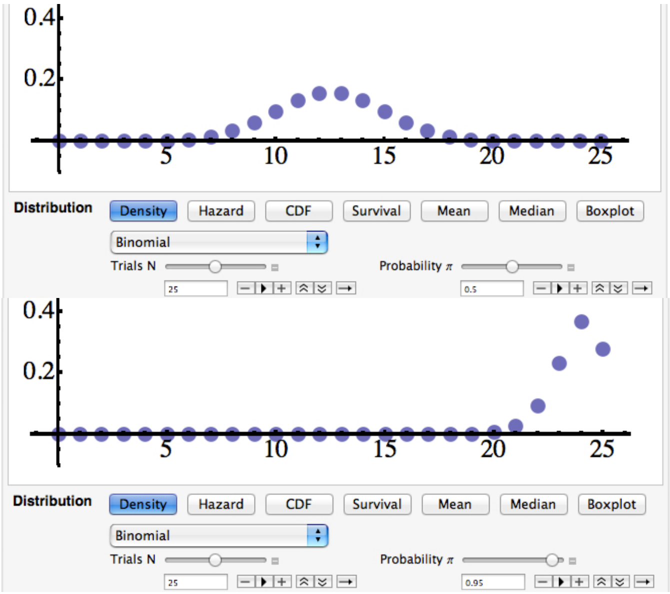
Histogram binwidth. In elementary applied statistics courses, histograms are one of the first topics discussed. I have noticed a significant improvement in student learning and understanding of concepts such as bin width and variability (through resampling) when using the module. The interactive manipulation of binwidth really helps to explain the importance of binwidth selection (Options > Histogram and density options > Simulated data). An example is contained in Figure 7. The real-time responsiveness of a slider governing the binwidth is a simple but vast Figure7 improvement over more commonly used applications such as JMP, where the user is forced to navigate a drop-down menu before manually typing in a different binwidth and recomputing the histogram every time she wants a different binwidth. Variability of the histogram for a given sample size can be seen through resampling, an insight that can be used to motivate normal quantile plots. Specifically, by setting the distribution to normal, turning on the simulated data histogram, setting the sample size to 10 or 20, and resampling, you can illustrate how small normal datasets may appear very non-normal using a histogram.

Figure 7. Interactive histogram binwidth selection on the tips dataset reveals that tippers usually round to convenient quantities.



The normal approximation to the binomial. In anticipation of the normal approximation to the binomial for the classic CLT confidence intervals and hypothesis tests for a proportion, selecting the binomial distribution, showing the density, and ramping up the number of trials N clearly shows the bell shaped behavior of the distribution, and shows how it breaks down as π gets close to 0 or 1. An example is contained in Figure 8.

Figure 8. Illustrating the normal approximation to the binomial (π = .50), and how it breaks down (π = .95), with N = 25 trials.



Binomial

Trials N 25

Probability π 0.5

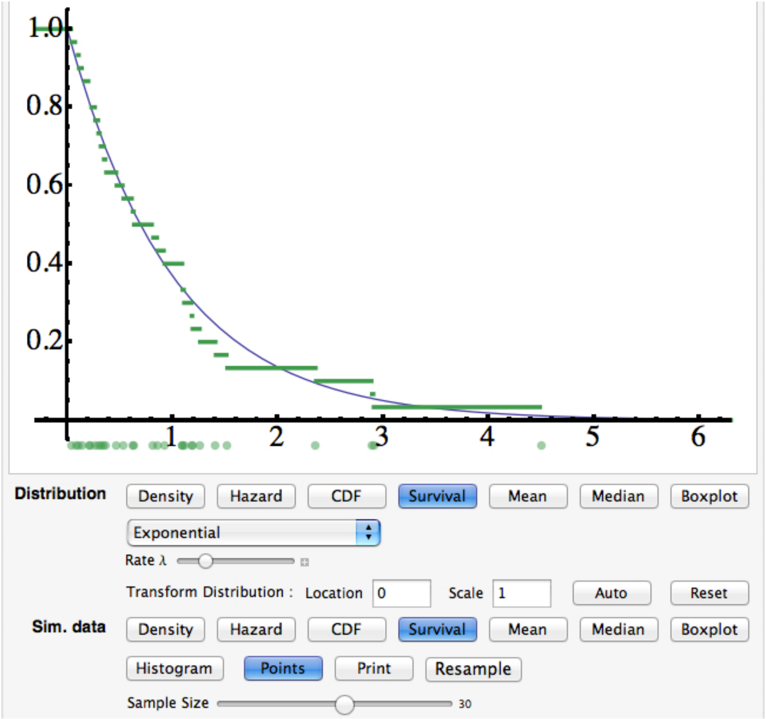
Binomial

Trials N 25

Probability π 0.95

Kolmogorov-Smirnov (KS) test statistic. While the module is really geared towards instruction at the introductory level, it is by no means limited to that level. One example can be seen when trying to illustrate the KS test statistic. As another application of the law of large numbers, the empirical distribution function $F_n(x)$ converges pointwise to the (population) distribution function F(x); similarly, the empirical survival function $1 – F_n(x)$ converges to the (population) survival function 1 – F(x). The test statistic of the Kolmogorov-Smirnov goodness-of-fit test is simply the largest vertical difference between the empirical distribution (or survival) function and the population distribution (or survival) function times the square root of the sample size; it rejects the proposed distribution when the test statistic is too large. While the module cannot compute the test statistic, the intuition can be easily communicated. This is illustrated in Figure 9, which compares survival functions instead of distribution functions. The test statistic would be the largest vertical distance between the empirical survival (in green) and the true survival (in blue), which appears to happen around x = 1.25 or 1.5, times the square root of the sample size. The distribution of the test statistic under the null hypothesis is, of course, very far outside the range of an introductory course and even some graduate courses. Increasing the sample size, we note that the maximum distance decreases.

Figure 9. The Kolmogorov-Smirnov test using the survivals.



Exponential

Rate λ

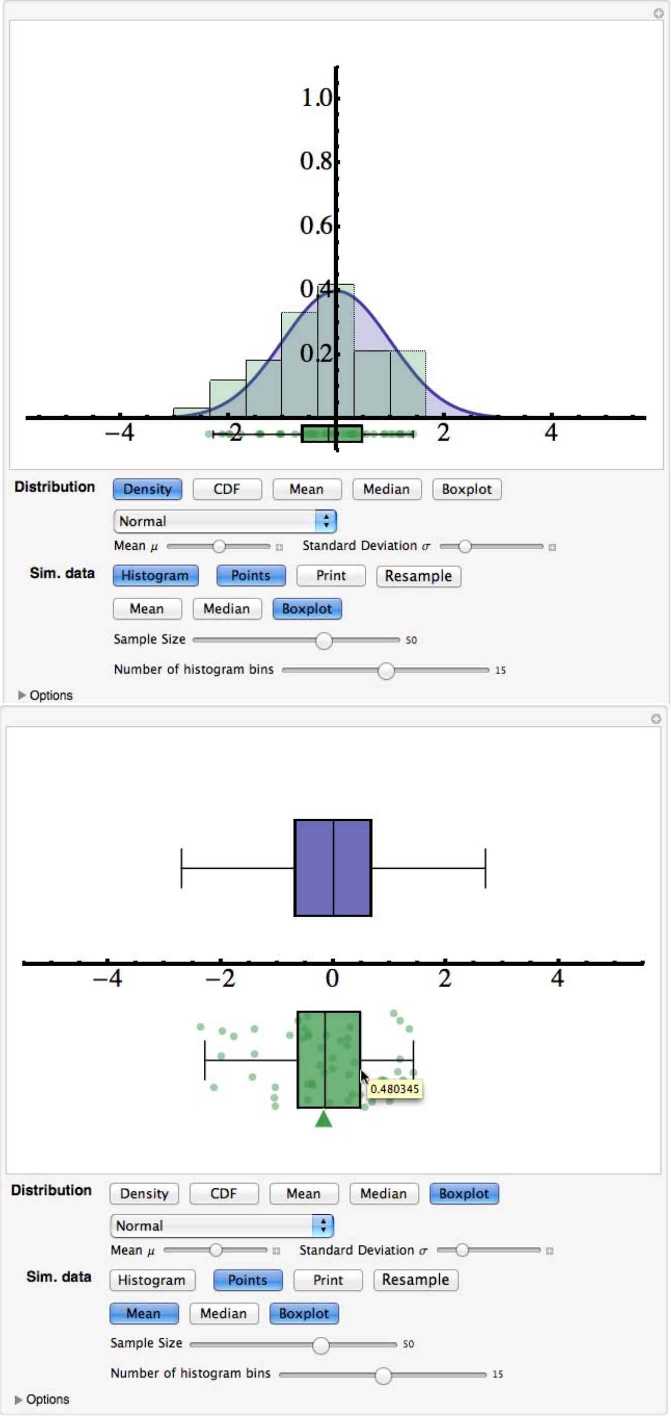
Transform Distribution : Location 0 Scale 1

Sample Size 30

5. A Simpler Module for Students

As most educators have experienced, students tend to have a better learning experience when applets are available at increasing levels of sophistication as opposed to a single high-level applet (Saw 2011). To meet this need for the probability distributions module, a simplified module is also provided. The module has significantly limited functionality when compared to the teaching module, but it still hits many of the key concepts of distributions and their samples. See Figure 10 for a screenshot.

Figure 10. Screenshots of the simplified student module.



Normal

Sample Size 50

Number of histogram bins 15

Normal

Sample Size 50

Number of histogram bins 15

While the simpler tool is not as flexible as the larger tool, it still provides a wealth of opportunity for the student both as an in-class tool and an out-of-class tool. Here is a simple exercise for the introductory level that can help students understand variability in boxplots and how to interpret skewness in a boxplot:

1. Set the distribution to Normal and vary the Mean μ slider and the Standard Deviation σ slider to observe their effect on the distribution. Describe the shape of a normal distribution.
2. Now change the distribution to Exponential, and vary the Rate λ slider. How does the rate parameter affect the exponential distribution? Describe the shape of an exponential distribution.

   Change the distribution back to Normal and, using the plus boxes to the right of the sliders, change μ to 0 and σ to 1. Click the Distribution Density button to turn it off, and then click Boxplot under Sim. Data. Click Resample several times to observe the variability of a boxplot created from 50 samples from a normal distribution. Now, change the distribution back to Exponential, and set λ to 1 and click Resample several times to note the variability of a boxplot created from 50 samples from the exponential distribution. Comparing the two types of boxplots, and using your answers to 1. and 2., how might you use boxplots to compare the shapes of distributions?

6. Conclusions

In this article, I have introduced a novel pedagogical module, freely available online, that can help teachers motivate and animate statistical concepts. Due to the complexity of the module, it is accompanied by a simplified module for student use. Both are created using the revolutionary .cdf technology of Wolfram Research, which affords developers a high-powered framework for interactive mathematical and statistical computation that is translated into a fluid user experience. It is hoped that the module will be helpful to instructors at both the high school and college levels for the conceptual understanding of probability distributions.

Acknowledgements

It is my pleasure to thank the Editor, Dr. Everson, as well as an anonymous Associate Editor and reviewers for several helpful comments that have made this work stronger.

References

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Witmer, J. (2005), *Guidelines for assessment and instruction in statistics education: College report*. The American Statistical Association. Available at http://www.amstat.org/education/gaise/.

Azzalini, A. and Bowman, A. (1990), "A look at some data on the old faithful geyser," *Applied Statistics*, 39, 357–365.

Berenson, M., Krehbiel, T., and Levine, D. (2006), *Basic business statistics: Concepts and applications*, 10th Edition, Upper Saddle River, NJ: Prentice Hall.

Cobb, G. (1992), Teaching Statistics. In L.A. Steen (Ed.) *Heeding the Call for Change: Suggestions for Curricular Action.* Washington: Mathematical Association of America.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing students conceptual understanding after a first course in statistics," *Statistics Education Research Journal*, 6, 28–58.

Dinov, I. (2006), "SOCR: Statistics Online Computational Resource," *Journal of Statistical Software*, 16, 1–16. Available at http://www.socr.ucla.edu/.

Dinov, I., Pearl, D., and Siegrist, K. (2008), "Distributome – an interactive web-based resource for probability distributions," [online]. Available at http://distributome.org.

Fisher, R. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, 179–188.

Lane, D. and others (2008), "Rice Virtual Lab in Statistics," [online]. Available at http://onlinestatbook.com/rvls.html.

Leemis, L., Luckett, D., Powell, A., and Vermeer, P. (2012), "Univariate probability distributions," *Journal of Statistics Education*, 20, 1–11. Available at http://www.math.wm.edu/~leemis/chart/UDR/UDR.html.

Mangano, S. (2010), *Mathematica Cookbook*, Sebastopol, CA: O'Reilly Media, Inc.

Peck, R., Olsen, C., and Devore, J. (2011), *Introduction to Statistics and Data Analysis*, 4th Edition, Boston, MA: Brooks/Cole.

Prince, M. (2004), "Does active learning work? A Review of the Research," *Journal of Engineering Education*, 93, 223–231.

Resnick, S. (2005), *A Probability Path*, Boston, MA: Birkhaüser.

Rossman, A. and Garfield, J. (2011), "Interview with Joan Garfield," *Journal of Statistics Education*, 19, 1–24.

Rossman, A., Chance, B., Garcia, F., Lima, C., Holmes, E., and Gill, R. (2011), *Rossman/Chance Applet Collection* [online]. Available at http://www.rossmanchance.com/applets/index.html.

Saw, A. (2011), "Learner Control, Expertise, and Self-Regulation: Implications for Web-Based Statistics Tutorials," Ph.D. Thesis, Claremont Graduate University, Dept. of Psychology.

Scott, D. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York, NY: John Wiley & Sons, Inc.

Siegrist, K. (2006), *Virtual Laboratories in Probability and Statistics* [online]. Available at http://www.math.uah.edu/stat/.

Stigler, S. (1977), "Do Robust Estimators Work with Real Data?" *The Annals of Statistics*, 5, 1055–1098.

Szpunar, K., Khan, N., and Schacter, D. (2013), "Interpolated Memory Tests Reduce Mind Wandering and Improve Learning of Online Lectures," Proceedings of the National Academy of Sciences, 110, 6313–6317.

Szpunar, K., McDermott, K., and Roediger III, H. (2008), "Testing During Study Insulates Against the Buildup of Proactive Interference," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1392–1399.

Weiss, N. (2012), *Introductory Statistics*, 9th Edition, Boston, MA: Pearson Education, Inc.

Wolfram, C. (2010), "Teaching kids real math with computers," *TED Talks* [online]. Available at http://www.ted.com/talks/conrad_wolfram_teaching_kids_real_math_with_computers.html.

Wolfram Blog Team (2012), "A Preview of CDF on iPad," [online]. Available at http://blog.wolfram.com/2012/02/17/a-preview-of-cdf-on-ipad/.

Wolfram Research Inc. (2010), "Probability and Statistics Solvers and Properties: New in Mathematica 8," [online]. Available at http://www.wolfram.com/mathematica/new-in-8/probability-and-statistics-solvers-and-properties/.

David J. Kahle, Ph.D.
Assistant Professor of Statistical Science
Department of Statistical Science
Baylor University
One Bear Place #97140
Waco, Texas, USA 76798
(254) 710-6102
david_kahle@baylor.edu