PROBABILITY DISTRIBUTOME – COMPUTING, VISUALIZATION & INSTRUCTION

Jared Tianyi Chai ^, Mark Bobrovnikov ^, Ivo D. Dinov *
Statistics Online Computational Resource (SOCR), University of Michigan, Ann Arbor, MI 48109, USA
https://socr.umich.edu, * statistics@umich.edu, ^ equally contributing authors

The concept of a probability distribution is in the core of all modern data-driven modeling, simulation, analytics, inference, prediction, and prognostication of observable natural phenomena. Despite the large number of possible probability distributions and their significant heterogeneity, only dozens of distributions are named and well understood, many are interconnected, and all of them span the universe of all possible physical processes. This manuscript presents the probability Distributome, which allows navigation, exploration and learning of many univariate distributions, and facilitate the computation and visualization associated with univariate, bivariate and trivariate probability distributions. We illustrate applications of three specific resources including an HTML5/JavaScript applications (http://Distributome.org), RShiny apps (https://socr.umich.edu/html/dist) and 2D and 3D Plotly apps (https://socr.umich.edu/HTML5/BivariateNormal/BVN2/). These resources are freely accessible, open-source, and platform agnostic. Both learners and instructors may utilize, expand, improve, and customize these tools for their specific research and education needs.

INTRODUCTION

Statistics is the study of data (ID Dinov, 2018). It can tell you when to plant crops for the best harvest, when to invest in the stock market, and a multitude of other things. Mathematical statistics and modelling form the core of modern data science, AI systems and Machine Learning forecasts for risk analysis. It takes previously collected data and through a method of mathematical formulations, computational algorithms and visual representations forms a complex story of the past, present, and potential future. For example, companies use statistical forecasting methods to estimate customer shopping patterns and increase their sales (Shang, McKie, Ferguson, & Galbreth, 2020). YouTube uses Machine Learning and AI algorithms to form a list of recommended videos to watch next for each viewer (Tanzil, Hoiles, & Krishnamurthy, 2017). And governments employ computational statistics to form both short term and long term macroeconomic policies (Kuziemski & Misuraca, 2020).

There are many different types of probability distributions that are used in the modern world, but they can all be divided into two broad categories (I. Dinov, Siegrist, K, Pearl, DK, Kalinin, A, Christou, N, 2015; Forbes, Evans, Hastings, & Peacock, 2011). Discrete distributions deal with variables and data that can be arranged into a limited number of classes or categorical labels. Continuous distributions deal with data that covers a continuous space, such as the set of all real numbers. As an example, a discrete distribution may model the event of rolling a die several times and predicting the individual outcomes, whereas a continuous distribution may model the distance Olympic javelin throwers can throw a javelin. These distributions are also studied based on the number of variables they try to model and account for. If a distribution only models a single variable, then it is called a univariate distribution. Modeling processes involving two variables requires bivariate distributions, three variables - trivariate distributions, and more variables correspond to multivariate distributions.

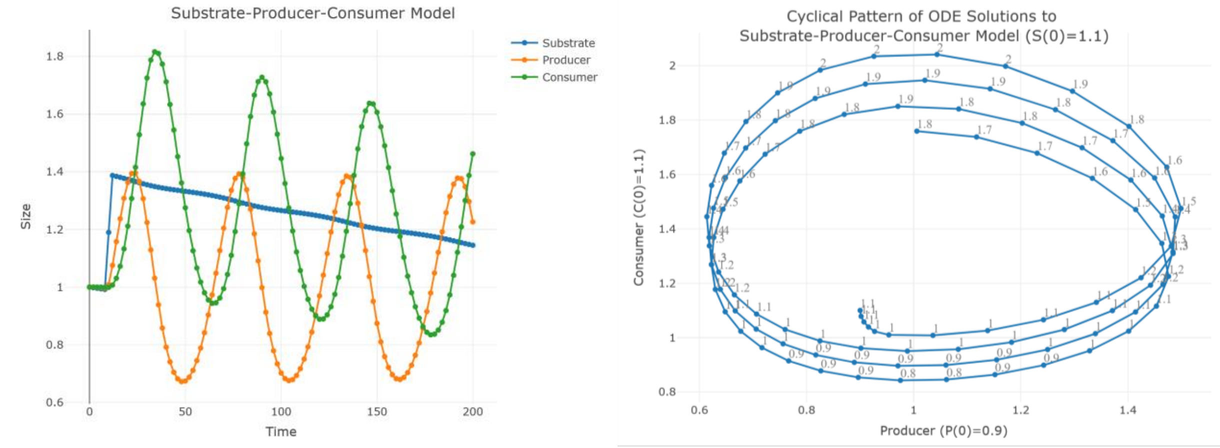
Distributions are further divided based on the processes and data they model, from coin tosses and dice rolls to radioactive decay, to frequencies of encounters, to name but a few. The most well-known and commonly seen distribution is the normal distribution, and under classical parametric assumptions is used as the default distribution. Furthermore, the central limit theorem (CLT) provides a justification for the popularity of the Normal distribution as the gravitational black hole in the universe of all “nice” probability distributions (I Dinov, Christou, & Sanchez, 2008). The normal distribution can be used as a model of heights of all people in a country, or the blood pressure in all hospital patients. Some notable other distributions are the Binomial distribution, which is a discrete distribution that is most often used for coin flips or other binary choices, the Poisson distribution, which can be used to describe radioactive decay and the number of mutations in genes, and the Geometric distribution, which studies the probability of a successful binary choice being encountered after k trials.

Unfortunately, univariate distributions have a limited scope of applications in the real world, as they only represent scalar processes, without considering the multiple factors influencing the state of events. This is where multivariate probability distributions come in. Seeing how different probability

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

distributions and variables interact with each other, subject to certain dependencies, may help to predict more complex and realistic world interactions. As an example, consider a producer-consumer pair subject to some environmental constraints (limited substrate). The relation between the two entities can be described in a trivariate statistical model of the sizes of the produced and consumed products relative to the available substrate. For instance, the prices of cars depend on raw materials, the manufacturing supply (production), and the consumer demand (consumption). Changes in either of these affect the other. This process forms a cycle that will center around a certain equilibrium point where the number of goods (produced and consumed) reaches a sustainable level based on the available environmental resources (substrate), **Figure 1**. The simulation in this figure is generated by solving the Volterra-Lotka ordinary differential equation model of species competition, under some specific initial parameters (Hsu & Zhao, 2012). This cycle can also be modelled as a bivariate probability problem, where a particular probability reflects the chances of each entity to be at certain level (size or number of goods). **Figure 2** illustrates an example of a bivariate probability density function.

Computing multivariate statistical distributions requires more than just a simple product of the univariate probabilities. Correspondingly, the mathematical formulations required to fully describe multivariate processes are naturally more intricate. Such mathematical models of naturally occurring or man-made phenomena are useful and can generally be enhanced with specific illustrations of their practical utilizations. Visualizing probabilities corresponding to univariate distributions can easily be accomplished using areas under planar density curves. **Figure 3** shows a univariate normal distribution as a 2D graph. The correspondence between critical values on the horizontal axis and probability values, areas under the density curve, can be explored with sliders or via numerical entries. For example, once we specify the distribution parameters, in this case the distribution mean and variance, we can enter either a critical value on the x-axis or a corresponding cumulative probability value as the shaded area under the density over the specified range.



Longitudinal processes (time-series)

Bivariate relation between production and consumption of goods subject to available resources

Figure 1. Relations between environmental factors (substrate), level of product production (producer) and the corresponding level of product consumption (consumer). Left and right images show the longitudinal changes of all 3 components (substrate, production, and consumption), and the cyclical relation of the levels of production and consumption of the product, respectively.

Probability visualization becomes considerably more challenging for multivariate distributions that may combine several univariate processes to generate a joint model of a multivariate process representing a vectorized outcome. A univariate distribution may be manipulated by specifying a data point (critical value) or by fixing the associated cumulative probability value (p-value) and recovering the corresponding critical value. The simplest case for specifying the domain range in a bivariate distribution case involves two degrees of freedom - two parameters for controlling the upper limit of the rectangular area (critical range) - to compute the corresponding cumulative probability value. Going a step further, the simplest models of trivariate processes, or more generally multivariate processes, require higher order of controls for navigating multivariate probability distributions, which increase linearly with the dimensionality of the process. Therefore, when computing or visualizing trivariate probability distributions, special mechanisms are necessary to facilitate the interaction, interpretation, and display of the corresponding relation between critical domain restrictions and the corresponding cumulative probability values. The Probability Distributome Project (www.distributome.org) provides simple strategies to visualize the universe of many univariate distributions, examine their interrelations, explore their individual properties, as well as display and compute critical and probability values (I. Dinov, Siegrist, K, Pearl, DK, Kalinin, A, Christou, N, 2015). However, in general, there are fewer mechanisms for visualization of bivariate, trivariate and higher-order variate probability distributions.

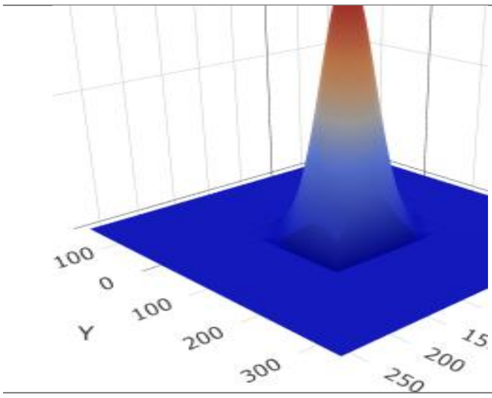
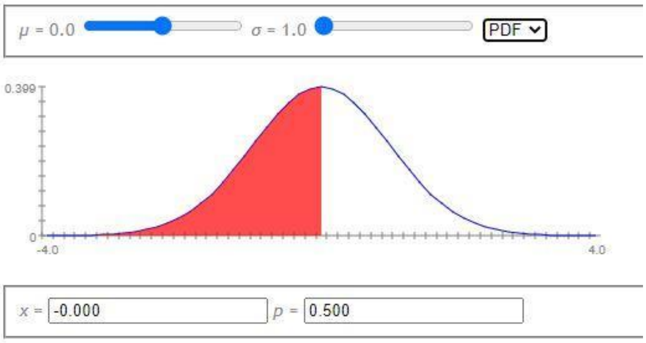


Figure 2. Bivariate probability of the deer (variable Y) and lynx (variable X) populations, each modeled as a normal distribution


μ = 0.0 σ = 1.0 PDF

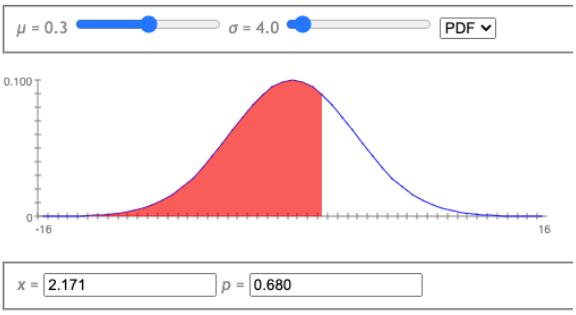
x = -0.000 p = 0.500

Figure 3. Univariate Normal distribution example

INTERACTIVE PROBABILITY RESOURCES

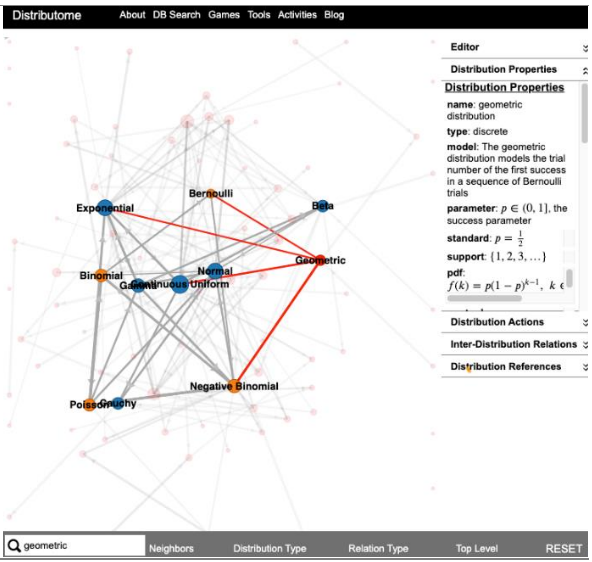
There are several open-sourced visualization solutions for probability distributions available. The Probability Distributome Project (www.distributome.org) mentioned earlier has a number of HTML5 based applications on its website. These web applications allow the visualization, calculation, and sampling of 47 different univariate probability distributions. A sample use case of the Probability Distributome Project has been shown in **Figure 3** illustrating a screenshot from the Normal Distribution Interactive Calculator application from the Distributome website. In this application, one can specify the parameters of the normal distribution, mean and standard deviation, and the calculator generates the corresponding PDF or CDF plot. Specifying a data point and the application will automatically compute the corresponding p-value, or vice versa. For example, to find the critical value that corresponds to a left-tail p-value of 0.68 in a normal distribution with mean of 0.3 and standard deviation of 4.0, we could use the sliders to specify the parameters and type 0.68 into the input box on the right. The calculator would then return a critical value 2.171, and the left-tail of the normal distribution becomes shaded. **Figure 4** below shows the output of the application from these actions. The Probability Distributome Project also provides a Navigator (http://distributome.org/V3/) that describes the properties of each probability distribution in-detail while visualizing the relationships among the distributions. In the Navigator, learners can simply select a node representing a specific probability distribution or type in the name of the distribution in the search bar to learn more it. For example, searching for “geometric” will yield the properties of the Geometric Distribution and navigate to its corresponding Distributome web application. **Figure 5** below shows an example using the Navigator.

Another open-source computation and visualization tool for univariate probability distributions is the SOCR Probability Distribution Calculator (https://socr.umich.edu/html/dist/) (Chu, Cui, & Dinov, 2009). This web application is developed using R and the RShiny package as an all-in-on tool for up to 75 univariate probability distributions. For each supported distribution, the app allows computing, visualizing, and interacting with its corresponding PDF and CDF plots, while also learning about its properties. Similar to applications in the Probability Distributome Project, learners can also specify the input parameters of the distribution and the application will compute critical values and p-values given the inputs. However, the SOCR Probability Distribution Calculator supports the computation between two data points, instead of only the left-tail probabilities. **Figure 6** below shows a use case of this application. In this example, to compute the p-value corresponding to the area between 3.4 and 5.8 in an exponential distribution model with an input parameter of 0.5 for lambda, we can select the plot type in the upper-left corner of the application and enter 0.5 to specify the distribution parameter. Using the sliders, we can select the range between 3.4 and 5.8 on the PDF plot of the exponential distribution, and the application computes the approximate probability (12.77% in the given range). By using the switch on top of the plot, we can also switch from sliders to textboxes for improved input accuracy. By clicking on the camera icon at the top-right corner of the plot, we can also save the generated plot locally as images.


μ = 0.3 σ = 4.0 PDF

x = 2.171 p = 0.680

Figure 4: Probability Distributome Normal Distribution Interactive Calculator.



Distribution Properties

name: geometric distribution

type: discrete

model: The geometric distribution models the trial number of the first success in a sequence of Bernoulli trials

parameter: $p \in (0, 1]$, the success parameter

standard: $p = \frac{1}{2}$

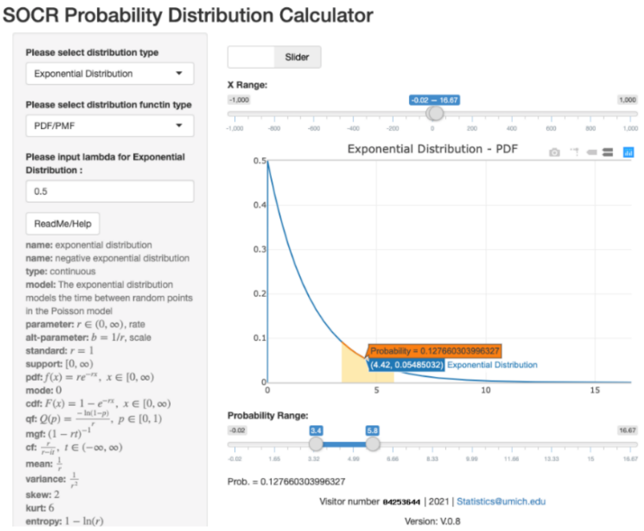
support: $\{1, 2, 3, \dots\}$

pdf: $f(k) = p(1-p)^{k-1},\ k \in$

geometric

Figure 5: Probability Distributome Project Navigator example.

These two tools are both designed only for computations using univariate probability distributions. However, in practice, the use of multivariate distributions is also very common. The following tools can be helpful in the computation of bivariate and trivariate normal probability distributions. The Bivariate Normal Distribution Interactive Calculator (3D) is an open-source computation and visualization tool developed using JavaScript. This tool allows the learners to input the parameters for two univariate normal distributions as well as the correlation coefficients between the two distributions (https://socr.umich.edu/HTML5/BivariateNormal/BVN2/). The app will automatically generate a 3D plot for the resulting joint normal distribution, which can be rotated and interactively manipulated.


SOCR Probability Distribution Calculator

Please select distribution type: Exponential Distribution

Please select distribution functin type: PDF/PMF

Please input lambda for Exponential Distribution : 0.5

ReadMe/Help

name: exponential distribution
name: negative exponential distribution
type: continuous
model: The exponential distribution models the time between random points in the Poisson model
parameter: $r \in (0, \infty)$, rate
alt-parameter: $b = 1/r$, scale
standard: $r = 1$
support: $[0, \infty)$
pdf: $f(x) = re^{-rx},\ x \in [0, \infty)$
mode: 0
cdf: $F(x) = 1 - e^{-rx},\ x \in [0, \infty)$
qf: $Q(p) = \frac{-\ln(1-p)}{r},\ p \in [0, 1)$
mgf: $(1 - rt)^{-1}$
cf: $\frac{r}{r-it},\ t \in (-\infty, \infty)$
mean: $\frac{1}{r}$
variance: $\frac{1}{r^2}$
skew: 2
kurt: 6
entropy: $1 - \ln(r)$

X Range: -0.02 – 16.67

Probability = 0.127660303996327
(4.42, 0.05485032) Exponential Distribution

Probability Range: 3.4 5.8

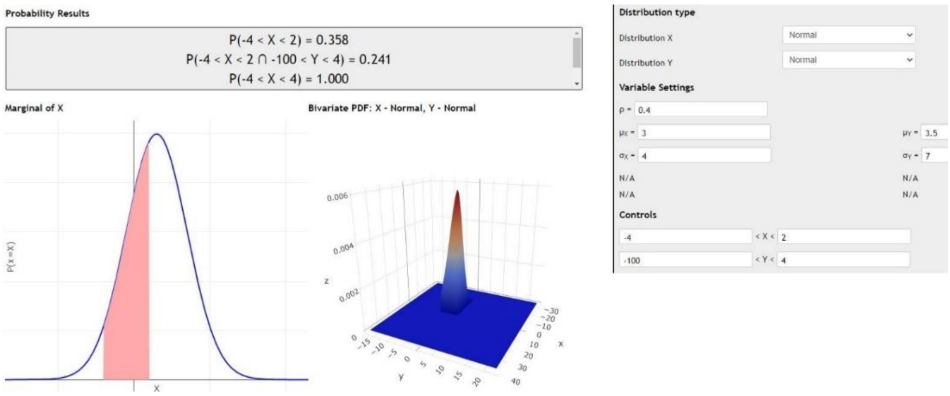
Prob. = 0.127660303996327

Visitor number 84253644 | 2021 | Statistics@umich.edu

Version: V.0.8

Figure 6. SOCR Probability Distribution Calculator example

The Bivariate calculator includes input controls on the ranges of the two normal distributions. After specifying these controls, the Bivariate calculator will compute and visualize the corresponding probability values for the individual univariate normal distributions as well as the joint distribution. Additionally, the calculator can also compute the conditional probabilities of one variable given the value of the other variable. **Figure 7** below shows a use case with the Bivariate calculator.



Probability Results

P(-4 < X < 2) = 0.358
P(-4 < X < 2 ∩ -100 < Y < 4) = 0.241
P(-4 < X < 4) = 1.000

Distribution type

Distribution X: Normal
Distribution Y: Normal

Variable Settings

ρ = 0.4
μX = 3 μY = 3.5
σX = 4 σY = 7
N/A N/A
N/A N/A

Controls

-4 < X < 2
-100 < Y < 4

Figure 7. Bivariate Normal Distribution Interactive calculator (3D) example

In this example, two univariate normal distributions are specified. Distribution X has a mean of 3 with a standard deviation of 4, and distribution Y has a mean of 3.5 with a standard deviation of 7. The correlation coefficient between the two distributions is -0.4. After the joint distribution PDF of the two univariate normal distributions is plotted on the left of **Figure 7**, we can modify the ranges of the two distributions: X has a range of values between -4 and 2, and Y has a range between negative infinity and 4. The app estimates the probability of the joint distribution and the X marginal distribution. Toggling on the "*Marginal of X*" option yields the joint distribution p-value (0.1404) and the corresponding marginal probability (0.361). The univariate normal distribution PDF for X with the range of interest is also plotted in a 2D graph on the right of **Figure 7**.

The Trivariate Distribution 3D Calculator is also developed using JavaScript. This Trivariate calculator (https://socr.umich.edu/HTML5/BivariateNormal/TVN/) is an open-source tool that allows the computation and visualization of the joint distribution among three univariate distributions. Different from the Bivariate calculator, the Trivariate calculator allows the learners to select among 40 distributions for each of the three univariate distributions. Following the univariate distributions selection, we can specify the parameters of the three marginal distributions as well as the correlational coefficients among the distributions. We can also specify the ranges of each variable, similar to the other tools discussed earlier. The corresponding probabilities are computed and displayed on the interactive 3D plots. The following is a use case example for the Trivariate calculator.

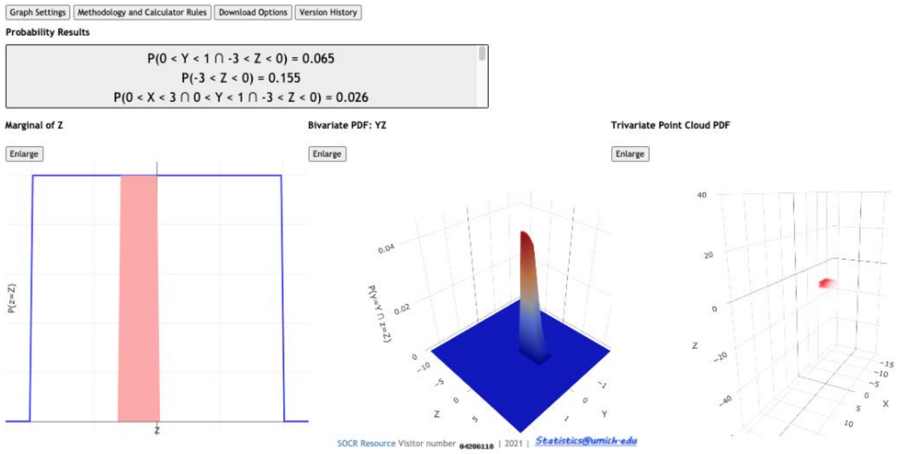
Let's consider a normal distribution with mean of 1 and standard deviation of 4 for the first marginal, another normal distribution with mean of 0 and standard deviation of 0.5 for the second marginal, and a uniform distribution with lower bound of -10 and upper bound of 10 for the last marginal distribution. The correlation coefficient is 0.3 between distributions X and Y, 0.5 between X and Z, and 0.4 between Y and Z. Then, we can use the Controls section of the graph setting to specify the joint distribution for X in the range between 0 and 3, Y between 0 and 1, and Z between -3 and 0.

With all the settings specified, the Trivariate calculator computes and displays the corresponding joint distribution probability. **Figure 8** below shows the output of the Trivariate calculator. In the Probability Results section, it shows that the p-value of the trivariate distribution is 0.026 with the given ranges. It also shows that the p-values of the bivariate joint distribution of Y and Z is 0.065 with the given ranges, and the univariate probability value for Z is 0.155.

The corresponding plots for these distributions are generated and displayed below the Probability Results section. The plot on the left of **Figure 8** shows the PDF for the univariate uniform distribution for distribution Z, where the range of Z values are in red. The 3D surface plot in the middle of **Figure 8** shows the PDF for the bivariate joint distribution between distributions Y and Z, and the 3D (volume) plot on the right shows the PDF of the trivariate joint distribution in a Point Cloud format.

CONCLUSION

Modern technology provides powerful instructional aides to enhance the traditional STEM curricula by integrating observational data, dynamic simulations, and Cloud services. These resources have potentials to improve K-16 and higher education achievements. This manuscript presents examples of interactive probability calculation resources that can be used for teaching and applications of univariate and multivariate processes modeled by specific probability distributions. The joint calculation of *marginal*, *joint* and *conditional probabilities* along with the interactive visualization of the corresponding regions and probability values may be useful in certain educational settings. These probability calculators, relevant learning modules and educational resources, multivariate datasets, and training materials are available on the SOCR website (https://SOCR.umich.edu).



Probability Results

P(0 < Y < 1 ∩ -3 < Z < 0) = 0.065
P(-3 < Z < 0) = 0.155
P(0 < X < 3 ∩ 0 < Y < 1 ∩ -3 < Z < 0) = 0.026

Figure 8. Trivariate Distribution 3D Calculator example

ACKNOWLEDGMENTS

This work was supported in part by NSF grants 1916425, 1734853, 1636840, 1416953, 0716055 and 1023115, and NIH grants P20 NR015331, U54 EB020406, UL1 TR002240, R01 CA233487, R01 MH121079, R01 MH126137, T32 GM141746. The funders had no role in the design, implementation, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Chu, A., Cui, J., & Dinov, I. (2009). SOCR Analyses: Implementation and Demonstration of a New Graphical Statistics Educational Toolkit. *JSS, 30*(3), 1-19.

Dinov, I. (2018). *Data Science and Predictive Analytics: Biomedical and Health Applications using R*: Springer International Publishing.

Dinov, I., Christou, N., & Sanchez, J. (2008). Central Limit Theorem: New SOCR Applet and Demonstration Activity. *Journal of Statistical Education, 16*(2), 1-12. doi:http://www.amstat.org/publications/jse/v16n2/dinov.html

Dinov, I., Siegrist, K, Pearl, DK, Kalinin, A, Christou, N. (2015). Probability Distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions. *Computational Statistics, 594*, 1-19. doi:10.1007/s00180-015-0594-6

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions* Wiley Online Library.

Hsu, S.-B., & Zhao, X.-Q. (2012). A Lotka–Volterra competition model with seasonal succession. *Journal of Mathematical Biology, 64*(1), 109-130. doi:10.1007/s00285-011-0408-6

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy, 44*(6), 101976.

Shang, G., McKie, E. C., Ferguson, M. E., & Galbreth, M. R. (2020). Using transactions data to improve consumer returns forecasting. *Journal of Operations Management, 66*(3), 326-348.

Tanzil, S. S., Hoiles, W., & Krishnamurthy, V. (2017). Adaptive scheme for caching YouTube content in a cellular network: Machine learning approach. *IEEE Access, 5,* 5870-5881.