Info Vis

Article

Information Visualization
1–7

DOI: 10.1177/1473871618756584
journals.sagepub.com/home/ivi
SAGE

Mixed-type distribution plots

Christopher Weld[1] and Lawrence Leemis[2]

Abstract

Plotting is among the most effective ways to quickly and accurately describe a probability distribution. It makes often complex information accessible, enabling intuition for respective outcomes at a glance. Matters complicate, however, for mixed-type distributions. Mixed-type distributions contain *both* continuous and discrete components, and accurately portraying those on a single axis can prove difficult—misleading intuition as a consequence of pulling two otherwise disjoint components into focus together. This article examines the challenges of maintaining the simple, concise, and accurate format of traditional probability distribution plots for mixed-type distributions. We illustrate issues arising within this plot classification paradigm, and why a secondary axis is uniquely suited to improve its communication. An algorithm is devised to consistently scale such plots so that they better coincide with intuition. National Football League football starting field position, meteorological data, and financial instruments provide examples demonstrating effectiveness of this plot technique.

Keywords

Algorithms, data analysis, data visualization, graphical perception, multidimensional scaling, plots, statistical graphics, visual data analysis

Introduction

Mixed-type distributions have *both* continuous and discrete components and are important to applications ranging from business and finance, to actuarial science, meteorology, sports analytics, and queuing. Their effective manipulation and communication is an important component within these fields. Capturing the shape of the probability of a mixed-type distribution in a single comprehensive plot, however, is nontrivial.

Plots are often best to quickly convey the nature of a dataset, but they can also deceive. Poor techniques—inappropriate choices for axes, labels, scale, color, and so on—are often to blame for deceptive illustrations, but when it comes to modeling mixed-type probability distributions, it is more so a matter of poor *circumstance* combined with little existing precedence or guidance. The marriage of otherwise disjoint continuous and discrete probability components lies at the heart of this graphic mischief.

Implications of a mixed-type distribution plot

Two examples will highlight challenges inherent in plotting mixed-type distributions. Each confronts the visual implications of pulling continuous and discrete components into focus under a common set of axes. The plot pairs in Figures 1 and 2 illustrate how at-a-glance intuition suffers within this paradigm. Mixed-type distribution plots can distort perception of the relative contribution of their continuous and discrete components.

[1]Department of Applied Science, College of William and Mary, Williamsburg, VA, USA
[2]Department of Mathematics, College of William and Mary, Williamsburg, VA, USA

Corresponding author:
Christopher Weld, Department of Applied Science, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795, USA.
Email: ceweld@email.wm.edu

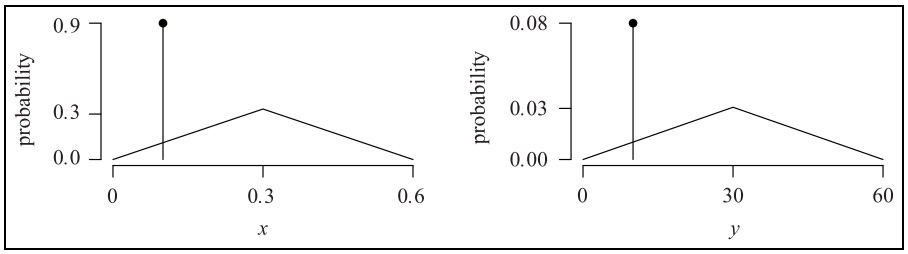

Figure 1. Similar plots, despite vastly different continuous and discrete proportions.

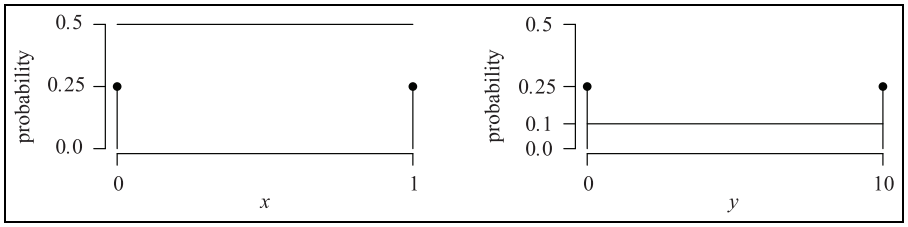

Figure 2. Dissimilar plots, despite identical relative continuous and discrete proportions.

Figure 1 concerns mixed-type random variables X and Y. The random variable X is deterministic(0.1) with probability 0.9 and triangular$(0, 0.3, 0.6)$ with probability 0.1. The random variable Y is deterministic(10) with probability 0.08 and triangular$(0, 30, 60)$ with probability 0.92. Ignoring axis scales, both plots appear identical. This does not, however, imply their relative contributions of continuous and discrete components are proportional. In fact, this example illustrates the counterpoint.

Figure 2 defines random variable X as Bernoulli $(1/2)$ with probability 0.5 and uniform$(0, 1)$ with probability 0.5. The random variable Y is the transformation $Y = g(X) = 10X$. This time, despite similar composition and identical ratios of continuous and discrete components, vastly different plots result.

Both of the preceding examples succeed in presenting information in a concise and accurate way; each completely captures the probability distribution it represents. Where they fail, however, is in their ability to communicate that information *effectively*. Mixed-type distributions unearth a paradigm unfamiliar to many—pulling continuous and discrete into focus together—and can deceive without sufficient inspection.

Background

Tufte[1] defines graphic excellence as communicating complex ideas with clarity, precision, and efficiency. He was among several prominent statistical graphics researchers—including Tukey[2] and Cleveland[3]—who brought focus to the topic in the late twentieth century, and the field of data visualization continues to grow. Its current popularity is not surprising given the recent rise of data-intensive scientific discovery, and the need to visualize information often too complex and/or cumbersome to make sense of numerically or through formulas, as highlighted by Chen and Zhang.[4] Despite growing interest in the field, the authors found no evidence of specific attention given to the visualization of mixed-type random variables. This work is, however, an extension of the authors' Winter Simulation Conference presentation, also available in its proceedings.[5]

Mixed-type random variables in the literature span back to Aitchison[6] who identified a dichotomy in some demand models, whereas a (potentially continuous) parametric distribution represents the purchase amount of a given commodity with exception of an uncharacteristic spike for those abstaining from purchase. Tweedie[7] is later credited with a similar distribution bearing his name, whereas a point mass probability at zero may complement a positive continuous component. Mullahy[8] introduced hurdle models, which use a sequenced Bernoulli trial and subsequent (potentially continuous) distribution outcome for those successfully completing the Bernoulli "hurdle." Zero-inflated data models—a term adopted in the early 1990s largely in conjunction with Lambert's[9] introduction of zero-inflated Poisson (ZIP) distribution—differ from hurdle models only in the their accounting of structural versus sampling zeros. Mixed-type distributions are common in environmental models such as measured rainfall as shown by Feuerverger[10] and

Table 1. Summary data for 2016 NFL regular season kickoff starting field position outcomes—measured as the distance from the return team's end zone. End zone results (touchdowns or safeties) in either direction are assigned to those respective goal lines.

Type	Category	Starting field position	Frequency	Probability
Continuous	Returned in the field of play	$(0, 100)$	1047	$\frac{1047}{2593} \cong 0.404$
Discrete	End zone (return team)	0	3	$\frac{3}{2593} \cong 0.001$
	Touchback	25	1518	$\frac{1518}{2593} \cong 0.585$
	Out of bounds	40	18	$\frac{18}{2593} \cong 0.007$
	End zone (kicking team)	100	7	$\frac{7}{2593} \cong 0.003$
		Total	2593	$\frac{2593}{2593} = 1.000$

contamination concentration levels as shown by Owen and DeRouen.[11] Rainfall is among the most frequently cited mixed-type distribution due to its important implications within the field of hydrology.

Football starting field position

Football starting field position after a kickoff is a mixed-type distribution. Returns within the field of play comprise its continuous outcomes, and discrete outcomes result from National Football League (NFL) rules as given by Goodell.[12] A summary of 2016 NFL regular season kickoff results as assembled by Horowitz[13] is given in Table 1.

Continuous kickoff data are modeled with a kernel density function having a Gaussian smoothing kernel with a bandwidth of 1.84. Figure 3 illustrates the resulting mixed-type probability distribution. Isolating its continuous component produces Figure 4. In total, two distinct modes are evident in Figure 4. The first—at the 22 yard line—represents distances most often attained before a returner is stopped. The second—just past mid-field—is a result of the 54 onside kick attempts during the season, as confirmed by the dotted line.

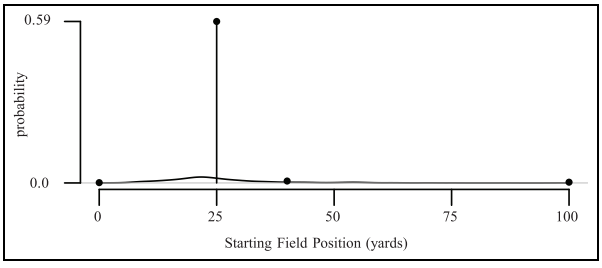

Figure 3. A mixed-type distribution representing 2016 NFL starting field position.

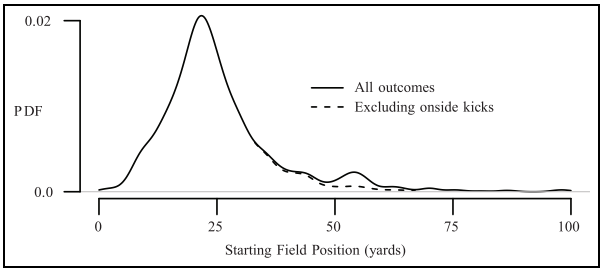

Figure 4. The continuous component of starting field position isolated.

Plot analysis and alternatives

Figure 3 is accurate but ineffective. Its continuous portion provides insufficient detail. The nuanced and noteworthy peaks and troughs evident in Figure 4 are indistinguishable within the low profile of Figure 3. The relative influence of its continuous component—representing 40% of all outcomes—also appears understated in contrast to its discrete outcomes. Finally, several infrequent discrete outcomes appear indistinguishable.

Before looking at a mixed-type probability function plot alternative, first consider its cumulative distribution plot, shown in Figure 5. Although it does well portraying relative discrete and continuous component sizes, it does have flaws. Small discrete outcomes at 0, 40, and 100 yards are nearly imperceptible as a jump in the cumulative distribution. Also, details regarding the peaks and troughs of its continuous component (see Figure 4) are difficult to extract from Figure 5. These issues now motivate the pursuit of a new

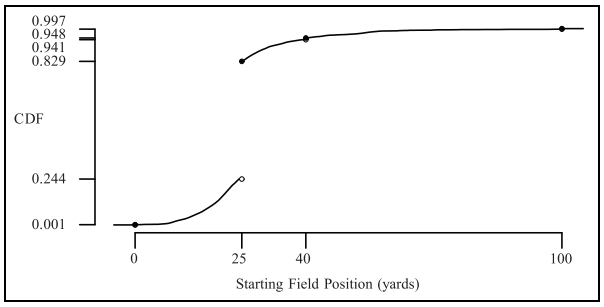

Figure 5. Cumulative distribution function of mixed-type distribution, *X*.

probability function plot, a single plot capable of providing this valuable perspective and intuition.

The methodology producing Figure 6 incorporates two major revisions to enhance readability: a secondary vertical axis and an alternate scale. Secondary axes often draw criticism for their ability to mislead, but mixed-type distributions appear uniquely apt to accommodate them considering their unfavorable graphic consequences under a single set of axes (Figures 1 and 2). The secondary vertical axis enables custom scaling to match intuition. This axis also accommodates relative comparison of small discrete spikes by accentuating them with its square root scale. It is chosen over a logarithmic scale to mitigate the risk of exaggerating the influence of these infrequent discrete events.

Two final touches complete Figure 6. First, midplot labels state respective continuous and discrete portion contributions. This is an effective convention applicable to any mixed-type distribution plot. Next, two silhouettes—a football kicker and a returner courtesy of Freepik[14]—are added; a subjective addition with adequate payoff for this particular example considering a kickoff can occur in either direction. They provide perspective and context by orienting to the underlying scenario responsible for the distribution.

Methodology

Scaling relative continuous and discrete component heights to align with intuition using a secondary axis—as seen in Figure 6—requires user calibration. One heuristic to facilitate this subjective scaling is to adjust the maximum height of the continuous component, relative to its discrete counterpart(s), according to the total probability it represents. Using this methodology, if the integrated probability under the continuous component equals that of a discrete spike in the mixed-type random variable, then those two components also share maximum plot heights. This equivalence arguably aligns with intuition and will, therefore, anchor our calibration of continuous and discrete plot heights. It implies that a discrete spike height exceeding the maximum height of its continuous counterpart also exceeds its probability and vice versa. For example, in Figure 6, the maximum height of the continuous component is roughly two-thirds the height of the maximum discrete spike, since their respective probability ratio is 0.404:0.585.

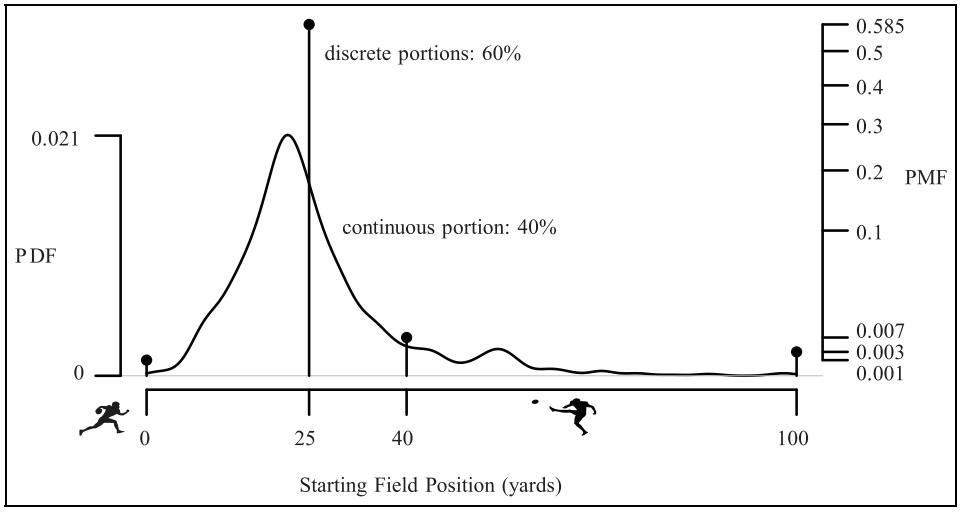

Figure 6. Secondary axes with discrete square root scale for mixed-type distribution, *X*.

Algorithm 1 scales the primary and secondary axes to attain the aforementioned results. It identifies vertical limits for each respective axis, which, in turn, dictate relative plot heights. Vertical limits are the minimum and maximum values visible in the pictured plot area and may differ from the range. For example, in Figure 6, the probability density function (PDF) axis maximum vertical limit is 0.031, whereas its PDF maximum range value is 0.021.

A descriptive summary of Algorithm 1 follows. Its conditional expression from lines 1–7 sets PDF and probability mass function (PMF) vertical limits based on its dominant component—the continuous component or discrete spike accounting for the greatest probability and, therefore, determining the highest point on its corresponding plot. In lines 1–3, its maximum PDF height dictates those limits. Lines 4–6 address when a discrete spike is the dominant plot value. In that case, line 4 scales its PDF vertical limit by the relative probability of that maximum discrete spike to its continuous component probability, $\max_i\{d_i\}/\int_{-\infty}^{\infty} f(x)dx$. Lines 8–12 then compare PDF and PMF vertical limits. The closer these limits are to each other, the lesser the impact of the secondary axis. At some point—if the scales are close enough—simplification to a single vertical axis may benefit plot clarity more than a nuanced adjustment in scale via a secondary axis. This threshold for *close enough* is subjective.

Algorithm 1 often improves mixed-type distribution illustrations; however, matching plot proportions to intuition remains a subjective endeavor. Pulling continuous and discrete components into focus together is an unfamiliar paradigm, so the at-a-glance impression of these plots will vary from individual to individual. Influential to its audience's plot perception is also the continuous distribution shape, whose countless possible permutations are not taken into account under this simple scaling algorithm. Nonetheless, Algorithm 1 holds up well for many circumstances and is a recommended starting point when plotting mixed-type distributions. Examples of its performance are given in the next section.

Additional examples

This section applies Algorithm 1 to three examples. The latter two are real-world systems, but first, we will revisit Figure 1. Figure 7 applies Algorithm 1 to the

Algorithm 1: Mixed-Type Distribution Plots Algorithm

input: $f(x) \leftarrow$ continuous component
$\{d_1, d_2, \ldots\} \leftarrow$ discrete PMF component(s) corresponding to the support values $\{x_1, x_2, \ldots\}$
output: PDF and PMF vertical limits such that the maximum height of the continuous component scales, relative to its discrete counterpart(s), according to the total probability it represents

1 **if** $\int_{-\infty}^{\infty} f(x)dx > \max_i\{d_i\}$ **then**
2 | set PDF vertical limits to $[0, \max_x\{f(x)\}]$;
3 | set PMF vertical limits to $[0, \int_{-\infty}^{\infty} f(x)dx]$;
4 **else**
5 | set PDF vertical limits to $[0, \max_x\{f(x)\} * (\max_i\{d_i\}/\int_{-\infty}^{\infty} f(x)dx)]$;
6 | set PMF vertical limits to $[0, \max_i\{d_i\}]$;
7 **end**
8 **if** PDF vertical limits $\simeq$ PMF vertical limits **then**
9 | use single vertical axis with limits $[0, \max\{\int_{-\infty}^{\infty} f(x)dx, \max_i\{d_i\}\}]$
10 **else**
11 | use respective PDF and PMF vertical axes limits for primary and secondary axes;
12 **end**

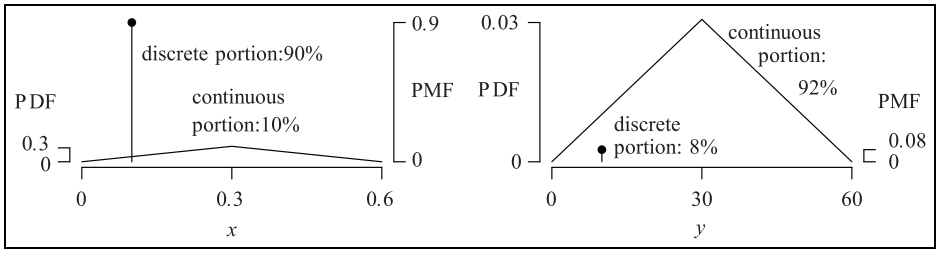

Figure 7. Plots from Figure 1 rescaled with Algorithm 1 scaling methodology.

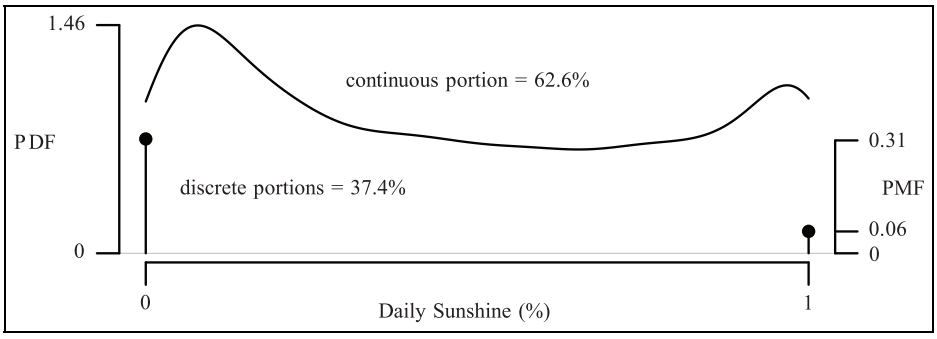

Figure 8. Daily Sunshine for Juneau, Alska, for calendar years 1966–1978.

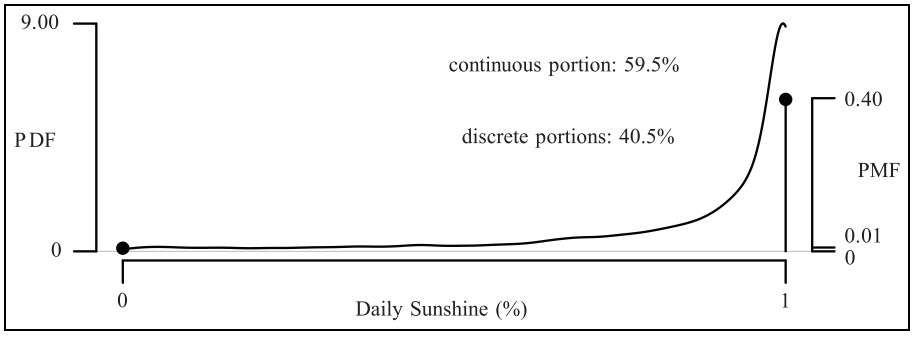

Figure 9. Daily Sunshine for Phoenix, Arizona, for calendar years 1966–1978.

distributions given in Figure 1, yielding results more in line with intuition.

Percentage sunshine is important to solar power generation and serves as the next example. Percentage sunshine is the fraction of possible daylight hours—regardless of length of-day—with direct sunlight (no cloud cover) as recorded at the regional airports of the respective cities illustrated in Figures 8 and 9 as reported by National Oceanic and Atmospheric Administration (NOAA).[15] It has discrete components at 0% and 100% and is one of many meteorological systems exhibiting mixed-type distribution behavior.

Striking contrast between the continuous portion shape in Figures 8 and 9 gives insight into the performance of Algorithm 1 under various circumstances, and some inconsistencies are noteworthy. Both feature roughly 60% continuous probability which accordingly dictates their respective highest plot points; however, the continuous portion of Figure 9 visually garners additional emphasis by maintaining relatively high values throughout its support, whereas the continuous profile for Figure 9 drops substantially at any distance from its mode. A similar circumstance hypothetically arises if the support range must extend to account for a discrete outlier or due to a broader—possibly piecewise—continuous support. Extending its support to include those values implicitly reduces the allotted graphic real estate (and perceived influence) of its continuous component. Despite the need to remain mindful of unique circumstance, results using Algorithm 1 remain a generally effective starting point for visualization.

Financial instruments provide a final example. These monetary contracts dictate terms agreed to by all parties and come in many varieties. Some offer investors an opportunity to mitigate risk of loss at the expense of unrealized gains should the market outperform expectations. A structured note is one type of financial instrument and will illustrate this hypothetical example: An investor agrees to a structured note terms, whereas their money is indexed against the S&P 500. Their profits for the upcoming year are capped according to its terms—say, at 12%—however, they are compensated with downside protection, meaning their original investment is protected from loss. In other words, if the S&P 500 has net losses for the year, they would get a return of their original investment. Assuming the S&P 500 historic geometric mean of 9.5% with standard deviation of 19.7%—as assembled by Damodaran[16]—yields the mixed-type distribution in Figure 10.

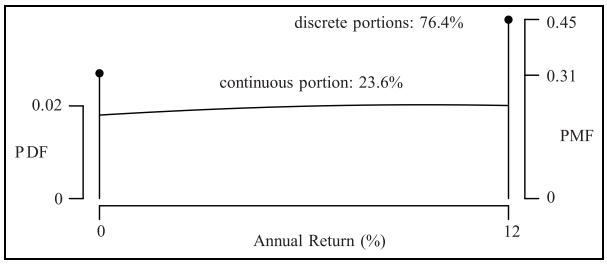

Figure 10. Structured note with downside protection, capped at 12% profit.

Summary

Statistical graphics are a tremendous vehicle to quickly communicate the nuanced landscape of a stochastic variable, but plotting mixed-type distributions is a nontrivial endeavor. Pulling components of continuous and discrete probability into focus together often occurs at the expense of equitable presentation of each. To counter these implications, it is necessary to balance the complexities inherent with changing the standard paradigm view of a probability distribution with its associated benefits. The relative portrayed sizes of its continuous and discrete components are at the center of this deliberation, with the goal of avoiding reader misinterpretation. The fickle nature of mixed-type distribution plots justifies consideration of a secondary axis, capable of adjusting relative continuous and discrete component plot heights to better align with intuition. An algorithm to customize the two components tunes the maximum height of the continuous component, relative to its discrete counterpart(s), according to the total probability it represents.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The first author was supported in part by the Omar Nelson Bradley Research Fellowship in Mathematics.

ORCID iD

Christopher Weld https://orcid.org/0000-0001-5902-9738

References

1. Tufte ER. *The visual display of quantitative information*. Cheshire, CT: Graphics Press, 1983.
2. Tukey JW. *Exploratory data analysis*. North Reading, MA: Pearson, 1977.
3. Cleveland WS. *The elements of graphing data*. Monterey, CA: Wadsworth Advanced Books and Software, 1985.
4. Chen CP and Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inform Sciences* 2014; 275: 314–347.
5. Weld C and Leemis L. Modeling mixed type random variables. In: *Proceedings of the 2017 winter simulation conference*, Las Vegas, NV, 3–6 December 2017, pp. 1595–1606. New York: *IEEE*.
6. Aitchison J. On the distribution of a positive random variable having a discrete probability mass at the origin. *J Am Stat Assoc* 1955; 50(271): 901–908.
7. Tweedie M. An index which distinguishes between some important exponential families. In: Ghosh JK and Roy J (eds) *Statistics: applications and new directions. Proceedings of the Indian statistical institute golden jubilee international conference*. Calcutta, India: Indian Statistical Institute, 1984, pp. 579–604.
8. Mullahy J. Specification and testing of some modified count data models. *J Econometrics* 1986; 33(3): 341–365.
9. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; 34(1): 1–14.
10. Feuerverger A. On some methods of analysis for weather experiments. *Biometrika* 1979; 66(3): 655–658.
11. Owen W and DeRouen T. Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants. *Biometrics* 1980; 36: 707–719.
12. Goodell R. 2016 official playing rules of the National Football League. *Natl Footb Leag* 2016; 1: 1–81.
13. Horowitz M. nflscrapR: R package for scraping NFL data off their JSON API, https://github.com/maksimhorowitz/nflscrapR (accessed January 2018).
14. Freepik. American football player silhouettes collection, www.freepik.com/free-vector/american-football-player-silhouettes-collection_722363.htm (accessed January 2018).
15. National Oceanic and Atmospheric Administration. National centers for environmental information, https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND (accessed January 2018).
16. Damodaran A. Annual returns on Stock, T.bonds and T.bills: 1928–current, http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html (accessed January 2018).