

Statistical Characterization, Pattern Identification, and Analysis of Big Data

Zhigang Wei
 Tenneco Inc.

Kamran Nikbin
 Imperial College London

ABSTRACT

In the Big Data era, the capability in statistical and probabilistic data characterization, data pattern identification, data modeling and analysis is critical to understand the data, to find the trends in the data, and to make better use of the data. In this paper the fundamental probability concepts and several commonly used probabilistic distribution functions, such as the Weibull for spectrum events and the Pareto for extreme/rare events, are described first. An event quadrant is subsequently established based on the commonality/rarity and impact/effect of the probabilistic events. Level of measurement, which is the key for quantitative measurement of the data, is also discussed based on the framework of probability. The damage density function, which is a measure of the relative damage contribution of each constituent is proposed. The new measure demonstrates its capability in distinguishing between the extreme/rare events and the spectrum events. Several case studies including vehicle reliability, vehicle road test score, warranty, salary distribution of an institution, the city population distribution in 3 countries, and the earthquake distribution worldwide and in the USA, are provided to demonstrate the role of the statistical and probabilistic approaches in the characterization and analysis of the big data.

CITATION: Wei, Z. and Nikbin, K., "Statistical Characterization, Pattern Identification, and Analysis of Big Data," *SAE Int. J. Mater. Manf.* 10(3):2017, doi:10.4271/2017-01-0236.

1. INTRODUCTION

Data-based modeling, prediction, forecasting, and associated decision-making is becoming an indispensable approach in many engineering applications, business operations, prevention and mitigation of natural disasters, and analysis of social events [1]. In the Big Data era, how to extract useful information from the sea of data and how to find the general trends and the basic structures of the data are the key to comprehend the world, to win business, and to survive. Statistics and probability are the foundation of Big Data analysis, and the occurrence of an event can often be appropriately described with a probabilistic distribution function.

A probabilistic distribution function roughly consists of two parts: the middle and the tails. Based on the contribution of the tails to the overall damage or impact, probabilistic events can be divided into two categories: (1) spectrum events, and (2) extreme/rare events [2]. The damage caused by the constituents in the spectrum events is comparable, oftentimes, dominated by the mean behaviors. The traditional continuous two-parameter normal, lognormal, and Weibull distribution functions are often used to describe the probabilistic distribution of the spectrum events such as fatigue life. By contrast, the damage caused by the extreme/rare events, such as 100-year flood or mega-earthquake or nuclear accident etc. is controlled by the tails. Probabilistic distribution functions such as the Pareto power law and exponential functions are more appropriate than the traditional functions in interpreting these extreme/rare events. It should be

emphasized here that many phenomena in both the natural and social sciences have power law statistics (Pareto distribution). These phenomena include city sizes, incomes, word frequencies, earthquake magnitudes, and many other natural and manmade engineering events [3]. A power law distribution implies that small occurrences are extremely common, whereas large instances are extremely rare. In extreme/rare events, the very large and rare events have more significant impact and consequence than the small and common events. It should be noted that the extreme/rare events are very complex. It has been found that despite the fact that a power law models the tails of the empirical distribution well, the largest events are significantly outlying, meaning much larger or smaller than what would be expected under the power law. Such events are interpreted as "Dragon Kings" as they indicate a departure from the generic process underlying the power law [4, 5]. The extreme/rare events are difficult to understand in their nature and are traditionally treated as "Black Swan" phenomena [6] meaning they rarely happen, with significant consequences when they happen, but are unpredictable. However, since these phenomena bear significant consequences, deep understandings of these "Dragon Kings" behaviors and substantial development of the related predictive tools would be highly demanded [4, 5]. These two categories of the probabilistic events are different in nature. How to characterize and categorize these probabilistic events and subsequently analyze the data are critical in engineering, natural, and social areas in which the analysis of Big Data is concerned.

In this paper, the general probabilistic description of events is given first, with emphasis on the Weibull and Pareto power law distributions. Level of measurement, which is fundamental in quantitatively assessment of probabilistic events is also reviewed and discussed. Events categorization based on commonality/rarity and impact/effect are described subsequently and an event quadrant is proposed for ease of events categorization. To facilitate the damage analysis and distinguish between extreme/rare events and the spectral events, the concept of a damage density function is proposed and some analysis results are discussed. Finally, statistical data from a wide variety of engineering, natural, and social events are analyzed to demonstrate the concepts and ideas developed in this paper.

2. CHARACTERIZATION OF PROBABILISTIC EVENTS

2.1. Description of a probabilistic event

The probabilistic behavior of an event can be simply and effectively described using the probability ($P-x$) diagram shown in Fig. 1. The horizontal axis in Fig. 1 represents the variable of interest while P along the vertical axis represents the cumulated distribution function (CDF), which relates the probabilistic density function (PDF) $p(x)$ through $P(x) = \int_a^x p(x)$ and $P(b) = \int_a^b p(x) = 1$. The lower limit a and the upper limit b are finite values for most of the real problems. The infinities $a = -\infty$ and $b = \infty$ are often a mathematical idealization.

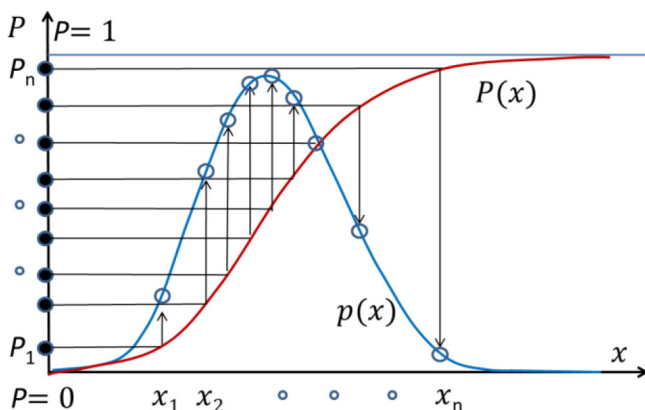


Fig. 1. Schematic of the probabilistic density function (PDF) and the cumulative distribution function (CDF) in the probability diagram.

For a sequence of n independent and random data points, there will be a corresponding discrete sequence of values of x , say $x_1, x_2, \dots, x_n$ (ranked in an ascending order). No matter what the $p(x)$ looks like, $P(x)$ can always be estimated based on the ranked data. The accuracy of the estimation depends on the sample size and the estimation model. For a given reasonably accurate model, the larger the sample size, the more accurate the estimated $P(x)$. The simplest estimation of $P(x)$ is i/n [7], which obviously has some issues. For example, when $i = n$, the formula results in $P(x_n) = 1$, which contradicts the fact that any randomly generated data can be close to but never reach $P = 1$. i is the rank order number, and again n is the total number of data points. Similarly, several other simple formulae such as $(i-1)/n$, $(i-1/2)/n$, and $i/(n+1)$ have similar issues [7]. Based on the cumulative

binomial distribution and median rank (MR) formula, the so called Benard's approximation, Eq. (1), has been developed and it is now widely used in industry to estimate $P(x)$ [7].

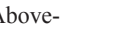
$$P(x_i) \approx MR = \frac{i - 0.3}{n + 0.4} \quad (1)$$

It should be noted that the randomly generated data as projected onto the vertical axis P as shown in Fig. 1 is uniformly distributed because of the equal opportunity of the occurrence of each event in terms of the accumulated function. Therefore, the binomial distribution function can be assumed regardless of the information about what the exact forms of $p(x)$ and $P(x)$ are. It should be emphasized here that the generation process of uniformly distributed random number is also a key step in Monte Carlo simulation for generating random number for any specific distributions, such as Weibull or Lognormal.

The working mechanism of creating a probability plot using the median rank approach is schematically illustrated in the $P-x$ diagram shown in Fig. 1. The distribution of $p(x)$ is determined by the location of the data points on x axis. $P(x)$ can be obtained by integrating $p(x)$, theoretically, or by using the approximation formula, Eq. (1). The data horizontally projected onto the vertical axis from the $P(x)$ curve is eventually uniformly distributed. In other words, corresponding to the uniform distributed data points on the vertical axis P , the intervals between the adjacent points on x axis are unequally spaced with the dense segments at the peak of the $p(x)$ curve. The segments become coarser as the data points go further away from the peak.

2.2. Level of Measurement

The exact forms of $P(x)$ and $p(x)$ also depend upon the level (unit) and the scale of the x axis. Level of measurement or scale of measure is a classification that describes the nature of information within the numbers assigned to variables. Some physical quantities, such as cycles to failure, energy, city population etc., can be directly used. The use of unit in log or exponential form and other objective and subjective measurements are also very common. In addition to the precise probabilistic descriptions as illustrated in Section 2.1, oftentimes, classification with 3-, 4-, 5-, 10-, 12-, 100-level measures are also frequently used for ease of communication and simplicity. For example, Beaufort wind force scale (0-12) is an empirical measure that relates wind speed to observed conditions at sea or on land. Richter magnitude scale for earthquake assigns a magnitude number to quantify the energy or moment released by an earthquake. Richter scale is a base-10 logarithmic scale, which defines magnitude as the logarithm of the ratio of the amplitude of the seismic waves to an arbitrary, minor amplitude. Earthquakes are also classified as moderate if their magnitude is in the range of 5-5.9, strong if the magnitude is in the range of 6-6.9, major if the magnitude is in the range of 7-7.9, and great if the magnitude is 8 or larger [8]. The decibel (dB) is also a logarithmic unit used to repress the ratio of two values of a physical quantity, often power or intensity. The Numeric Rating Scale (NRS-11) is an 11-point scale for patient self-reporting of pain. The FICO credit scores are designed to measure the risk of default by taking into account various factors in a person's financial history. The generic or classic FICO score is between 300 and 850.

There are some other measures without numbering system but with colors and graph symbols (ideogram) to represent severity level. For example, Consumer Reports [9] uses five modified Harvey Balls to assess reliability performance of products: Open-Red-Circle: , Half-Red-Circle , Open-Black-Circle , Half-Black-Circle , and Filled-Black-Circle , in order from better to worse. J.D. Power [10] uses Power Circle ratings of five, four, three, and two: (1) Among-The-Best , Better-Than-Most , Above-Average , and the Rest . J.D. Power does not publish a rating lower than two Power Circles. Similar examples also include: the US Homeland Security Advisory System is a color-code terrorism threat advisory scale: Green (Low), Blue (Guarded), Yellow (Elevated), Brown (High), and Red (Severe). Clearly, different measures of the variables result in different statistical characterization methods and associated probabilistic distributions. Therefore, the empirical interpretation of the probabilistic behavior must be consistently based on a certain level of measurement.

2.3. Two Types of Probabilistic Distribution Functions

The commonly used Weibull distribution for spectrum events and the Pareto distribution for extreme/rare events are described below for events that can be described using a continuous probabilistic distribution function.

2.3.1. Weibull Distribution Functions

In many engineering applications, Weibull distribution functions are often used to describe a spectrum event, with which all of the spectra under normal conditions are evaluated so that all major contributions can be properly considered [1]. These spectra include the loads experienced by the components in nuclear power plants, ground vehicles, and landing gears of aircrafts. The three-parameter Weibull distribution, Eq. (2), is such a probabilistic distribution function to describe the spectrum events.

$$p(x) = \left(\frac{\beta}{\eta}\right) \left(\frac{x-a}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{x-a}{\eta}\right)^{\beta}\right] \quad (2)1$$

$$P(x) = 1 - \exp\left\{-\left[\frac{(x-a)}{\eta}\right]^{\beta}\right\}; 0 < a \leq x < \infty, \eta, \beta > 0 \quad (2)2$$

When $a = 0$ in Eq. (2), the three-parameter Weibull distribution can be further reduced to the conventional two-parameter Weibull distribution. Both three-parameter and two-parameter Weibull functions have been implemented into commercial software, such as Minitab [11], and therefore, the detailed analysis will not be provided here.

2.3.2. Pareto (Power Law) Distribution Function

When dealing with crises and extremes, power law tails are “normal” [3]. The unique property of power law is that they are scale-invariant/self-similar/fractal. This property implies that all events, both large and small, are generated by the same mechanism. A continuous variable with a power law distribution has a probability $p(x)dx$ of

taking a value in the interval from x to $x + dx$, where $p(x) = Hx^{-\alpha}$ with $\alpha > 0$ [3]. The constant in the power law is given by the normalization requirement that $\int_{x_{\min}}^{\infty} p(x) = 1$. Then, $H = (\alpha - 1)x_{\min}^{\alpha-1}$ must be held. $x_{\min}$ is the lower bound of the distribution, above which the power law is obeyed. Finally, the Pareto PDF and CDF can be expressed in Eq. (3).

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} \quad (3)1$$

$$P(x) = 1 - \left(\frac{x_{\min}}{x}\right)^{\alpha-1} \quad (3)2$$

The fit parameter α of the Pareto distribution can be derived by applying the least squares method or the maximum likelihood method. However, results show that the maximum likelihood method provides a better data correlation than the least squares method [3]. The natural log likelihood function is shown in Eq. (4).

$$L(\alpha) = Ln \left[\prod_{i=1}^n \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}}\right)^{-\alpha} \right] \quad (4)$$

The maximum likelihood is found by differentiating $L(\alpha)$ in Eq. (4) with respect to parameter α , setting the result equal to zero. Upon rearrangement, this yields the estimator equation

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^n \ln \left(\frac{x_i}{x_{\min}}\right) \right]^{-1} \quad (5)$$

where x_i ($i = 1, \dots, n$) are the n data points for $x_i \geq x_{\min}$. In practical situations $x_{\min}$ usually corresponds not to the smallest value of x measured but to the smallest for which the power law behavior holds [3].

Identifying power law behavior is important but the process can be tricky [3]. The standard strategy makes use of a histogram of a quantity with a power law distribution appears as straight line when plotted on logarithmic scales. However, bin size selection is always an issue for this treatment and noise can be generated that hinders the data interpretation.

Another, and in many ways a superior, method of plotting the data is to calculate a CDF [3]. Instead of plotting a histogram of the data, a plot of the CDF, called complementary CDF (C-CDF or CP) here, has a value greater than or equal to x : $CP(x) = \int_x^{\infty} f(x') dx'$. If the distribution follows a power law, then

$$CP(x) = \frac{H}{\alpha - 1} x^{-(\alpha-1)}$$

 Documents
Document Sharing Hub

thus, the C-CDF also follows a power law, but with an exponent of $\alpha - 1$, indicating a straight line on logarithmic scales, but with a shallower slope if the power law distribution is held [3]. This can be used as a check if a distribution follows the power law distribution. But notice that there is no need to bin the data at all to calculate $CP(x)$. By its definition, $CP(x)$ is well-defined for every value of x and so can be plotted without binning.

2.4. Events Quadrant

In the natural, social, and engineering world, some events may be very common, such as low amplitude fatigue vibration as experienced by a vehicle and the low amplitude earthquakes, but their impact might not be remarkable, whereas some events might be rare, but their consequence might be significant, such as a mega-earthquake. Based on the combination of the commonality/rarity and the impact/effect an event can be roughly categorized into one of the four events as shown in the events quadrant shown in Fig. 2.

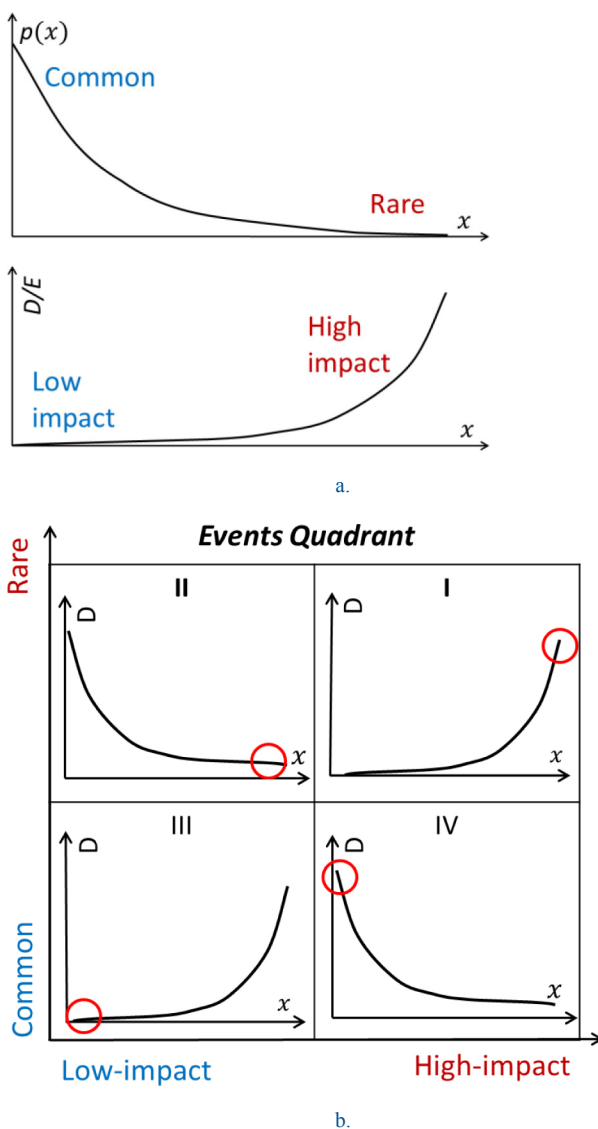


Fig. 2. Schematic of (a) a PDF function in Pareto power law form and the corresponding damage per event (D/E), and (b) the events quadrant based on the combination of commonality/rarity and impact/effect.

The first plot in Fig. 2(a) schematically shows a monotone decreasing probabilistic function, which could be a Pareto type. Therefore, common events occur on the left side and the rare events occur on the right side. It is assumed that the curve of damage per event (D/E) shown in the second plot of Fig. 2(a) is a monotone increasing function of the magnitude, which is often the case in fatigue and earthquake. Clearly, an event that happens on the right side of the D/E curve shows higher impact than that on the left side. Fig. 2 (b) shows four typical combinations of the events in terms of total possible damage (D): Quadrant-I: Rare-High Impact, Quadrant-II: Rare-Low Impact, Quadrant-III: Common-Low Impact, and Quadrant-IV: Common-High Impact.

2.5. Event Characterization with Damage Density Function

How to quantitatively discriminate between spectrum events and extreme/rare events is an interesting and important topic. Here in this paper, a damage density function is developed as a parameter to characterize the two events. The damage density function is proposed and expressed in Eq. (7)

$$g(S) = p(S)D(S) \quad (7)$$

$g(S)$ in Eq.(7) consists of two parts: (1) the occurrence probability $p(S)$ at level S , earthquake amplitude for example, and (2) the damage $D(S)$ made at the level S . Clearly, $g(S)$ shown in Eq.(7) measures the damage probability per event at a level S , so called the damage density function. The total damage $E[D]$ is an important parameter and valid as long as all of the damaging events are properly considered. However, $E[D]$ alone cannot determine which event is dominating and which event is negligible in terms of damage. In these cases, the damage density function $g(S)$, Eq. (7), provides some unique advantages. Several examples are provided below to demonstrate the properties and the effectiveness of $g(S)$.

Without losing generality, the damage per event (D/E) can be written as Eq. (8).

$$D(S) = (S/C)^m \quad (8)$$

Eq. (8) can be derived from a power law relationship between load amplitude and characteristic life or damage. A specific example is the fatigue S-N curve [2], Eq. (9).

$$S = CN^{-1/m} \quad (9)$$

Eq. (8) can be obtained from Eq. (9) with a linear damage assumption $D = 1/N$. Three different $p(S)$ are considered to demonstrate the power of the density function $g(S)$: (1) the Pareto power law, (2) the exponential distribution, and (3) the two-parameter Weibull distribution.