Rare, outlier and extreme: beyond the Gaussian model and measures

Paul C. Nystrom and Ehsan S. Soofi*

Sheldon B. Lubar School of Business,
University of Wisconsin-Milwaukee,
P.O. Box 742, Milwaukee, WI 53201, USA
E-mail: nystrom@uwm.edu
E-mail: esoofi@uwm.edu
*Corresponding author

Abstract: Probability models for rare, outlier and extreme outcomes are different than the Gaussian (normal) distribution commonly used in management research. This paper illustrates the theoretical basis and implementation of these concepts. One example uses data on organisational size and compensation of CEOs in large US corporations in order to illustrate rare and outlier outcomes. Models that fit the data on these variables are very different than the Gaussian distribution, so mean, standard deviation and correlation are useless here. Another example uses data collected from business executives' economic forecasts shortly after the 9/11 terrorist attacks to illustrate how to identify extreme outcomes and a Bayesian approach for inferring relationship between extreme outcomes and strategy type. Differentiations between extremes and rares are illustrated using data simulated by a Monte Carlo method. Visualisation of fit of a model for data by Q-Q plot and discussion of distributional testing precedes some concluding remarks.

Keywords: rare outcomes; outliers; Gaussian model; Bayesian approach; CEO compensation; organisational size; business strategy; economic jolt; extremes.

Reference to this paper should be made as follows: Nystrom, P.C. and Soofi, E.S. (2012) 'Rare, outlier and extreme: beyond the Gaussian model and measures', *Int. J. Complexity in Leadership and Management*, Vol. 2, Nos. 1/2, pp.6–38.

Biographical notes: Paul C. Nystrom is Professor Emeritus of Management in the Lubar School of Business at the University of Wisconsin-Milwaukee, where he had been the Robert and Sally Manegold Professor of Organisations and Strategic Planning. Some of his research has appeared in *Administrative Science Quarterly*, *Academy of Management Journal*, *Journal of Applied Psychology*, *Management Science*, *Managerial & Decision Economics*, *Journal of International Management*, *Organisational Research Methods*, and *Journal of Management*. He served as an Associate Editor of *Management Science*.

Ehsan S. Soofi is a University of Wisconsin-Milwaukee Distinguished Professor at the Lubar School of Business and a Research Associate of the Center for Research on International Economics at the university. He is a Fellow of the American Statistical Association. His research areas include information theoretic and Bayesian statistics, distribution theory and model fit, and their applications in business and economic problems. His research has appeared in *Journal of American Statistical Association (JASA)*, *Journal of Royal Statistical Society*, *Biometrika*, *Statistical Science*, *International*

Statistical Review, *Journal of Multivariate Analysis*, *Journal of Econometrics*, *IEEE Transactions on Information Theory*, *Operations Research*, *Journal of Applied Probability*, *Naval Research Logistics*, *Marketing Science*, *Journal of Management Information Systems* and *Organizational Research Methods*. He served as an associate editor of *JASA* and is currently serving as the Guest Editor of *Econometric Reviews*, *Special Issue on Bayesian Inference and Information: In Memory of Arnold Zellner*.

This paper is a revised and expanded version of a paper entitled 'Identification and analysis of extremes with two examples: executives' forecasting after 9/11 and CEO salaries' presented at the 14th Organization Science Winter Conference (OSWC-XIV), The Resort, Olympic Valley, Squaw Creek, California, USA, 6–10 February 2008.

1 Introduction

Statistical analyses of management phenomena usually employ the *Gaussian* model of the bell-shaped normal curve for the data. The dominance of Gaussian statistics for nearly a century has deeply penetrated our thinking and shaped scholars' views on the notions of typical, outlier, rare, and extreme data points, as well as the relationships between variables. A great scientist, Carl Friedrich Gauss, in 1809 discovered the normal distribution as a way to rationalise the method of least squares for estimating a quantity measured with errors. Later, it was discovered that under certain conditions the distributions of variables constituted by sums (and averages) of a large number of variables tend to the Gaussian distributions (central limit theorems). These formed the foundations for statistical measures and methods that are useful for applications when measurements are like Gauss's error and the conditions of the central tendency hold.

In the Gaussian world, the frequency of measurements resembles a nicely behaved distribution having a unique symmetric shape, a unique centre where the centre of gravity of the distribution (mean), the 50%–50% divide (median), and the most frequent outcome (mode) are the same point, a measure of spread (standard deviation, SD) that determines the proportions of distribution around the centre, and no marked gap is expected between the data points. With a Gaussian model, any data point beyond three SD or so is suspiciously different and usually thought to be recorded incorrectly or else generated from a different distribution such as another Gaussian having a different mean or different SD. In the Gaussian world, one variable is either related to another variable only through its mean (regression function) linearly or the variables are independent – one variable has no predictive power about the other. Since no other relationship is possible, a single parameter (correlation) can map the strength of the only possible relationship between the variables. Yet the underlying phenomena of interest might actually be generated by probability distributions very different from the Gaussian model (McKelvey and Andriani, 2005; Andriani and McKelvey, 2007; O'Boyle and Aguinis, 2012), so that the usual measures such as mean, variance and correlation are not useful or may not be defined (being infinite). In this paper, we illustrate several different distributions including heavy-tail ones that can generate outcomes with marked gaps between them, and for which the mean, variance and correlation are not defined.

The three terms of rare, outlier and extreme are often used interchangeably. For example, a special issue of *Organization Science* focusing on rare events notes that "Rare events are often set aside as statistical outliers" [Lampel et al., (2009), p.835]. This paper provides readers with a way to think about the meaning of rare, outlier and extreme data points in a statistical analysis. Recall that for a statistical inference the data are viewed as being a random sample of observations generated from a probability distribution. This view is key to this thinking about the probability distribution. *Rare* and *outlier observations or data points* can be visually seen on plots as those outcomes that are substantially distant from the mass of the data. If such data points are not erroneously recorded, then they are outcomes generated from a probability distribution. We will illustrate that a seemingly uncommon observation can be produced along with the mass of data by a single probability distribution markedly different from a Gaussian model. A seemingly unusual rare outcome can be a legitimate outcome of a probability distribution whose tail decays slowly, referred to as a heavy-tail distribution. A seemingly unusual outlier can be a legitimate outcome of a probability distribution that is a weighted average of two or more distributions, where the weight of one of the distributions relative to the others is very high, which is the underlying distribution of the mass of outcomes. Such distributions are referred to as mixture distributions, sometime referred to as a contaminated distribution. Some heavy-tail distributions that can produce a rare outcome are also mixture distributions in a mathematical sense. This duality provides a plausible explanation for the common feature of rare and outlier as being outcomes substantially distant from the mass of the data. In this paper we illustrate the distributional issues pertaining to rare and outlier outcomes using data on some typical variables of management research. We include an Appendix for a brief visual comparison of several distributions; more details and references can be found at Wikipedia (Heavy-tailed distribution and Mixture distribution).

Extreme outcomes in a sample refer to outcomes that are farthest from the centre (median) of the data, namely the sample minimum and maximum. Thus, an extreme outcome is not necessarily far distant from the mass of the data. The data can be a random sample from a Gaussian or any other distribution. Yet the distribution of a sample extreme point, say maximum, is related but very different than the distribution that has generated the entire sample. For example, the distribution of the maximum of a Gaussian sample is not a Gaussian distribution. Consequently, the traditional Gaussian methods are not applicable for making inferences about extremes. In this paper, through a management research example, we illustrate data points that may be considered as plausible outcomes of the distribution of sample minimum or sample maximum, hence, inferred as extremes. We also illustrate making inference about the relationship between extremes of one variable with a covariate variable. Details for these procedures are given in Nystrom et al. (2010). Methods for statistical analyses involving rare, outlier and extremes are abundant and lie beyond the scope of this paper; for examples, see Beirlant et al. (2004), Coles (2001), Resnick (2007), and Schwertman and de Silva (2007). Also see McKelvey and Andriani (2005), Andriani and McKelvey (2007), and Schwab et al. (2011).

The next section illustrates rare and outlier outcomes through an example of organisational size and compensation data for top-paid CEOs of large US companies. This Section 2 also presents summary measures suitable for data containing rare and outlier outcomes. Section 3 presents models for extremes and it illustrates our method for

computing thresholds that identify extremes using a management example that focuses on the economic forecast accuracy of business executives after the 9/11 jolt. This management example also shows how to analyse an association with a covariate (strategy) by using a Bayesian approach and it also illustrates differences between extreme and rare outcomes by using a Monte Carlo simulation method. Section 4 presents an extension for considering order statistics – such as deciles or quartiles – illustrated with data from a study on the organisational implementation of new technologies. The penultimate Section 5 presents ways to visualise and test model fit. Section 6 provides some concluding remarks. An Appendix provides examples of probability models suitable for data containing rare and outlier outcomes in contrast with the Gaussian mode. A second Appendix shows measurement of forecast inaccuracy that we use in Section 3.

2 Organisational size and CEO compensation

This section illustrates rare and outlier outcomes using data on two organisational size variables and two CEO compensation variables for the 50 top-paid CEOs of large US companies in 2006. The organisational size data are from Fortune magazine's annual report on the 500 largest US companies. The compensation data are from Forbes magazine's annual report on top-paid 500 CEOs of US corporations. The size variables studied are total sales and total assets. Forbes defines CEOs' total compensation as the sum of salary, bonuses, perks (such as company-paid club memberships), vested stock grants, stock gains, and the value realised by exercised stock options during the year. Exercised stock options now usually account for most of the 'additional compensation' beyond the 'salary and bonus' component. The amounts of compensation received by many CEOs of US's largest companies have been characterised as being excessive, insufficiently performance-based, unfair and dysfunctional (e.g., Bebchuk and Fried, 2004; Dittmann and Maug, 2007; Dow and Raposo, 2005; Harris and Bromiley, 2007; Jensen et al., 2004; Lie, 2005; Siegel and Hambrick, 2005; Tosi et al., 2000; Wade et al., 2006).

Table 1 Summary statistics for organisational size and compensation variables for the 50 top-paid CEOs in 2006

	n	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*	*Mean*	*S.D.*	*Skewness*
Organisational size($billions)									
Sales	47	1.5	5.79	15.7	46.71	167.6	28.3	34.0	2.19
Assets	47	1.5	8.60	32.4	104.30	1459.7	158.2	329.4	2.88
Compensation ($millions)									
Salary and bonus	50	0.000001	3.08	5.0	9.04	32.2	7.7	7.2	1.69
Additional comp	50	9.2	27.30	38.4	65.70	646.6	71.9	103.9	4.08

Table 1 shows the summary statistics for the data. Q1 and Q3 are the first and third quartiles. Note that the Q1 and median of each variable is substantially closer to the minimum than the Q3 and median to the maximum. These measures together indicate that the distributions of these variables are markedly skewed, which is also confirmed by the high coefficients of skewness. Thus, unlike for Gaussian data, the mean and SD reported in the table are useless and possibly meaningless (as shown in Section 2.2) as descriptive measures for the underlying distributions of these variables.

Management researchers often perform log transformation on the data having skewed distributions as a way to perform analyses using methods that require the Gaussian model assumption. Figure 1 shows dot plots for the organisational size variables and their natural logarithm in Panels (a) to (d) and shows dot plots for the components of the CEO compensation and their natural logarithm in Panels (e) to (h). *Dot plot* is a histogram-type display that shows individual data points for moderate sample sizes, hence is suitable for visualisation of rare, outlier, and extreme outcomes. The horizontal scales of the graphs are different due to the ranges of the sampled data.

Figure 1 Panels (a) to (d) Dot plots of organisational size variables ($ billions) and their natural logarithm for organisations led by the 50 top-paid CEOs in 2006 (see online version for colours)

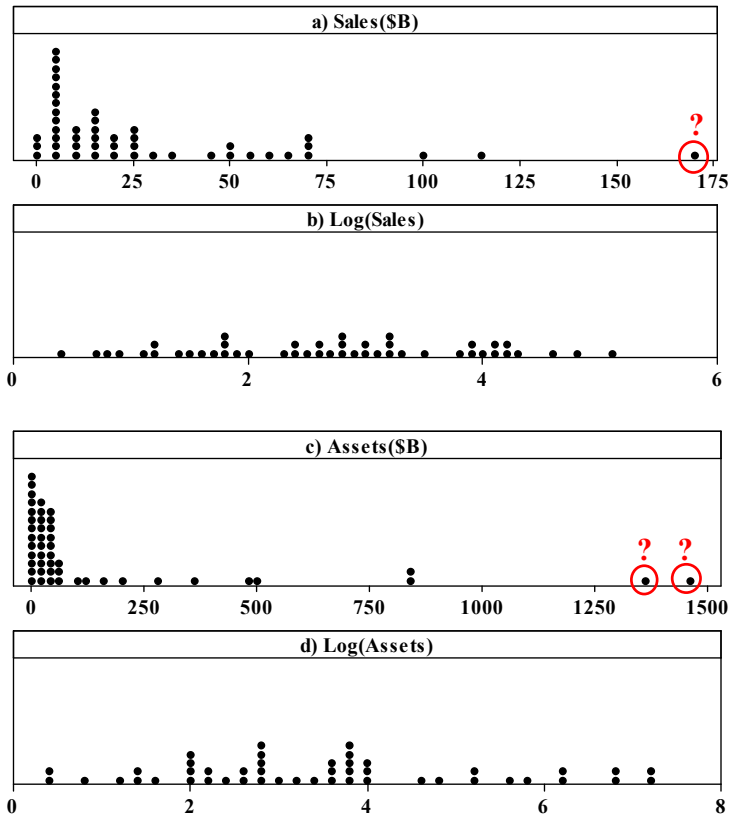
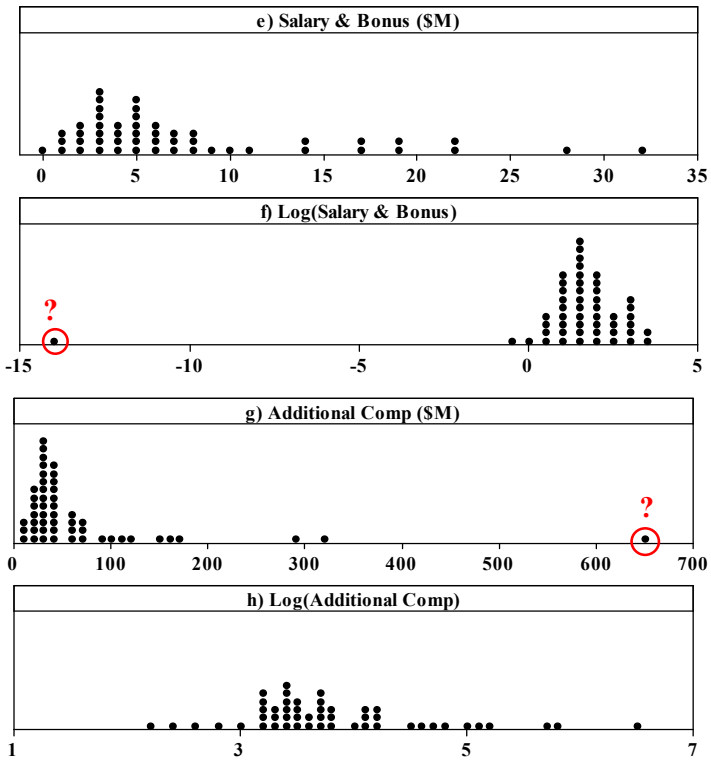


Figure 1 Panels (e) to (h) Dot plots of compensation variables ($ millions) and their natural logarithm for organisations led by the 50 top-paid CEOs in 2006 (see online version for colours) (continued)



In Figure 1, the plot of raw data for each variable is juxtaposed with the plot of its log-transformed version below it. The raw data plots show that the distributions of all four variables are skewed to various degrees. The points circled in red in Panels (a), (c), (f) and (g), are far from the mass of data, which a researcher might think are rare or outlier observations. The juxtaposition clearly highlights that log transformation can resolve some problems of data points that look like outliers as raw data in Panels (a), (c) and (g), and can uncover an outlier in log transformed data [Panel (f)] that was not apparent in raw data [Panel (e)]. While the distributions of the raw data for these four variables are highly skewed, the distributions of the log-transformed variables seem fairly symmetric for three of them [Panels (b), (d) and (h)], but the distribution of the log of Salary and bonus [Panel (f)] is symmetric for the mass of outcomes except for a single point which stands alone from the mass. Although the points circled in red in Panels (a), (c), (f) and (g) are distant from the mass of data, the circled points in Panels (a), (c) and (g) look to be parts of the overall patterns of the respective plots. In each of these three panels, we see an interval of high frequency points followed by clusters of a few points that appear in decreasing sizes and increasing spread from each other progressively. But the circled point in Panel (f) is not part of any pattern of separation from the mass of points. This contrast illustrates the difference between samples that might be generated from a single probability distribution [Panels (a), (c) and (g)] and a sample that might be generated from a mixture of two distributions [Panel (f)], see Appendix 1.

The point circled in red in Panel (f) represents Apple Computer, Inc. As seen in Panel (f), Apple is far distantly smaller than the other 49 data points, those being tightly clustered resembling a bell-shaped distribution. In 2006, Steve Jobs as CEO of Apple Computers, Inc. received one dollar of salary: Is $1 an outlier when compared with the other CEOs amongst the top-paid 50? The dot circled in red in Panel (g) is also Apple. Steve Jobs received the highest amount of 'additional compensation' ($646.6 million) whereas the #2 ranked CEO received about half that amount: Is $646 million an outlier when compared with the other CEOs included in the top-paid 50? If a researcher was to arbitrarily conclude that Apple is an outlier and then exclude Apple, it would limit inferences that can be made. We think that it would be preferable to be able to make inferences about Forbes top-paid CEOs by including Apple rather than excluding it.

2.1 Models for the variables

Are the circled points in Figure 1 outlier or rare outcomes? The answer to this question depends on whether a single probability distribution could describe variations of the entire data points in a panel or not. For this purpose, we test whether a probability distribution does or does not provide an acceptable fit to the data in each panel. The standard assumption for statistical analysis is that a sample of measurements $x_1, x_2,\ldots, x_n$ are observations on random variables $X_1,\ldots, X_n$ generated independently from a probability distribution function F_X for all $i = 1,\ldots, n$. Strictly speaking, our sample is not such a random sample, but it dramatically illustrates some important features of typical data on top-paid CEOs of large US companies. We proceed under the above standard assumption for the purpose of illustration. Various tests of distributional fit are available. We use the Anderson-Darling (A-D) test, which nowadays is one of the most common goodness-of-fit tests; more details are given in Section 5.

Table 2 shows the model-fitting results for the organisational size and CEO compensation variables. The upper panel of Table 2 gives the results for the *log-normal* model. The distribution of a variable X is said to be log-normal if the distribution of its log transformation $Y = \log X$ is *normal*. In other words, a researcher is fitting a normal model to the log-transformed data. The log-normal model has two parameters, referred to as the location and scale parameters which are respectively the mean and SD of the log-transformed data. The lower panel of Table 2 shows the results for log-logistic model. Analogous to the log-normal model, the distribution of a variable X is said to be *log-logistic* if the distribution of its log transformation $Y = \log X$ is *logistic*. The log-logistic model is also a type of Pareto distribution called *Type III Pareto* where the Pareto exponent parameter equals inverse of the log-logistic scale. (The models are discussed in Appendix 1). The columns of Table 2 give the following information. The 'Model fit test' column gives the A-D statistic for each model, the A-D being a measure of discrepancy between the empirical distribution of the data and the distribution of the model (see Figure 6 in Section 5). This is a lack-of-fit statistic, so a large value of the A-D statistic rejects the fit of the model. The critical value of the A-D test for rejecting a log-normal is different than that for a log-logistic model. In the 'Model fit acceptable' column, 'Yes' indicates that the lack-of-fit (discrepancy between the empirical and model's distributions) is not statistically significant at 10% level. That is, the test does not reject the model, so the model is acceptable at 10% level. A 'No' in this column indicates that the lack-of-fit (discrepancy between the empirical and model's distributions) is statistically significant at 1% level. That is, the test rejects the model, so the model is not

acceptable at 1% level. A 'Barely' indicates that the P-value of the test is barely above 1% so the model is acceptable at 1% level. An 'Okay' indicates that the P-value of the test is slightly below 5% so the model is acceptable at 4% level. The last two columns give estimates of the model parameters.

Table 2 Model tests for organisational size and CEO compensation variables

	Model fit	*Model fit*	*Model parameters*	
	Test (A–D)	*Acceptable (Level)*	*Location*	*Scale (Pareto exponent)*
Log-normal				
Sales	0.298	Yes (10%)	2.71	1.19
Assets	0.564	Yes (10%)	3.49	1.78
Salary and bonus	6.890	No (1%)	1.40	2.40
Data + $0.82 million	0.349	Yes (10%)	1.85	0.78
Additional compensation	1.380	No (1%)	3.82	0.84
Data – $6.57 million	1.010	Barely (1%)	3.59	1.02
Log-logistic (Pareto)				
Salary and bonus	1.299	No (1%)	1.66	0.71 (1.41)
Data + $0.43 million	0.257	Yes (10%)	1.75	0.48 (2.08)
Additional compensation	0.975	No (1%)	3.76	0.45 (2.22)
Data – $7.73 million	0.685	Okay (4%)	3.50	0.57 (1.75)

As seen in Table 2, a log-normal model is acceptable for each of the size variables. That is, the log-transformed data for both sales and assets do have acceptable fits with the Gaussian model. In other words, these transformed data on organisational size [Panels (b) and (d) of Figure 1] do not contain rare or outlier observations. However, this cannot be said about the original (untransformed) size data that include observations markedly distant from the mass of the data [Panels (a) and (c) of Figure 1]. But we find that the variation of the entire data points in each of Panels (a) and (c) of Figure 1 can be described by a single log-normal model, so the circled points in these panels are not outliers. A log-normal distribution allows such occurrence with some low but non-negligible probabilities since its tail decays slowly (a heavy-tail distribution). Such observations may be classified as rare outcomes. Consequences of the Gaussian model fitting the log-transformed data will be discussed in sequel (see Section 2.2 and Point 1 in Section 2.3).

As seen in Table 2, the log-normal model does not fit the data on the compensation variables for all 50 organisations. Next we consider the log-normal model which includes a third parameter. This parameter shifts the start of the probability distribution from zero to a positive or negative value. For instance, as seen in Table 2, after a shift from zero to +$0.82 million a log-normal model fits the entire data on the Salary and Bonus with Apple included. This means that a log-normal distribution which starts at $0.82 million fits the Salary and Bonus. That is, the variation of the entire data points in Panel (e) of Figure 1 can be described by a single log-normal model, so the circled points in these panels are not outliers for this distribution. Such observations may be classified as rare outcomes. However, this transformation is consequential for interpretation of the results

of analysis based on such a model. For the Additional Compensation, a log-normal model does not fit the data even with inclusion of a negative shift from zero to –6.57 million, a log-normal model only barely fits the entire data (lack-of-fit is not significant at 1%). Thus, we proceed on with examining other models.

As seen in Table 2, for the salary and bonus data [Figure 1(e)], the fit of log-logistic is also rejected. Note that the Pareto exponent is 1.41 = 1/0.71. As in the log-normal case, including a shift parameter allows the start of the log-logistic distribution to take a positive or negative value rather than being set at zero. Table 2 shows the results for a shift from zero to +$0.43 million. With this shift, the fit of a log-logistic model for all 50 Salary and Bonus data points becomes excellent. For the additional compensation data, the fit of log-logistic model with two parameters is also rejected. With a shift from zero to –$7.73 million, the fit of a log-logistic model for all 50 additional compensation data points becomes acceptable (lack-of-fit is not significant at 4%). We examined numerous other models and found that some more complex models (Burr distributions with four and three parameters, Dagum distributions with four and three parameters, and Generalised Extreme Value distribution) fit the data better than the log-logistic with three parameters. Thus both components of Apple's CEO compensation package may be viewed as plausible outcomes of the respective log-logistic distributions that fit all 50 data points. That is, the variation of the entire data points in each of Panels (e) and (g) of Figure 1 can be described by a single log-logistic (Pareto) model. As such, the circled points in these panels are not outliers for the shifted log-logistic (Pareto) model. However, in order to include Apple in an analysis based on such models, we need to first add $43 thousand to each of the Salary and Bonus of each CEO and subtract $7.73 million from the additional compensation of each CEO.

Table 3 shows the model-fitting results for the compensation variables when Apple is excluded [point circled in red in Panels (f) and (g) of Figure 1]. Then both the log-normal model and log-logistic model fit the salary and bonus component equally good. Hence, the Salary and bonus component of the Apple's CEO compensation is an outlier for these models. However, these models barely fit (lack-of-fit is not significant at 1%) the additional compensation component of the Apple's CEO. Table 3 also shows that the log-normal model with a negative shift from zero to -$5.47 million is barely acceptable and the log-logistic model with a negative shift from zero to -$7.05 million provides an okay fit.

Table 3 Model tests for CEO compensation variables (Apple excluded)

	Model fit	*Model fit*	*Model parameters*	
	Test (A-D)	*Acceptable (Level)*	*Location*	*Scale (Pareto exponent)*
Log-normal				
Salary and bonus	0.266	Yes (10%)	1.71	0.86
Additional compensation	1.080	No (1%)	3.77	0.76
Data– $5.47 million	0.848	Barely (2.5%)	3.59	1.02
Log-logistic (Pareto)				
Salary and bonus	0.258	Yes (10%)	1.69	0.49 (2.04)
Additional compensation	0.820	Barely (1%)	3.70	0.41 (2.44)
Data – $7.05 million ($n$ = 49)	0.604	Okay (7.5%)	3.49	0.51 (1.96)

2.2 Beyond Gaussian measures and methods

The traditional measures (the mean, SD, and correlation coefficient) and methods such as the ordinarily least squares (OLS) regression are highly sensitive to data points that are far distant from the mass of data. The traditional summary measures generally reported by management researchers do not provide useful information much beyond the Gaussian model and are not applicable for some heavy-tail distributions, such as Pareto with exponent parameter 2. Thus, inferences based on the OLS are not applicable to such data distributions. Descriptions of methods that are suitable for analysis of such data are beyond the scope of this paper. Here we provide a brief presentation of some measures suitable for such data.

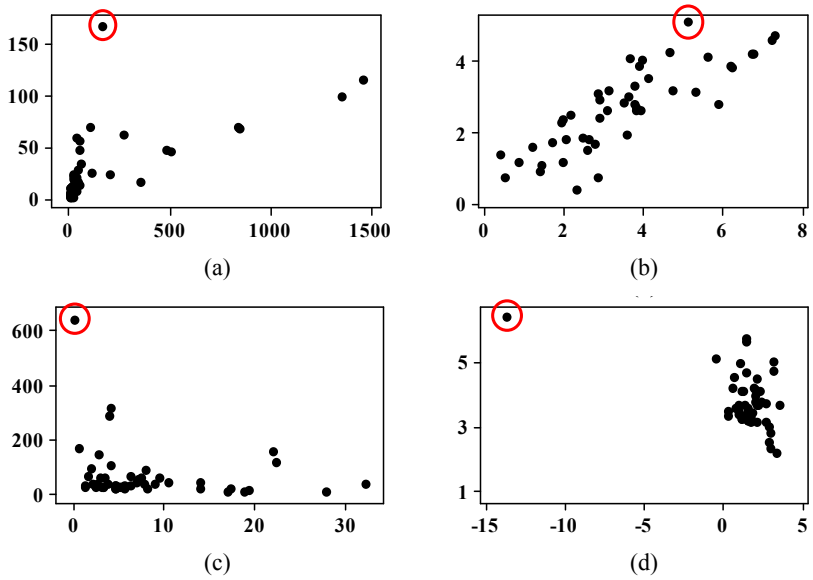
Measures that are based on rank and percentile such as median, quartiles, range, and interquartile range (the range between the first and third quartiles) are applicable to all distributions. A useful property of these latter measures is their properties under monotone transformations of the data such as log-transformation. For example, when data on a variable X is log-transformed to $Y = \text{Log } X$, the median of $Y = \text{Log}$ (median of X). However, this does not hold for the mean and SD.

Like the mean and SD, the correlation coefficient does not provide useful information much beyond the Gaussian model and is not applicable for some heavy-tail distributions. Recall that the correlation coefficient measures how tightly points on a scatter plot are clustered along a line. Also recall that when data are generated by a bivariate normal distribution, the points on the scatter plot resemble an ellipse. Now, consider the scatter plots of data on the organisational size variables shown in Panel (a) of Figure 2. The points are not clustered along a line nor do they resemble an ellipse; furthermore, an observation (circled in red) is far from the mass of points. Consequently, the correlation coefficient for the data is $\text{Corr}(X, Y) = 0.66$, but this does not provide information about a linear relationship as often interpreted. Panel (b) of Figure 2 shows the scatter plot for the log-transformations of these size variables. The points are clustered along a line and the seemingly outlying observation (circled in red) of the original data is now close to the mass of points. The correlation coefficient for the transformed data is $\text{Corr}(\text{Log } X, \text{Log } Y) = 0.83$, which now is measuring the degree of tightness of the transformed data points along a line. Here, we see that the correlation coefficient changes when the variables are transformed non-linearly such as log-transformation. However, since correlation coefficient is not invariant under all monotone increasing transformations, the strength of linear relationship ($\text{Corr}(\text{Log } X, \text{Log } Y) = 0.83$) in Panel (b) does not translate into the same strength of linear relationship between the sales and assets ($\text{Corr}(X, Y) = 0.66$). In general, one correlation could be statistically significant while the other is not. Measures of the strength of association between two variables are invariant under all monotone increasing transformations. Examples include Kendall's tau and Spearman's rho. For example, Spearman's rho is a correlation coefficient between the ranks of two variables. Since log-transformation is a monotone increasing transformation, Spearman's rho for plots in Panels (a) and (b) of Figure 2 is the same (Spearman's rho(X, Y) = Spearman's rho(Log X, Log Y) = 0.88). Thus we can conclude that there is a positive association between organisational sales and assets.

Panels (c) and (d) of Figure 2 show scatter plots of the two components of CEO compensation data and their natural logarithm with an observation (circled in red) far from the mass of points. As it was shown in Table 2, the model that fit the additional compensation variable is a heavy tail (Pareto with exponent parameter 1.75), and thus the

correlation coefficient for the bivariate model for these variables is not defined. That is, the correlation coefficient computed from the data is meaningless. However, association measures such as Kendall's tau and Spearman's rho are defined for all distributions. For plots in Panels (c) and (d) of Figure 2, (Spearman's rho(X, Y) = Spearman's rho(Log X, Log Y) = –0.25), which indicates a negative association between the two components of CEO compensation.

Figure 2 Scatter plots of organisational size and CEO compensation variables, (a) sales vs assets, (b) log (sales) vs log (assets), (c) additional comp vs salary and bonus, (d) log (additional comp) vs log (salary and bonus) (see online version for colours)



2.3 Implications

From our model-fitting analysis, we can conclude as follows:

1 For the organisational size variables (Sales and Assets), log-normal distributions with two parameters are suitable. Hence, the Gaussian methods are not directly applicable to these variables. The log-transformation provides an option for the use of Gaussian methods. A researcher could proceed to use conventional Gaussian statistical methods to analyse these log-transformed data further. However, the results of such methods must be interpreted accordingly, as discussed in Section 2.2.

2 For salary and bonus, the log-normal distribution with two parameters is not suitable. Hence, the Gaussian methods are not directly or indirectly applicable to this variable. Apple is an outlier for the log-normal distribution. With removing Apple, Gaussian methods become applicable to the log-transformed data. However, a) the results of such methods must be interpreted in terms of the transformed variables, and b) the inferences should be limited to a subset of top-paid CEOs, not to the entire 50 top-paid CEOs. With a further transformation (first adding $426 thousand to the salary and bonus of each CEO and then taking log), Gaussian methods become

applicable to the transformed data, although the results of such methods must be interpreted accordingly.

3 For the additional compensation, log transformation even with Apple removed cannot make Gaussian methods applicable. Log transformation of the additional compensation minus \$7.73 million barely makes Gaussian methods applicable to analysis of this variable. However, the results of such methods must be interpreted accordingly and also very cautiously.

4 Overall, the results in Tables 2 and 3 indicate that the data generating distribution with the Apple data included is too complex even to the extent that the fit of a Pareto model requires data distortions (adding and subtracting the components of the CEO compensation). Without such distortions or moving to a more complex model, we can conclude that Apple might be generated from a different distribution other than the mass of the data. That is, the data generating distribution is a mixture of two distributions, which makes Apple an outlier with respect to the distribution of the other 49 top-paid CEOs. Even so, researchers who seek to make inferences based on the entire data set ought to avoid the common, simplifying action of trimming out data that initially look like outliers. Implementing some inferences based on a model that fits the entire data may require more complicated and advanced inferential methods. Preferable options are non-parametric and robust methods of inference (Schwab et al., 2011). Examples include bootstrap standard error and interval estimates, robust regression methods such as the least absolute deviation, non-parametric regression and ANOVA, and generalised least squares fit functions, which are available in statistical packages.

3 Executives' forecast accuracy after 9/11 jolt

Organisations' environments occasionally deliver a jolt that requires executives to make sense out of the new situation before they decide what actions to take in response (Meyer, 1982; Meyer et al., 1990; Weick, 1995). A company's strategy could shape environmental scanning and sense making (Garg et al., 2003; Yasai-Ardekani and Nystrom, 1996). Moreover, the type of strategic milieu inhabited by executives could affect their attentiveness to environmental changes as well as their ability to respond to changes in appropriate ways (Kiesler and Sproull, 1982). Attentiveness to environmental changes could lead to greater accuracy in forecasts, especially during periods of greater uncertainty. We focus here on whether organisations' strategies (proactive versus reactive) are associated with the ability of their executives to forecast their future economic environments.

We use data on the accuracy of economic forecasts made by a sample of executives shortly after the September 11, 2001 terrorist attacks on the World Trade Centre in New York City and on the Pentagon near Washington, D.C. The event and its aftermath constituted a widespread jolt to the national economy. Recall that this extraordinary event occurred after the US economy had ended a long period of high growth and had already lapsed into recession. Many executives and governmental leaders were expressing deep concerns about the uncertainty generated by this rare event. Executives were unsure about how these unprecedented attacks would affect the national economy. Executives

were also unsure about how the suddenly changed economic outlook would affect their firms. A disturbing scenario held that a pervasive increase in fear amongst the population could reduce consumer confidence and could alter household spending patterns – which could adversely impact future economic growth.

Business executives could react to a predicted economic downturn by anticipating a reduction in demand for their company's products or services, and might react by reducing their expenditures. In the aftermath of the 9/11 jolt, business executives were confronted with the task of quickly predicting potential changes in the economy and then assessing how those changes would affect their own organisations. Yasai-Ardekani and Nystrom surveyed executives regarding their companies' reactions five months after the 9/11 jolt. Data were collected from executives in 93 companies throughout the US; see Soofi et al. (2009) for details about the organisations included in the study. That study of the 9/11 aftermath included eliciting business executives' forecasts for future economic conditions. A survey questionnaire provided executives with a paragraph describing the US economy's GDP (Gross Domestic Product) and its rates of change in recent years and quarters, so that all respondents began with the same basic information. Each executive then forecasted the future economy one year after the attacks. Each executive also provided his or her likelihood estimates (expressed as percentages that totalled to 100%) of three possible future states of economy: growth, stagnation, decline. Their responses were the input to computation of a measure of forecast inaccuracy, which we based on the commonly used mean square error (MSE) of each executive's forecast distribution; for details on measurement and examples characterising four executives, see Appendix 2.

3.1 Extreme accurate and inaccurate forecasts

Similarly to rare and outlier, the concept of extreme value involves comparison of magnitudes of data points. An *extreme value* refers to an outcome farthest away from the centre (median) of a set of measurements and need not to be far distant from the mass of data points. The set of sample outcomes, $x_1, x_2,\ldots, x_n$, arranged into an array of ascending order, $y_1 \le y_2 \le \ldots \le y_n$ are referred to as the *order statistics* of the sample. The sample minimum y_1 and maximum y_n are referred to as extreme values.

As noted in Section 2.2, a sample of measurements $x_1, x_2,\ldots, x_n$ are observations on random variables $X_1,\ldots, X_n$ generated independently from a probability distribution function F_X for all $i = 1,\ldots, n$. The distribution is not restricted to be any specific type (Gaussian, heavy tail, or mixture, see Appendix 1). Each sample of n observations drawn from a distribution F_X gives a set of different order statistics. This sampling process leads to probability distributions for order statistics: $y_1 \le y_2 \le \ldots \le y_n$ are observations from a set of random variables $Y_1 \le Y_2 \le \ldots \le Y_n$. Their distributions $G_j(y_j)$, $j = 1,\ldots, n$, are derived from the data-generating distribution F_X, which is sometimes referred to as the *parent distribution* for $G_j(y_j)$. Note that the subscript j is the rank of Y_j in increasing order. Unlike random variables $X_1,\ldots, X_n$, the order statistics $Y_1,\ldots, Y_n$ are neither independent nor identically distributed because of the inequalities among them. The distribution $G_j(y_j)$ is a function of its parent distribution F_X, the sample size n, and the rank of Y_j (the subscript $_j$). Hence, unlike the rare and outlier cases, the distribution of extreme value does not solely depend on F_X.

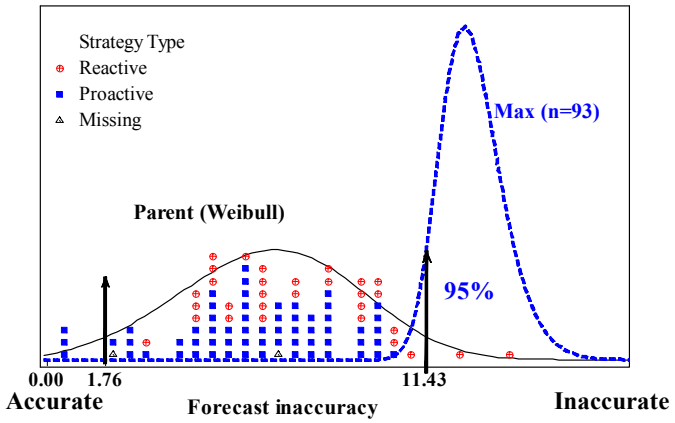
Probability models for extreme values are distributions of Y_1 and Y_n, which can be found using direct probability arguments [see, e.g., DeGroot and Schervish, (2002), p.166]. The extreme value theory refers to approximations of the distributions of the

sample extremes when the sample size n is large (theoretically when n approaches infinity); see, for example, Beirlant et al. (2004) and Coles (2001) for theoretical developments and Baum and McKelvey (2006) for management applications. In a recent paper (Nystrom et al., 2010), we propose that a measurement x_i be inferred as an extreme when it falls in a high probability interval under the distribution of one of the sample extrema, minimum or maximum; all data points falling within a high probability interval for an extreme value will be inferred indistinguishable with the respective extreme. When the parent distribution F_X can be approximated by a known parametric distribution, computation of thresholds for identifying extremes is rather simple.

In order to identify the forecast inaccuracies that can be inferred as extremes, we first find models that are plausible for the distribution of the data. That is, we find plausible models for the parent distribution F_X, where X denotes the forecast inaccuracy variable (MSE). We find that a Weibull distribution with three parameters (Location = –2.45, Scale = 10.14, Shape = 3.93, see Appendix 1) provides an excellent fit to this data (A–D test = 0.493). A Weibull distribution is commonly used in extreme-value theory, reliability engineering and survival analysis.

Figure 3 shows the dot plot of the 93 executives' forecast inaccuracy (MSEs) superimposed by two density curves. The solid curve is for the Weibull distribution. The dashed curve in Figure 3 is the density curve for the distribution of sample maximum for n = 93 (the distribution of minimum is not shown because the height of its density distorts the graph). Vertical arrows flag the 95% thresholds for the extremes for the minimum (1.76) and for the maximum (11.43). These thresholds are computed using the MINITAB software codes given in Nystrom et al. (2010, pp.801–802).

Figure 3 Dot plot of forecast MSE data superimposed by density functions of a three-parameter Weibull distribution and distribution of its maximum for n = 93 (see online version for colours)



As seen in Figure 3, three observations can be inferred as extreme low inaccuracy (that is, highly accurate forecasts) and two observations can be inferred as extreme high inaccuracy in forecasting.

We also find that three other models fit the data satisfactorily: a normal distribution, a three-parameter log-normal distribution and a three-parameter gamma distribution. All four models that fit the data lead to the same results: three observations below the 95%

thresholds for the minimum and two observations above the 95% thresholds for the maximum. Furthermore, in the terminology of extreme-value theory, the Weibull parent distribution is in the domain of attraction of Gumbel extreme-value distribution; see, for example, Beirlant et al. (2004). The Gumbel model also confirms the same results. Hereafter, the three data points in the lower side will be referred to as extreme accurate (minimum MSE) forecasts and the two data points in the upper side will be referred to as extreme inaccurate (maximum MSE) forecasts.

3.2 Forecast accuracy and company's strategy

Information about each company's strategy is also indicated in Figure 3. Squares (blue) symbols represent those executives in companies with proactive strategies; circles (red) symbols represent those executives in companies with reactive strategies; and triangles represent the two executives with missing data on strategy. A *proactive* strategy (prospector or analyser types as described by Miles and Snow, 1978) focuses on monitoring conditions in the company's environments, adapting quickly when market conditions shift, and introducing new products before its competitors. In contrast, a *reactive* strategy (defender or reactor types) emphasises a company's present market niches, seeks to protect its current domain and often ignores or discounts the relevance of changes taking place in its industry. The data were collected as follows. Each strategy type is described by one of four paragraphs developed by Miles and Snow (1978). An executive check marks the one paragraph that best describes that executive's organisation; none of the paragraphs show the type labels such as defender. As seen in Figure 3, all three extreme accurate forecasts are made by executives in companies with proactive strategies (blue squares) and both of the extreme inaccurate forecasts are made by executives in companies with reactive strategies (red circles).

We use a Bayesian approach to infer more formally about the relationship between a company's strategy and the probabilities that an executive's forecast is extreme accurate or is extreme inaccurate.

Table 4 Data for company strategy and economic forecasting accuracy

Strategy type	*Economic forecasting accuracy*			
	Extreme accurate	*Not extreme*	*Extreme inaccurate*	*Total*
Proactive	3	58	0	61
Reactive	0	28	2	30

Briefly, a Bayesian approach is used to update a prior distribution into a posterior distribution of outcome probabilities. A prior distribution reflects beliefs about the distribution of a parameter before using the current data. The belief can be formed based on all available information, including the past empirical findings. Readers interested in a more detailed explanation of a Bayesian approach are referred to Zellner (1971) and Lee (1997). For examples of management analyses using a Bayesian approach, see Hansen et al. (2004), Hahn and Doh (2006), Soofi et al. (2009), and Nystrom et al. (2010).

For implementation of the Bayesian, we cross-classify the executives according to their strategy type and the category of their forecast accuracy (extreme accurate, not extreme, extreme inaccurate). Table 4 shows the data. The sample proportion of *extreme accurate* forecasts made by executives of companies having a *proactive strategy* is

(3/61 = 0.049) and the proportion of *extreme inaccurate* forecasts made by these executives is (0/61 = 0). The corresponding sample proportions for executives of companies having a *reactive strategy* are (0/30 = 0) and (2/30 = 0.067). These sample proportions are the maximum likelihood estimate (MLE) of the following four *conditional probabilities*:

- π_1 = The probability of an *extreme accurate* forecast being made by an executive, given that this executive's company has a *proactive strategy*, (MLE of $\pi_1 = 0.049$).
- π_2 = The probability of an *extreme accurate* forecast being made by an executive, given that this executive's company has a *reactive strategy*, (MLE of $\pi_2 = 0$).
- π_3 = The probability of an *extreme inaccurate* forecast being made by an executive, given that this executive's company has a *proactive strategy*, (MLE of $\pi_3 = 0$).
- π_4 = The probability of an *extreme inaccurate* forecast being made by an executive, given that this executive's company has a *reactive strategy,* (MLE of $\pi_4 = 0.067$).

We are interested in the following comparisons: π_1 versus π_2 and π_3 versus π_4. We do these comparisons by inferring about the following differences:

$$DAccurate = \pi_1 - \pi_2 \text{ and } DInaccurate = \pi_3 - \pi_4.$$

A positive *DAccurate* implies that the probability of making an extreme accurate forecast for an executive of a company with proactive strategy (π_1) is higher than the probability of making an extreme accurate forecast for an executive of a company with a reactive strategy (π_2). A negative *DInaccurate* implies that the probability of making an extreme inaccurate forecast for an executive of a company with proactive strategy (π_3) is lower than the probability of making an extreme inaccurate forecast by an executive of a company with a reactive strategy (π_4). We compute the posterior odds in favour of *DAccurate* being positive against *DAccurate* being negative.

In order to make inference about the above assertions, we follow the same Bayesian procedure that we used in Nystrom et al. (2010) for inferring about the probability of an extreme CEO compensation and the probability of an extreme performance of the firm. Here, we compute posterior odds in favour of $\pi_1 > \pi_2$ (for an executive in a company with a proactive strategy, the probability of extreme accurate forecast is higher than for an executive in a company with a reactive strategy) against $\pi_1 < \pi_2$ (for an executive in company with a proactive strategy, the probability of extreme accurate forecast is lower than for an executive in a company with a reactive strategy). For the inference about inaccuracy of prediction, we do similarly.

The standard prior for Bayesian inference about a probability is a Beta distribution. Beta family of distributions have two parameters, say a and b, which provide various shapes (symmetric ($a = b$), left-skewed ($a > b$), right-skewed ($a < b$), U-shape ($a = b < 1$), the uniform ($a = b = 1$), and more). This flexibility allows choosing priors that reflect various beliefs about a probability of interest. For each probability π_1, π_2, π_3, and π_4 we use two different prior distributions:

a a right-skewed Beta distribution (mode at zero), which reflects a belief that the probability of an extreme is relatively lower than the probability of a non-extreme outcome

b the uniform distribution, which reflects a belief that connotes ignorance about the probabilities of extreme accurate and extreme inaccurate forecasts.

The beta family is also mathematically convenient in that the data only update the parameters of a Beta prior distribution by the sample size and the counts to produce a Beta posterior distribution.

For our purpose of calculating $P(\pi_1 > \pi_2)$ and $P(\pi_3 < \pi_4)$ with a Beta distribution for each parameter, a formula is available (Lee, 1997) which was used by Nystrom et al. (2010). But the formula involves gamma functions and is tedious to evaluate. Instead, here we use a Monte Carlo simulations technique that gives very accurate results. Table 5 shows the posterior probabilities based on the two different priors examined: a right-skewed beta distribution and a uniform distribution. All posterior probabilities are high. The ratio of $P(\pi_1 > \pi_2)$ to $P(\pi_1 < \pi_2)$ gives the *posterior odds* in favour of the hypothesis that $\pi_1 > \pi_2$. Since we are using the same prior for π_1 and π_2, the prior probability is $P(\pi_1 > \pi_2) = P(\pi_1 < \pi_2) = .5$; in other words, we take an agnostic stance about the assertions and let the data decide. In this case, the prior odds = one and the posterior odds = Bayes factor. Interpretations of the grade of evidence for the Bayes factor are given in Kass and Raftery (1995). We can infer as follows:

a The data provide 'substantial' evidence in favour of a higher probability of an extreme accurate forecast by an executive in a company with a proactive strategy as compared with an executive in a company with a reactive strategy;

b The data provide 'strong' evidence in favour of a lower probability of an extreme inaccurate forecast by an executive in a company with a proactive strategy as compared with an executive in a company with a reactive strategy.

Table 5 Posterior probabilities and posterior odds for comparison of probabilities of extreme accurate and extreme inaccurate forecasts for different strategies

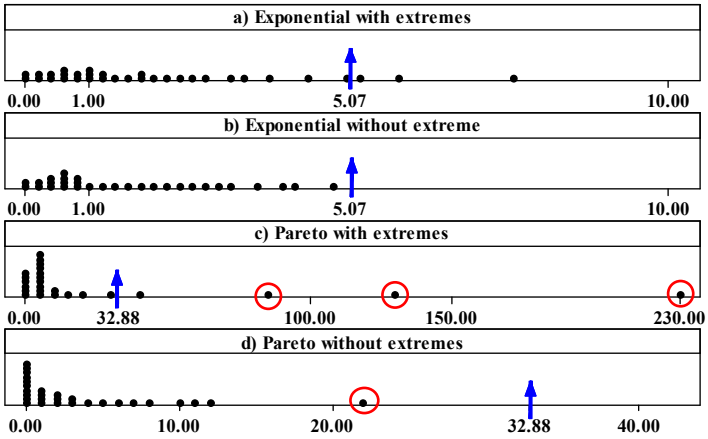
	$\pi_1 > \pi_2$	$\pi_3 < \pi_4$
Posterior probability		
Right-skewed beta prior	.844	.945
Uniform prior	.789	.962
Odds in favour		
Right-skewed beta prior	5.41 (Substantial)	17.15 (Strong)
Uniform prior	3.73 (Substantial)	25.32 (Strong)

3.3 Extreme versus rare outcomes

Our data on forecast inaccuracy did not contain any observation far from the mass of data that could be viewed as a rare or outlier outcome. Yet, from Figure 3 we inferred five data points as extremes (3 extreme low inaccurate and 2 high extreme inaccurate). In other kinds of samples, however, a rare observation could be inferred as being extreme or as not extreme; similarly, an extreme observation could be inferred as being rare or as not rare. We illustrate these distinctions between inferred extreme and rare by using two samples of size $n = 100$ simulated from the exponential distribution (0, 1) and two samples of size $n = 100$ simulated from the Pareto (0, 1, 1); see Panel (b) of Figure 9 and Table 6 in Appendix 1. Figure 4 shows dot plots of these four samples. Arrows mark the

95% thresholds of the upper extremes (UEs) calculated from the distributions of the maximums for these models. Panels (a) and (b) show plots for two samples from the exponential distribution. The exponential sample in Panel (a) includes three data points above thresholds for maximum (5.07), but the exponential sample in Panel (b) includes no observation above 5.07 that can be inferred as extreme high. The probability of obtaining a sample without inferred extremes based on 95% thresholds is 0.05. Thus by the law of large numbers, in the long run about 5% of samples are expected to include no extreme. (We simulated 10,000 samples of n = 100 and found 53 samples with no extreme observation).

Figure 4 Dot plots of samples simulated from four distributions (n = 100; arrow shows 95% threshold for extremes and circle shows rare outcome) (see online version for colours)



Panels (c) and (d) in Figure 4 show plots for two samples from the Pareto distribution. The horizontal scales for these two panels are different. The Pareto sample in Panel (c) includes four data points above the threshold for the maximum (32.88), but the Pareto sample in Panel (d) includes no observation above 32.88 that can be inferred as extreme high. Yet both Pareto samples include data points (circled in red) that are distinctly apart from the others, which can be considered as rare of different degrees.

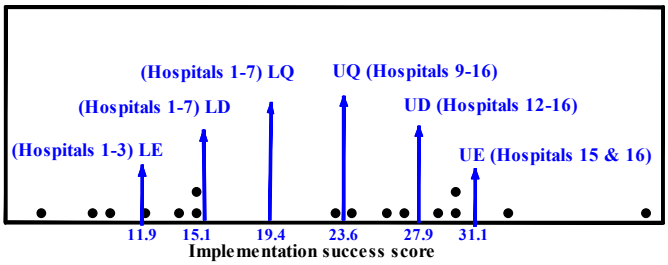
In summary, Figure 4 shows instances where the maximum of a sample is an inferred extreme [Panel (a)], but may not be an inferred extreme due to the 5% error rate [Panel (b)]. When the data distribution is heavy tail, an inferred extreme might also be a rare outcome [Panel (c) highest three points] but an inferred extreme might also not be a rare outcome [also in Panel (c), the fourth point is relatively close to the fifth point]. Moreover, a rare outcome may not be an inferred extreme [Panel (d)]. We conclude that the labels of extreme and rare ought not be used interchangeably.

4 Extension from extremes to percentiles

Researchers interested in high-performing organisations and/or low-performing ones usually select organisations for inclusion by using sample deciles, quartiles, or other sample splits. Examples include Harris and Katz's (1991) choice of quartiles for selecting

the highest performing firms and the lowest ones in a study of information technology in the insurance industry; Edmondson et al. (2001) choice of the seven highest scoring hospitals and seven lowest scoring ones arrayed along their implementation success index; and Siegel and Hambrick's (2005) choice of the ten highest performing firms and the ten lowest ones in a study of pay disparity within top-management teams.

Figure 5 Dot plot of hospitals' implementation success index data and 95% thresholds for upper and lower extremes, deciles and quartiles (see online version for colours)



Thresholds for identifying outcomes in upper and lower percentiles, deciles, quartiles, and so forth can be calculated by using the same method discussed in Section 3.1 for identifying extreme outcomes. The distribution of the corresponding order statistic is substituted in place of the sample maximum or minimum when calculating the threshold for a specific percentile.

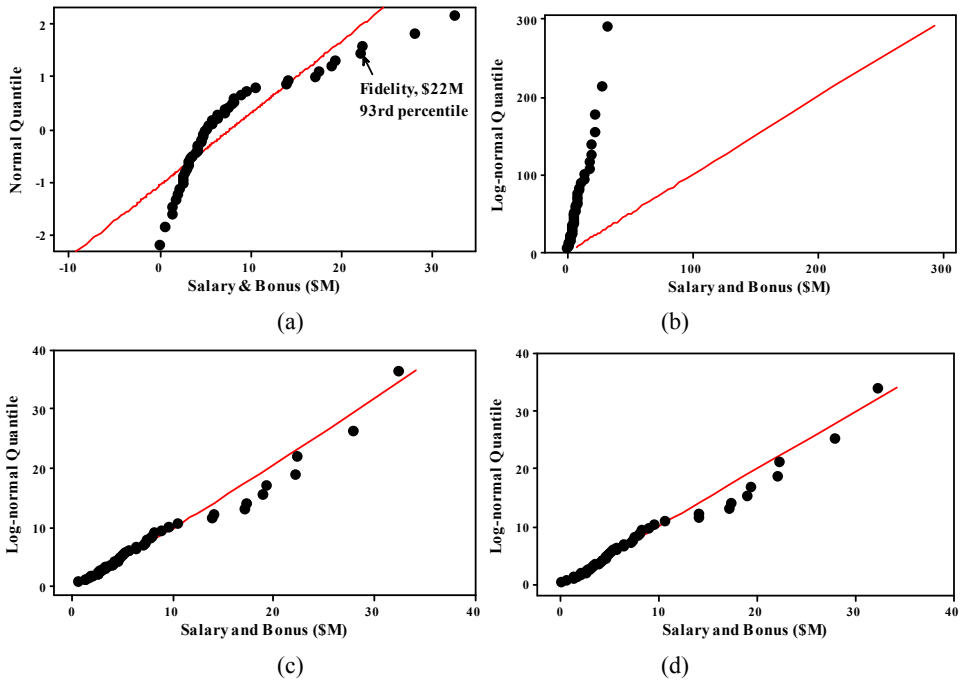
In their study of 16 hospitals implementing new technologies, Edmondson et al. (2001) computed an *implementation success index*. This index ranges from 6 to 41 and the two middle scores are 23 and 24 (see Figure 5). They classified the upper seven hospitals as high implementation success and the lower seven hospitals as low implementation success. They excluded the two middle cases "both to reflect the location of step changes in the implementation success index and to avoid an arbitrary distinction between two adjacent sites in the middle" [Edmondson et al., (2001), p.693]. We show that researchers can use the thresholds based on high probability intervals under the distributions of the appropriate order statistics in order to provide cut-off points that assist those researchers who want to avoid making arbitrary choices during sample selection.

Suppose a researcher seeks to select those organisations included in a sample that are plausibly in the upper 10% of the underlying population from which this sample was drawn as well as those in the lower 10% of that population. We begin by finding a model that fits the sample data. In this case, the normal distribution fits the data (mean = 21.5, SD = 10.1, A–D test AD = 0.377). Figure 5 shows the dot plot of the data and the 95% thresholds for upper extreme (UE) and lower extreme (LE). The figure also shows results from extending the analyses by using order statistics that identify the upper decile (UD) and lower decile (LD) as well as the upper quartile (UQ) and lower quartile (LQ). These thresholds enable a researcher to classify the hospitals under study according to the upper and lower relative positions that can be inferred for each. As can be seen in Figure 5, the threshold for the UD is a score of 27.9, which would lead to inclusion of five hospitals; the threshold for the LD is a score of 15.1, which would lead to inclusion of seven hospitals.

5 Quantile-quantile plot and distributional tests

For checking the compatibility of a model with the data, statisticians often use a highly effective visualisation tool such as Q-Q (quantile-quantile) plot, which is available in statistical packages. (Quantile is synonymous with percentile). Figure 6 shows the Q-Q plots of salary and bonus data with the normal and three log-normal models of Table 2 for this variable. On these Q-Q plots we have included the line $y = x$ for use as the reference. When the model is correct, the points are clustered close to the line $y = x$. The coordinates of the points are data values (horizontal axis) and their expected values if the model were correct (vertical axis), i.e., the corresponding quantiles of the model. For example, the salary and bonus of the CEO of Fidelity National Financial is \$22 million; its rank in increasing order is 46, which makes \$22 million the 93rd quantile of the sample, approximately. (We computed this quantile using a simple formula $p_{46} = (47–0.5)/50 = 0.93$; several other methods are available). In the normal Q-Q plot [Panel (a) of Figure 6], the model 93rd quantile is $z = 1.47$ given in the normal table. The vertical scale may also be set as the probabilities instead of the quantiles. In that case, the plot is referred to as a probability plot. The Q-Q plots in Panels (a) and (b) of Figure 6 show that there are substantial discrepancies between the points and the line for the normal and log-normal models. In Panel (c), Apple is excluded, so the points are clustered along the reference line. In Panel (d), the model is a log-normal that includes a shift parameter (like adding 0.82 \$million to each data point), and the points are clustered along the reference line similar to Panel (c).

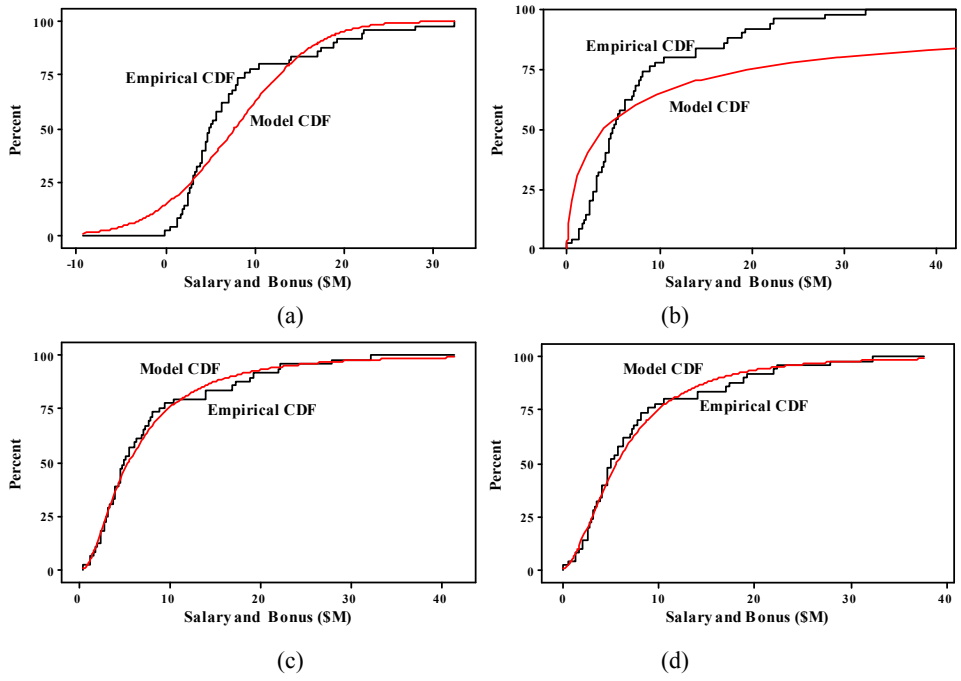
Figure 6 Q-Q plots of salary and bonus variable, (a) all data points, normal (7.74, 7.23), (b) all data points, log normal (1.40, 2.37), (c) Apple excluded, log normal (1.71, 0.86), (d) all data points, (data+.82) log normal (1.85, 0.77) (see online version for colours)



Some packages include the parameter estimates, a fit statistic, and the P-value for the Q-Q plot. For example, MINITAB gives A-D, Kolmogrov-Smirnov (K-S), and Wilks-Shapiro (W-S) tests for the normal model and the A-D test for several other distributions. Our model-fitting results of the A-D tests are confirmed by the K-S test.

The A-D, K-S, and W-S statistics are in the class of tests referred to as empirical distribution function (EDF) tests (Stephens, 1974, 1979). Recall that the empirical distribution function is also a CDF in the step-function form where each step begins at a data point and the height of each step is $1/n$. Figure 7 exhibits the empirical CDF and the model CDF curves corresponding to the panels of Figure 6. Visually it is clear that there are substantial discrepancies between the empirical CDF and the normal CDF [Panel (a)] and between the empirical CDF and log-normal CDF in Panel (b). In Panel (c), Apple is excluded and the discrepancy between the empirical CDF and log-normal CDF is substantially reduced. The model in Panel (d) is a log-normal that includes a shift parameter (like adding 0.82 $million to each data point). This shift also substantially reduces the discrepancy between the empirical CDF and log-normal CDF.

Figure 7 Empirical CDF (step-function) of salary and bonus variable and CDFs of four models (curves), (a) all data points, normal (7.74, 7.23), (b) all data points, log normal (1.40, 2.37), (c) Apple excluded, log normal (1.71, 0.86), (d) all data points, (data+.82) log normal (1.85, 0.77) (see online version for colours)



The A-D and K-S statistics are two measures of discrepancy between the model (CDF) and the empirical CDF in terms of the vertical distances at the data points. The K-S statistic measures the discrepancy in terms of the maximum of these vertical distances. The A-D statistic measures the discrepancy in terms of a weighted average of these vertical distances. The weight for each data point is inversely related to the height of the model CDF at that point, hence A-D gives more weight to the points toward the tail.

The A-D, K-S, and W-S tests use individual data points and the corresponding expected points if the model were true. These tests are developed as alternatives to the classic chi-square test, which requires grouping the data into categories. Unlike the chi-square goodness-of-fit test, these three tests require neither

a grouping data into categories nor

b a large number of observations (the chi-square test requires more than five data points in each bin).

In addition, Stephens (1974) shows that the powers of chi-square test for testing the normal and exponential models against several alternatives are the lowest as compared with the EDF tests. For these reasons, the classic chi-square test is seldom used in statistics literature since the 1980s and statisticians have been developing distributional tests that use individual data points instead of categories. The goodness-of-fit tests are sensitive to large sample size, but the chi-square test is even considerably more sensitive to the sample size. It is well known that a very large sample generally leads to large chi-square statistics, hence statistically significant lack-of-fit and rejection of any model for the data. When a lack-of-fit test gives statistically significant results for various models, the conclusion can be only a comparison: one model is worse than the others [although often stated positively in terms of one model 'better fits' than the others; see, for example, O'Boyle and Aguinis, (2012), pp.94, 101].

6 Concluding remarks

Gaussian statistics has dominated management scholars' thinking and shaped their views on notions of typical, outlier, rare and extreme. Often these terms are used interchangeably. Yet they can be distinguished in terms of probability distributions beyond the traditional Gaussian (normal) distribution. In this paper, we illustrate the distinctions using data on three management research examples, as well as using distributional graphs and simulated data.

In a management example on organisational size and CEO compensation variables for top-paid CEOs of large US companies, we illustrate that the common practice of log transformation can

a resolve some problems of data points which look like outliers as raw data

b uncover an outlier in the transformed data which wasn't apparent in raw data.

This example underscores the importance of moving beyond the Gaussian model and methods for analyses of organisational size and CEO compensation variables. All four variables that we considered (Sales, Assets, Salary and Bonus, and Additional Compensation) have non-Gaussian highly skewed distributions. In particular, we find that a heavy-tail distribution is needed as a model for the additional compensation component of CEO pay in order to capture variation that includes outcomes such as Apple's CEO additional compensation ($646.6 million in 2006) far distant from the mass of data. The Gaussian methods and measures are not applicable for making inferences about the additional compensation of top-paid CEOs. A heavy-tail model such as log-logistic (Pareto) enables such inferences with the retention of the most highly paid CEO rather

than to delete it as being an outlier. More generally stated, when a researcher wants to use Gaussian statistical methods, the researcher should first examine whether a Gaussian model does or does not fit the data. When a Gaussian model does not fit the data, a researcher ought not to arbitrarily assume that some bits of the data are outliers and then use this untested assumption to eliminate the troublesome bits from the sample in order arbitrarily to force the Gaussian model to fit the truncated data. Instead, other viable options exist: a researcher could consider using non-parametric and robust methods of inference or could examine the fits of other models even though going beyond Gaussian methods would likely necessitate using inferential methods which might be more complicated.

In a recent paper (Nystrom et al., 2010), we propose that a data point can plausibly be inferred as an extreme when it falls within a high probability interval under the distribution of one of the sample extrema, whether the minimum or maximum. This method for identifying extremes is described and illustrated by another of our management examples, in which we analyse executives' economic forecast accuracy after the 9/11 jolt. We identify plausible extreme accurate forecasts and extreme inaccurate forecasts, and then examine the relationship between forecast accuracy and company strategy. Using a Bayesian analysis, we find that the evidence (odds) is 'substantial' in favour of a higher probability for an extreme accurate forecast by an executive in a company with a proactive strategy as compared with an executive in a company with a reactive strategy. We also find that the evidence (odds) is 'strong' in favour of a lower probability for an extreme inaccurate forecast by an executive in a company with a proactive strategy as compared with an executive in a company with a reactive strategy.

Our forecast inaccuracy data included inferred extreme points, but did not include any point far distant from the mass of points. Further differentiation between inferred extremes and rare outcomes are illustrated by using samples simulated from two probability models: exponential and Pareto. For each model, plots of two samples are displayed: one with and a second without data points that are inferred as being extreme. Erroneous inference about a sample maximum or minimum not being extreme occurs at a small error rate specified by the researcher. Moreover, we show instances where an inferred extreme is not a rare outcome and other instances where a rare outcome cannot be inferred as an extreme.

This paper also extends our method for identifying extremes by examining percentiles, such as the top 10%. We present an application of the high-probability thresholds for identifying technological implementation outcomes in UD and LD and UQ and LQ for a sample of 16 organisations.

Visualisation of model fit through Q-Q plot and distributional testing are illustrated using the data on CEO salary and bonus. The Anderson-Darling, Kolmogrov-Smirnov and Wilks-Shapiro tests, unlike the chi-square goodness-of-fit test, use individual data points and do not require a large number of observations. The aforementioned four tests of model fit all confirm our results.

Finally, Appendix 1 compares several probability models in contrast with the Gaussian model by using plots of their density curves and their survival functions. Plot of density curve, which gives probability in terms of the area, provides useful visualisation of some features such as the shape and spread of the distribution and intervals of high and low probabilities. Plot of the survival function, where the vertical axis gives the

probability of an outcome occurring beyond a given point, provides useful visualisation for the tail probabilities.

Acknowledgements

We began this line of research in collaboration with Masoud Yasai-Ardekani in response to the Call for the Fourteenth Organisation Science Winter Conference (OSWC-XIV) 2008 'On the Organisation Science of Extreme Events'. The main points of this paper were presented at the plenary panel entitled 'Methods: How Better to Study Extremes'. We thank the conference participants whose questions and comments were most persuasive for us to pursue these topics. We acknowledge Bill McKelvey with thanks for the opportunity that he provided us to present at OSWC-XIV, for his invitation to participate in this special issue, and for his comments and suggestions on the first draft of this paper which led us to improve the exposition. We are indebted to Masoud Yasai-Ardekani for his contributions at various stages of this research. Ehsan Soofi's research was partially supported by a Sheldon B. Lubar School Dean's Research Fellowship.

References

Andriani, P. and McKelvey, B. (2007) 'Beyond Gaussian averages: redirecting international business and management research toward extreme events and power laws', *Journal of International Business Studies*, Vol. 38, No. 7, pp.1212–1230.

Baum, J.A.C. and McKelvey, B. (2006) 'Analysis of extremes in management studies', in Ketchen, D.J. and Bergh, D.D. (Eds.): *Research Methodology in Strategy and Management*, Vol. 3, pp.123–196, Elsevier JAI, Oxford.

Bebchuk, L.A. and Fried, J.M. (2004) *Pay Without Performance: The Unfulfilled Promise of Executive Compensation*, Harvard University Press, Cambridge, MA.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*, Wiley, West Sussex, UK.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer, New York.

DeGroot, M.H. and Schervish, M.J. (2002) *Probability and Statistics*, 3rd ed., Addison-Wesley, Reading, MA.

Dittmann, I. and Maug, E. (2007) 'Lower salaries and no options? On the optimal structure of executive pay', *Journal of Finance*, Vol. 62, No. 1, pp.303–343.

Dow, J. and Raposo, C.C. (2005) 'CEO compensation, change, and corporate strategy', *Journal of Finance*, Vol. 60, No. 6, pp.2701–2727.

Edmondson, A.C., Bohmer, R.M. and Pisano, G.P. (2001) 'Disrupted routines: team learning and new technology implementation in hospitals', *Administrative Science Quarterly*, Vol. 46, No. 4, pp.685–716.

Garg, V.K., Walters, B.A. and Priem, R.L. (2003) 'Chief executive scanning emphases, environmental dynamism, and manufacturing firm performance', *Strategic Management Journal*, Vol. 24, No. 8, pp.725–744.

Hahn, B.J. and Doh, J.P. (2006) 'Using Bayesian methods in strategy research: an extension of Hansen et al.', *Strategic Management Journal*, Vol. 27, No. 8, pp.783–798.

Hansen, M.H., Perry, L.T. and Reese, C.S. (2004) 'A Bayesian operationalization of the resource-based view', *Strategic Management Journal*, Vol. 25, No. 13, pp.1279–1295.

Harris, J. and Bromiley, P. (2007) 'Incentives to cheat: the influence of executive compensation and firm performance on financial misrepresentation', *Organization Science*, Vol. 18, No. 3, pp.350–367.

Harris, S.E. and Katz, J.L. (1991) 'Organizational performance and information technology investment intensity in the insurance industry', *Organization Science*, Vol. 2, No. 3, pp.263–295.

Jensen, M.C., Murphy, K.J. and Wruck, E.G. (2004) 'Remuneration: where we've been, how we got to here, what are the problems, and how to fix them', Harvard NOM Working Paper No. 04-28, available at http://ssrn.com/abstract=561305 (accessed on 20 January 2011).

Kass, R.E. and Raftery, A.E. (1995) 'Bayes factors', *Journal of the American Statistical Association*, Vol. 90, No. 430, pp.773–795.

Kiesler, S. and Sproull, L. (1982) 'Managerial response to changing environments: perspectives on problem sensing from social cognition', *Administrative Science Quarterly*, Vol. 27, No. 4, pp.548–570.

Lampel, J., Shamsie, J. and Shapira, Z. (2009) 'Experiencing the improbable: rare events and organizational learning', *Organization Science*, Vol. 20, No. 5, pp.835–845.

Lee, P.M. (1997) *Bayesian Statistics: An Introduction*, 2nd ed., Wiley, New York.

Lie, E. (2005) 'On the timing of CEO stock option awards', *Management Science*, Vol. 51, No. 5, pp.802–812.

McKelvey, B. and Andriani, P. (2005) 'Why Gaussian statistics are mostly wrong for strategic organization', *Strategic Organization*, Vol. 3, No. 2, pp.219–228.

Meyer, A.D. (1982) 'Adapting to environmental jolts', *Administrative Science Quarterly*, Vol. 27, No. 4, pp.515–537.

Meyer, A.D., Brooks, G.R. and Goes, J.B. (1990) 'Environmental jolts and industry revolutions: organizational responses to discontinuous change', *Strategic Management Journal*, Special issue, Summer, Vol. 11, pp.93–110.

Miles, R.E. and Snow, C.C. (1978) *Organizational Strategy, Structure, and Process***,** McGraw-Hill, New York.

Nystrom, P.C., Soofi, E.S. and Yasai-Ardekani, M. (2010) 'Identifying and analyzing extremes: illustrated by CEOs' pay and performance', *Organizational Research Methods*, Vol. 13, No. 4, pp.782–805.

O'Boyle, Jr., E. and Aguinis, H. (2012) 'The best and the rest: revisiting the normality of individual performance', *Personnel Psychology*, Vol. 65, No. 1, pp.79–119.

Resnick, S.I. (2007) *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Verlag, New York.

Schwab, A., Abrahamson, E., Fidler, F. and Starbuck, W.H. (2011) 'Researchers should make thoughtful assessments instead of null-hypothesis significance tests', *Organization Science*, Vol. 22, No. 4, pp.1105–1120.

Schwertman, N.C. and de Silva, R. (2007) 'Identifying outliers with sequential fences', *Computational Statistics and Data Analysis*, Vol. 51, No. 8, pp.3800–3810.

Siegel, P.A. and Hambrick, D.C. (2005) 'Pay disparities within top management groups: evidence of harmful effects on performance of high-technology firms', *Organization Science*, Vol. 16, No. 3, pp.259–274.

Soofi, E.S., Nystrom, P.C. and Yasai-Ardekani, M. (2009) 'Executives' perceived environmental uncertainty shortly after 9/11', *Computational Statistics and Data Analysis*, Vol. 53, No. 9, pp.3502–3515.

Starbuck, W.H. (2009) 'Perspective: cognitive reactions to rare events: perceptions, uncertainty, and learning', *Organization Science*, Vol. 20, No. 5, pp.925–937.

Stephens, M.A. (1974) 'EDF statistics for goodness-of-fit and some comparisons', *Journal of the American Statistical Association*, Vol. 69, No. 347, pp.730–737.

Stephens, M.A. (1979) 'Tests of fit for the logistic distribution based on the empirical distribution function', *Biometrka*, Vol. 66, No. 3, pp.591–595.

Tosi, H.L., Werner, S., Katz, J.P. and Gomez-Mejia, L.R. (2000) 'How much does performance matter? A meta-analysis of CEO pay studies', *Journal of Management*, Vol. 26, No. 2, pp.301–339.

Wade, J.B., O'Reilly, C.A. and Pollock, T.G. (2006) 'Overpaid CEOs and underpaid managers: fairness and executive compensation', *Organization Science*, Vol. 17, No. 5, pp.527–544.

Weick, K.E. (1995) *Sensemaking in Organizations*, Sage, Thousand Oaks, CA.

Yasai-Ardekani, M. and Nystrom, P.C. (1996) 'Designs for environmental scanning systems: tests of a contingency theory', *Management Science*, Vol. 42, No. 2, pp.187–204.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York (Reprinted in 1996 in Wiley Classic Series).

Appendix 1 Probability distributions

Under the standard assumption for statistical analysis, a sample of measurements x_1, x_2,…, x_n are observations generated independently according to the same probability distribution. That is, x_1,…, x_n are statistical replicates of a random variable X with a probability distribution $F(x) = \Pr(X_i \leq x)$. This presentation of a probability distribution in terms of the CDF is applicable to both discrete and continuous random variables. This function increases from 0 to 1, depicting accumulation of the probability.

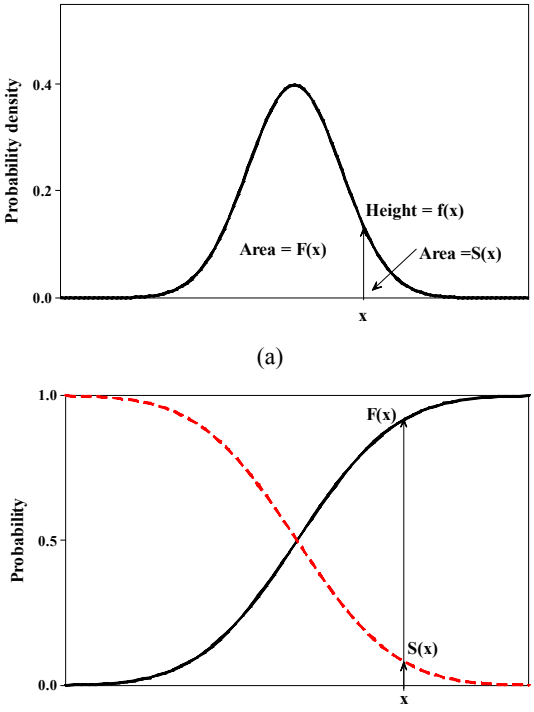
In the continuous case, the derivative of $F(x)$ is called the probability density function denoted by the lower case letter as $f(x)$. This is the more familiar density curve (often shown above z and other tables), which gives the probability of outcomes falling in an interval by the area under the curve. A density curve clearly shows some features of the distribution such as the shape and spread of the distribution, describes the relative likelihood for a random variable to take on a value in an interval, but a density curve does not provide clear comparison between distributions in terms of probability and percentile.

Figure 8 illustrates the relationship between the normal density curve and the normal CDF, $F(x)$. In Panel (a), $F(x) = \Pr(X \leq x)$ is the area under the curve up to and including the point x. The same area is given by the height of the solid curve in Panel (b), which is the normal CDF, $F(x)$. A third important representation of a probability distribution is the survival function defined as

$$S(x) = \Pr(X > x) = 1 - F(x).$$

The survival function gives the areas under the density curve for an outcome exceeding beyond a point x. In Panel (a) of Figure 8, $S(x) = \Pr(X > x)$ is the area under the curve to the right of point x. The same area is given by the height of the dashed curve in Panel (b), which is the normal survival function, $S(x)$. When a probability distribution represents a model for a population variable, the height of $S(x)$ gives the percentage of elements in the population greater than the value of x. Since $F(x)$ increases from 0 to 1, $S(x)$ decreases from 1 to 0.

Figure 8 Three representations of a probability distribution, (a) probability density function, (b) cumulative distribution and survival functions (see online version for colours)



A1.1 Some well-known families of distributions

Table 6 gives the formulas for the survival functions of several well-known distributions. These distributions are available at Wikipedia; Table 6 presents them with common notations and some remarks that are most relevant to the content of this paper, such as mean and variance undefined, heavy tail, fat tail, representations, special cases, and transformational relationships.

The first three distributions in Table 6 are well known *symmetric* distributions: Gaussian (normal), logistic, and Cauchy. These distributions have mound-shape density curves shown in Panel (a) of Figure 9. These distributions include two parameters (θ, λ), referred to as the location and scale parameters respectively. For these three distributions, the location parameter is the centre (median) of the density curve as well as the mode, where nearby outcomes occur with relatively higher probability than distant outcomes. The scale parameter determines the spread of distribution. When $\theta = 0$ and $\lambda = 1$, the distribution is said to be in the standard form (standardised). We consciously have avoided the traditional measures, the mean and SD, because these measures are not defined for all distributions, as noted in the last column of Table 6. The density curves in Panel (a) of Figure 9 are for the standard Gaussian and Cauchy distributions and the logistic distribution with the scale parameter as $\lambda = 0.57$. (The variances of the normal and logistic distributions in Figure 9 are nearly equal and their scale parameters are the

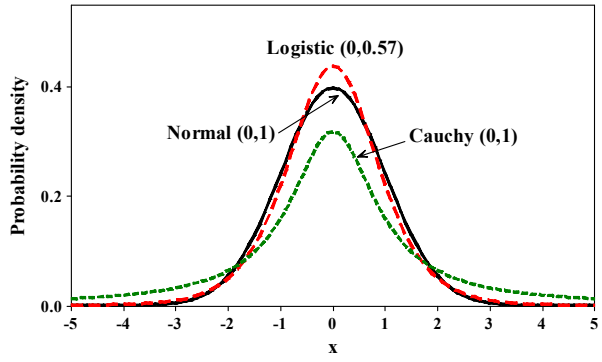
same as those for the log-normal and log-logistic models for the additional compensation reported in Table 2). Note that at the tail, the area under the Cauchy density curve is relatively higher than the other two distributions.

Table 6 Examples of probability distributions

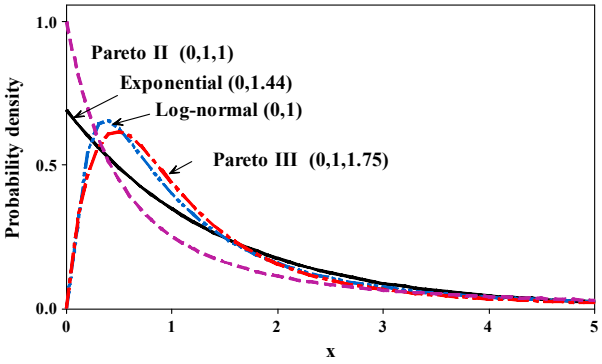
Model (parameters)	*Survival function* $S(x) = Pr(Outcome > x)$	*Remarks*
Symmetric distributions		
Normal (θ, λ)	$S(x) = 1 -$ entries in the normal z table, $z = \frac{x-\theta}{\lambda}, \lambda > 0$	Mean = Median = Mode = θ Variance = λ^2
Cauchy (θ, λ)	$S(x) = \frac{1}{\pi}\tan\left(\frac{x-\theta}{\lambda}\right) - \frac{1}{2}$	Median = Mode = θ Interquartile range = 2λ Fat tail, mean and variance undefined
Logistic (θ, λ)	$S(x) = \frac{e^{-\frac{x-\theta}{\lambda}}}{1+e^{-\frac{x-\theta}{\lambda}}}, \quad \lambda > 0$	Mean = Median = Mode = θ Variance = $\frac{\pi^2}{3}\lambda^2$
Skewed distributions		
Log-normal (θ, λ)	$S(x) = 1 -$ entries in the normal z table, $z = \frac{\log x - \theta}{\lambda}, \quad x \geq 0, \quad \lambda > 0$	Shape determined by both parameters Heavy tail $y = \log x$ Normal
Log-Cauchy (θ, λ)	$S(x) = \frac{1}{\pi}\tan^{-1}\left(\frac{\log x - \theta}{\lambda}\right) - \frac{1}{2}, x \geq 0$ $\lambda > 0$	Mean and variance undefined Shape determined by both parameters Supper-heavy tail $y = \log x$ Cauchy
Log-logistic (θ, λ) Pareto III representation: $x^a = \theta^{\frac{1}{\lambda}\log x}, \quad a = \frac{1}{\lambda}$	$S(x) = \frac{e^{-\frac{\log x-\theta}{\lambda}}}{1+e^{-\frac{\log x-\theta}{\lambda}}}, \quad x \geq 0, \quad \lambda > 0$	Mean undefined when $\lambda > 1$ Fat tail when $\lambda > 2$, variance undefined $y = \log x$ Logistic
Exponential (θ, λ)	$S(x) = e^{-\frac{x-\theta}{\lambda}}, \quad x \geq \theta, \quad \lambda > 0$	Mean = λ, Median = λ log2, Mode = 0; Variance = λ^2
Weibull (θ, λ, α) Special case: Exponential, α = 1	$S(x) = e^{-\left(\frac{x-\theta}{\lambda}\right)^{\alpha}}, \quad x \geq \theta, \quad \lambda > 0$	α, shape parameter; λ, Scale parameter Heavy tail when $\alpha < 1$ $y = x^{\alpha}$ exponential
Pareto II (θ, λ, α) Special case: Pareto I, θ = λ	$S(x) = \frac{1}{\left(1+\frac{x-\theta}{\lambda}\right)^{\alpha}}, \quad x \geq \theta, \quad \alpha, \lambda > 0$	Mean undefined when $\alpha < 1$ Fat tail when $\alpha < 2$, variance undefined
Pareto III (θ, λ, α)	$S(x) = \frac{1}{1+\left(\frac{x-\theta}{\lambda}\right)^{\alpha}}, \quad x \geq \theta, \quad \alpha, \lambda > 0$	α, shape parameter Mean undefined when $\alpha < 1$ Fat tail when $\alpha < 2$, variance undefined

Table 6 includes several other distributions, all of which have *skewed* density curves. These models are used in various fields for distributions of variables with non-negative outcomes such as income, sales, and duration. The log-transformed distributions usually start at $x = 0$, but can also include a shift parameter that sets the point where the area under the density curve begins. For these log-transformed distributions, the two parameters (θ, λ) are the location θ and scale λ of the distributions of the natural log of the variable; for example, for log-normal, θ is the mean and λ is the SD of the corresponding normal model for the natural log of the variable. For other distributions, $\theta \geq 0$ is the threshold parameter and $F(x) = 0$ for $x \leq \theta$ (the distribution starts at $x = \theta$), and $\lambda > 0$ is the scale parameter. The Weibull distribution includes a third parameter α that appears as exponent and determines the shape of its density curve. When $\alpha = 1$, the distribution is exponential. When $\alpha > 1$, the density curve has a mode and when $\alpha < 1$, the density curve shoots up along an asymptote at zero. Pareto II and Pareto III distributions also include a third parameter α that appears as exponent. Pareto I is a special case of Pareto II when $\theta = \lambda$. The third parameter α of the Pareto III distribution determines the shape of its density curve similar to the Weibull, but for the case of $\alpha = 1$ Pareto II and Pareto III are identical distributions. (Two other Pareto distributions, known as Pareto IV and generalised Pareto, are not discussed in this paper, hence not shown in Table 6).

Figure 9 Probability density functions of symmetric and skewed distributions, (a) symmetric density curves, (b) skewed density curves (see online version for colours)



(a)



(b)

Panel (b) of Figure 9 shows density curves of four of these *skewed* distributions: Exponential, Pareto II, Pareto III, and log-normal distributions. Here, the parameters are set such that the median=1 for all distributions. For all values of its parameters, the shape of Pareto II density curve is the same as shown in here. Since the exponent parameter of the Pareto III is $\alpha = 1.75 > 1$, its density curve has a mode.

As noted in Table 6, the mean for these distributions is undefined when the exponent parameter is less than or equal to 1 and the variance is undefined when it is less than or equal to 2. As mentioned in Section 2.2, Pareto III and log-logistic distributions are two representations of the same distribution. This is because of the following distributional relationship between two random variables: if the distribution of a variable X is Type III Pareto, then the distribution of its log transformation $Y = \log X$ is logistic, which is the same relationship between the log-logistic and logistic distributions, and is analogous to the relationship between the log-normal and normal distributions. Panel (b) of Figure 9 also includes a log-normal density curve. As can be seen, the log-normal (0, 1) and Pareto III (0, 1, 1.75) are similar.

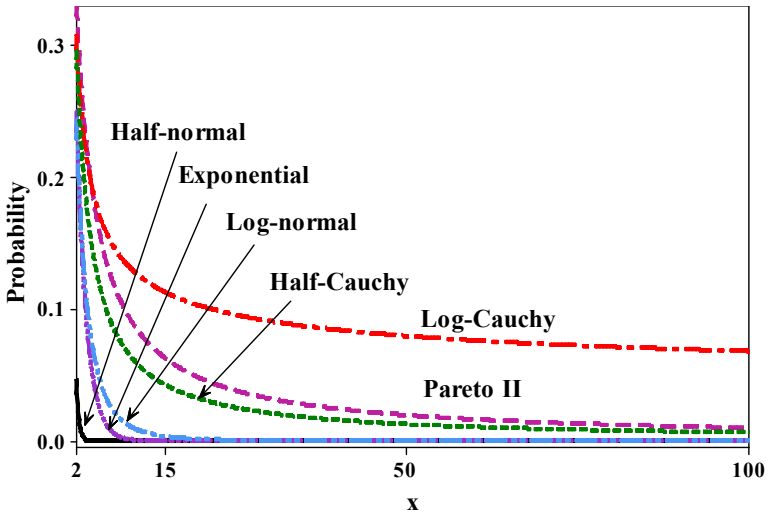
A1.2 Heavy-tail and mixture distributions

The class of heavy-tail distributions is defined in terms of a decaying property of $S(x)$; the fat-tail and long-tail distributions are sub-classes. The tails of heavy-tail distribution decay very slowly which make them capable of generating some data points far from the mass of the data. The Cauchy distribution is an example of symmetric heavy-tail distribution. The Cauchy distribution is a special case of the t family (t with degree of freedom = 1), all of which are heavy tail.

Figure 10 shows the survival functions $S(x)$ for $x > 2$, which depict the tail probabilities of the distributions in a more direct form. The vertical axis (height of a curve) in Figure 10 gives the tail probability (area under the respective density curve in Figure 9). The Half-normal and Half-Cauchy are the right halves of the respective distributions, normalised for the area under the curve to be one (these distributions are also used for duration variables such as lifetime). This figure also shows the survival function of the Log-Cauchy distribution, which is sometimes called super-heavy-tail distribution. As seen, Gaussian (normal) tail decays very quickly, followed by the exponential tail. (The logistic tail, not shown here, decays slower than the normal tail but faster than the exponential tail). Among the heavy-tail distributions in Figure 10, the log-normal's tail probabilities stand visible till about 15 (the Pareto III tail, not shown here, decays similarly), and the tails of the Log-Cauchy, Pareto I and Half-Cauchy continue well beyond, remaining visible at 100. Thus, unlike the normal distribution under which an outcome more than 4 is almost impossible, some heavy-tail distributions can produce outcomes tens and hundreds of units away from the median. Figure 10 shows that a data point of larger than 100 is a rare outcome, and yet it can be a legitimate outcome of the standard Cauchy, Log-Cauchy, or Pareto distributions, whereas for the standard normal distribution, having an outcome of larger than 5 can safely be judged as impossible. Thus, a researcher who simply assumes that a data set is distributed normally would conclude that a data point beyond a few SDs is an outlier and might decide to delete it from further analyses. But suppose that a test of that data set shows that a

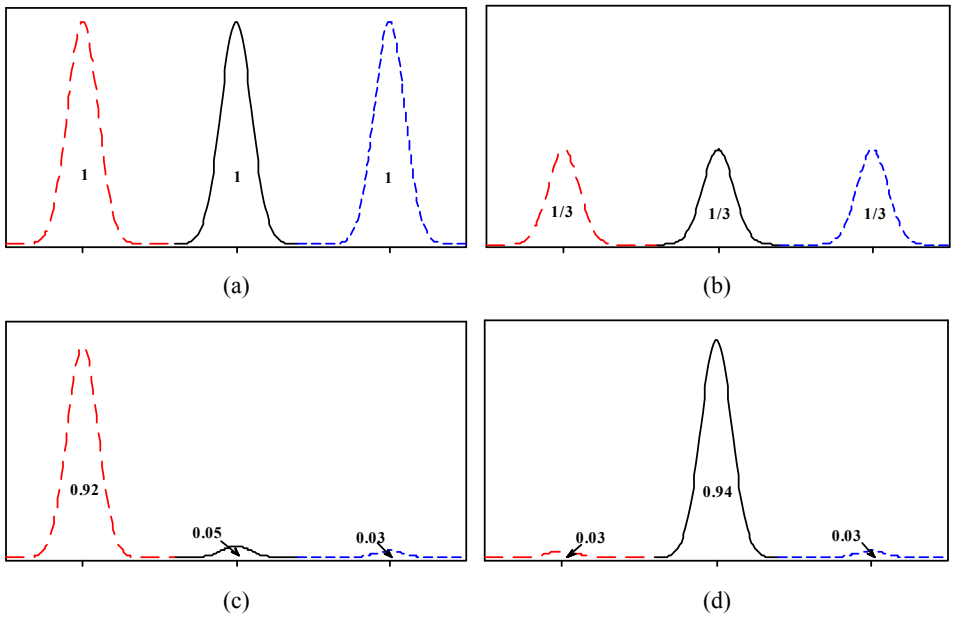
Cauchy actually fits the data better than a normal distribution. In that case, the correct interpretation of such a data point far away from the mass could be that it is a rare outcome and not an outlier.

Figure 10 Probability of an outcome greater than a value of x for six distributions [Survival function $S(x)$, $x>2$] (see online version for colours)



The probability model for data containing outliers is usually a *mixture*, meaning that it is a weighted average of two or more distributions. Mixtures are encountered in statistical problems that involve measurements for more than one group (two-sample problems, Analysis of Variance (ANOVA), clustering, to name a few). For example, the standard ANOVA assumption is that the distribution of measurements for each group is normal, the variances of distributions of different groups are the same, but their means can be different. Panel (a) of Figure 11 depicts three density plots for an ANOVA problem with three groups. When the ANOVA test rejects the equality of the means, then the distribution of all measurements for the three groups is a Gaussian mixture with three components. If the sizes of three groups in the population are equal, the population of measurements has a mixture distribution whose density curve is depicted in Panel (b) of Figure 11. In the case of outliers, the weights of one group is drastically different from the weights of others: the weight of one of the distributions in the mixture (the underlying distribution of the mass of outcomes) is very high relative to the others. (Think of the distribution of heights of buildings in a metropolitan area; mass of one and two story houses, a smaller proportion of several-story buildings, and a very small proportion of skyscrapers.) Panel (c) of Figure 11 depicts such a mixture. Panel (b) of Figure 11 depicts a mixture that generates outliers at either side of the mass. In general, the distributions in the mixture can take any form; none of them must be a normal distribution. The underlying distribution of data containing an outlier is sometimes referred to as a contaminated distribution [DeGroot and Schervish, (2002), pp.576–577].

Figure 11 Density functions of three normal distributions and various mixtures with three Gaussian components, (a) three normal distributions, (b) to (d) various mixtures of three normal distributions (see online version for colours)



Some heavy-tail distributions can be derived as continuous mixtures of infinitely many distributions. For example, Cauchy (in fact any *t*) distribution can be obtained as a continuous mixture of the normal distributions with varying variances (scale parameters); Type II Pareto distribution can be obtained as a continuous mixture of the exponential distributions with varying scale parameter; Type III Pareto distribution can be obtained as a continuous mixture of the Weibull distributions with varying scale parameter. (A continuous mixture is found by giving weight to each value of a parameter according to a probability density function). This derivation can provide a plausible explanation for the common feature of rare and outlier as being outcomes substantially distant from the mass of the data. Yet it is conceivable that a single heavy-tail model may fit the mass of the data points, but not all. In such case, one can draw a distinction between a rare outcome of a heavy-tail model and an outlier relative to that model which is generated from a different model; hence, capturing the entire set of data points would require a mixture.

Bivariate and multivariate versions of all univariate models discussed above are also available. For the heavy-tail distributions (Cauchy, Pareto), the correlation coefficient is not defined. However, Kendall's tau and Spearman rank measures are applicable to measure association between variables for all distributions. By their invariance property under monotone increasing transformations, each of these measures gives the same strength of association for bivariate log-normal as for bivariate normal. Each measure also gives the same strength of association for bivariate log-logistic (Pareto) as bivariate logistic and bivariate exponential.

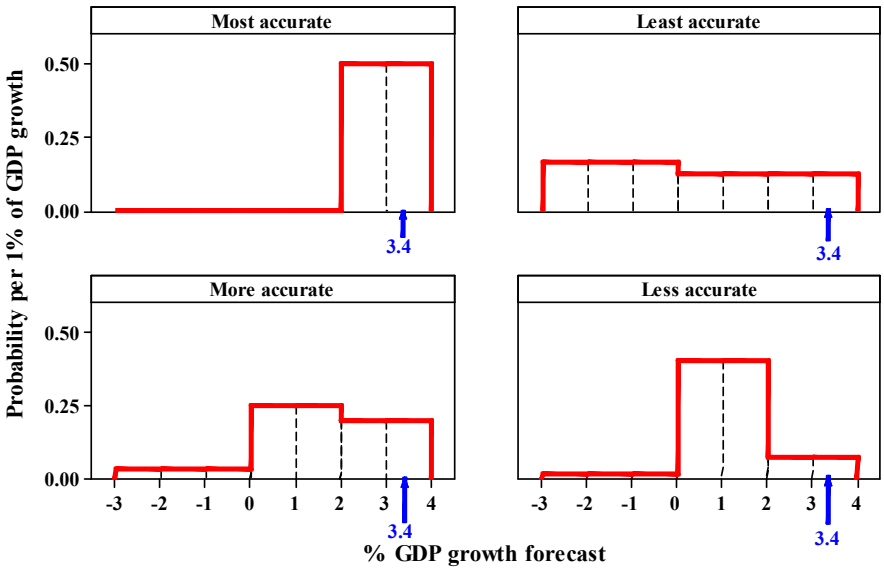
Appendix 2 Measure of forecasting inaccuracy

For the forecast W_i of each executive i = 1,…, 93, we construct a piece-wise uniform distribution in the range of –3% to +4%. This range was chosen according to the paragraph describing the GDP rates. Figure 12 shows the forecast distributions for four executives. The arrows point to the actual GDP growth for the forecast period – which turned out to be a surprisingly health 3.4%. Most executives made unduly pessimistic forecasts. Many executives no doubt dreaded the prospect of the US economy sliding back into another recession and research has found that "people overestimate the likelihoods of events that they dread" [Starbuck, (2009), p.931]. The upper panels in Figure 12 show distributions for the most accurate forecast (left) and the least accurate forecast (right) from amongst the 93 executives. The lower panels in Figure 12 show forecast distributions for another two executives, where the one at the left is relatively more accurate than the one at the right.

We measured the forecast inaccuracy of each executive using the *mean squared error* (MSE) of the executive's forecast distribution. An executive's forecast is scored as more inaccurate when her or his estimated probabilities are greater for outcomes that are farther away from what turns out to be the actual outcome (3.4% GDP growth in this case). The MSE of the forecast W_i for each executive is computed using the following relationship:

$$\mathrm{MSE}(W_i) = E\left[(W_i - 3.4)^2\right] = \mathrm{Var}(W_i) + \left[\mathrm{Mean}(W_i) - 3.4\right]^2$$

Figure 12 Economic forecast distributions for four respondents in the 9/11 study (see online version for colours)



Our results are robust; we obtained similar results when we tried some other procedures for constructing the forecast distribution and other loss functions such as the mean absolute error.