Thai Journal of Mathematics : (2016) 51–61
Special Issue on Applied Mathematics : Bayesian Econometrics

http://thaijmath.in.cmu.ac.th
ISSN 1686-0209

Empirically Successful Transformations from Non-Gaussian to Close-to-Gaussian Distributions: Theoretical Justification

Thongchai Dumrongpokaphan[†,1] Pedro Barragan[‡] and Vladik Kreinovich[‡]

[†]Department of Mathematics, Faculty of Science
Chiang Mai University, Chiang Mai, Thailand
e-mail : tcd43@hotmail.com
[‡]Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA
e-mail : pabarraganolague@miners.utep.edu (T. Dumrongpokaphan)
vladik@utep.edu (V. Kreinovich)

Abstract : A large number of efficient statistical methods have been designed for a frequent case when the distributions are normal (Gaussian). In practice, many probability distributions are not normal. In this case, Gaussian-based techniques cannot be directly applied. In many cases, however, we can apply these techniques *indirectly* – by first applying an appropriate transformation to the original variables, after which their distribution becomes close to normal. Empirical analysis of different transformations has shown that the most successful are the power transformations $X \to X^h$ and their modifications. In this paper, we provide a symmetry-based explanation for this empirical success.

This work was supported by Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation. The authors are thankful to Peter Moschopoulos for valuable discussions.

[1] Corresponding author.

Keywords : non-Gaussian distributions; transformations to Gaussian; symmetry.
2010 Mathematics Subject Classification : 62P20; 62G99; 60F17.

1 Formulation of the Problem

Many statistical techniques have been designed for normal distributions. A significant number of statistical techniques have been designed and tested for the case when the distributions are normal; see, e.g., [1].

Normal distributions are indeed ubiquitous in applications. Their abundance comes from the Central Limit Theorem, according to which, if we have many independent small random variables, then the distribution of their sum is close to normal. Thus, if the observed phenomenon is the joint effect of many independent random factors, its distribution is close to normal.

In practice, many distributions are not normal. While many real-life phenomena have close-to-normal distributions, many other distributions are different from norma. For example, an experimental analysis of measuring instruments showed that only about 60% have close-to-normal distributions; for others, the distribution is not normal; see, e.g., [2, 3].

In economics, non-Gaussian distributions are also widely spread; see, e.g., [4–11].

Need for transformations to close-to-normal distributions. Since many statistical methods are based on the assumption that the actual distribution is normal (or close to normal), but the actual distribution is often not normal, a natural idea to apply some transformation $Y = f(X)$ to the original random variable X so that for the new variable Y, the distribution will be close to normal.

Which transformations have been used? In order to describe which transformations have been used, let us take into account in many cases, non-normality comes from the fact that we consider a function $X = g(X_1, \ldots, X_n)$ of several independent variables $X_1, \ldots, X_n$ each of which is normally distributed. For linear functions $g(X_1, \ldots, X_n) = a_0 + \sum\limits_{i=1}^{n} a_i \cdot X_i$, the distribution is still normal. However, when we consider the next approximation – quadratic functions $g(X_1, \ldots, X_n)$, the distribution of X stops being normal.

The simplest possible case of a quadratic function is $X = g(X_1) = X_1^2$. In this case, if we take the square root $Y = f(X) = \sqrt{X}$, we get back the normally distributed variable X_1. In view of this example, the very first idea of transformation to a close-to-normal distribution was to use the function $f(X) = \sqrt{X}$. This idea was first proposed by R. A. Fisher; see, e.g., [12], pp. 96–97.

In some cases, this transformation works well, but, somewhat surprisingly, it turned out that other transformations work better that the square root. The first such transformation $f(X) = X^{1/3}$, was proposed by Wilson and Hilferty in 1931 [13].

In 1972, it was shown that a more general transformation $f(X) = \left(\frac{X}{c}\right)^h$ for general c and h results in variables y whose distribution is closer to normal that $\sqrt{X}$. This power-law transformation has been efficiently used in statistics; see, e.g., [14–18].

An even more general transformation was proposed in [15]:

$$f(X) = \left(\frac{X+a}{c}\right)^h.$$

Main problem: these empirically successful transformations are heuristic. While Fisher's square root transformation has some justification, all the other transformations – starting with the cubic root – are pure heuristics: let us try this, and see what happens.

As a result, it is not clear if these are the transformations that we should use – or there are some other transformations which are more adequate for our goal.

Bayesian approach. According to decision making theory (see, e.g., [19–22]), a rational decision maker always maximizes the expected value of the objective appropriate function called *utility*. In our case, the utility function $u(f,\rho)$ depends on how close, for the variable X with the original probability distribution ρ, the distribution of $f(X)$ is close to Gaussian.

The main idea behind the Bayesian approach is that in situations in which several different alternatives are consistent with our knowledge, we select a *prior* probability distribution on the set of all possible alternatives.

In our problem, alternatives are different probability distributions. Thus, for our problem, the Bayesian approach means that we select a prior distribution on the set of all possible probability distributions $\rho(z)$. Then, we find the function $f(X)$ that maximizes the expected utility $\int u(f,\rho)\,d\rho$.

What we do in this paper. The solution to the above optimization problem depends on the selection of the prior distribution. A usual way to come up with such a prior distribution is to require that this distribution is invariant with respect to some natural symmetries. In this case, if the utility function is also invariant, the resulting transformation should also be invariant with respect to the same symmetries.

This is what we do in this paper. The symmetries that we use in this paper come from the fact that the numerical values of real-life quantities are not absolute: they depend on the choice of the measuring unit and on the choice of the starting point (see, e.g., [23, 24]).

It is therefore reasonable to consider transformations that do not depend on these choices. It turns out that, as a result, we get exactly the empirical transformations listed above.

Comment. In our theoretical explanation, we will use symmetry ideas and derivations which are mathematically similar to ideas and derivations that we used in a different problem: the problem of estimating stiffness for unbound aggregate materials (e.g., for the road's subbase); see [25].

2 Symmetries: General Idea

Symmetry idea: general case. Computers process numerical values of different quantities. A numerical value of a quantity depends on the choice of a measuring unit and – in many cases – also on the choice of the starting point.

For example, depending on the choice of a measuring unit, we can describe the height of the same person as 1.7 m or 170 cm. Similarly, we can describe the same moment of time as 2 pm (14.00) if we use El Paso time or 3 pm (15.00) if we use Austin time – the difference is caused by the fact that the starting points for these two times – namely midnight (00.00) in El Paso and midnight (00.00) in Austin – differ by one hour.

The choice of a measuring unit is rather arbitrary. For example, we can measure length in meters or in centimeters or in feet. Similarly, the choice of the starting point is arbitrary: when we analyze a cosmic event, it does not matter the time of what location we use to describe it. It is therefore reasonable to require that the fundamental physical formulas not depend on the choice of a measuring unit and – if appropriate – on the choice of the starting point. We do not expect that, e.g., Newton's laws look differently if we use meters or feet.

Of course, if we change the units in which we measure one of the quantities, then we may need to adjust units of related quantities. For example, if we replace meters with centimeters, then for the formula $v = d/t$ (that describes velocity v as a ratio of distance d and time t) to remain valid, we need to replace meters per second with centimeters per second when measuring velocity. However, once the appropriate adjustments are made, we expect the formulas to remain the same.

Symmetry idea: economic case. We can measure income in US dollars or in Euros or in Thai Bahts. If we change the measuring unit, the amount of income remains the same, but its numerical value changes.

It is reasonable to require that all relationships – in particular, the transformation that transforms the original variable into a close-to-Gaussian one – should not depend on what exactly unit we use.

Change of a starting point also makes perfect sense for some economic quantities. For example, we can measure unemployment in absolute units, or we can measure it by considering the difference $X - X_0$ between the actual unemployment level X and the ideal level $X_0 > 0$ which, in the opinion of the economists, corresponds to full employment; see, e.g., [24] and references therein.

How to describe the corresponding symmetries in precise terms. If we replace the original measuring unit with a new unit which is a times smaller, then all numerical values of the measured quantity get multiplied by a: $x' = a \cdot x$.

For example, if we replace dollars with cents – which are $a = 100$ times smaller

– then the original amount of $X = \$1.7$ becomes

$$X' = a \cdot X = 100 \cdot 1.7 = 170 \text{ cents.}$$

Similarly, if we replace the original starting point by a new one which is b earlier (or smaller), then to all numerical values of the measured quantity the value b is added: $X' = X + b$. For example, if the current unemployment level is $X = 8\%$, and the ideal unemployment level is $X_0 = 3\%$ (which corresponds to $b = -3$), then the re-scaled value of unemployment is $X + b = 8 + (-3) = 5\%$.

In general, we can change both the measuring unit and the starting point. If we first change the measuring unit and the starting point, then:

- first, the original value X first gets multiplied by a, resulting in $X' = a \cdot X$, and
- then the value b is added to the new value X', resulting in

$$X'' = X' + b = a \cdot X + b.$$

Thus, in general, when we change both the measuring unit and the starting point, we get a linear transformation $x \to a \cdot x + b$.

3 Using Symmetries: First Try

Idea. We want a transformation $Y = f(X)$ to be independent on the scale used to describe X. In other words, if we re-scale X, then after appropriately re-scaling Y, we should get the exact same transformation.

In the original scale, we had values X for which the transformation leads to $Y = f(X)$. A general re-scaling has the form $X \to a \cdot X + b$, for some $a > 0$. So, what was X in the original units becomes $X' = a \cdot X + b$ in the new units. If we now apply the same transformation f to the new numerical value X', we get a new value $Y = f(X') = f(a \cdot X + b)$.

The desired invariance means that this new dependence $Y = f(a \cdot X + b)$ should have exactly the same form as $Y = f(X)$ – if we also appropriately re-scale Y, i.e, if we apply a transformation $Y' = c \cdot Y + d$ for some $c > 0$ and d.

Thus, we arrive at the following definition.

From the idea to its precise description. We would like to find all functions $f(X)$ for which, for every pair (a, b) with $a > 0$, there exists a pair (c, d) with $c > 0$ for which, for all X, we have

$$f(a \cdot X + b) = c(a, b) \cdot f(X) + d(a, b). \tag{1}$$

What we show in this section: that such a requirement is not possible.

Continuity and smoothness of the transformation $f(X)$. It is reasonable to require that the transformation $f(X)$ preserve continuity, i.e., that the function $f(X)$ be continuous.

It is well known that on any interval, every continuous function can be approximated, with any given accuracy, by a smooth (differentiable) function – it can even be approximated by a polynomial. Thus, without losing generality, we can assume that the transformation $f(X)$ is differentiable.

Now, we are ready to start the mathematical analysis of the above problem.

From smoothness of the transformation $f(X)$, it follows that the auxiliary functions $c(a,b)$ and $d(a,b)$ are also differentiable. Let us fix two different values $X_1 \neq X_2$ for which $f(X_1) \neq f(X_2)$. Then, for every a and b, by applying the formula (1) with $X = X_1$ and $X = X_2$, we have a system of two linear equations with two unknowns $c(a,b)$ and $d(a,b)$:

$$f(a \cdot X_1 + b) = c(a,b) \cdot f(X_1) + d(a,b); \tag{2}$$

$$f(a \cdot X + b) = c(a,b) \cdot f(X_2) + d(a,b). \tag{3}$$

Subtracting equation (3) from equation (2) and dividing both sides of the resulting equality by $f(X_1) - f(X_2)$, we conclude that

$$c(a,b) = \frac{f(a \cdot X_1 + b) - f(a \cdot X_2 + b)}{f(X_1) - f(X_2)}. \tag{4}$$

Similarly, if we first multiply the equation (2) by $f(X_2)$ and the equation (3) by $f(X_1)$ and then subtract the results, we get

$$d(a,b) = \frac{f(X_2) \cdot f(a \cdot X_1 + b) - f(X_1) \cdot f(a \cdot X_2 + b)}{f(X_2) - f(X_1)}. \tag{5}$$

The function $f(X)$ is smooth. Thus, the right-hand sides of the formulas (4) and (5) are also smooth – as compositions of smooth functions. Thus, the auxiliary functions $c(a,b)$ and $d(a,b)$ are indeed smooth.

Since all the functions are differentiable, let us differentiate the above equality. Since all the functions $f(X)$, $c(a,b)$, and $d(a,b)$ involved in the equality (1) are differentiable, we can differentiate both sides of this equality.

Let us first differentiate both sides with respect to b, and take $a = 1$ and $b = 0$. As a result, we get the following equality:

$$f'(X) = c_2 \cdot f(X) + d_2, \tag{6}$$

where:

- $f'(X)$ denotes the derivative of the function $f(X)$,
- c_2 denotes the value of the derivative of the functions $c(a,b)$ with respect to its second argument b when $a = 1$ and $b = 0$: $c_2 = \dfrac{\partial c}{\partial b}(1,0)$; and
- d_2 denotes the value of the derivative of the functions $d(a,b)$ with respect to its second argument b when $a = 1$ and $b = 0$: $d_2 = \dfrac{\partial d}{\partial b}(1,0)$.

Similarly, by differentiating with respect to a, we get the equality

$$X \cdot f'(X) = c_1 \cdot f(X) + d_1, \tag{7}$$

where $c_1 = \dfrac{\partial c}{\partial a}(1, 0)$ and $d_1 = \dfrac{\partial c}{\partial a}(1, 0)$.

Dividing both sides of (7) by both sides of (6), we conclude that

$$X = \frac{c_1 \cdot f(X) + d_1}{c_2 \cdot f(X) + d_2}, \tag{8}$$

i.e., we conclude that X can be obtained from $f(X)$ by a fractional-linear transformation. Thus, $f(X)$ can be obtained from X by an inverse to a fractional-linear transformation – and we know that such inverse transformations are also fractional-linear. So,

$$f(X) = \frac{p \cdot X + q}{r \cdot X + s}, \tag{9}$$

for some values p, q, r, and s.

If $r = 0$, then we simply get a linear function $f(X)$. It is well known, however, that a linear transformation of a Gaussian distribution is still Gaussian, so such a transformation will not help us transform a non-Gaussian distribution into a close-to-Gaussian one. Thus, for our purposes, we can safely assume that $r \neq 0$. In this case, we can divide both numerator and denominator of the expression (8) by $r \neq 0$, and get a simplified formula

$$f(X) = \frac{P \cdot X + Q}{X + S}, \tag{10}$$

where we denoted $P = \dfrac{p}{r}$, $Q = \dfrac{q}{r}$, and $S = \dfrac{s}{r}$.

Substituting the expression (10) into the formula (6), we get

$$\frac{P \cdot (X + S) - (P \cdot X + Q)}{(X + S)^2} = c_2 \cdot \frac{P \cdot X + Q}{X + S} + d_2. \tag{11}$$

Multiplying both sides of this formula by $(X + S)^2$, we get

$$P \cdot (X + S) - (P \cdot X + Q) = c_2 \cdot (P \cdot X + Q) \cdot (X + S) + d_2 \cdot (X + S)^2. \tag{12}$$

For $X = -S$, this formula leads to $P \cdot (-S) + Q = 0$, hence $Q = P \cdot S$,

$$P \cdot X + Q = P \cdot X + P \cdot S = P \cdot (X + S), \tag{13}$$

and thus,

$$f(X) = \frac{P \cdot X + Q}{X + S} = P = \text{const}. \tag{14}$$

So, only constant functions $f(X)$ are invariant with respect to all possible transformations.

4 Using Symmetries: Final Result

Main idea: since we cannot require all the invariances, let us require only some of them. Since we cannot require invariance with respect to *all* possible re-scalings, we should require invariance with respect to *some* family of re-scalings.

It is sufficient to consider infinitesimal transformations. If a formula does not change when we apply each transformation, it will also not change if we apply them one after another, i.e., if we consider a composition of transformations. Each shift can be represented as a superposition of many small (infinitesimal) shifts, i.e., shifts of the type $X \to X + B \cdot dt$ for some B.

Similarly, each re-scaling can be represented as a superposition of many small (infinitesimal) re-scalings, i.e., re-scalings of the type $X \to (1 + A \cdot dt) \cdot X$.

Thus, it is sufficient to consider invariance with respect to an infinitesimal transformation, i.e., a linear transformation of the type

$$X \to X' = (1 + A \cdot dt) \cdot X + B \cdot dt.$$

From the idea to exact formulas. Invariance means that the value $f(X')$ has the same form as $f(X)$, i.e., that $f(X')$ is obtained from $f(X)$ by an appropriate (infinitesimal) re-scaling $Y \to (1 + C \cdot dt) \cdot Y + D \cdot dt$. In other words, we require that

$$f((1 + A \cdot dt) \cdot X + B \cdot dt) = (1 + C \cdot dt) \cdot f(X) + D \cdot dt, \tag{14}$$

i.e., that

$$f(X + (A \cdot X + B) \cdot dt) = f(X) + C \cdot f(X) \cdot dt + D \cdot dt. \tag{14a}$$

Here, by definition of the derivative, $f(X + q \cdot dt) = f(X) + f'(X) \cdot q \cdot dt$. Thus, from (14a), we conclude that

$$f(X) + (A \cdot X + B) \cdot f'(X) \cdot dt = f(X) + C \cdot f(X) \cdot dt + D \cdot dt. \tag{15}$$

Subtracting $f(X)$ from both sides of this formula and dividing the resulting equality by dt, we conclude that

$$(A \cdot X + B) \cdot f'(X) = C \cdot f(X) + D. \tag{16}$$

Since $f'(X) = \dfrac{df}{dX}$, we can separate the variables by moving all the terms related to f to one side and all the terms related to X to another side. As a result, we get

$$\frac{df}{C \cdot f + D} = \frac{dX}{A \cdot X + B}. \tag{17}$$

Degenerate cases when $A = 0$ and/or $C = 0$ can be approximated, with any given accuracy, by cases when A or C is small but non-zero. So, without losing generality, we can safely assume that $A \neq 0$ and $C \neq 0$.

In this case, for $x \stackrel{\text{def}}{=} X + a$ and $y \stackrel{\text{def}}{=} f + \ell$, where $a \stackrel{\text{def}}{=} \dfrac{B}{A}$ and $\ell \stackrel{\text{def}}{=} \dfrac{D}{C}$, we have

$$\frac{dy}{y} = h \cdot \frac{dx}{x}, \tag{18}$$

where $h \stackrel{\text{def}}{=} \dfrac{C}{A}$. Integration leads to $\ln(y) = h \cdot \ln(x) + C_0$ for some constant C_0, thus $y = C_1 \cdot x^h$ for $C_1 \stackrel{\text{def}}{=} \exp(C_0)$, hence $y = C_1 \cdot (X + a)^h$ and

$$f(X) = y - \ell = C_1 \cdot (X + a)^h - \ell. \tag{19}$$

Our objective is to transform a distribution into a close-to-Gaussian form. Adding or subtracting a constant ℓ does not change the Gaussian character or a random variable, so the use of the transformation (19) is equivalent to using a simpler transformation $f(X) = C_1 \cdot (X + a)^h$. By representing C_1 as c^{-h} for an appropriate $c = C_1^{-1/h}$, we arrive at the following conclusion.

Conclusion. If we want to transform a non-Gaussian distribution into a close-to-Gaussian one, then the only invariant transformations are exactly the transformations

$$f(X) = \left(\frac{X + a}{c}\right)^h$$

that have been empirically shown to be most efficient.

So, we have come up with the desired theoretical justification for the empirically successful transformations.

References

[1] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[2] P.V. Novitskii, I.A. Zograph, Estimating the Measurement Errors, Energoatomizdat, Leningrad, 1991 (in Russian).

[3] A.I. Orlov, How often are the observations normal?, Industrial Laboratory 57 (7) (1991) 770-772.

[4] B.K. Chakrabarti, A. Chakraborti, A. Chatterjee, Econophysics and Sociophysics: Trends and Perspectives, Wiley-VCH, Berlin, 2006.

[5] A. Chatterjee, S. Yarlagadda, B.K. Chakrabarti, Econophysics of Wealth Distributions, Springer-Verlag Italia, Milan, 2005.

[6] J. . Farmer, T. Lux (eds.), Applications of statistical physics in economics and finance, a special issue of the Journal of Economic Dynamics and Control 32 (1) (2008) 1-320.

[7] B. Mandelbrot, The variation of certain speculative prices, J. Business 36 (1963) 394-419.

[8] B. Mandelbrot, R.L. Hudson, The (Mis)behavior of Markets: A Fractal View of Financial Turbulence, Basic Books, 2006.

[9] J. McCauley, Dynamics of Markets, Econophysics and Finance, Cambridge University Press, Cambridge, Massachusetts, 2004.

[10] S.V. Stoyanov, B. Racheva-Iotova, S.T. Rachev, F.J. Fabozzi, Stochastic models for risk estimation in volatile markets: a survey, Annals of Operations Research 196 (2010) 293-309.

[11] P. Vasiliki, H.E. Stanley, Stock return distributions: tests of scaling and universality from three distinct stock markets, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 77 (3, Pt. 2) (2008) Publ. 037101.

[12] R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, London, UK, 1928.

[13] E.B. Wilson, M.M. Hilferty, The distribution of chi-square, Proceedings of the US National Academy of Sciences 17 (1931) 684-688.

[14] A. Mathai, S. Provost, Quadratic Forms in Random Variables, Marcel Dekker, New York, Basel, Hong Kong, 1992.

[15] P.G. Moschopoulos, On a new transformation to normality, Communications in Statistics 12 (16) (1983) 1873-1875.

[16] P.G. Moschopoulos, G.S. Mudholkar, A likelihood-ratio-based normal approximation for the non-null distribution of the multiple correlation coeffcient, Communications in Statistics – Simulation and Computation 12 (3) (1983) 355-371.

[17] G.S. Mudholkar, M.C. Trivedi, A normal approximation for the distribution of the likelihood ratio statistic in multivariate analysis of variance, Biometrika 67 (2) (1980) 485-488.

[18] J.G. Staniswalis, T.A. Severini, P.G. Moschopoulos, On a data-based power transformation for reducing skewness, Journal of Statistical Computation and Simulation 46 (1993) 91-100.

[19] H.T. Nguyen, O. Kosheleva, V. Kreinovich, Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction, International Journal of Intelligent Systems 24 (1) (2009) 27-47.

[20] H. Raiffa, Decision Analysis, Addison-Wesley, Reading, Massachusetts, 1970.

[21] P.C. Fishburn, Utility Theory for Decision Making, John Wiley & Sons Inc., New York, 1969.

[22] R.D. Luce, R.Raiffa, Games and Decisions: Introduction and Critical Survey, Dover, New York, 1989.

[23] M. Koshelev, Standard statistical transformations (logarithm and logit) are uniquely determined by the corresponding symmetries, Journal of Uncertain Systems 9 (2) (2015) 103-112.

[24] V. Kreinovich, O. Kosheleva, H.T. Nguyen, S. Sriboonchitta, Invariance explains multiplicative and exponential skedactic functions, In: V.N. Huynh, V. Kreinovich, and S. Sriboonchitta (eds.), Causal Inference in Econometrics, Springer Verlag, Cham, Switzerland (2016), 119-131.

[25] P. Barragan Olague, S. Nazarian, V. Kreinovich, A. Gholamy, M. Merani, How to estimate resilient modulus for unbound aggregate materials: a theoretical explanation of an empirical formula, Proceedings of the 2016 World Conference on Soft Computing, Berkeley, California, May 22-25, (2016) 203-207.

(Received 27 August 2016)
(Accepted 29 October 2016)