Why Some Families of Probability Distributions Are Practically Efficient: A Symmetry-Based Explanation

Vladik Kreinovich, Olga Kosheleva, Hung T. Nguyen, and Songsak Sriboonchitta

Abstract Out of many possible families of probability distributions, some families turned out to be most efficient in practical situations. Why these particular families and not others? To explain this empirical success, we formulate the general problem of selecting a distribution with the largest possible utility under appropriate constraints. We then show that if we select the utility functional and the constraints which are invariant under natural symmetries – shift and scaling corresponding to changing the starting point and the measuring unit for describing the corresponding quantity x. then the resulting optimal families of probability distributions indeed include most of the empirically successful families. Thus, we get a symmetry-based explanation for their empirical success.

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Olga Kosheleva
University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA,
e-mail: olgak@utep.edu

Hung T. Nguyen
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico 88003, USA, and Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand, e-mail: hunguyen@nmsu.edu

Songsak Sriboonchitta
Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand,
e-mail: songsak@econ.chiangmai.ac.th

1 Formulation of the Problem

Some families of probability distributions are empirically successful. Theoretically, we can have infinite many different families of probability distributions, but in practice, only a few families have been empirically successful; see, e.g., [20].

In some cases, there is a clear theoretical explanation for the families' success, but not always. For some of these families, there is a good theoretical explanation for their success. For example, the Central Limit theorem explains the ubiquity of normal distributions in situations where we have a joint effect of numerous small factors.

However, for many other empirically successful families of distributions, there is no clear theoretical explanation for their empirical success.

What we do in this paper. In this paper, for many empirically successful families of 1-D distributions, we provide a theoretical explanation of their success.

2 Our Main Idea

What we want: reminder. In general, we are looking for a family which is the best among all the families that satisfy appropriate constraints.

Natural symmetries. In selecting appropriate objective functions (which describe what is the best) and appropriate constraints, we use the fact that in practical applications, the numerical value of the corresponding quantity x depends:

- on the choice of the starting point for its measurement and
- on the choice of the measuring unit.

If we change the starting point to the one which is x_0 units smaller, then all the values shift by x_0: $x \to x + x_0$. Similarly, if we change the original measuring unit by a one which is λ times smaller, then all the values are scaled by λ: $x \to \lambda \cdot x$.

For example, if we replace a meter with a centimeter, a 100 times smaller measuring unit, then all numerical values multiply by 100: 2 m becomes 200 cm. Another example: shift and scaling describe the transition between Celsius to Fahrenheit.

Invariance. Since these shifts and scaling do not change the corresponding quantities – just change the numbers that represent their values – it is therefore reasonable to require that the appropriate objective functions and constraints do not change (= are invariant) under these transformations.

What we do in this paper: an idea. Since it is reasonable to restrict ourselves to invariant objective functions and invariant constraints, we describe all such objective functions and constraints. We then describe the distributions which are optimal for thus selected objective functions and constraints.

It turns out that the resulting optimal families indeed include many empirically successful families of distributions. Thus, our approach explains the empirical success of many such families.

Comments. The fact that natural symmetries explain the empirical success of families of probability distributions is in good accordance with modern physics, where symmetries are one of the main ways to generate new physical theories; see, e.g., [11]. This fact is also in good accordance with the fact that many empirically successful formulas from neural networks, expert systems, etc., can also be explained by the corresponding symmetries [31].

It should be noted that in this paper, we only consider 1-D families. Our preliminary results show that a symmetry-based explanation of empirically successful families can be extended to the multi-D case as well. Indeed, one of the ways to describe a multi-D distribution $F(x_1, \ldots, x_n) = \text{Prob}(X_1 \leq x_1 \& \ldots \& X_n \leq x_n)$ is to describe the corresponding marginal distributions $F_i(x_i) = \text{Prob}(X_i \leq x_i)$ and the corresponding *copula*, i.e., a function $C(u_1, \ldots, u_n)$ for which $F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))$ [25, 26, 30]. In [24], we have shown that symmetries can explain the empirically successful families of copulas.

3 Which Objective Functions Are Invariant?

We should maximize utility. According to decision theory, decisions of a rational agent are equivalent to maximizing a certain objective function known as *utility*; see, e.g., [12, 27, 29, 40].

Localness property. Based on partial information about the probability distribution, we want to reconstruct the values $\rho(x)$ corresponding to all possible x. It is reasonable to require that if have two distribution which differ only in some local region, and the first distribution is better, then if we replace a common distribution outside this region by another common distribution, the first distribution will still be better.

It is known (see, e.g., [13]) that each utility function with this property is either a sum or a product of functions depending only on the local value $\rho(x)$. Since maximizing the product is equivalent to maximizing its logarithm, and logarithm of the product is equal to the sum of logarithms, we can thus conclude, without losing generality, that the utility function is a sum of functions of $\rho(x)$. In the continuous case, with infinitely many variables $\rho(x)$, we have the limit of the sums, i.e., an integral. Thus, the general expression of an objective function with the localness property is $\int A(\rho(x), x) \, dx$.

Shift-invariance. We want the resulting criterion not to change if we simply shift x, i.e., replace each numerical value x with the shifted value $x + x_0$. Thus, the above integral expression should not change – which means that there should be no explicit dependence on x, i.e., that we should have $\int A(\rho(x)) \, dx$.

Scale-invariance: formulation of the requirement. We also want the resulting comparison not to change if we simply re-scale x, i.e., replace each numerical value x with the re-scaled value $y = \lambda \cdot x$. In terms of the re-scaled values, the pdf changes to $\lambda^{-1} \cdot \rho(\lambda^{-1} \cdot y)$, so the new objective function has the form

$$\int A(\lambda^{-1} \cdot \rho(\lambda^{-1} \cdot y)) \, dy.$$

By changing the variable to $x = \lambda^{-1} \cdot y$, we get $\lambda \cdot \int A(\mu \cdot \rho(x)) \, dx$, where we denoted $\mu \stackrel{\text{def}}{=} \lambda^{-1}$.

Scale-invariance: analysis of the requirement. Scale-invariance means, in particular, that if we add a small deviation $\delta\rho(x)$ to the original distribution in such a way that the value of the objective function does not change, then the value of the re-scaled objective function should not change either. The fact that we still get a pdf means that $\int \delta\rho(x) \, dx = 0$. For small deviations, $A(\rho(x) + \delta\rho) = A(\rho(x)) + A'(\rho(x)) \cdot \delta\rho(x)$. Thus, the fact that the value of the re-scaled objective function does not change means that $\int A'(\rho(x)) \cdot \delta\rho(x) \, dx = 0$. Similarly, the fact that the value of the original objective function does not change means that

$$\int A'(\mu \cdot \rho(x)) \cdot \delta\rho(x) \, dx = 0.$$

So, we arrive at the following requirement: for every function $\delta\rho(x)$ for which $\int \delta\rho(x) \, dx = 0$ and $\int A'(\rho(x)) \cdot \delta\rho(x) \, dx = 0$, we should have

$$\int A'(\mu \cdot \rho(x)) \cdot \delta\rho(x) \, dx = 0.$$

Functions form a Hilbert space – an infinite-dimensional analog of the Euclidean space, with the scalar (dot) product $\langle a, b \rangle \stackrel{\text{def}}{=} \int a(x) \cdot b(x) \, dx$. In these terms, the condition $\int A'(\rho(x)) \cdot \delta\rho(x) \, dx = 0$ can be written as $\langle A'(\rho), \delta\rho \rangle = 0$, i.e., as the condition that the functions $A'(\rho(x))$ and $\delta\rho(x)$ are orthogonal: $\delta\rho \perp A'(\rho)$. In these terms, the invariance requirement means that any function $\delta\rho$ which is orthogonal to 1 and to $A'(\rho(x))$ should also be orthogonal to the function $A'(\mu \cdot \rho(x))$. From this geometric reformulation, one can see that the function $A'(\mu \cdot \rho(x))$ should belong to the linear space spanned by 1 and $A'(\rho(x))$, i.e., that we should have

$$A'(\mu \cdot \rho(x)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x)) \tag{1}$$

for some constants $a(\mu, \rho)$ and $b(\mu, \rho)$.

Let us show that the values $a(\mu, \rho)$ and $b(\mu, \rho)$ do not depend on the pdf $\rho(x)$. Indeed, if we plug in two different values x_1 and x_2 into the formula (1), we get a system of two linear equations for two unknowns $a(\mu, \rho)$ and $b(\mu, \rho)$:

$$A'(\mu \cdot \rho(x_1)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x_1));$$

$$A'(\mu \cdot \rho(x_2)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x_2)).$$

From this system, we conclude that

$$b(\mu,\rho) = \frac{A'(\mu \cdot \rho(x_2)) - A'(\mu \cdot \rho(x_1))}{A'(\rho(x_2)) - A'(\rho(x_1))}$$

and $a(\mu,\rho) - b(\mu,\rho) \cdot A'(\rho(x_1))$. From these formulas, we see that the values of $a(\mu,\rho)$ and $b(\mu,\rho)$ depend only on the values $\rho(x_1)$ and $\rho(x_2)$ and thus, do not depend on any other value $\rho(x)$ for $x \neq x_1, x_2$.

If we start with some other values x_1', x_2' which are different from x_1 and x_2, we conclude that $a(\mu,\rho)$ and $b(\mu,\rho)$ do not depend on the values $\rho(x_1)$ and $\rho(x_2)$ either. Thus, a and b do not depend on $\rho(x)$ at all: $a(\mu,\rho) = a(\mu)$, $b(\mu,\rho) = b(\mu)$, and the equation (1) takes the form

$$A'(\mu \cdot \rho(x)) = a(\mu) + b(\mu) \cdot A'(\rho(x)). \tag{2}$$

It is reasonable to consider the case when the function $A(\rho)$ is twice differentiable – we can do that since any continuous function can be approximated, with any given accuracy, by twice differentiable functions. In this case, the derivative A' is differentiable. From the above expression for a and b in terms of A', we conclude that the functions $a(\mu)$ and $b(\mu)$ are also differentiable. Differentiating both side of the equality (2) with respect to μ and taking $\mu = 1$, we get

$$\rho \cdot \frac{dA'}{d\rho} = a'(1) + b'(1) \cdot A'(\rho).$$

We can separate A and ρ if we multiply both sides by $d\rho$ and divide both sides by ρ and by the right-hand side; we then get

$$\frac{dA'}{a'(1) + b'(1) \cdot A'} = \frac{d\rho}{\rho}.$$

Let us consider two possible cases: $b'(1) = 0$ and $b'(1) \neq 0$.

When $b'(1) = 0$, then integrating this equation leads to the following expression for the derivative A' of the desired function $A(\rho)$: $A' = a'(1) \cdot \ln(\rho) + \text{const}$. Now, the second integration leads to $A(\rho) = a'(1) \cdot \rho \cdot \ln(\rho) + c_1 \cdot \rho + c_2$. Since for the term $c_1 \cdot \rho$, the integral is always constant $\int (c_1 \cdot \rho(x)) \, dx = c_1 \cdot \int \rho(x) \, dx = c_1$, optimizing the expression $\int A(\rho(x)) \, dx$ is equivalent to optimizing the entropy $- \int \rho(x) \cdot \ln(\rho(x)) \, dx$.

When $b'(1) \neq 0$, then for $B \stackrel{\text{def}}{=} A' + \frac{a'(1)}{b'(1)}$, we get $\frac{dB}{b'(1) \cdot B} = \frac{d\rho}{\rho}$, so integration leads to $\ln(B) = b'(1) \cdot \ln(\rho) + \text{const}$ and $B = C \cdot \rho^\beta$ for $\beta \stackrel{\text{def}}{=} b'(1)$. Hence, $A'(\rho) = B - \text{const} = C \cdot \rho^\beta + \text{const}$. Integrating one more time, when $\beta \neq -1$, we get $A(\rho) = \text{const} \cdot \rho^{\beta+1} + c_1 \cdot \rho + c_2$. Similarly to the above case, optimizing the expression $\int A(\rho(x)) \, dx$ is equivalent to optimizing *generalized entropy* $\int (\rho(x))^\alpha \, dx$, for $\alpha = \beta + 1$.

When $\beta = -1$, integration leads to $A(\rho) = \mathrm{const} \cdot \ln(\rho) + c_1 \cdot \rho + c_2$, so optimizing $\int A(\rho(x)) \, dx$ is equivalent to optimizing another case of *generalized entropy* $\int \ln(\rho(x)) \, dx$.

Conclusion: which objective functions are invariant. There are exactly three shift- and scale-invariant objective functions: entropy $- \int \rho(x) \cdot \ln(\rho(x)) \, dx$ and generalized entropy $\int \ln(\rho(x)) \, dx$ and $\int (\rho(x))^{\alpha} \, dx$, for $\alpha \neq 1$.

Comment. From the purely mathematical viewpoint, this result is similar to a classification of invariant objective functions for selecting the best image [21].

4 Which Constraints Are Invariant?

Shift-invariance: formulation of the problem. We say that the constraints

$$\int f_i(x) \cdot \rho(x) \, dx = c_i$$

corresponding to the functions $f_i(x)$ $(1 \le i \le n)$ are *shift-invariant* if the values of the corresponding quantities $\int f_i(x) \cdot \rho(x) \, dx$ uniquely determine the values of these quantities for a shifted distribution.

To be more precise, after the shift, for the same quantity, the original numerical value x is replaced by the new value $y = x + x_0$. In the new scale, the probability density $\rho_y(y)$ has the form $\rho_y(y) = \rho(y - x_0)$. Thus, when we compute the constraint-related quantities based on shifted values, we get the integrals $\int f_i(y) \cdot \rho(y - x_0) \, dy$. Our requirements is that the values of all these new integrals should be uniquely determined based on the values of n original integrals.

Analysis of the problem. In the new integral, we can consider $x = y - x_0$ as the new variable; in this case, each new integral takes the form

$$\int f_i(x + x_0) \cdot \rho(x) \, dx.$$

Thus, our requirement is that if for every pdf, we know the values $\int f_i(x) \cdot \rho(x) \, dx$, then we can uniquely determine the values $f_i(x + x_0) \cdot \rho(x) \, dx$. So, for every small change $\delta\rho(x)$ for which:

- $\rho(x) + \delta\rho(x)$ remains a pdf, i.e., $\int \delta\rho(x) = 0$ and
- the values of the original integrals do not change, i.e., $\int f_i(x) \cdot \delta\rho(x) = 0$,

the values of the new integrals shall also remain unchanged, i.e., we should have $\int f_i(x + x_0) \cdot \delta\rho(x) \, dx = 0$.

In geometric terms, this means that any function $\delta\rho$ which is orthogonal to all the functions $1, f_1, \ldots, f_n$, should also be orthogonal to the function $f_i(x + x_0)$. Thus, similarly to the case of invariant objective functions, we conclude that the function

$f_i(x+x_0)$ should belong to the linear space spanned by 1 and f_i, i.e., that we should have

$$f_i(x+x_0) = c_i(x_0) + \sum_{j=1}^{n} c_{ij}(x_0) \cdot f_j(x) \qquad (3)$$

for some constants $c_i(x_0)$ and $c_{ij}(x_0)$.

It is reasonable to consider the case of differentiable functions $f_i(x)$. In this case, by selecting sufficiently many values x_k, we can get a system of linear equations from which we can uniquely determine $c_i(x_0)$ and $c_{ij}(x_0)$:

$$f_i(x_k+x_0) = c_i(x_0) + \sum_{j=1}^{n} c_{ij}(x_0) \cdot f_j(x_k).$$

By the Cramer's rule [35], a solution to a system of linear equations is a differentiable function of its parameters $f_i(x_k+x_0)$. Since the functions f_i are differentiable, we conclude that the functions $c_i(x_0)$ and $c_{ij}(x_0)$ are differentiable as well – as compositions of differentiable functions.

Differentiating both sides of the formula (3) with respect to x_0 and taking $x_0 = 0$, we conclude that

$$\frac{df_i}{dx} = c_i'(0) + \sum_{j=1}^{n} c_{ij}'(0) \cdot f_j(x).$$

Thus, the functions $f_i(x)$ together with a constant function 1 are solutions to a system of linear differential equations with constant coefficients. Solutions to such systems are known (see, e.g., [38]): they are linear combinations of functions of the type $x^k \cdot \exp(a \cdot x) \cdot \sin(\omega \cdot x + \varphi)$, where $k \geq 0$ is a natural number and $a + \omega \cdot \mathrm{i}$ is an eigenvalue of the corresponding matrix. So, we arrive at the following conclusion:

Conclusion: shift-invariant constraints. The functions $f_i(x)$ corresponding to shift-invariant constraints are linear combinations of the functions of the type $x^k \cdot \exp(a \cdot x) \cdot \sin(\omega \cdot x + \varphi)$.

Scale-invariance: formulation of the problem. We say that the constraints

$$\int f_i(x) \cdot \rho(x) \, dx = c_i$$

corresponding to the functions $f_i(x)$ ($1 \leq i \leq n$) are *scale-invariant* if the values of the corresponding quantities $\int f_i(x) \cdot \rho(x) \, dx$ uniquely determine the values of these quantities for a shifted distribution.

After the shift, for the same quantity, the original numerical value x is replaced by the new value $y = \lambda \cdot x$. In the new scale, the probability density $\rho_y(y)$ has the form $\rho_y(y) = \lambda^{-1} \cdot \rho(\lambda^{-1} \cdot y)$. Thus, when we compute the constraint-related quantities based on shifted values, we get the integrals $\int \lambda^{-1} \cdot f_i(y) \cdot \rho(\lambda^{-1} \cdot y) \, dy$. Our requirement is that the values of all these new integrals should be uniquely determined based on the values of n original integrals.

Analysis of the problem. In the new integral, we can consider $x = \lambda^{-1} \cdot y$ as the new variable; in this case, each new integral takes the form

$$\int f_i(\lambda \cdot x) \cdot \rho(x) \, dx.$$

Thus, our requirement is that if for every pdf, we know the values $\int f_i(x) \cdot \rho(x) \, dx$, then we can uniquely determine the values $f_i(\lambda \cdot x) \cdot \rho(x) \, dx$. Similar to the shift-invariant case, we can conclude that

$$f_i(\lambda \cdot x) = c_i(\lambda) + \sum_{j=1}^{n} c_{ij}(\lambda) \cdot f_j(x)$$

for some differentiable function $c_i(\lambda)$ and $c_{ij}(\lambda)$. Differentiating both sides of this equality relative to λ and taking $\lambda = 1$, we conclude that

$$x \cdot \frac{d f_i(x)}{dx} = c_i'(1) + \sum_{j=1}^{n} c_{ij}'(1) \cdot f_j(x).$$

Here, $\dfrac{dx}{x} = dz$ for $z = \ln(x)$. Thus, if we express all the functions $f_i(x)$ in terms of z, i.e., consider $f_i(x) = F_i(\ln(x))$, with $F_i(z) \overset{\text{def}}{=} f_i(\exp(z))$, then for the new functions $F_i(z)$, we get a system of linear differential equations with constant coefficients:

$$\frac{dF_i(z)}{dz} = c_i'(1) + \sum_{j=1}^{n} c_{ij}'(1) \cdot F_j(z).$$

We already know that each solution is a linear combination of functions of the type $z^k \cdot \exp(a \cdot z) \cdot \sin(\omega \cdot z + \varphi)$. Substituting $z = \ln(x)$ into this formula, we conclude that each function $f_i(x)$ is a linear combination of functions the type

$$(\ln(x))^k \cdot x^a \cdot \sin(\omega \cdot \ln(x) + \varphi).$$

Comment. Note that scaling only related values of the same sign, so we may have two different expressions for $x < 0$ and for $x > 0$.

If instead of scaling relative to 0, we have scaling relative to some other value x_0, i.e., transformations $x - x_0 \to \lambda \cdot (x - x_0)$, then we get expressions

$$(\ln(x - x_0))^k \cdot (x - x_0)^a \cdot \sin(\omega \cdot \ln(x - x_0) + \varphi).$$

In this case, we may have different expressions for $x \leq x_0$ and for $x \geq x_0$.

Conclusion: scale-invariant constraints. Each function $f_i(x)$ corresponding to scale-invariant constraints is a linear combinations of functions

$$(\ln(x - x_0))^k \cdot (x - x_0)^a \cdot \sin(\omega \cdot \ln(x - x_0) + \varphi),$$

where k is a natural number. Note that we may have different expressions for $x < x_0$ and for $x > x_0$.

Which constraints are both shift- and scale-invariant? To answer this question, let us check when shift-invariant constraints are also scale-invariant. We cannot have $a \neq 0$, since then the corresponding function grows too fast for scale constraints. We similarly cannot have $b \neq 0$, so the only remaining terms are monomials x^k. Thus, each function corresponding to shift- and scale-invariant constraints is a linear combination of monomials, i.e., a polynomial:

Conclusion: shift- and scale-invariant constraints. Each function $f_i(x)$ corresponding to shift- and scale-invariant constraints is a polynomial.

5 Invariant Objective Functions and Constraints: Summary

Let us summarize our results.

Symmetry-based criteria. There are three possible symmetry-based criteria: entropy $-\int \rho(x) \cdot \ln(\rho(x)) \, dx$ and generalized entropy $\int \ln(\rho(x)) \, dx$ and $\int (\rho(x))^\alpha \, dx$, for $\alpha \neq 1$.

Constraints which are both shift- and scale-invariant. The only such constraints correspond to polynomials $P(x)$.

Shift-invariant constraints. A function $f(x)$ corresponding to each such constraint is a linear combination of functions of the type $x^k \cdot \exp(a \cdot x) \cdot \sin(\omega \cdot x + \varphi)$, where $k \geq 0$ is a natural number.

Scale-invariant constraints. For scaling around a point x_0, a function $f(x)$ corresponding to each such constraint is a linear combinations of functions

$$(\ln(x - x_0))^k \cdot (x - x_0)^a \cdot \sin(\omega \cdot \ln(x - x_0) + \varphi),$$

where k is a natural number. Note that we may have different expressions for $x < x_0$ and for $x > x_0$.

In particular, we can have scale-invariant constraint $\int f(x) \cdot \rho(x) = 0$ with $f(x) = 1$ for $x \geq x_0$ and $f(x) = 0$ for $x < x_0$. Since $\rho(x) \geq 0$, this constraint implies that $\rho(x) = 0$ for all $x \geq x_0$. Similarly, we can have a constraint implying that $\rho(x) = 0$ for all $x \leq x_0$.

By combining two such constraints, we get a restriction of a distribution to an interval.

Optimizing an invariant objective function under invariant constraints: general formulas. In general, we optimize an invariant objective function $J(\rho)$ under the constraints $\int \rho(x) \, dx = 1$ and $\int f_i(x) \cdot \rho(x) \, dx = c_i$ for several invariant constraints $f_i(x)$. For this constraint optimization problem, the Lagrange multiplier methods results in an unconditional optimization of a functional

$$J(\rho) + \lambda \cdot \left(\int \rho(x)\,dx - 1 \right) + \sum_i \left(\int f_i(x) \cdot \rho(x)\,dx - c \right),$$

where λ and λ_i are the corresponding Lagrange multipliers. Differentiating this expression with respect to $\rho(x)$ and equating the resulting derivative to 0, we get the following equations:

$$\ln(\rho(x)) = -1 + \lambda + \sum_i \lambda_i \cdot f_i(x) \text{ for the usual entropy;}$$

$$-(\rho(x))^{-1} = \lambda + \sum_i \lambda_i \cdot f_i(x) \text{ for } J(\rho) = \int \ln(\rho(x))\,dx; \text{ and}$$

$$(-\alpha) \cdot (\rho(x))^{\alpha-1} = \lambda + \sum_i \lambda_i \cdot f_i(x) \text{ for } J(\rho) = \int (\rho(x))^{\alpha}\,dx.$$

6 Resulting Distributions

Let us now list effective families of distributions which result from our approach, i.e., which are optimal with respect to some symmetry-based criterion under some invariant constraints. In our listing:

- we start with the case when all constraints are both shift- and scale-invariant,
- then we consider the case when all constraints are scale-invariant, with respect to the same value x_0,
- then we consider the case when all constraints are shift-invariant,
- finally, we will consider the case when different constraints are invariant relative to different transformations.

For each of these cases, we consider first situations when the objective function is the usual entropy, and then situations when a generalized entropy is used.

Comments.

- Some distributions have several different symmetry-based justifications. For example, among all distributions located on an interval, the uniform distribution has the largest possible entropy and also the largest possible generalized entropy.
- For most distributions, we mention one or two practical situations in which this particular distribution has been effectively used. Of course, our list of examples does not exhaust all efficient applications of the distribution: for many distributions, the number of practical applications can fill a book (and such books have been published; see, e.g., [22]).
- While our symmetry-based approach explains many empirically successful probability distributions, not all such distributions can be thus described. For example, many infinitely divisible distributions do not have an analytical representations and thus, cannot be represented in this form. It should be mentioned that this

omissions is not so bad, since most of such distributions already have a theoretical explanation for their success.

6.1 All Constraints Are Both Shift- and Scale-Invariant, Objective Function is Entropy

Towards a general formula. In this case, we optimize entropy $-\int \rho(x) \cdot \ln(\rho(x)) \, dx$ under constraints $\int \rho(x) \, dx$ and constraints of the type $\int P_i(x) \cdot \rho(x) \, dx = c_i$ for some polynomials $P_i(x)$. For this constraint optimization problem, the Lagrange multiplier method leads to optimizing the expression

$$-\int \rho(x) \cdot \ln(\rho(x)) \, dx + \lambda \cdot \left(\int \rho(x) \, dx - 1 \right) + \sum_i \lambda_i \cdot \left(\int P_i(x) \cdot \rho(x) \, dx - c_i \right),$$

where λ and λ_i are Lagrange multipliers.

Differentiating this expression with respect to $\rho(x)$ and equating the derivative to 0, we conclude that

$$-\ln(\rho(x)) - 1 + \lambda + \sum_i \lambda_i \cdot P_i(x) = 0,$$

hence

$$\ln(\rho(x)) = -1 + \lambda + \sum_i \lambda_i \cdot P_i(x).$$

The right-hand side of this formula is a linear combination of polynomials and is, thus, also a polynomial. We will denote this polynomial by $P(x)$. From $\ln(\rho(x)) = P(x)$, we conclude that the corresponding pdf has the form $\rho(x) = \exp(P(x))$ for some polynomial $P(x)$. This is a general formula for such optimal probability distributions.

Example of a successful probability distribution of this type. The most widely used distribution, the normal distribution with probability density

$$\rho(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{\sigma^2} \right),$$

is exactly of this type. It is a well-known fact that of all the distributions with given mean and given variance, the normal distribution has the largest possible entropy.

6.2 All Constraints Are Both Shift- and Scale-Invariant, Objective Function is Generalized Entropy

Towards a general formula. In this case, we optimize generalized entropy $\int \ln(\rho(x))\,dx$ or $(\rho(x))^{\alpha}\,dx$ under constraints $\int \rho(x)\,dx$ and constraints of the type $\int P_i(x) \cdot \rho(x)\,dx = c_i$ for some polynomials $P_i(x)$. For this constraint optimization problem, we optimize

$$\int \ln(\rho(x))\,dx + \lambda \cdot \left(\int \rho(x)\,dx - 1 \right) + \sum_i \lambda_i \cdot \left(\int P_i(x) \cdot \rho(x)\,dx - c_i \right) \text{ or}$$

$$\int (\rho(x))^{\alpha}\,dx + \lambda \cdot \left(\int \rho(x)\,dx - 1 \right) + \sum_i \lambda_i \cdot \left(\int P_i(x) \cdot \rho(x)\,dx - c_i \right).$$

Differentiating this expression with respect to $\rho(x)$ and equating the derivative to 0, we conclude that

$$-\alpha \cdot (\rho(x))^{\alpha-1} = \lambda + \sum_i \lambda_i \cdot P_i(x) = 0,$$

hence

$$(\rho(x))^{\alpha-1} = -\frac{1}{\alpha} \cdot \left(\lambda + \sum_i \lambda_i \cdot P_i(x) \right).$$

The right-hand side $P(x)$ of this formula is a polynomial, so we get the following general formula for such optimal probability distributions: $\rho(x) = (P(x))^{\beta}$, where $\beta = \dfrac{1}{\alpha - 1}$.

Example of a successful probability distribution of this type. An example of such a distribution is *Cauchy distribution*, with probability density

$$\rho(x) = \frac{\Delta}{\pi} \cdot \frac{1}{1 + \dfrac{(x-\mu)^2}{\Delta^2}}.$$

This distribution is actively used in physics, to describe resonance energy distribution and the corresponding widening of spectral lines; see, e.g. [19]. It is also used to estimate the uncertainty of the results of data processing [23].

Comment. It should be mentioned that while formally, we get the Cauchy distribution, the above derivation is not fully straightforward: it includes a constraint of the type $\int x^2 \cdot \rho(x)\,dx = $ const, but for the Cauchy distribution, the corresponding integral is infinite. So, to make the above derivation mathematically correct, we should first consider the problem limited to distributions located on an interval $[-T, T]$ and then tend T to infinity.

6.3 All Constraints Are Scale-Invariant Relative to the Same Value x_0, Objective Function is Entropy

Half-normal distribution. By combining the constrains on mean and second moment with the constraint implying that $\rho(x) = 0$ for $x \leq 0$, we get a *half-normal* distribution, i.e., a distribution whose pdf for $x \leq 0$ is 0, and for $x > 0$ is twice that of the normal distribution.

Generalized Gamma distribution. For $f_1(x) = \ln(x)$, $f_2(x) = x^\alpha$, and for a scale-invariant constraint corresponding to $x \geq 0$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x) + \lambda_2 \cdot x^\alpha$, i.e., to the *Generalized Gamma distribution*

$$\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot \exp(\lambda_2 \cdot x^\alpha)$$

which is efficiently used in survival analysis in social sciences [4].

Several probability distributions are particular cases of this general formula. Let us list some of them in alphabetic order.

Particular cases of Generalized Gamma: chi-square distribution. When λ_1 is a natural number and $\alpha = 2$, we get the *chi-square distribution* used to check how well the model fits the data. Under the name of *Nakagami distribution*, this distribution is also used to model attenuation of wireless signals traversing multiple paths [34].

Particular cases of Generalized Gamma: inverse Gamma distribution. When $\alpha = -1$, then we get the *inverse Gamma distribution* which is often used as a prior distribution in Bayesian analysis [3, 14], e.g., to describe the prior distribution of the variance. In particular, when $2\lambda_1$ is a negative integer, we get the *scaled-inverse chi-square distribution*, and for specific values of α_2, we get the *inverse chi-square distribution*.

Particular cases of Generalized Gamma: exponential distribution. When $\lambda_1 = 0$ and $\alpha = 1$, we get the *exponential distribution* $\rho(x) = \text{const} \cdot \exp(-k \cdot x)$. This distribution describe the time between consecutive events, e.g., in queuing theory, in radioactive decay, etc. (It should be noted that the exponential distribution can also be obtained by using generalized entropy.)

Particular cases of Generalized Gamma: Gamma distribution. When $\alpha = 1$, we get the *Gamma distribution* which is often used as a prior distribution in Bayesian analysis. In particular, when $\lambda_1 = k$ is a natural number, we get the *Erlang distribution* that describe the time during which k consecutive events occur [6].

Particular cases of Generalized Gamma: Fréchet distribution. When $\lambda_1 = 0$, we get the *Fréchet distribution* which describes the frequency of extreme events, such as the yearly maximum and minimum stock prices in economics [10] and yearly maximum rainfalls in hydrology [5].

Particular cases of Generalized Gamma: half-normal distribution. When $\lambda_1 = 0$ and $\alpha = 2$, we get the above-described *half-normal distribution*.

Particular cases of Generalized Gamma: inverse Gamma distribution. When $\alpha = -1$, we get the *inverse Gamma distribution* which is used to describe a prior distribution for variance in Bayesian analysis.

Particular cases of Generalized Gamma: Rayleigh distribution. When $\lambda_1 = 1$ and $\alpha = 2$, we get the *Rayleigh distribution* which is used to describe the length of random vectors – e.g., the distribution of wind speed in meteorology.

Particular cases of Generalized Gamma: type-2 Gumbel (Weibull) distribution. When $\lambda_1 = \alpha - 1$, we get the *type-2 Gumbel (Weibull) distribution* which is used to describe the frequency of extreme events and time to failure.

Further generalization of Generalized Gamma: 3-parametric Gamma distribution. For $f_1(x) = \ln(x - \mu)$, $f_2(x) = (x - \mu)^{\alpha}$, and for a scale-invariant constraint corresponding to $x \geq \mu$, optimization leads to

$$\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x - \mu) + \lambda_2 \cdot (x - \mu)^{\alpha},$$

i.e., to the *3-parametric Gamma distribution*

$$\rho(x) = \text{const} \cdot (x - \mu)^{\lambda_1} \cdot \exp(\lambda_2 \cdot (x - \mu)^{\alpha})$$

which is efficiently used in hydrology [41, 42].

Inverse Gaussian (Wald) distribution. For scale-invariant constraints $f_1(x) = \ln(x)$, $f_2(x) = x$, $f_3(x) = x^{-1}$, and for a scale-invariant constraint leading to $x > 0$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x) + \lambda_2 \cdot x + \lambda_4 \cdot x^{-1}$, i.e., to $\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot \exp(\lambda_2 \cdot x + \lambda_3 \cdot x^{-1})$. In particular, for $\lambda_1 = -1.5$, we get the *inverse Gaussian (Wald) distribution*. This distribution describes the time a Brownian Motion with positive drift takes to reach a fixed positive level.

Laplace distribution. For a scale-invariant constraint $f_1(x) = |x - \mu|$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot |x - \mu|$, so we get *Laplace distribution*

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot |x - \mu|).$$

This distribution has many applications [22]. For example, it is used:

- in *speech recognition*, as a prior distribution for the Fourier coefficients [9];
- in *databases*, where, to preserve privacy, each record is modified by adding a Laplace-generated noise [8].

Lévy (van der Waals) distribution. For scale-invariant constraints

$$f_1(x) = \ln(x - \mu),$$

$f_2(x) = (x - \mu)^{-1}$, and for a scale-invariant constraint equivalent to $x - \mu > 0$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x - \mu) + \lambda_2 \cdot (x - \mu)^{-1}$, i.e., to $\rho(x) = \text{const} \cdot (x - \mu)^{\lambda_1} \cdot \exp(\lambda_2 \cdot (x - \mu)^{-1})$. In particular, for $\lambda_1 = -1.5$, we get the *Lévy*

(van der Waals) distribution. This distribution is used in spectroscopy, to describe different spectra [7].

Log-normal distribution. For scale-invariant constraints $f_1(x) = \ln(x)$ and $f_2(x) = (\ln(x))^2$ and for a scale-invariant constraint equivalent to $x > 0$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x) + \lambda_2 \cdot (\ln(x))^2$, i.e., to the *log-normal distribution* $\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot \exp(\lambda_2 \cdot (\ln(x))^2)$. This distribution describes the product of several independent random factors. It has many applications. In particular, it is used in econometrics to describe:

- the compound return of a sequence of multiple trades,
- a long-term discount factor, etc.

6.4 All Constraints Are Shift-Invariant, Objective Function Is Entropy

Gumbel distribution. For a shift-invariant constraint $f_1(x) = \exp(k \cdot x)$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \exp(k \cdot x)$, i.e., to the *Gumbel distribution* $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot \exp(k \cdot x))$ which is used to describe the frequency of extreme events.

Type I Gumbel distribution. For shift-invariant constraints $f_1(x) = x$ and $f_2(x) = \exp(k \cdot x)$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot x + \lambda_2 \cdot \exp(k \cdot x)$ and thus, to $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x + \lambda_2 \cdot \exp(k \cdot x))$. In particular, for $\lambda_1 = k$, we get *type I Gumbel distribution* which is used to decrease frequencies of extreme values.

6.5 All Constraints Are Shift-Invariant, Objective Function Is Generalized Entropy

Hyperbolic secant distribution. When we use the objective function $\int \ln(\rho(x)) \, dx$ and a shift-invariant constraint $f_1(x) = \exp(k \cdot x) + \exp(-k \cdot x)$, optimization leads to $(\rho(x))^{-1} = -\lambda - \lambda_1 \cdot (\exp(k \cdot x) + \exp(-k \cdot x)) = -\lambda + c \cdot \cosh(k \cdot x)$. Thus, we get $\rho(x) = \text{const} \cdot \dfrac{1}{-\lambda + c \cdot \cosh(k \cdot x)}$. The requirement that $\int \rho(x) \, dx = 1$ leads to $\lambda = 0$, so we get a *hyperbolic secant distribution*. This distribution is similar to the normal one, but it has a more acute peak and heavier tails, so it is used when we have a distribution which is close to normal but has heavier tails.

6.6 Different Constraints Have Different Symmetries, Objective Function Is Entropy

In some cases, to get the desired distribution, we need to combine constraints with symmetries of different type. Let us give examples of the resulting distributions.

Uniform distribution. If we impose constraints leading to $x \geq a$ and $x \leq b$, then the largest values of the entropy is attained on the uniform distribution on the interval $[a, b]$.

Comment. It should be noted that the same result holds if we use generalized entropy.

Beta and arcsine distribution. Constraints $\int \ln(x) \cdot \rho(x) \, dx$ and $\int \ln(a - x) \cdot \rho(x) \, dx$ are both scale-invariant, but the first one is scale-invariant relative to $x_0 = 0$, while the second one is scale-invariant relative to $x_0 = a$. Optimizing entropy under these constraints – and under similarly scale-invariant constraints implying that $0 \leq x \leq a$, results in $\rho(x) = A \cdot x^{\alpha} \cdot (a - x)^{\beta}$ for some A, α, and β.

This formula describes a *Beta distribution* on the interval $[0, a]$. This distribution has numerous practical applications in many areas including agriculture [18], epidemiology [43], geosciences [16], meteorology [39], population genetics [1], and project management [28].

In particular, for $a = 1$ and $\alpha = \beta = 0.5$, we get the *arcsine distribtion*, with probability density $\rho(x) = \dfrac{1}{\pi \cdot \sqrt{x \cdot (1 - x)}}$. This distribution describes, for example, the measurement error caused by an external sinusoidal signal coming at a random moment of time [37].

Beta prime (F-) distribution. For scale-invariant constraints $f_1(x) = \ln(x)$ and $f_2(x) = \ln(x + a)$ and for a constraint leading to $x > 0$, optimizing entropy leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x) + \lambda_2 \cdot \ln(x + a)$, i.e., to the *Beta prime (F-) distribution* $\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot (x + a)^{\lambda_2}$.

Log distribution. Let us impose scale-invariant constraints $f_1(x) = x$ and $f_2(x) = \ln(x)$, and constraints leading to $x \geq a$ and $x \leq b$. Then the largest entropy occurs when for $x \in [a, b]$, we have $\ln(\rho(x)) = \lambda + \lambda_1 \cdot x + \lambda_2 \cdot \ln(x)$, hence $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x) \cdot x^{\lambda_2}$. For $\lambda_1 = -1$, we get the *log distribution*.

Generalized Pareto distribution. For scale-invariant constraint $f_1(x) = \ln(x + x_0)$ and a scale-invariant constraint leading to $x > x_m$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x + x_0)$, hence to the *Generalized Pareto distribution*

$$\rho(x) = \text{const} \cdot (x + x_0)^{\lambda_1}.$$

This distribution describes the frequency of large deviations in economics, in geophysics, and in other applications areas [10]. The case $x_0 = 0$ is known as the *Pareto distribution*.

Comment. The Generalized Pareto distribution can also be derived by using generalized entropy.

Gompertz distribution. For shift-invariant constraints $f_1(x) = \exp(b \cdot x)$ and $f_2(x) = x$, and a scale-invariant constraint leading to $x > 0$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot x + \lambda_2 \cdot \exp(b \cdot x)$, hence to *Gompertz distribution*

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x) \cdot \exp(\lambda_2 \cdot \exp(b \cdot x)).$$

This distribution describes aging and life expectancy [2, 36]; it is also used to in software engineering, to describe the "life expectancy" of software [33].

Reciprocal and U-quadratic distribution. For a scale-invariant constraint $f_1(x) = \ln(x - \beta)$ and scale-invariant constraints corresponding to $x \geq a$ and $a \leq b$, optimization leads to $\rho(x) = A \cdot x^{\alpha}$ for $x \in [a, b]$. In particular:

- for $\alpha = -1$ and $\beta = 0$, we get the *reciprocal distribution* $\rho(x) = \text{const} \cdot x^{-1}$. This distribution is used in *computer arithmetic*, to describe the frequency with which different numbers occur [17, 32];
- for $\alpha = 2$, we get the *U-quadratic distribution* $\rho(x) = \text{const} \cdot (x - \beta)^2$; this distribution is often effectively used to describe random quantities with a bimodal distribution.

Comment. Please note that, as we show later in this paper, both distributions can also be obtained by using generalized entropy.

Truncated normal distribution. By combining the constrains on mean and second moment with the constraint implying that $\rho(x) = 0$ for $x \leq a$ and for $x \geq b$, we get the *truncated normal distribution*, i.e., a normal distribution limited to the interval $[a, b]$. This distribution is actively used in econometrics, to model quantities about which we only know lower and upper bounds [15].

von Mises distribution. For a shift-invariant constraint $f_1(x) = \cos(x - \mu)$ and for scale-invariant criteria corresponding to $x \geq -\pi$ and $x \leq \pi$, optimization leads to $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \cos(x - \mu)$, i.e., to the *von Mises distribution*

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot \cos(x - \mu))$$

which is frequently used to describe random angles $x \in [-\pi, \pi]$.

6.7 Different Constraints Have Different Symmetries, Objective Function is Generalized Entropy

Uniform distribution. If we impose constraints leading that $x \geq a$ and $x \leq b$, then the largest values of the generalized entropy is attained on the uniform distribution on the interval $[a, b]$.

Comment. It should be noted that the same result holds if we use the usual entropy.

Exponential and Erlang distribution. For the objective function $\int (\rho(x))^2 \, dx$, for the shift-invariant constraint $f_1(x) = x^k \cdot \exp(-a \cdot x)$, and a scale-invariant constraint corresponding to $x > 0$, optimization leads to the *Erlang distribution*

$$\rho(x) = \text{const} \cdot x^k \cdot \exp(-a \cdot x);$$

in particular, for $k = 0$, we get an *exponential distribution* $\rho(x) = \text{const} \cdot \exp(-a \cdot x)$.

Comment. Exponential distribution can also be obtained by using the usual entropy.

Generalized Pareto distribution. For the objective function $\int (\rho(x))^\alpha \, dx$ and for the scale- and shift-invariant constraint $f_1(x) = x$ and a scale-invariant constraint leading to $x > 0$, optimization leads to $-\alpha \cdot (\rho(x))^{\alpha-1} = \lambda + \lambda_1 \cdot x$, hence to the *Generalized Pareto distribution* $\rho(x) = \text{const} \cdot (x + x_0)^{-\gamma}$, where $\gamma = -1/(\alpha - 1)$.

Comment. The Generalized Pareto distribution can also be derived by using the usual entropy.

Raised cosine distribution. For the objective function $\int (\rho(x))^2 \, dx$, for a shift-invariant constraint $f_1(x) = \cos(\omega \cdot x + \varphi)$ and scale-invariant constraints corresponding to $x \geq a$ and $x \leq b$, optimization leads to the *raised cosine distribution* $\rho(x) = c_1 + c_2 \cdot \cos(\omega \cdot x + \varphi)$.

Reciprocal distribution. For the generalized entropy $\int \ln(\rho(x)) \, dx$, a scale- and shift-invariant constraint $f_1(x) = x$ and scale-invariant constraints corresponding to $x \geq a$ and $a \leq b$, optimization leads to the reciprocal distribution $\rho(x) = A \cdot x^{-1}$ for $x \in [a, b]$.

Comment. Please note that this distribution can also be obtained by using the usual entropy.

U-quadratic distribution. For the generalized entropy $\int (\rho(x))^2 \, dx$, a scale- and shift-invariant constraint $f_1(x) = (x - \beta)^2$ and scale-invariant constraints corresponding to $x \geq a$ and $a \leq b$, optimization leads to the U-quadratic distribution $\rho(x) = \text{const} \cdot (x - \beta)^2$ for $x \in [a, b]$.

Comment. Please note that this distribution can also be obtained by using the usual entropy.

7 Conclusion

In the previous section, we listed numerous families of distributions which are optimal if we optimize symmetry-based utility functions under symmetry-based constraints. One can see that this list includes many empirically successful families of

distributions – and that most empirically successful families of distributions are on this list. Thus, we indeed provide a symmetry-based explanation for the empirical success of these families.

Acknowledgments

References

1. D. J. Balding and R. A. Nichols, A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, *Genetica*, 1995, Vol. 96, No. 1–2, pp. 3–12.
2. B. Benjamin, H. W. Haycocks, and J. Pollard, *The Analysis of Mortality and Other Actuarial Statistics*, Heinemann, London, 1980.
3. J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, New York, 1993.
4. J. M. Box-Steffensmeier and B. S. Jones, *Event History Modeling: A Guide for Social Scientists*, Cambridge University Press, New York, 2004.
5. S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, Berlin, Heidelberg, New York, 2001.
6. D. R. Cox, *Renewal Theory*, Wiley, New York, 1967.
7. C. A. Croxton, *Statistical Mechanics of the Liquid Surface*, Wiley, New York, 1980.
8. C. Dwork, F. McSherry, K. Nissim, and A. Smith, Calibrating noise to sensitivity in private data analysis, In: *Proceedings of the Theory of Cryptography Conference TCC'2006*, Springer Verlag, Berlin, Heidelberg, New York, 2006.
9. T. Eltoft, K. Taesu, and T.-W. Lee, On the multivariate Laplace distribution, *IEEE Signal Processing Letters*, 2006, Vol. 13, No. 5, pp. 300–303.
10. P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer Verlag, Berlin, Heidelberg, New York, 2013.
11. R. P. Feynman, R. B. Leighton, and M. Sands, *Feynman Lectures on Physics*, Basic Books, New York, 2011.
12. P. C. Fishburn, *Utility Theory for Decision Making*, Wiley, New York, 1969.
13. P. C. Fishburn, *Nonlinear Preference and Utility Theory*, The John Hopkins Press, Baltimore, Maryland, 1988.
14. A. Gelman, J. B. Carlin, H. S. Stern, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, Boca Raton, Florida, 2013.
15. W. H. Greene, *Econometric Analysis*, Prentice Hall, Upper Saddle River, New Jersey, 2011.
16. R. S. Gullco and M. Anderson, Use of the Beta distribution to determine well-log shale parameters, *SPE Reservoir Evaluation & Engineering*, 2009, Vol. 12, No. 6, pp. 929–942.
17. R. W. Hamming, On the distribution of numbers, *The Bell System Technical Journal*, 1970, Vol. 49, No. 8, pp. 1609–1625.
18. J. D. Haskett, Y. A. Pachepsky, and B. Acock, Use of the beta distribution for parameterizing variability of soil properties at the regional level for crop yield estimation, *Agricultural Systems*, 1995, Vol. 48. No. 1, pp. 73–86.

19. E. Hecht, *Optics*, Addison-Wesley, New York, 2001.

20. N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, Wiley, New York, Vol. 1, 1994; Vol. 2, 1995.

21. O. Kosheleva, Symmetry-group justification of maximum entropy method and generalized maximum entropy methods in image processing, In: G. J. Erickson, J. T. Rychert, and C. R. Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 1998, pp. 101–113.

22. S. Kotz, T. J. Kozubowski, and K. Podgórski, *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering and Finance*, Birkhauser, Boston, Massachusetts, 2001.

23. V. Kreinovich and S. Ferson, A new Cauchy-based black-box technique for uncertainty in risk analysis, *Reliability Engineering and Systems Safety*, 2004, Vol. 85, No. 1–3, pp. 267–279.

24. V. Kreinovich, H. T. Nguyen, and S. Sriboonchitta, Why Clayton and Gumbel copulas: a symmetry-based explanation, In: V.-N. Huynh, V. Kreinovich, S. Sriboonchitta, and K. Suriya (eds.), *Uncertainty Analysis in Econometrics, with Applications*, Springer Verlag, Berlin, Heidelberg, 2013, pp. 79–90.

25. D. Kurowicka and R. M. Cooke, *Uncertainty Analysis with High Dimensional Dependence Modelling*, Wiley, New York, 2006.

26. D. Kurowicka and H. Joe (Eds.), *Dependence Modeling: Vine Copula Handbook*, World Scientific, Singapore, 2010.

27. D. R. Luce and H. Raiffa, *Games and Decisions, Introduction and Critical Survey*, Wiley, New York, 1957.

28. D. G. Malcolm, J. H. Roseboom, C. E. Clark, and W. Fazar, Application of a technique for research and development program evaluation, *Operations Research*, 1958, Vol. 7, No. 5, pp. 646–669.

29. R. B. Myerson, *Game theory: Analysis of Conflict*, Harvard University Press, Cambridge, Massachusetts, 1991.

30. R. B. Nelsen, *An Introduction to Copulas*, Springer-Verlag, New York, 2007.

31. H. T. Nguyen and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.

32. H. T. Nguyen, V. Kreinovich, and L. Longpré, Dirty pages of logarithm tables, lifetime of the Universe, and (subjective) probabilities on finite and infinite intervals, *Reliable Computing*, 2004, Vol. 10, No. 2, pp. 83–106.

33. K. Ohishi, H. Okamura, and T. Dohi, Gompertz software reliability model: estimation algorithm and empirical validation. *Journal of Systems and Software*, 2009, Vol. 82, No. 3, pp. 535–543.

34. J. D. Parsons, *The Mobile Radio Propagation Channel*, Wiley, New York, 1992.

35. D. Poole, *Linear Algebra: A Modern Introduction*, Cengage Learning, Boston, Massachusetts, 2014.

36. S. H. Preston, P. Heuveline, and M. Guillot, *Demography: Measuring and Modeling Population Processes*, Blackwell, Oxford, UK, 2001.

37. S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, Berlin, Heidelberg, New York, 2005.

38. J. C. Robinson, *An Introduction to Ordinary Differential Equations*, Cambridge University Press, Cambridge, UK, 2004.

39. M. Y. Sulaiman, W. M. H. Oo, M. A. Wahab, and A. Zakaria, Application of beta distribution model to Malaysian sunshine data, *Renewable Energy*, 1999, Vol. 18, No. 4, pp. 573–579.

40. P. Suppes, D. M. Krantz, R. D. Luce, and A. Tversky, *Foundations of Measurement. Vol. II. Geometrical, Threshold, and Probabilistic Representations*, Academic Press, San Diego, California, 1989.

41. R. van Nooijen, T. Gubareva, A. Kolechkina, and B. Gartsman. Interval analysis and the search for local maxima of the log likelihood for the Pearson III distribution, *Geophysical Research Abstracts*, 2008, Vol. 10, p. EGU2008-A05006.

42. R. van Nooijen and A. Kolechkina, Two applications of interval analysis to parameter estimation in hydrology, In: M. Nehmeier (ed.), *Abstracts of the 16th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics SCAN'2014*, Würzburg, Germany, September 21–26, 2014, p. 161.

43. J. A. Wiley, S. J. Herschkorn, and N. S. Padian, Heterogeneity in the probability of HIV transmission per sexual contact: The case of male-to-female transmission in penile-vaginal intercourse, *Statistics in Medicine*, 1989, Vol. 8, No. 1, pp. 93–102.