Why Skew Normal: A Simple Pedagogical Explanation

José Guadalupe Flores Muñiz[1], Vyacheslav V. Kalashnikov[2,3], Nataliya Kalashnykova[1,4], Olga Kosheleva[5], and Vladik Kreinovich[6]
[1]Department of Physics and Mathematics
Universidad Autoónoma de Nuevo León
San Nicolás de los Garza, México
jose_guadalupe64@hotmail.com
nkalash2009@gmail.com
[2]Department of Systems and Industrial Engineering
Tecnológico de Monterrey ITESM, Campus Monterrey
Monterrey, Mexico, kalash@itesm.mx
[3]Department of Experimental Economics
Central Economics and Mathematics Institute (CEMI)
Moscow, Russian Federation
[4]Department of Computer Science
Sumy State University
Sumy, Ukraine
Departments of [5]Teacher Education and [6]Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA, olgak@utep.edu, vladik@utep.edu

Abstract

In many practical situations, we only know a few first moments of a random variable, and out of all probability distributions which are consistent with this information, we need to select one. When we know the first two moments, we can use the Maximum Entropy approach and get normal distribution. However, when we know the first three moments, the Maximum Entropy approach does not work. In such situations, a very efficient selection is a so-called skew normal distribution. However, it is not clear why this particular distribution should be selected. In this paper, we provide an explanation for this selection.

1 Formulation of the Problem

General problem: need to select a probability distribution under uncertainty. Most traditional statistical techniques assumes that we know the corresponding probability distribution – or at least that we know a finite-parametric family of distributions that contains the given distribution; see, e.g., [5].

However, often, the only information that we have about the probability distribution of a quantity X is its few first moments $M_k \stackrel{\text{def}}{=} E[X^k]$. In such a situation, there are many possible distributions consistent with this information. To apply the traditional statistical techniques to such situations, it is therefore necessary to select, out of all possible distributions, one single distribution (or a finite-parametric family of distributions).

Ideally, we should select the distribution which is either, in some sense, the most realistic for a given situation, and/or leads to the simplest data processing techniques.

The simplest case when we know the first two moments. In uncertain situations when we the only information that we have are the first two moments, then, out of all possible distributions with these two moments, it is reasonable to select the distribution that maximally preserves uncertainty – i.e., the one for which the entropy $S = -\int \rho(x) \cdot \ln(\rho(x))\,dx$ is the largest possible, where $\rho(x)$ is the probability density function; see, e.g., [3].

By applying the Lagrange multiplier method to the corresponding constraint optimization problem of maximizing S under the constraints $\int \rho(x)\,dx = 1$, $\int x \cdot \rho(x)\,dx = M_1$ and $\int x^2 \cdot \rho(x)\,dx = M_2$, we can reduce this problem to the unconstrained optimization problem of maximizing the expression

$$-\int \rho(x) \cdot \ln(\rho(x))\,dx + \lambda_0 \cdot \left(\int \rho(x)\,dx - 1\right) + \lambda_1 \cdot \left(\int x \cdot \rho(x)\,dx - M_1\right) + \lambda_2 \cdot \left(\int x^2 \cdot \rho(x)\,dx - M_2\right).$$

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

$$-\ln(\rho(x)) - 1 + \lambda_0 + \lambda_1 \cdot x + \lambda_2 \cdot x^2 = 0,$$

hence

$$\ln(\rho(x)) = (\lambda_0 - 1) + \lambda_1 \cdot x + \lambda_2 \cdot x^2,$$

and thus, $\rho(x) = \exp(-Q(x))$ for some quadratic expression $Q(x)$.

This is the well-known Gaussian (normal) distribution.

What if we also know the third moment? What if, in addition to the first two moments M_1 and M_2, we also know the third moment

$$M_3 = \int x^3 \cdot \rho(x)\,dx?$$

At first glance, it may seem that in this case, we can also select, out of all possible distributions with these three moments, the distribution with the largest possible value of the entropy. In this case, the corresponding constraint optimization problem of maximizing S under the constraints $\int \rho(x)\,dx = 1$, $\int x\cdot\rho(x)\,dx = M_1$, $\int x^2\cdot\rho(x)\,dx = M_2$, and $\int x^3\cdot\rho(x)\,dx = M_3$ can be reduced to the unconstrained optimization problem of maximizing the expression

$$-\int \rho(x)\cdot\ln(\rho(x))\,dx + \lambda_0\cdot\left(\int \rho(x)\,dx - 1\right) + \lambda_1\cdot\left(\int x\cdot\rho(x)\,dx - M_1\right) +$$

$$\lambda_2\cdot\left(\int x^2\cdot\rho(x)\,dx - M_2\right) + \lambda_3\cdot\left(\int x^3\cdot\rho(x)\,dx - M_3\right).$$

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

$$-\ln(\rho(x)) - 1 + \lambda_0 + \lambda_1\cdot x + \lambda_2\cdot x^2 + \lambda\cdot x^3 = 0,$$

hence

$$\ln(\rho(x)) = (\lambda_0 - 1) + \lambda_1\cdot x + \lambda_2\cdot x^2 + \lambda_3\cdot x^3,$$

and thus,

$$\rho(x) = \exp(C(x)),$$

where we denoted

$$C(x) \stackrel{\text{def}}{=} (\lambda_0 - 1) + \lambda_1\cdot x + \lambda_2\cdot x^2 + \lambda_3\cdot x^3.$$

The problem with this formula is that:

- when $\lambda_3 > 0$, we get $C(x) \to \infty$ when $x \to \infty$, thus

  $$\rho(x) = \exp(C(x)) \to +\infty$$

  and therefore, we cannot have $\int \rho(x)\,dx = 1$;
- similarly. when $\lambda_3 < 0$, we get $C(x) \to \infty$ when $x \to -\infty$, thus

  $$\rho(x) = \exp(C(x)) \to +\infty$$

  and therefore, we also cannot have $\int \rho(x)\,dx = 1$; see, e.g., [2].

So, the only possible case when we have $\int \rho(x)\,dx = 1$ is when $\lambda_3 = 0$. However, in this case, we simply get a normal distribution, and normal distributions are uniquely determined by the first two moments and thus, do not cover all possible combinations of three moments.

So what do we do? In the case of three moments, there is a widely used selection, called a *skew normal* distribution (see, e.g., [1, 4]), when we choose a distribution with the probability density function

$$\rho(x) = \frac{1}{2\omega}\cdot\phi\left(\frac{x-\eta}{\omega}\right)\cdot\Phi\left(\alpha\cdot\frac{x-\eta}{\omega}\right),$$

where:

- $\phi(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$ is the pdf of the standard Gaussian distribution, with mean 0 and standard deviation 1, and
- $\Phi(x)$ is the corresponding cumulative distribution function

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt.$$

Comment. For this distribution,

- the first moment M_1 is equal to $M_1 = \mu = \eta + \omega \cdot \delta \cdot \sqrt{\frac{2}{\pi}}$, where

$$\delta \stackrel{\text{def}}{=} \frac{\alpha}{\sqrt{1+\alpha^2}},$$

- the second central moment $\sigma^2 = E[(X-\mu)^2]$ is equal to

$$\sigma^2 = \omega^2 \cdot \left(1 - \frac{2\delta^2}{\pi}\right),$$

and

- the third central moment $m_3 = E[(X-\mu)^3]$ is equal to

$$m_3 = \frac{4-\pi}{2} \cdot \sigma^3 \cdot \frac{(\delta \cdot \sqrt{2/\pi})^3}{(1-2\delta^2/\pi)^{3/2}}.$$

Why? The skew normal distribution has many applications, but it is not clear why it is selected.

What we do in this paper. In this paper, we provide a pedagogical explanation for the skew normal distribution.

2 Why Skew Normal: Analysis of the Problem and the Resulting Selection

Meaning of probability density function: reminder. In the above formula, the skew normal distribution is described in terms of the probability density function. To see how we can explain the above formula, let us recall the meaning of the probability density function.

By definition, the probability density is equal to the limit

$$\rho(x) = \lim_{\overline{x} \to x, \underline{x} \to x} \frac{\text{Prob}(X \in [\underline{x}, \overline{x}])}{\overline{x} - \underline{x}}.$$

Limit means that when the width $\overline{x}-\underline{x}$ of the corresponding interval is small, we have

$$\rho(x) \approx \frac{\text{Prob}(X \in [\underline{x}, \overline{x}])}{\overline{x}-\underline{x}},$$

and the smaller the width, the more accurate this formula.

In particular, for small $\varepsilon > 0$, we have

$$\rho(x) \approx \frac{\text{Prob}(X \in [x-\varepsilon, x+\varepsilon])}{2\varepsilon},$$

i.e.,

$$\text{Prob}(X \in [x-\varepsilon, x+\varepsilon]) \approx \rho(x) \cdot 2\varepsilon.$$

Thus, if we interpret $X \in [x-\varepsilon, x+\varepsilon]$, or, equivalently, $|X-x| \le \varepsilon$ as "X and x are ε-equal" – and denote it by $X =_\varepsilon x$ – then

$$\text{Prob}(X =_\varepsilon x) \approx \rho(x) \cdot 2\varepsilon.$$

What does such ε-equality mean? In practice, all the values are only measured with some accuracy δ. Thus, even if two values x_1 and x_1 are absolutely equal, all we get is their δ-approximate value $\widetilde{x}_1$ and $\widetilde{x}_2$, for which $|\widetilde{x}_1 - x_1| \le \delta$ and $|\widetilde{x}_2 - x_2| \le \delta$ imply that $|\widetilde{x}_1 - \widetilde{x}_2| \le |\widetilde{x}_1 - x_1| + |x_2 - \widetilde{x}_2| \le 2\delta$. Vice versa, if $|\widetilde{x}_1 - \widetilde{x}_2| \le 2\delta$, then it is possible that the values $\widetilde{x}_1$ and $\widetilde{x}_2$ come from measuring the same value $x_1 = x_2$: namely, if we take $x_1 = x_2 = \dfrac{\widetilde{x}_1 + \widetilde{x}_2}{2}$, we get $|\widetilde{x}_1 - x_1| \le \delta$ and $|\widetilde{x}_2 - x_2| \le \delta$.

From this viewpoint, the ε-equality is the practically checkable version of equality. Thus, modulo a multiplicative factor, the probability density $\rho(x)$ is the probability that the random value X is practically equal to x.

Let us start with the normal distribution. Let us start with the case when we know the first two moments and thus, get a normal distribution, with probability density $\rho_0(x)$. For simplicity, we can consider the case when the mean of the normal distribution is 0.

We want asymmetry. Normal distribution with 0 mean is symmetric with respect to change of sign $x \to -x$. As a result, for the normal distribution, the third moment M_3 is 0. To cover possible non-zero values of M_3, we thus need to "add" asymmetry to the normal distribution.

What we mean by "adding". A natural interpretation of adding is that, instead of considering a simple condition $X = x$, we consider a modified condition "$X = x$ *and* ..."

How can we describe the probability of such a combined statement? We have no reason to believe that the newly added condition is positively or negatively correlated with the event $X = x$. Thus, it is reasonable to consider these events to be independent – this is, by the way, what the maximum entropy principle implies in such a situation; see, e.g., [3].

Examples of such "adding". One possibility is to add, to the original condition $X = x$, a somewhat modified condition $X = \alpha \cdot x$, for some constant α.

Interestingly, this addition does not change much. Indeed, the original probability density function – corresponding to $X = x$ – has the form $\text{const} \cdot \exp\left(-\dfrac{x^2}{\sigma^2}\right)$. Thus, the additional condition $X = \alpha \cdot x$ has the form $\text{const} \cdot \exp\left(-\dfrac{(\alpha \cdot x)^2}{\sigma^2}\right)$, and the product of these two probabilities has the form $\text{const} \cdot \exp\left(-\dfrac{(1+\alpha^2) \cdot x^2}{\sigma^2}\right)$, i.e., the form $\text{const} \cdot \exp\left(-\dfrac{x^2}{(\sigma')^2}\right)$, where $\sigma' \stackrel{\text{def}}{=} \dfrac{\sigma}{\sqrt{1+\alpha^2}}$. So, we still get a normal distribution.

What is a natural asymmetric version of equality. As we have mentioned, the probability density function of the normal distribution describes the probability of equality $X = x$ (or, as we have just learned, $X = \alpha \cdot x$). A natural way to get asymmetry is to consider a natural asymmetric version of equality: inequality $X \le \alpha \cdot x$.

The probability of this inequality is equal to the corresponding cumulative distribution function. So, if we interpret "adding" this additional condition as multiplying, we get the product of:

- the original probability (i.e., the probability density function) and
- the new additional probability – which is described by the cumulative distribution function.

So, we get exactly the above formula for the skew normal distribution. Thus, we have indeed explained this formula.

Acknowledgments

This work was supported by grant CB-2013-01-221676 from Mexico Consejo Nacional de Ciencia y Tecnología (CONACYT). It was also partly supported by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence).

This work was partly performed when José Guadalupe Flores Muñiz visited the University of Texas at El Paso.

References

[1] A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge, Massachusetts, 2013.

[2] T. Dumrongpokaphan and V. Kreinovich, "Why cannot we have a strongly consistent family of skew normal (and higher order) distributions", In:

V. Kreinovich, S. Sriboonchitta, and V. N. Huynh (eds.), *Robustness in Econometrics*, Springer Verlag, Cham, Switzerland, 2017, pp. 69–78.

[3] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[4] B. Li, D. Shi, and T. Wang, "Some applications of one-sided skew distributions", *International Journal Intelligent Technologies and Applied Statistics*, 2009. Vol. 2, No. 1.

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