A Natural Simple Model of Scientists' Strength Leads to Skew-Normal Distribution

Komsan Suriya[1], Tatcha Sudtasan[1], Tonghui Wang[2],
Octavio Lerma[3], and Vladik Kreinovich[3]
[1]Faculty of Economics, Chiang Mai University, Thailand,
suriyakomsan@gmail.com, miracle.zaza@gmail.com
[2]Department of Mathematical Sciences, New Mexico State University
Las Cruces, NM 80003, USA, twang@nmsu.edu
[3]Computational Science Program, University of Texas at El Paso
El Paso, TX 79968, USA, lolerma@episd.org, vladik@utep.edu

Abstract

In many practical situations, we have probability distributions which are close to normal but skewed. Several families of distributions were proposed to describe such phenomena. The most widely used is *skew-normal* distribution, whose probability density (pdf) is equal to the product of the pdf of a normal distribution and a cumulative distribution function (cdf) of another normal distribution. Out of other possible generalizations of normal distributions, the skew-normal ones were selected because of their computational efficiency, and not because they represent any real-life phenomena. Interestingly, it turns out that these distributions do represent a real-life phenomena: namely, in a natural simple model of scientists' strength, this strength is skew-normally distributed. We also describe what happens if we consider more complex models of scientists' strength.

1 Introduction

Normal distributions are ubiquitous. In practice, many quantities – ranging from the distribution of measurement errors [3] to the distribution of blood pressure in humans – are normally distributed. The probability density of a normal distribution has the form

$$f(x) = \frac{1}{\sigma} \cdot f_0\left(\frac{x-\mu}{\sigma}\right),$$

where μ is the mean, σ is the standard deviation, and

$$f_0(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right)$$

is the probability density function of the "standard" normal distribution, with mean 0 and standard deviation 1. Correspondingly, the cumulative distribution function (cdf) $F(x) \stackrel{\text{def}}{=} \text{Prob}(\xi \le x)$ of the normal distribution has the form

$$F(x) = F_0\left(\frac{x-\mu}{\sigma}\right),$$

where $F_0(x) \stackrel{\text{def}}{=} \int_{-\infty}^{x} f_0(t)\,dt$ is the cdf of the standard normal distribution.

The ubiquity of normal distribution can be explained by the fact that in many cases, the value of the quantity is caused by many independent factors, and the known Central Limit Theorems states, crudely speaking, that the distribution of the sum of large number small independent factors is close to normal; see, e.g., [4].

Some distributions are close to normal but skewed. Normal distribution is symmetric. In practice, we sometimes encounter distributions which are close to normal but "skewed" (asymmetric).

Skew-normal distributions. Several different families of distributions have been proposed to described skewed distributions which are close to normal. The most widely used distributions of this type are *skew-normal* distributions, whose probability density function has the form

$$f(x) = \text{const} \cdot f_0\left(\frac{x-\mu}{\sigma}\right) \cdot F_0\left(\frac{x-\mu'}{\sigma'}\right)$$

for some parameters μ, σ μ', and σ'. For $\sigma' = \infty$, we get the usual normal distribution; see, e.g., [1, 2, 6].

Why this form. Out of other possible approximations, skew-normal distributions were introduced because they are, for many problems, computationally efficient [1], they were not intended as exactly describing some special class of phenomena.

What we show in this paper. In this paper, we show that skew-normal distributions actually describe an important phenomena – namely, that a natural simple model of scientists' strength leads to this distribution.

We also show that a more detailed description of scientists' strength leads to a natural generalization of skew-normal distributions.

Comment. Some preliminary results were described in [5].

2 A Natural Simple Model of Scientists' Strength

Toward a simple model: idea. To become a professional scientist, one has to defend his/her PhD. Not all students who start their PhD studies end up with a dissertation: some students succeed, but many don't. Crudely speaking, a student succeeds if his/her strength is sufficient to solve the corresponding problem, i.e., in other words, if his/her strength is larger than or equal to the complexity of the selected problem.

Transforming the above idea into a precise model. It is reasonable to assume that the strength x of students entering a PhD program is normally distributed, with some mean μ and standard deviation σ. (The strength is caused by many different factor, so it is reasonable to apply the Central Limit Theorem.) It is similarly reasonable to assume that the complexity y of a problem is normally distributed, with some mean μ' and some standard deviation σ'.

Because of the above assumptions, the number of students of strength x who enter the PhD program is proportional to $f_0\left(\frac{x-\mu}{\sigma}\right)$. It is also reasonable to assume that a student picks a problem at random. Thus, out of the incoming students of strength x, the proportion of those who succeed is equal to the probability $\text{Prob}_y(y \le x)$ that the randomly selected problem has complexity $\le x$ – i.e., to the value $F(x)$ of the corresponding cdf. Since the complexities are normally distributed, this probability is equal to $F_0\left(\frac{x-\mu'}{\sigma'}\right)$.

The resulting number of scientists of strength x can be obtained by multiplying the number $\text{const}\cdot f_0\left(\frac{x-\mu}{\sigma}\right)$ of incoming students of strength x by the proportion $F_0\left(\frac{x-\mu'}{\sigma'}\right)$ of those who successfully get their PhD degrees. Thus, the probability density function that described scientists with PhDs is equal to

$$\text{const}\cdot f_0\left(\frac{x-\mu}{\sigma}\right)\cdot F_0\left(\frac{x-\mu'}{\sigma'}\right).$$

This is exactly the skew-normal distribution!

3 More Detailed Models and Resulting Distributions

Towards a more detailed model: idea. In the above analysis, to determine whether a student succeeds or not in solving the corresponding problem, we only took into account the student's strength and the problem's complexity. In practice, often, there is an additional factor affecting the student's success: the presence of competition.

In a well-organized university department, students' topics are distributed in such a way that an unproductive competition between students from the same university be avoided. However, since students from different universities handle largely the same problems, competition between students from different university is inevitable.

If we take this competition into account, then we see that for a student to succeed, it is not enough that this student's strength is larger than or equal to the complexity of the problem, it is also important to make sure that the student solves the problem ahead of the competition, i.e., that his/her strength is larger than than the strengths of students from other departments who select the same problem.

Transforming the above idea into a precise model. For a student of strength x to succeed, this strength must be larger than or equal to the complexity y of the selected problem and also great then the strengths $x_1, \ldots, x_n$ of students from competing universities who handle the same problem. In this case, the probability of a student succeeding is equal to the probability that

$$y \le x \text{ and } x_1 \le x \text{ and } \ldots \text{ and } x_n \le x.$$

It is reasonable to assume that the corresponding distributions are independent, so this probability is equal to the product of the corresponding probabilities

$$\text{Prob}_y(y \le x) \cdot \text{Prob}_1(x_1 \le x) \cdot \ldots \cdot \text{Prob}_n(x_n \le x).$$

For each university i, the strengths x_i are normally distributed with mean μ_i and standard deviation σ_i. Thus, the probability that a student of strength x succeeds is equal to the product

$$F_0\left(\frac{x-\mu'}{\sigma'}\right) \cdot F_0\left(\frac{x-\mu_1}{\sigma_1}\right) \cdot \ldots \cdot F_0\left(\frac{x-\mu_n}{\sigma_n}\right).$$

As a result, the probability density function that describes scientists with PhDs is equal to

$$\text{const} \cdot f_0\left(\frac{x-\mu}{\sigma}\right) \cdot F_0\left(\frac{x-\mu'}{\sigma'}\right) \cdot F_0\left(\frac{x-\mu_1}{\sigma_1}\right) \cdot \ldots \cdot F_0\left(\frac{x-\mu_n}{\sigma_n}\right).$$

This is a generalization of the skew-normal distribution, in which the original pdf is multiplied not by one normal cdf, but possible by many normal cfds.

Acknowledgments. This work was supported in part by the Chiang Mai University, Thailand, and by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence), DUE-0926721, and OCI-1135525 (Cyberinfrastructure-Team Diffusion Project: The Virtual Learning Commons).

The authors are thankful to all the participants of the 7th International Conference of Thailand Econometric Society TES'2014 (Chiang Mai, Thailand, January 8–10, 2014) for valuable discussions.

References

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

[2] 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.

[3] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.

[4] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[5] K. Suriya and T. Sudtasan, "Fighter ants algorithm with application to efficient team building", *Abstracts of the 7th International Conference of Thailand Econometric Society TES'2014*, Chiang Mai, Thailand, January 8–10, 2014.

[6] T. Wang, B. Li, and A. K. Gupta, "Distribution of quadratic forms under skew normal settings", *Journal of Multivariate Analysis*, 2009, Vol. 100, pp. 533–545.