Casper Albers

Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column   **Casper sees a chance**

De Moivre–Gauss–Laplace: extraordinarily normal

The normal distribution is well-known for its versatility. You can use it for almost all types of statistical analyses. The mathematical underpinnings of this distribution are less well-known, which is a pity as these are an example of mathematical beauty.

Probably the most-used statistical distribution is the normal distribution, also known as the Gaussian distribution, after its 'inventor', Carl Friedrich Gauss. This distribution, classified with mean μ and variance σ has density

$$\phi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

which has the familiar bell shape (see Figure 1).

Normality of residuals is the standard assumption when performing basic statistical analyses such as the t-test, ANOVA or linear regression. The normal distribution appears in statistical mechanics. The normal distribution appears as approximating distribution of the binomial, Poisson, x^2 and Student-t-distribution. Furthermore, the Central Limit Theorem dictates that whatever exotic shape the population distribution has, the distribution of the sample mean will always converge to a normal distribution as the sample size goes up.

De Moivre
Thus, normality is everywhere. But why does this distribution has the shape it has? For that, we have to look at the history of the normal distribution, which is nicely outlined in [5]. The first instances of the normal distribution, attributed to De Moivre [1] in the early eighteenth century, occur as the limiting distribution of the binomial distribution.

The normal distribution, however, is more often used as an error curve. Even the old Greeks realised that their measurements

were not free of error and used some ad hoc ways to improve their measurement. For instance, Hipparchus used the middle of the range of measurements. At the end of the eighteenth century, Thomas Simpson and Pierre Laplace both developed a distribution for the error. Other 'bell curves' have also been considered.

Looking at Figure 1, these curves have a few things in common: (i) small errors are more likely than large errors, (ii) the distribution is symmetrical around 0 (or, in general, around μ): the probability of an error of value ε is equivalent to that of value $-\varepsilon$. These are very sensible properties for an error distribution.
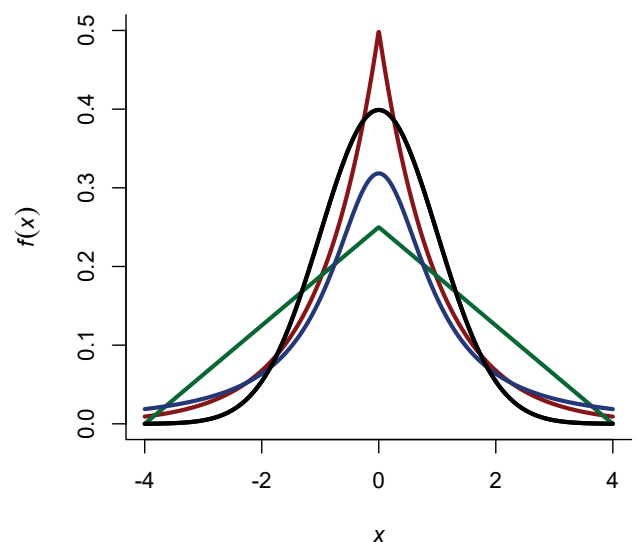

Figure 1   Four bell curves: the standard normal distribution (black), the Cauchy distribution (blue), Simpson's distribution (green) and Laplace's distribution (red).

Gauss

In 1809, Gauss derived 'his' curve by extending these two properties with: (iii) the error distribution should be differentiable (thus, excluding Simpson's and Laplace's suggestions), (iv) having several measurements of the same quantity, the most likely value of the quantity being measured is their average.

Strikingly, these four properties alone are sufficient to reach the normal distribution: no other distribution satisfies these four properties.

Proof

Gauss provided a remarkably elegant proof for this claim. Translated into 'modern terminology', the essence of the proof is as follows (based on [5, pp. 104–105]).

Let $\phi(x)$ be the probability density function of the random error, let μ be the true (unknown) value of the measured quantity, and let n independent observations yield estimates $x_1, \ldots, x_n$. Property (i) implies $\phi(x)$ is maximal at $x = 0$, and property (ii) implies $\phi(-x) = \phi(x)$. Using property (iii) we can define $f(x) = \phi'(x)/\phi(x)$, then $f(-x) = -f(x)$.

Since we assume independence of the errors $(x_i - \mu)$, the joint density of the n errors is given by

$$\Phi = \prod_{i=1}^{n} \phi(x_i - \mu).$$

Property (iv) states that the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the maximum likelihood estimate, which means that Φ is maximised at the value $\bar{x}$: $\partial\Phi/(\partial\mu \mid \mu = \bar{x}) = 0$. Thus,

$$
\begin{aligned}
0 =& -\phi'(x_1 - \mu)\phi(x_2 - \mu)\cdots\phi(x_n - \mu) \\
& -\phi(x_1 - \mu)\phi'(x_2 - \mu)\cdots\phi(x_n - \mu) \\
& \vdots \\
& -\phi(x_1 - \mu)\phi(x_2 - \mu)\cdots\phi'(x_n - \mu) \\
=& -\left(\frac{\phi'(x_1 - \mu)}{\phi(x_1 - \mu)} + \frac{\phi'(x_2 - \mu)}{\phi(x_2 - \mu)} + \cdots + \frac{\phi'(x_n - \mu)}{\phi(x_n - \mu)}\right)\Phi.
\end{aligned}
$$

Using our function $f(x)$, we can write this as

$$f(x_1 - \bar{x}) + f(x_2 - \bar{x}) + \cdots + f(x_n - \bar{x}) = 0.$$

Suppose now that $x_1 = x$ and $x_2 = x_3 = \cdots = x_n = x - nN$ for arbitrary values x and N. Then the formula above can be rewritten as

$$f((n-1)N) = (n-1)f(N).$$

From the theory of differential equations, we know that this (recall the continuity of f) implies that for some $k \in \mathbb{R}$: thus, $\phi'(x)/\phi(x) = kx$. Integration with respect to x provides

$$\ln\phi(x) = \frac{k}{2}x^2 + c \quad \rightarrow \quad \phi(x) = Ce^{kx^2/2}.$$

As $\phi(x)$ must have its maximum at 0, k must be negative. Substituting $k = -1/\sigma^2$, we obtain $\phi(x) \propto e^{-x^2/(2\sigma^2)}$. As

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} = \sqrt{2\pi\sigma^2},$$

we obtain

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{x^2}{\sigma^2}},$$

thus concluding our proof.

Gauss also provided a second proof, based on the least-squares properties. As for a normal distribution, the least squares estimator and maximum likelihood estimator coincide, this proof follows much of the same reasoning.

Laplace

It would only have taken a small twist of fate, and we would have known the Gauss distribution as the Laplace distribution. Only a year after Gauss' publication (and based on work he did some 25 years earlier), Laplace published [3] a mathematical underpinning of the normal distribution based on the central limit theorem. In what is now known as the De Moivre–Laplace theorem, he proved that, if $x_n \sim \text{Bin}(n,p)$ then

$$\lim_{n\to\infty} P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq x\right) = \int_{-\infty}^{x} \phi(t)\,dt.$$

For this reason, statisticians such as Karl Pearson called what we now call the normal distribution the Gauss–Laplace distribution (unfortunately ignoring the work by De Moivre nearly a century earlier). Pearson propagated against the use of the terminology 'normal distribution' as this might "lead people to believe all other distributions of frequency are in one sense or another abnormal" [4]. Given the vast statistical heritage of Pearson, we don't have to feel sorry that this advice of his hasn't been followed. Given the sheer range of applications, no other distribution than that by De Moivre, Gauss and Laplace would be worthy of the name 'normal'.   ⸫⸱⸱

References
1   A. De Moivre, *Approximatio ad Summam Terminorum Binomii* $(a + b)^n$ *in Seriem Expansi* (1733).
2   C. F. Gauss, *Theoria Motus Corporum Celestium*, Perthes et Besser, Hamburg, 1809.
3   P.S. Laplace, Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités, *Mémoires de l'Academie des sciences de Paris* (1810).
4   K. Pearson, Notes on the history of correlation. *Biometrika* (1920).
5   S. Stahl, The evolution of the normal distribution, *Mathematics Magazine* 79(2) (2006) 96–113.