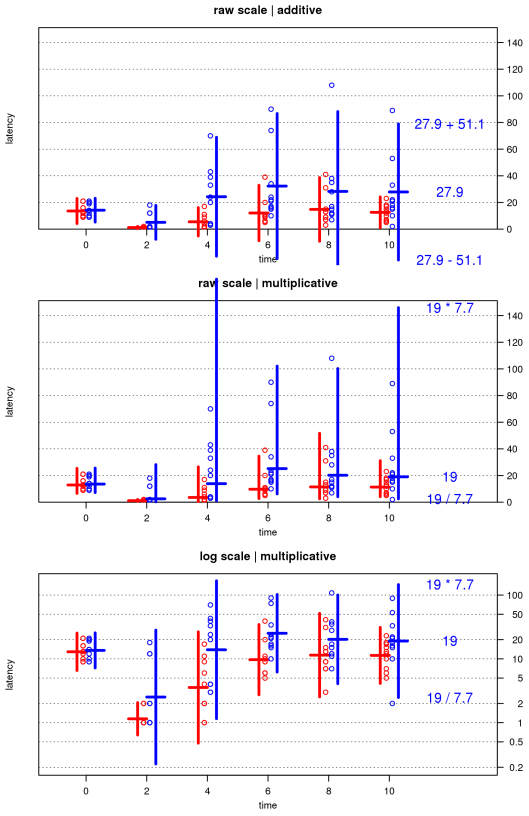
Reconsidering the Normal Distribution – Benefits from Replacing Plus/Minus by Times/Divide

Eckhard Limpert* and Werner A Stahel
Seminar for Statistics, ETH Zurich, CH-8092 Zurich, Switzerland



The fundamental benefit of reconsidering the normal distribution is to gain an adequate comprehension of the nature of quantitative variation. To this aim, the „95 % range check“[1] provides an easy tool. Too frequently data fail the check and prove that distributions across plant protection and the sciences most often are clearly skewed. The standard way of characterizing data using the (additive) normal distribution is then inefficent or misleading.

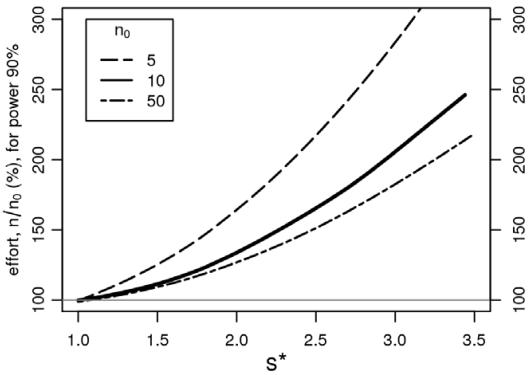
Chance 1: Improved recognition of variation.
The example shows the evolution of latencies of wild type *C. elegans* for two different treatments (data courtesy Raizen, see Nature **451**: 569-573)

a) Data characterized by symmetric bars - originally depicted by SEMs here only - suggest an approximately (additive) normal distribution.
Frequently, data are characterized at this original, additive scale with the mean and standard deviation (SD), or standard error (SEM). The skewed nature of the data becomes obvious with the „95 % range check“.
Checking these ranges - mean ± two standard deviations – shows that most of them extend to negative values that are, of course, impossible.

b) The same data, again shown at the original scale, but by ranges based on the geometric mean and the multiplicative standard deviation.
Now, the ranges fit the data, and do not include negative values any more. A disadvantage remains, as the original scale is not well suited for visualizing variations of data in the lower part of the figure.

c) The same data shown at the multiplicative or logarithmic scale. Again, the data do fit well the 95 % ranges indicated and appear normally distributed at this scale. Now, interestingly, both treatments exhibit similar variation and the nature of the data can be nicely recognized.

Chance 2: Improved efficiency, and ethic responsibility.
Two samples of log-normal data with skewness parameter s* have been simulated repeatedly. The ratio of medians m* are chosen to achieve 90% power with a t-test on the log scale (the appropriate test in this case) to distinguish the two groups. If a t-test on the raw scale is used, larger samples (size $n>n_0$) are needed to obtain the same power. The ratio n/n_0 is a measure of inefficiency. The extra effort needed if the inadequate method is used, with a median s*=2.3 and n=10 amounts to 50%. -
With humans or animals involved, this avoidable effort is an ethical issue.



Replacing plus/minus by times/divide: examples from plant protection

Example	Description, original		95 % range	Description, recommended		95 % range	abbreviations
	m ± SEM (n)	m ± SD		m* ˣ/ SEM*	m* ˣ/ s*		n = number investig.
Bacteria in rhizosphere, 15d x 10^3, [59cf2]	*55 ± 13 (10)*	55 ± 41.1	-27.2 to 137.2	**44.1 ˣ/ 1.23**	**44.1 ˣ/ 1.95**	**11.6** to **167**	m = mean
Ps. savastanoi, CFU x 10^6, Tab. 2, Bagno, [60cf2]	*61 ± 59 (8*	61 ± 170	-279 to 401	**20.6 ˣ/ 1.68**	**20.6 ˣ/ 4.36**	**1.08** to **392**	m* = m multiplicative
Sensitivity wheat p. mildew, mg l^{-1}, [9cf2]		*25* ± 26.4	-27.8 to 77.8		**17.2 ˣ/ 2.38**	**3.04** to **97.1**	SD = standard dev.
EC50 of MO..129, mg l^{-1} [Phytopathol. **105**:292]		14.2 ± 17.1	-2.9 to 31.3		**9.07 ˣ/ 2.58**	**1.37** to **60.2**	s* = SD multiplicative
H. annosum-caused gaps in forests, m^2, [63cf2]		2898 ± 1898	-898 to 6794		**2424 ˣ/ 1.82**	**734** to **8008**	SEM = SD of the mean

Conclusions: Quantitative data most often follow skewed distributions, which can be approximated by the multiplicative or log-normal law[1-3]. Consequently, data should be described in the form of m* ˣ/ s* rather than mean ± SD. In graphics, log axes are recommended. Using the appropriate versions of t-tests and, more generally, regression methods and ANOVA, leads to more appropriate models and gains statistical efficiency, that is, requires less observations. - For further information see our poster on „Plant protection and data science – the normal distribution is the log-normal distribution“.

*Present Address: ELI-o-Research, Scheuchzerstr. 210, CH-8057 Zürich; +41-76-369 2132 (handy); eckhard.limpert@bluewin.ch

References
1. Limpert E, Stahel WA, 2011, Problems with Using the Normal Distribution – and Ways to Improve Quality and Efficiency of Data Analysis. PLoS ONE 6(7):e21403. doi:10.1371/journal.pone.0021403
2. Limpert E, Stahel WA, Abbt M., 2001, Log-normal distributions across the sciences - keys and clues. BioScience 51, 341-352.
3. Stahel WA, Limpert E, The normal distribution is the log-normal distribution. Talk Leibniz-Inst. Magdeburg, Dec 2, 2014, http://stat.ethz.ch/~stahel/talks/lognormal.pdf