An Example of Questionable Modeling

As mentioned, the normal distribution is a very popular model for statistical modeling. There is a good reason for that: the Central Limit Theorem provides a theoretical support for this choice in many situations.

However, even a good thing can be pushed too far. While it may not always be obvious when a normal distribution may be inappropriate, an obvious case is when the quantity under consideration cannot take negative values, and yet this could happen with positive probability under a normal model. This dilemma has generated some curious pseudo-solutions. Let us look at this situation closer.

Case 1: The CLT is reasonably applied

For instance, this is the case in many binomial cases (e.g., polling): the number of positive responses cannot be less than zero, yet the normal approximation is practically always used. This is not a real problem: suppose, to argue through a concrete example, that you are polling 1000 people, and that the proportion of positive responses has been 0.7. That suggests a normal approximation for the average of, approximately, mean 0.7, and variance $\frac{0.7 \cdot 0.3}{1000} = 0.00021$. Under such a model, the probability of a negative average is

$$P[X < 0] = P\left[\frac{X - 0.7}{\sqrt{2.1 \cdot 10^{-4}}} \lesssim -48\right]$$

which is ridiculously small (whatever software you use, it will almost surely evaluate to 0).

More generally, as long as the probability of the negative tail is negligible, there clearly is no harm done with a normal approximation: after all, we know it is an *approximation*, and a minute glitch like this is on par with the impossibility of outrageously large values too.

Case 2: No Theorem Was Applied, and we just took the normal because we like it

This strange attitude is much more present than one would imagine, and it is particularly active in the non-mathematical, non-statistical "practical" research literature (a couple of examples are quoted at the end). In these circumstances, the use of a normal distribution is not justified by any specific argument, and the unfortunate fact that negative values of a most surely positive quantity have a non negligible probability is handled with an amazing ad-hoc tool: the distribution is arbitrarily truncated.

This means that, instead of the usual normal distribution function $P[X \leqslant x]$, a *conditional distribution function* is used: $\frac{P[X \leqslant x]}{P[X \geqslant 0]}$ (more generally, you may find a conditioning on $\{a \leqslant X \leqslant b\}$, for some values a, b). The problem here is that there is no theoretical argument, and the use of this truncated distribution is simply motivated by habit and, possibly, convenience.

Fact is, if we are looking at a positive quantity, there are many choices for surely positive distributions, and some better arguments (such as suitable limits, or other theoretical arguments) should be produced to support our choice. The truncated normal, however, has no reasonable argument in its favor at all, since it does not appear "naturally" in any significant modeling theorem!

One counter-argument could be the following. Since we may not have a solid argument for any distribution, we might as well use one that is easy to use, and does not describe things too badly. This might be acceptable, but, then again, it all depends on what you are going to do with your model. Let us look at an artificial simple example that illustrates one of many possible pitfalls.

A Simple (a little naive) "Counter-Example": Comparing an Exponential Variable and a "Truncated Normal" One

Consider a set of data, which only takes positive values. Suppose also that a histogram shows that higher values are less present in the sample than lower values. Compare now the consequence of modeling this experiment with an exponential random variable or a truncated normal. Suppose also, for simplicity, that it turned out that the "right" truncated normal model was a standard normal conditioned on being positive, which has density (for $x>0$) $\sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2}}$. Let's call such a Random Variable Y.

Remark 1. Actually, it is not hard to see that, the distribution of Y is the same as that of the absolute value of a standard normal variable. In certain circumstances, this could actually be a reasonable model. The choice here is motivated by the ease of computing mean and variance of this distribution, as opposed to a more generic truncation (the details that follow are meaningful if you took at least a second calculus class—if you didn't, you will just have to take my word for it).

Indeed, by substituting $u=\frac{x^2}{2}, du=xdx$ in the calculation for the mean, and by integrating by parts, with $u=x, dv=xe^{-\frac{x^2}{2}}dx=-d\left(e^{-\frac{x^2}{2}}\right)$ in the calculation of EY^2, we have

$$\int_0^\infty xe^{-\frac{x^2}{2}}dx=\int_0^\infty e^{-u}du=1$$

$$\int_0^\infty x^2e^{-\frac{x^2}{2}}dx=\int_0^\infty x\cdot xe^{-\frac{x^2}{2}}dx=-x\cdot e^{-\frac{x^2}{2}}\Big|_0^\infty+\int_0^\infty e^{-\frac{x^2}{2}}dx=0+\frac{1}{2}\sqrt{2\pi}=\sqrt{\frac{\pi}{2}}$$

The expected value of this distribution happens to be $\sqrt{\frac{2}{\pi}}\approx 0.8$, and its variance 1. Hence, an exponential distribution of expected value 1 (and, consequently, standard deviation and variance also equal to 1) would have similar indexes. For this exponential distribution, $P[X>x]=e^{-x}$.

Now, choosing one or the other has big implications in terms of the likelihood of observing large values. For example, using the normal tables, we see that $P[Y>3]\approx 0.002$, while $P[X>3]\approx 0.05$, and the discrepancy grows fast if you go further out. In other words, the likelihood of observing a value larger than 3 is 25 times larger if we adopt an exponential model. Incidentally, using a "fat tail" model (where the density would decrease, as x increases, no faster than a power of x) would cause an even greater discrepancy.

As a related anecdote, among the many causes behind the market crash of 2007 a small contribution was provided by the use of normal models to asses the risk financial institutions were exposed to. This made big losses extremely unlikely, yet that's precisely what happened (a 5% probability is considered seriously in risk management).

References

This is not a serious bibliography: it is only a casual choice of peer-reviewed papers implementing truncated normal models

1. F. Xu, R.C. Mittlehammer, L.A. Torell: Modeling Nonengativity via Truncated Logistic and Normal Distributions: An Application to Ranch Land Price Analysis. *Journal of Agricultural and resource Economics.* 19: 102–114, 1994

2. A.C. Johnson, N.T. Thomopoulos: Use of the Left-Truncated Normal Distribution for Improving Achieved Service Levels. *Decision Sciences Institute 2002 Annual Meeting Proceedings.* 2033–2041