Errors, Regress Arguments, and the Underestimation of Small Probabilities

Nassim N Taleb
NYU-Poly Institute

April 16,2011 (Draft version-contains errors -cannot be cited as yet)

PAPER TO BE PRESENTED AT BENOIT MANDELBROT'S MEMORIAL
Yale University, APRIL 29, 2011

Abstract

Reapplying measurements of uncertainty about the errors of the standard deviation for a conventional distribution in the Gaussian basin (i.e. stemming from the standard central limit theorem) leads to infinite variance -merely for epistemic reasons. A mere .01% error rate about the STD, and .01% error rate about that error rate, etc. (recursing all the way) results in explosive higher moments and convergence to power laws with infinite variance (but with no effect on the mean, meaning finite expected absolute deviations). The paper states the conditions under which the problem occurs, shows its explanation of the chronic bias leading to the underestimation of small probabilities (a standard example of which is Fukushima), and examines the consequences about acceptable metrics.

KEYWORDS: Fukushima, Risk management, Epistemology of probability, Model errors

(THIS IS DRAFT FORMAT AS I AM LOOKING FOR ERRORS - I APOLOGIZE FOR THE TEMPORARY MATHEMATICA LABEL)

Error Rate About the Error Rate

An error rate can be measured. The measurement, in turn, will have an error rate. The measurement of the error rate will have an error rate. The measurement of the error rate will have an error rate. What is called a regress argument by philosophers can be used to put some scrutiny on quantitative methods or risk and probability.

Background: The Regress Argument

The main problem behind *The Black Swan* problem is the limited understanding of model error, and, for those who get it, a lack of understanding of second order errors (about the methods used to compute the errors) and by a regress argument, continuously reapplying the thinking all the way to its limit. Few get the point that the skepticism in *The Back Swan* it does not invalidate all measurements of probability; its value lies in showing a map of what is vulnerable and what is not, defining such domains based on their propensity fat-tailedness, and building robustness by appropriate decision-making rules.

Epistemic not statistical approach: Previous derivations of power laws have been statistical (cumulative advantage, preferential attachment, winner-take-all effects, criticality), and the properties derived by Yule, Mandelbrot, Zipf, Simon, Bak, and others result from structural conditions or breaking the independence assumptions in the sums of random variables allowing for the application of the central limit theorem. This one is entirely epistemic, based on standard philosophical regress arguments.

Philosophers and Regress Arguments: I was having a conversation with the philosopher X about errors in the assumptions of a model (or its structure) not being customarily included back into the model itself when I realized that only a philosopher can understand a problem to which the entire quantitative field seems blind. For instance, probability professionals do not include in the probabilistic measurement itself an error rate about, say, the measument of a parameter provided by an expert, or other uncertainties attending the computations.

Unlike philosophers, quantitative scientists don't seem to get regress arguments; questioning all the way (without making a stopping assumption) is foreign to them (what I've called scientific autism). Just reapplying layers of uncertainties may show convexity biases, and, fortunately, it does not necessarily kill probability theory; it just disciplines the use of some distributions,

at the expense of others --distributions in the $\mathcal{L}^2$ norm (i.e., square integrable) are no longer valid, for epistemic reasons. This is does not mean we cannot have paremetric distributions; it just means that when there is no structure to the error rates we need to stay in the power law domains, even if the data does not give us reasons for that.

The bad news is that this recursion of the error rate invalidates all common measures of ***small*** probabilitites --and has the effect of raising them.

Indeed, the conversation with the philosopher was quite a relief as I had a hard time discussing with autistic quants and statisticians the point that ***without understanding errors, a measure is nothing and one should take the point to its logical consequence that any measure of error needs to have its own error taken into account***.

The epistemic aspect of standard deviations: One frustrating conversation took place a decade ago with another academic, a professor of risk management who writes papers on Value at Risk: he could not get that the standard deviation of a distribution ***for future outcomes*** (and not the sampling of some properties of existing population), the measure of dispersion, needs to be interpreted as the measure of uncertainty, hence ***epistemic***, and that it should necessarily have uncertainties (errors) attached to it. One needs to look at the standard deviation -or other measures of dispersion -as a degree of ignorance about the future realizations of the process, and, mostly, as uncertainty about the mean. The higher the uncertainty, the higher the measure of dispersion (variance, mean deviation, etc.)

Such uncertainty, by Jensen's inequality, creates non-negligible convexity biases. So far this is well known in places in which subordinated processes have been used --for instance stochastic variance models --but I have not seen the layering of uncertainties taken into account.

Betting on Rare Events: Finally, this note explains my main points about betting on rare events. Do I believe that the probability of the event is higher? No. I believe that any layer of uncertainty raises such probability (because of convexity effects) --and this note has the derivations to show it.

Main Results

1. Proof that under proportional constant (or increasing) recursive layers of uncertainty about rates of uncertainty, the distribution converges to an infinite variance power law, even when one starts with a standard Gaussian.
2. When the errors are decreasing (proportionally) for higher order, the ending distribution becomes fat-tailed but retains its finite variance.
3. So, in both cases the use of the Gaussian is not warranted unless higher order errors can be eliminated a priori.

Methods

Layering Uncertainties

Take a standard probability distribution, say the Gaussian. The measure of dispersion, here σ, is estimated, and we need to attach some measure of dispersion around it. The uncertainty about the rate of uncertainty, so to speak, or higher order parameter, similar to what called the "volatility of volatility" in the lingo of option operators (see Taleb, 1997, Dupire, x, Derman, x) --here it would be "uncertainty rate about the uncertainty rate". And there is no reason to stop there: we can keep nesting these uncertainties into higher orders, with the uncertainty rate of the uncertainty rate of the uncertainty rate, and so forth. There is no reason to have certainty anywhere in the process.

Now, for that very reason, this paper shows that, in the absence of knowledge about the structure of higher orders of deviations, we are forced to use a power-law tails. Most derivations of power law tails have focused on processes (Zipf-Simon preferential attachment, cumulative advantage, entropy maximization under constraints, etc.) Here we just derive using lack of knowledge about the rates of knowledge.

Higher order integrals in the Standard Gaussian Case

Define $\phi(\mu,\sigma,x)$ as the Gaussian density function for value x with mean μ and standard deviation σ.

N^{th} order stochastic standard deviation

A 2^{nd}order stochastic standard deviation is the integral of ϕ across values of $\sigma \in]0,\infty[$, under the measure $f(\overline{\sigma}, \sigma_1, \sigma)$, with σ_1 its coefficient of variation, not necessarily its standard deviation; the expected value of $\sigma_1$is $\overline{\sigma_1}$.

Generalizing to the N^{th} order:

$$\int_0^\infty , , , \int_0^\infty \phi(\mu, \sigma, x)\, f(\overline{\sigma}, \sigma_1, \sigma)\ f(\overline{\sigma_1}, \sigma_2, \sigma_1)\, f(\overline{\sigma_2}, \sigma_3, \sigma_2) \ldots d\sigma\, d\sigma_1\, d\sigma_2 \ldots d\sigma_N$$

The problem is that it is parameter-heavy and requires the specifications of the subordinated distributions. We would need to specify a measure *f* for each layer of error rate (and extract the Wiener Chaos terms to be able to add the integrals in place of summing them). Instead this can be approximated by using its mean deviation, as we will see next.

Simplification using nested series of two-states for σ

A quite effective simplification to capture the convexity, the ratio of (or difference between) $\phi(\mu,\sigma,x)$ and $\int_0^\infty \phi(\mu, \sigma, x)\, f(\overline{\sigma}, \sigma_1, \sigma)\ d\sigma$ would be to use a weighted average of values of σ, say, for a simple case of one-order stochastic volatility $\sigma(1+a[1])$, and $\sigma(1-a[1])$, $0 \leq a[1] < 1$ where *a[1] is* the proportional mean absolute deviation for σ. In other word a is a[1] measure of the absolute error rate for σ.

Thus the distribution using the first order stochastic standard deviation can be expressed as:

$$f(x) = \frac{1}{2}\left\{\phi(\mu,\ \sigma(1\ +\ a[1]),\ x)\ +\ \phi(\mu, \sigma(1 - a[1]), x)\right\}$$

Now assume uncertainty about the error rate a[1], expressed by a[2]. Thus in place of a[1] we have a[1](1+a[2]) and a[1] (1-a[2]) .

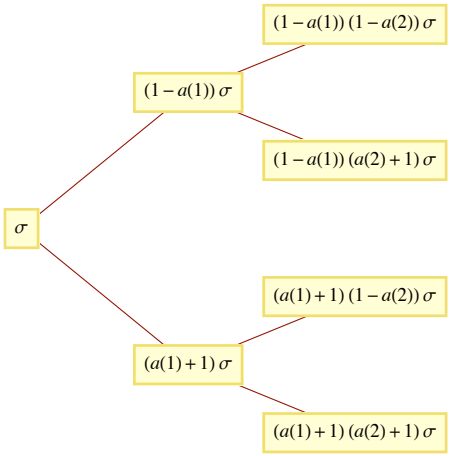

Figure 1- two levels of error rates

The second order stochastic standard deviation:

$$f(x) = \frac{1}{4}\left\{\phi(\mu, \sigma(1 + a[1]\,(1 + a[2])), x) + \phi(\mu, \sigma(1 - a[1]\,(1 + a[2])), x) + \right.$$
$$\left.\phi(\mu, \sigma\,(1 + a[1]\,(1 - a[2])), x) + \phi\,(\mu, \sigma\,(1 - a[1]\,(1 - a[2])), x)\right\}$$

and the $\mathtt{N}^{\mathtt{th}}$ order:

$$f(x)\ = \frac{1}{2^N}\ \sum_{i=1}^{2^N} \phi\left(\mu, \sigma\, M_i^N, x\right)$$

where M_i^N is the $\mathtt{i}^{\mathtt{th}}$ scalar (line) of $\mathtt{M}^{\mathtt{N}}$ $(1 \times 2^{\mathtt{N}})$

$$M^N = \left\{ \prod_{j=1}^{N} (a\,(j)\;T[\![i,\,j]\!] + 1) \right\}_{i=1}^{2^N}$$

and T[[i,j]] the element of i^{th}line and j^{th}column of the matrix of the exhaustive combination of N-Tuples of (-1,1),that is the N-dimentionalvector {1,1,1,...} representing all combinations of 1 and -1.

for N=3

$$T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{pmatrix} \qquad \text{and } M^3 = \begin{pmatrix} (1-a(1))\,(1-a(2))\,(1-a(3)) \\ (1-a(1))\,(1-a(2))\,(a(3)+1) \\ (1-a(1))\,(a(2)+1)\,(1-a(3)) \\ (1-a(1))\,(a(2)+1)\,(a(3)+1) \\ (a(1)+1)\,(1-a(2))\,(1-a(3)) \\ (a(1)+1)\,(1-a(2))\,(a(3)+1) \\ (a(1)+1)\,(a(2)+1)\,(1-a(3)) \\ (a(1)+1)\,(a(2)+1)\,(a(3)+1) \end{pmatrix}$$

$$\text{so } M_1^3 = \{(1-a[1])\,(1-a[2])\,(1-a[3])\},\ \text{etc.}$$

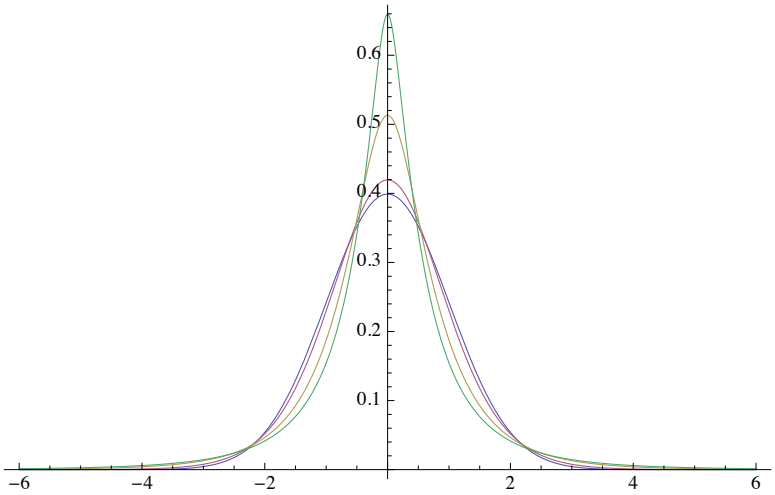

Figure x, Thicker tails (higher peaks) for higher values of N; here N=0,5,25,50

Characteristic Functions & Infinite Variance when N→∞

Assume μ=0. By summing characteristic functions, C(t) the characteristic function of order t:

$$C(t,\ M,\ N) = \frac{\sum_{i=1}^{2^N} e^{\exp\left(-\frac{1}{2}\,t^2\,\left(\sigma\,M_i^N\right)^2 + i\,m\,t\right)}}{2^N}$$

Assume that a[1]=a[2]=...a[N]=a, i.e. the case of flat proportional error rate a. The dispersion becomes a binomial tree.

$$C(t,\,a,\ N) = \sum_{j=0}^{N} 2^{-N} \binom{N}{j} \exp\left(-\frac{1}{2}\left(t^2\left(\sigma\,(a+1)^j\,(1-a)^{N-j}\right)^2\right)\right)$$

This allows us to calculate the moments.

Table of Moments

Order	Moment
1	0
2	$\left(1+a^2\right)^N \sigma^2$
3	0
4	$3\left(1+6a^2+a^4\right)^N \sigma^4$
5	0
6	$15\left(1+15a^2+15a^4+a^6\right)^N \sigma^6$
7	0
8	$105\left(1+28a^2+70a^4+28a^6+a^8\right)^N \sigma^8$

Consequences

We can see here how, for a constant a > 0, and in the more general case with variable a where a[n] ≥ a[n-1], how the moments explode.

We can thus prove the following:

A- Even the smallest value of a >0, since $\left(1+a^2\right)^N$ is unbounded, leads to the second moment going to infinity (though not the first) when N→∞. So something as small as a .001% error rate will still lead to explosion of moments and invalidation of the use of the class of $\mathcal{L}^2$ distributions.

B- We need the use of power laws for epistemic reasons. [I haven't tried to prove it yet, just numerical methods]

Convergence to Power Laws

We can see on the example next Log-Log plot (Figure 1) how, at higher orders of stochastic volatility, with equally proportional stochastic coefficient, (where a(1)=a(2)=...=a(N)= $\frac{1}{5}$) how the density converges to that of a power law, as shown in flatter density on the LogLog plot.

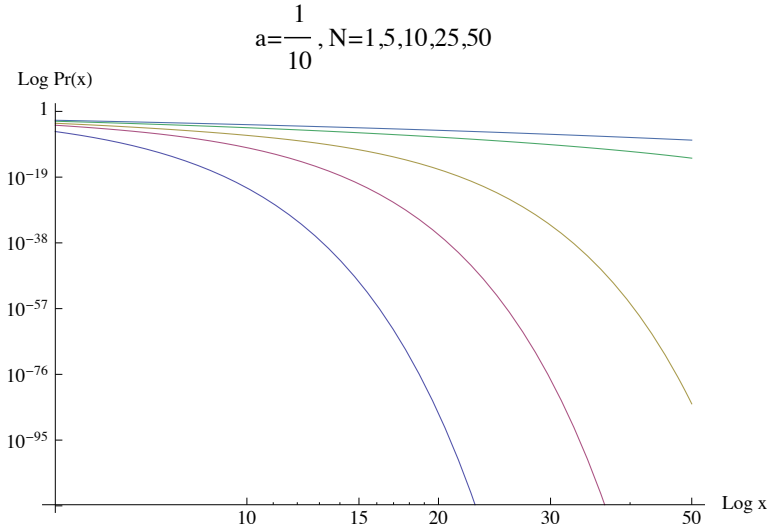

Figure x - LogLog Plot of the convergence to a power laws as N rises. Here all a= 1/10

Effect on Small Probabilities

Next we measure the effect on the thickness of the tails. The obvious effect is the rise of small probabilities.

Take the probability of exceeding K, given N, for a constant :

$$P > K \mid N = \sum_{j=0}^{N} 2^{-N-1} \binom{N}{j} \text{erfc}\left(\frac{K}{\sqrt{2}\ \sigma\ (a+1)^j\ (1-a)^{N-j}}\right)$$

where erfc(.) is the complementary of the error function, 1-erf(.), $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$

a = 1/100

N	P>3,N=0 / P>3,N	P>5,N=0 / P>5,N	P>10,N=0 / P>10,N
5	1.01724	1.155	7
10	1.0345	1.326	45
15	1.05178	1.514	221
20	1.06908	1.720	922
25	1.0864	1.943	3347

a = 1/10

N	P>3,N=0 / P>3,N	P>5,N=0 / P>5,N	P>10,N=0 / P>10,N
5	2.74	146	1.09×10^{12}
10	4.43	805	8.99×10^{15}
15	5.98	1980	2.21×10^{17}
20	7.38	3529	1.20×10^{18}
25	8.64	5321	3.62×10^{18}

Case of Finite Variance

When the higher order of a[i] decline, then the moments tend to be capped.

Take a "bleed" of higher order errors at the rate λ, $0 \le \lambda < 1$, so a[N] = λ a[N-1], hence a[N] = λ^N a[1], with a[1] the conventional intensity of stochastic standard deviation.

With N=2 , the second moment becomes:

$\sigma^2 \left(1 + a[1]^2\right) \left(1 + \lambda^2\, a[1]^2\right)$

With N=3,

$\sigma^2 \left(1 + a[1]^2\right) \left(1 + \lambda^2\, a[1]^2\right) \left(1 + \lambda^4\, a[1]^2\right)$

finally, for the general case:

$\sigma^2 \left(1 + a[1]^2\right) \prod_{i=1}^{N-1} \left(1 + \lambda^{2i}\, a[1]^2\right)$

Using the Q – Pochhammer symbol $(a; q)_N = \prod_{i=1}^{N-1} \left(1 - a q^i\right)$

$$\text{M2}(N) = \sigma^2 \left(-a[1]^2; \lambda^2\right)_N$$

Which allows us to get to the limit

$$\text{Limit M2}\,(N)_{N\to\infty} = \sigma^2 \frac{\left(\lambda^2; \lambda^2\right)_2 \left(a(1)^2; \lambda^2\right)_\infty}{\left(\lambda^2 - 1\right)^2 \left(\lambda^2 + 1\right)}$$

So the limiting second moment for λ=.9 and a[1]=.2 is just 1.28 σ^2, a significant but relatively benign convexity bias.

As to the fourth moment:

By recursion:

$$3\,\sigma^4 \prod_{i=0}^{N-1} \left(1 + 6\,\lambda^{2\,i}\,a[1]^2 + \lambda^{4\,i}\,a[1]^4\right)$$

$$\text{M4}(N) = 3\,\sigma^4 \left(\left(2\sqrt{2} - 3\right) a(1)^2; \lambda^2\right)_N \left(-\left(3 + 2\sqrt{2}\right) a(1)^2; \lambda^2\right)_N$$

$$\text{Limit M4}\, N_{N\to\infty} = 3\,\sigma^4 \left(\left(2\sqrt{2} - 3\right) a(1)^2; \lambda^2\right)_\infty \left(-\left(3 + 2\sqrt{2}\right) a(1)^2; \lambda^2\right)_\infty$$

So the limiting fourth moment for λ=.9 and a[1]=.2 is just $9.88\,\sigma^4$, more than 3 times the Gaussian's ($3\,\sigma^4$), but still finite fourth moment. For small values of a and values of λ close to 1, the fourth moment collapses to the level of a Gaussian.

Conclusion

So far we examined two regimes, one in which the higher order errors are proportionally constant, the other one in which we can allow them to decline. The difference between the two is easy to spot: naturally thin-tailed domains, something very rare on mother earth. Outside of these very special situations (say in some engineering applications), the Gaussian and its siblings (along with the measures such as STD, correlation, etc.) should be completely abandoned, along with any attempt to measure small probabilities. So the power law distributions are more prevalent than initially thought.

- Question: can we separate the two domains along the rules of tangibility/subjectivity of the probabilistic measurement? Daniel Kahneman had a saying about measuring future states: how can one "measure" something that does not exist?
- Should we use time as a dividing criterion: anything that has time in it (meaning involves a forecast of future states) needs to fall into the first regime of non-declining proportional uncertainty parameters a[i]?