THE PINKER FALLACY SIMPLIFIED [DRAFT]

(Pinker is saying in his book: "Violence has dropped by .0001 standard deviations. Let us see WHY."

Almost four years after I started reporting the problem to S. Pinker, he kept promoting his theory of the long peace, while spinning, going to the press, attacking, fake claims that "he misread my book" and fiercely defending his reputation instead of correcting his errors –and had recourse to bloggers to win a public argument *instead of qualified statistical venues – which is why I call it the Pinker fallacy. The scholarly approach isn't going to the blogo-press, but writing a TECHNICAL document.*)

Science never lets you produce a theory on the basis of insufficient departure from random. Otherwise it is patently "fooled by randomness". And scientific papers are written under the basis that:

> *You shall not produce a theory, or even produce the slightest explanation from statistical noise.*

We have established that science is what distinguishers itself from "anecdote", "BS", "charlatanism", or something similar. Further:

> *Science doesn't discuss anecdotes, only estimators.*

And an estimator is something that unlike the anecdote is non-noise, can be generalized to allow for theory. Strangely,

> *If the variance is zero, the anecdote and the estimation become the same thing. But only if the variance is zero.*

Processes with low variance (of the sort we see in physics) can be treated as deterministic, which is why we are spoiled in the physical world.

Now, for estimation what you need something called the *law of large numbers*: as your sample gets larger, the mean becomes more stable. You double your sample size, the error is reduced by a square root of two.

Now when you study statistics in college, you are focusing on processes in the thin tailed class what I call Mediocristan --alas, statistical inference wasn't concerned with other classes. So I focused my initial twenty years of trading, where markets are very fat tailed, on heuristic fixes to the problem and had to struggle initially without much help, and now I am doing it more formally.

> *Fat tailed processes need* more *information 1) because the law of large numbers is slower than square root of observations, 2) the true variance is likely to be much, much higher than what you see in your sample.*

Given that much of what's around us is from Mediocristan (thin tailed), we are spoiled because information is Gaussian, or the variance is low so anecdotes match estimators. In it *All You See is What is There*, statistically speaking. But if someone says: "violence has dropped" and violence is fat tailed, you cannot make the same statement –it is exactly like reporting what happened in Las Vegas on a roulette table after ten minutes of observation.

As we saw in (chapter x), someone saying "the market dropped one point" and produced an "explanation" would be considered a BS vendor. You need significant departure from past observation to be able to give one.

> *The Pinker fallacy is theorizing or spreading anecdotal inferences from statistical noise –that is, the gambler's fallacy —under fat tailed domains using metrics from thin tailed domains and producing theories and fooled by randomness style explanations.*

What is disappointing is that people –because of mechanistic but shallow training in statistics –can detect the gambler's fallacy in casinos (Pinker himself is aware of it since I have seen him use the term) but more dangerously, generalize and apply thin-tailed methods to the Extremistan, fat tailed domains.

It is strange how the spread of social science is causing a degradation of scientific rigor. But there is worse aspect of the Pinker fallacy: large deviations from the norm do not require large samples. But for that we need to understand the various concentration measures.

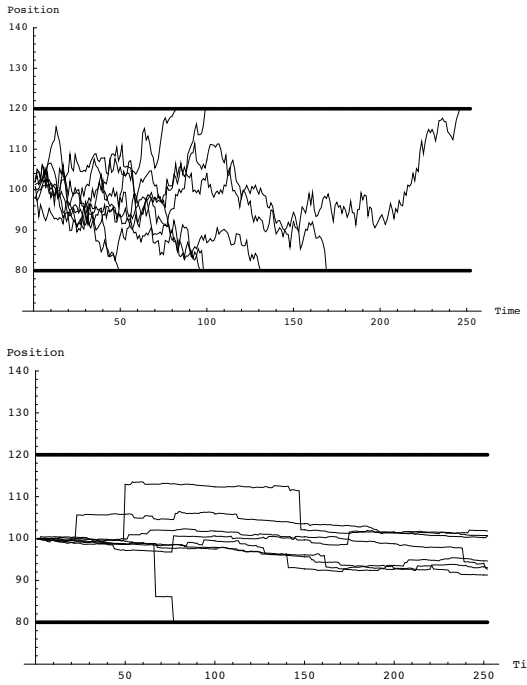
This includes 3 articles, the first a nontechnical discussion of the book by science writer S. Pinker, the second a technical discussion of the flaw in Pinker's book published in *Physica A: Statistical Mechanics and Applications*, the third a technical discussion of what I call the Pinker Problem, a corruption of the law of large numbers.

The "Long Peace" is a Statistical Illusion

Nassim Nicholas Taleb

When I finished writing *The Black Swan*, in 2006, I was confronted with ideas of "great moderation", by people who did not realize that the process was getting fatter and fatter tails (from operational and financial, leverage, complexity, interdependence, etc.), meaning *fewer but deeper* departures from the mean. The fact that nuclear bombs explode less often that regular shells does not make them safer. Needless to say that with the arrival of the events of 2008, I did not have to explain myself too much. Nevertheless people in economics are still using the methods that led to the "great moderation" narrative, and Bernanke, the protagonist of the theory, had his mandate renewed.

I had argued that we were undergoing a switch between the top graph (continuous low grade volatility) to the next one, with the process moving by jumps, with less and less variations outside of jumps.



My idea of current threats outside finance:

The Pinker Argument

{Technical Appendix}

Pinker's Rebuttal of This Note

Pinker has written a rebuttal (*ad hominem* blather, if he had a point he would have written something $\frac{1}{3}$of this, not 3 x the words). He still does not understand the difference between probability and expectation (drop in observed volatility/fluctualtion ≠ drop in risk) or the incompatibility of his claims with his acceptance of fat tails (he does not understand asymmetries-- from his posts on FB and private correspondence). Yet it was Pinker who said "what is the actual risk for any individual? It is approaching zero".

Second Thoughts on The Pinker Story: What Can We Learn From It

It turned out, the entire exchange with S. Pinker was a *dialogue de sourds*. In my correspondence and exchange with him, I was under the impression that he simply misunderstood the difference between inference from symmetric, thin-tailed random variables an one from asymmetric, fat-tailed ones --the 4th Quadrant problem. I thought that I was making him aware of the effects from the complications of the distribution. But it turned out things were worse, a lot worse than that.

Pinker doesn't have a clear idea of the difference between science and journalism, or the one between rigorous empiricism and anecdotal statements. Science is not about making claims about a sample, but using a sample to make general claims and discuss properties that apply outside the sample.

Take M^* the observed arithmetic mean from the realizations (a sample path) for some process, and M the "true" mean. When someone says: "Crime rate in NYC dropped between 2000 and 2010", the claim is about M^* the observed mean, not M the true mean, hence the claim can be deemed merely journalistic, not scientific, and journalists are there to report "facts" not theories. No scientific and causal statement should be made from M^* on "why violence has dropped" unless one establishes a link to M the true mean. M^* cannot be deemed "evidence" by itself. Working with M^* cannot be called "empiricism".

What I just wrote is at the foundation of statistics (and, it looks like, science). Bayesians disagree on how M^* converges to M, etc., never on this point. From his statements, Pinker seems to be aware that M^* may have dropped (which is a straight equality) and sort of perhaps we might not be able to make claims on M which might not have really been dropping.

Now Pinker is excusable. The practice is widespread in social science where academics use mechanistic techniques of statistics without understanding the properties of the statistical claims. And in some areas not involving time series, the differnce between M^* and M is negligible. So I rapidly jot down a few rules before showing proofs and derivations (limiting M to the arithmetic mean). Where E is the expectation operator under "real-world" probability measure P:

1. **Tails Sampling Property**: $E[|M^*-M|]$ increases in with fat-tailedness (the mean deviation of M* seen from the realizations in different samples of the same process). In other words, fat tails tend to mask the distributional properties.
2. **Counterfactual Property:** Another way to view the previous point, $\mu[M^*]$, The distance between different values of M^* one gets from repeated sampling of the process (say counterfactual history) increases with fat tails.
3. **Survivorship Bias Property**: $E[M^*-M]$ increases under the presence of an absorbing barrier for the process. (Casanova effect)
4. **Left Tail Sample Insufficiency:** $E[M^*-M]$ increases with negative skeweness of the true underying variable.
5. **Asymmetry in Inference**: Under both negative skewness and fat tails, negative deviations from the mean are more informational than positive deviations.
6. **Power of Extreme Deviations (N=1 is OK)**: Under fat tails, large deviations from the mean are vastly more informational than small ones. They are not "anecdotal". (The last two properties corresponds to the black swan problem).

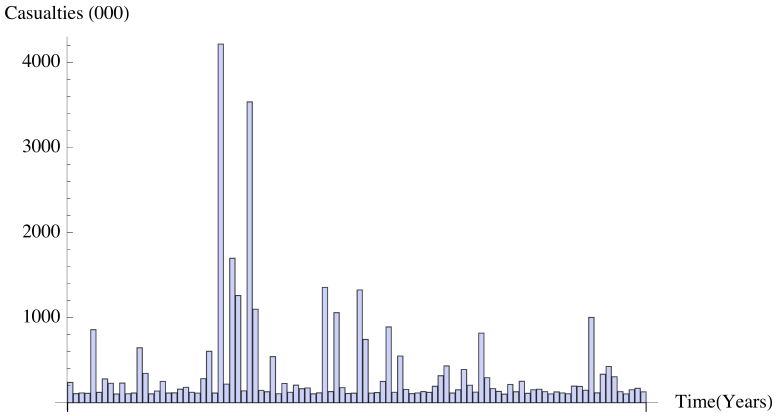


Fig1 **First 100 years (Sample Path):** A Monte Carlo generated realization of a process of the "80/20 or 80/02 style" as described by Pinker's in his book, that is tail exponent α= 1.1, dangerously close to 1

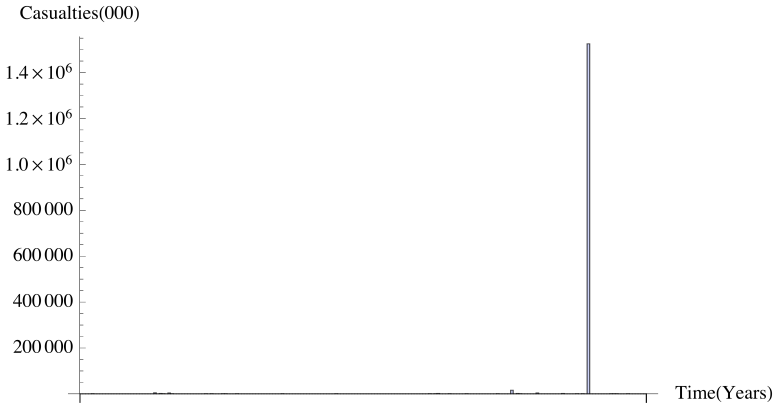


Fig 2: **The Turkey Surprise**: Now 200 years, the second 100 years dwarf the first; these are realizations of the exact same process, seen with a longer widow.

At the core, the estimator of the mean is NOT the mean, but rather the tail exponent (+ scale)

On the Super-Additivity and Estimation Biases of Quantile Contributions

Nassim Nicholas Taleb*, Raphael Douady†
*School of Engineering, New York University
†Riskdata & C.N.R.S. Paris, Labex ReFi, Centre d'Economie de la Sorbonne

***Abstract*—Sample measures of top centile contributions to the total (concentration) are downward biased, unstable estimators, extremely sensitive to sample size and concave in accounting for large deviations. It makes them particularly unfit in domains with power law tails, especially for low values of the exponent. These estimators can vary over time and increase with the population size, as shown in this article, thus providing the illusion of structural changes in concentration. They are also inconsistent under aggregation and mixing distributions, as the weighted average of concentration measures for A and B will tend to be lower than that from $A \cup B$. In addition, it can be shown that under such fat tails, increases in the total sum need to be accompanied by increased sample size of the concentration measurement. We examine the estimation superadditivity and bias under homogeneous and mixed distributions.**

Final version, Nov 11 2014. Accepted *Physica A:Statistical Mechanics and Applications*

This work was achieved through the Laboratory of Excellence on Financial Regulation (Labex ReFi) supported by PRES heSam under the reference ANR-10-LABX-0095.

I. Introduction

Vilfredo Pareto noticed that 80% of the land in Italy belonged to 20% of the population, and vice-versa, thus both giving birth to the power law class of distributions and the popular saying 80/20. The self-similarity at the core of the property of power laws [1] and [2] allows us to recurse and reapply the 80/20 to the remaining 20%, and so forth until one obtains the result that the top percent of the population will own about 53% of the total wealth.

It looks like such a measure of concentration can be seriously biased, depending on how it is measured, so it is very likely that the true ratio of concentration of what Pareto observed, that is, the share of the top percentile, was closer to 70%, hence changes year-on-year would drift higher to converge to such a level from larger sample. In fact, as we will show in this discussion, for, say wealth, more complete samples resulting from technological progress, and also larger population and economic growth will make such a measure converge by increasing over time, for no other reason than expansion in sample space or aggregate value.

The core of the problem is that, for the class one-tailed fat-tailed random variables, that is, bounded on the left and unbounded on the right, where the random variable $X \in [x_{\min}, \infty)$, the in-sample quantile contribution is a biased estimator of the true value of the actual quantile contribution.

Let us define the *quantile contribution*

$$\kappa_q = q\frac{\mathbb{E}[X|X > h(q)]}{\mathbb{E}[X]}$$

where $h(q) = \inf\{h \in [x_{min}, +\infty), \mathbb{P}(X > h) \leq q\}$ is the exceedance threshold for the probability q.

For a given sample $(X_k)_{1\leq k\leq n}$, its "natural" estimator $\widehat{\kappa}_q \equiv \frac{q^{th}\text{percentile}}{total}$, used in most academic studies, can be expressed, as

$$\widehat{\kappa}_q \equiv \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i > \hat{h}(q)} X_i}{\sum_{i=1}^{n} X_i}$$

where $\hat{h}(q)$ is the estimated exceedance threshold for the probability q :

$$\hat{h}(q) = \inf\{h : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x>h} \leq q\}$$

We shall see that the observed variable $\widehat{\kappa}_q$ is a downward biased estimator of the true ratio κ_q, the one that would hold out of sample, and such bias is in proportion to the fatness of tails and, for very fat tailed distributions, remains significant, even for very large samples.

II. Estimation For Unmixed Pareto-Tailed Distributions

Let X be a random variable belonging to the class of distributions with a "power law" right tail, that is:

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{1}$$

where $L : [x_{\min}, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x\to+\infty} \frac{L(kx)}{L(x)} = 1$ for any $k > 0$.

There is little difference for small exceedance quantiles (<50%) between the various possible distributions such as Student's t, Lévy α-stable, Dagum,[3],[4] Singh-Maddala distribution [5], or straight Pareto.

For exponents $1 \leq \alpha \leq 2$, as observed in [6], the law of large numbers operates, though *extremely* slowly. The problem is acute for α around, but strictly above 1 and severe, as it diverges, for $\alpha = 1$.

A. Bias and Convergence

1) Simple Pareto Distribution: Let us first consider $\phi_\alpha(x)$ the density of a α-Pareto distribution bounded from below by $x_{\min} > 0$, in other words: $\phi_\alpha(x) = \alpha x_{\min}^{\alpha} x^{-\alpha-1} \mathbb{1}_{x\geq x_{\min}}$, and

$\mathbb{P}(X > x) = \left(\frac{x_{\min}}{x}\right)^{\alpha}$. Under these assumptions, the cutpoint of exceedance is $h(q) = x_{\min} q^{-1/\alpha}$ and we have:

$$\kappa_q = \frac{\int_{h(q)}^{\infty} x\,\phi(x)dx}{\int_{x_{min}}^{\infty} x\,\phi(x)dx} = \left(\frac{h(q)}{x_{\min}}\right)^{1-\alpha} = q^{\frac{\alpha-1}{\alpha}} \tag{2}$$

If the distribution of X is α-Pareto only beyond a cut-point x_{cut}, which we assume to be below $h(q)$, so that we have $\mathbb{P}(X > x) = \left(\frac{\lambda}{x}\right)^{\alpha}$ for some $\lambda > 0$, then we still have $h(q) = \lambda q^{-1/\alpha}$ and

$$\kappa_q = \frac{\alpha}{\alpha - 1}\frac{\lambda}{\mathbb{E}[X]} q^{\frac{\alpha-1}{\alpha}}$$

The estimation of κ_q hence requires that of the exponent α as well as that of the scaling parameter λ, or at least its ratio to the expectation of X.

Table I shows the bias of $\widehat{\kappa}_q$ as an estimator of κ_q in the case of an α-Pareto distribution for $\alpha = 1.1$, a value chosen to be compatible with practical economic measures, such as the wealth distribution in the world or in a particular country, including developped ones.[1] In such a case, the estimator is extemely sensitive to *"small"* samples, *"small"* meaning in practice 10^8. We ran up to a trillion simulations across varieties of sample sizes. While $\kappa_{0.01} \approx 0.657933$, even a sample size of 100 million remains severely biased as seen in the table.

Naturally the bias is rapidly (and nonlinearly) reduced for α further away from 1, and becomes weak in the neighborhood of 2 for a constant α, though not under a mixture distribution for α, as we shall se later. It is also weaker outside the top 1% centile, hence this discussion focuses on the famed "one percent" and on low values of the α exponent.

TABLE I: Biases of Estimator of $\kappa = 0.657933$ From 10^{12} Monte Carlo Realizations

$\widehat{\kappa}(n)$	Mean	Median	STD across MC runs
$\widehat{\kappa}(10^3)$	0.405235	0.367698	0.160244
$\widehat{\kappa}(10^4)$	0.485916	0.458449	0.117917
$\widehat{\kappa}(10^5)$	0.539028	0.516415	0.0931362
$\widehat{\kappa}(10^6)$	0.581384	0.555997	0.0853593
$\widehat{\kappa}(10^7)$	0.591506	0.575262	0.0601528
$\widehat{\kappa}(10^8)$	0.606513	0.593667	0.0461397

In view of these results and of a number of tests we have performed around them, we can conjecture that the bias $\kappa_q - \widehat{\kappa}_q(n)$ is "of the order of" $c(\alpha, q)n^{-b(q)(\alpha-1)}$ where constants $b(q)$ and $c(\alpha, q)$ need to be evaluated. Simulations suggest that $b(q) = 1$, whatever the value of α and q, but the rather slow convergence of the estimator and of its standard deviation to 0 makes precise estimation difficult.

2) General Case: In the general case, let us fix the threshold h and define:

$$\kappa_h = P(X > h)\frac{\mathbb{E}[X|X > h]}{\mathbb{E}[X]} = \frac{\mathbb{E}[X\mathbb{1}_{X>h}]}{\mathbb{E}[X]}$$

[1]This value, which is lower than the estimated exponents one can find in the literature – around 2 – is, following [7], a lower estimate which cannot be excluded from the observations.

so that we have $\kappa_q = \kappa_{h(q)}$. We also define the n-sample estimator:

$$\widehat{\kappa}_h \equiv \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i>h} X_i}{\sum_{i=1}^{n} X_i}$$

where X_i are n independent copies of X. The intuition behind the estimation bias of κ_q by $\widehat{\kappa}_q$ lies in a difference of concavity of the concentration measure with respect to an innovation (a new sample value), whether it falls below or above the threshold. Let $A_h(n) = \sum_{i=1}^{n} \mathbb{1}_{X_i>h} X_i$ and $S(n) = \sum_{i=1}^{n} X_i$, so that $\widehat{\kappa}_h(n) = \frac{A_h(n)}{S(n)}$ and assume a frozen threshold h. If a new sample value $X_{n+1} < h$ then the new value is $\widehat{\kappa}_h(n+1) = \frac{A_h(n)}{S(n) + X_{n+1}}$. The value is convex in X_{n+1} so that uncertainty on X_{n+1} increases its expectation. At variance, if the new sample value $X_{n+1} > h$, the new value $\widehat{\kappa}_h(n+1) \approx \frac{A_h(n)+X_{n+1}-h}{S(n)+X_{n+1}-h} = 1 - \frac{S(n)-A_h(n)}{S(n)+X_{n+1}-h}$, which is now concave in X_{n+1}, so that uncertainty on X_{n+1} reduces its value. The competition between these two opposite effects is in favor of the latter, because of a higher concavity with respect to the variable, and also of a higher variability (whatever its measurement) of the variable conditionally to being above the threshold than to being below. The fatter the right tail of the distribution, the stronger the effect. Overall, we find that $\mathbb{E}[\widehat{\kappa}_h(n)] \leq \frac{\mathbb{E}[A_h(n)]}{\mathbb{E}[S(n)]} = \kappa_h$ (note that unfreezing the threshold $\hat{h}(q)$ also tends to reduce the concentration measure estimate, adding to the effect, when introducing one extra sample because of a slight increase in the expected value of the estimator $\hat{h}(q)$, although this effect is rather negligible). We have in fact the following:

Proposition 1. *Let $\boldsymbol{X} = (X)_{i=1}^{n}$ a random sample of size $n > \frac{1}{q}$, $Y = X_{n+1}$ an extra single random observation, and define: $\widehat{\kappa}_h(\boldsymbol{X} \sqcup Y) = \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i>h} X_i + \mathbb{1}_{Y>h} Y}{\sum_{i=1}^{n} X_i + Y}$. We remark that, whenever $Y > h$, one has:*

$$\frac{\partial^2 \widehat{\kappa}_h(\boldsymbol{X} \sqcup Y)}{\partial Y^2} \leq 0.$$

This inequality is still valid with $\widehat{\kappa}_q$ as the value $\hat{h}(q, \boldsymbol{X} \sqcup Y)$ doesn't depend on the particular value of $Y > \hat{h}(q, \boldsymbol{X})$.

We face a different situation from the common small sample effect resulting from high impact from the rare observation in the tails that are less likely to show up in small samples, a bias which goes away by repetition of sample runs. The concavity of the estimator constitutes a upper bound for the measurement in finite n, clipping large deviations, which leads to problems of aggregation as we will state below in Theorem 1. In practice, even in very large sample, the contribution of very large rare events to κ_q slows down the convergence of the sample estimator to the true value. For a better, unbiased estimate, one would need to use a different path: first estimating the distribution parameters $\left(\hat{\alpha}, \hat{\lambda}\right)$ and only then, estimating the theoretical tail contribution $\kappa_q(\hat{\alpha}, \hat{\lambda})$. Falk [7] observes that, even with a proper estimator of α and λ, the convergence is extremely slow, namely of the order of $n^{-\delta}/\ln n$, where the exponent δ depends on α and on the

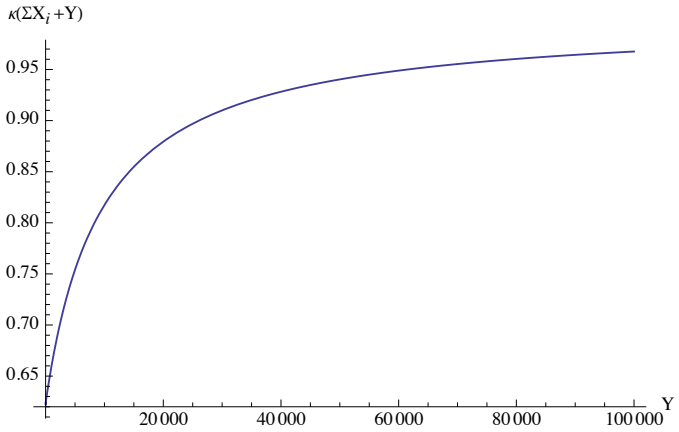


Fig. 1: Effect of additional observations on κ

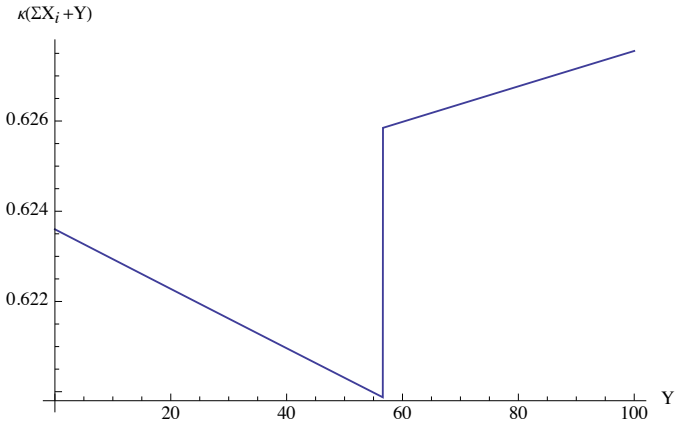


Fig. 2: Effect of additional observations on κ, we can see convexity on both sides of h except for values of no effect to the left of h, an area of order $1/n$

tolerance of the actual distribution vs. a theoretical Pareto, measured by the Hellinger distance. In particular, $\delta \to 0$ as $\alpha \to 1$, making the convergence really slow for low values of α.

III. An Inequality About Aggregating Inequality

For the estimation of the mean of a fat-tailed r.v. $(X)_i^j$, in m sub-samples of size n_i each for a total of $n = \sum_{i=1}^{m} n_i$, the allocation of the total number of observations n between i and j does not matter so long as the total n is unchanged. Here the allocation of n samples between m sub-samples does matter because of the concavity of κ.[2] Next we prove that global concentration as measured by $\widehat{\kappa}_q$ on a broad set of data will appear higher than local concentration, so aggregating European data, for instance, would give a $\widehat{\kappa}_q$ higher than the average measure of concentration across countries – an *"inequality about inequality"*. In other words, we claim that the estimation bias when using $\widehat{\kappa}_q(n)$ is even increased when dividing the sample into sub-samples and taking the weighted average of the measured values $\widehat{\kappa}_q(n_i)$.

[2]The same concavity – and general bias – applies when the distribution is lognormal, and is exacerbated by high variance.

Theorem 1. *Partition the n data into m sub-samples $N = N_1 \cup \ldots \cup N_m$ of respective sizes $n_1, \ldots, n_m$, with $\sum_{i=1}^{m} n_i = n$, and let $S_1, \ldots, S_m$ be the sum of variables over each sub-sample, and $S = \sum_{i=1}^{m} S_i$ be that over the whole sample. Then we have:*

$$\mathbb{E}\left[\widehat{\kappa}_q(N)\right] \geq \sum_{i=1}^{m} \mathbb{E}\left[\frac{S_i}{S}\right] \mathbb{E}\left[\widehat{\kappa}_q(N_i)\right]$$

If we further assume that the distribution of variables X_j is the same in all the sub-samples. Then we have:

$$\mathbb{E}\left[\widehat{\kappa}_q(N)\right] \geq \sum_{i=1}^{m} \frac{n_i}{n} \mathbb{E}\left[\widehat{\kappa}_q(N_i)\right]$$

In other words, averaging concentration measures of sub-samples, weighted by the total sum of each subsample, produces a downward biased estimate of the concentration measure of the full sample.

Proof. An elementary induction reduces the question to the case of two sub-samples. Let $q \in (0,1)$ and $(X_1, \ldots, X_m)$ and $(X'_1, \ldots, X'_n)$ be two samples of positive i.i.d. random variables, the X_i's having distributions $p(dx)$ and the X'_j's having distribution $p'(dx')$. For simplicity, we assume that both qm and qn are integers. We set $S = \sum_{i=1}^{m} X_i$ and $S' = \sum_{i=1}^{n} X'_i$. We define $A = \sum_{i=1}^{mq} X_{[i]}$ where $X_{[i]}$ is the i-th largest value of $(X_1, \ldots, X_m)$, and $A' = \sum_{i=1}^{mq} X'_{[i]}$ where $X'_{[i]}$ is the i-th largest value of $(X'_1, \ldots, X'_n)$. We also set $S'' = S + S'$ and $A" = \sum_{i=1}^{(m+n)q} X''_{[i]}$ where $X''_{[i]}$ is the i-th largest value of the joint sample $(X_1, \ldots, X_m, X'_1, \ldots, X'_n)$.

The q-concentration measure for the samples $\boldsymbol{X} = (X_1, ..., X_m)$, $\boldsymbol{X}' = (X'_1, ..., X'_n)$ and $\boldsymbol{X}'' = (X_1, \ldots, X_m, X'_1, \ldots, X'_n)$ are:

$$\kappa = \frac{A}{S} \qquad \kappa' = \frac{A'}{S'} \qquad \kappa'' = \frac{A''}{S''}$$

We must prove that he following inequality holds for expected concentration measures:

$$\mathbb{E}\left[\kappa''\right] \geq \mathbb{E}\left[\frac{S}{S''}\right] \mathbb{E}\left[\kappa\right] + \mathbb{E}\left[\frac{S'}{S''}\right] \mathbb{E}\left[\kappa'\right]$$

We observe that:

$$A = \max_{\substack{J \subset \{1,\ldots,m\} \\ |J| = \theta m}} \sum_{i \in J} X_i$$

and, similarly $A' = \max_{J' \subset \{1,\ldots,n\}, |J'| = qn} \sum_{i \in J'} X'_i$ and $A'' = \max_{J'' \subset \{1,\ldots,m+n\}, |J''| = q(m+n)} \sum_{i \in J''} X_i$, where we have denoted $X_{m+i} = X'_i$ for $i = 1 \ldots n$. If $J \subset \{1, ..., m\}, |J| = \theta m$ and $J' \subset \{m+1, ..., m+n\}, |J'| = qn$, then $J'' = J \cup J'$ has cardinal $m+n$, hence $A + A' = \sum_{i \in J''} X_i \leq A''$, whatever the particular sample. Therefore $\kappa'' \geq \frac{S}{S''}\kappa + \frac{S'}{S''}\kappa'$ and we have:

$$\mathbb{E}\left[\kappa''\right] \geq \mathbb{E}\left[\frac{S}{S''}\kappa\right] + \mathbb{E}\left[\frac{S'}{S''}\kappa'\right]$$

Let us now show that:

$$\mathbb{E}\left[\frac{S}{S''}\kappa\right] = \mathbb{E}\left[\frac{A}{S''}\right] \geq \mathbb{E}\left[\frac{S}{S''}\right]\mathbb{E}\left[\frac{A}{S}\right]$$

If this is the case, then we identically get for κ' :

$$\mathbb{E}\left[\frac{S'}{S''}\kappa'\right] = \mathbb{E}\left[\frac{A'}{S''}\right] \geq \mathbb{E}\left[\frac{S'}{S''}\right]\mathbb{E}\left[\frac{A'}{S'}\right]$$

hence we will have:

$$\mathbb{E}\left[\kappa''\right] \geq \mathbb{E}\left[\frac{S}{S''}\right]\mathbb{E}\left[\kappa\right] + \mathbb{E}\left[\frac{S'}{S''}\right]\mathbb{E}\left[\kappa'\right]$$

Let $T = X_{\lfloor mq \rfloor}$ be the cut-off point (where $\lfloor mq \rfloor$ is the integer part of mq), so that $A = \sum_{i=1}^{m} X_i \mathbb{1}_{X_i \geq T}$ and let $B = S - A = \sum_{i=1}^{m} X_i \mathbb{1}_{X_i < T}$. Conditionally to T, A and B are independent: A is a sum if $m\theta$ samples constarined to being above T, while B is the sum of $m(1-\theta)$ independent samples constrained to being below T. They are also independent of S'. Let $p_A(t, da)$ and $p_B(t, db)$ be the distribution of A and B respectively, given $T = t$. We recall that $p'(ds')$ is the distribution of S' and denote $q(dt)$ that of T. We have:

$$\mathbb{E}\left[\frac{S}{S''}\kappa\right] = \iint \frac{a+b}{a+b+s'}\frac{a}{a+b} p_A(t,da)\, p_B(t,db)\, q(dt)\, p'(ds')$$

For given b, t and s', $a \to \frac{a+b}{a+b+s'}$ and $a \to \frac{a}{a+b}$ are two increasing functions of the same variable a, hence conditionally to T, B and S', we have:

$$\mathbb{E}\left[\frac{S}{S''}\kappa \middle| T, B, S'\right] = \mathbb{E}\left[\frac{A}{A+B+S'} \middle| T, B, S'\right] \geq \mathbb{E}\left[\frac{A+B}{A+B+S'} \middle| T, B, S'\right]\mathbb{E}\left[\frac{A}{A+B} \middle| T, B, S'\right]$$

This inequality being valid for any values of T, B and S', it is valid for the unconditional expectation, and we have:

$$\mathbb{E}\left[\frac{S}{S''}\kappa\right] \geq \mathbb{E}\left[\frac{S}{S''}\right]\mathbb{E}\left[\frac{A}{S}\right]$$

If the two samples have the same distribution, then we have:

$$\mathbb{E}\left[\kappa''\right] \geq \frac{m}{m+n}\mathbb{E}\left[\kappa\right] + \frac{n}{m+n}\mathbb{E}\left[\kappa'\right]$$

Indeed, in this case, we observe that $\mathbb{E}\left[\frac{S}{S''}\right] = \frac{m}{m+n}$. Indeed $S = \sum_{i=1}^{m} X_i$ and the X_i are identically distributed, hence $\mathbb{E}\left[\frac{S}{S''}\right] = m\mathbb{E}\left[\frac{X}{S''}\right]$. But we also have $\mathbb{E}\left[\frac{S''}{S''}\right] = 1 = (m+n)\mathbb{E}\left[\frac{X}{S''}\right]$ therefore $\mathbb{E}\left[\frac{X}{S''}\right] = \frac{1}{m+n}$. Similarly, $\mathbb{E}\left[\frac{S'}{S''}\right] = \frac{n}{m+n}$, yielding the result.

This ends the proof of the theorem. □

Let X be a positive random variable and $h \in (0,1)$. We remind the theoretical h-concentration measure, defined as:

$$\kappa_h = \frac{P(X > h)\mathbb{E}\left[X \,|X > h\right]}{\mathbb{E}\left[X\right]}$$

whereas the n-sample θ-concentration measure is $\widehat{\kappa}_h(n) = \frac{A(n)}{S(n)}$, where $A(n)$ and $S(n)$ are defined as above for an n-sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of i.i.d. variables with the same distribution as X.

Theorem 2. *For any $n \in \mathbb{N}$, we have:*

$$\mathbb{E}\left[\widehat{\kappa}_h(n)\right] < \kappa_h$$

and

$$\lim_{n \to +\infty} \widehat{\kappa}_h(n) = \kappa_h \quad \textit{a.s. and in probability}$$

Proof. The above corrolary shows that the sequence $n\mathbb{E}\left[\widehat{\kappa}_h(n)\right]$ is super-additive, hence $\mathbb{E}\left[\widehat{\kappa}_h(n)\right]$ is an increasing sequence. Moreover, thanks to the law of large numbers, $\frac{1}{n}S(n)$ converges almost surely and in probability to $\mathbb{E}\left[X\right]$ and $\frac{1}{n}A(n)$ converges almost surely and in probability to $\mathbb{E}\left[X\mathbb{1}_{X>h}\right] = P(X > h)\mathbb{E}\left[X \,|X > h\right]$, hence their ratio also converges almost surely to κ_h. On the other hand, this ratio is bounded by 1. Lebesgue dominated convergence theorem concludes the argument about the convergence in probability. □

IV. Mixed Distributions For The Tail Exponent

Consider now a random variable X, the distribution of which $p(dx)$ is a mixture of parametric distributions with different values of the parameter: $p(dx) = \sum_{i=1}^{m} \omega_i p_{\alpha_i}(dx)$. A typical n-sample of X can be made of $n_i = \omega_i n$ samples of X_{α_i} with distribution p_{α_i}. The above theorem shows that, in this case, we have:

$$\mathbb{E}\left[\widehat{\kappa}_q(n, X)\right] \geq \sum_{i=1}^{m} \mathbb{E}\left[\frac{S(\omega_i n, X_{\alpha_i})}{S(n, X)}\right]\mathbb{E}\left[\widehat{\kappa}_q(\omega_i n, X_{\alpha_i})\right]$$

When $n \to +\infty$, each ratio $\dfrac{S(\omega_i n, X_{\alpha_i})}{S(n, X)}$ converges almost surely to ω_i respectively, therefore we have the following convexity inequality:

$$\kappa_q(X) \geq \sum_{i=1}^{m} \omega_i \kappa_q(X_{\alpha_i})$$

The case of Pareto distribution is particularly interesting. Here, the parameter α represents the tail exponent of the distribution. If we normalize expectations to 1, the cdf of X_α is $F_\alpha(x) = 1 - \left(\frac{x}{x_{\min}}\right)^{-\alpha}$ and we have:

$$\kappa_q(X_\alpha) = q^{\frac{\alpha-1}{\alpha}}$$

and

$$\frac{d^2}{d\alpha^2}\kappa_q(X_\alpha) = q^{\frac{\alpha-1}{\alpha}}\frac{(\log q)^2}{\alpha^3} > 0$$

Hence $\kappa_q(X_\alpha)$ is a convex function of α and we can write:

$$\kappa_q(X) \geq \sum_{i=1}^{m} \omega_i \kappa_q(X_{\alpha_i}) \geq \kappa_q(X_{\bar{\alpha}})$$

where $\bar{\alpha} = \sum_{i=1}^{m} \omega_i \alpha$.

Suppose now that X is a positive random variable with unknown distribution, except that its tail decays like a power low with unknown exponent. An unbiased estimation of the

exponent, with necessarily some amount of uncertainty (i.e., a distribution of possible true values around some average), would lead to a downward biased estimate of κ_q.

Because the concentration measure only depends on the tail of the distribution, this inequality also applies in the case of a mixture of distributions with a power decay, as in Equation 1:

$$\mathbb{P}(X > x) \sim \sum_{j=1}^{N} \omega_i L_i(x) x^{-\alpha_j} \tag{3}$$

The slightest uncertainty about the exponent increases the concentration index. One can get an actual estimate of this bias by considering an average $\bar{\alpha} > 1$ and two surrounding values $\alpha^+ = \alpha + \delta$ and $\alpha^- = \alpha - \delta$. The convexity inequaly writes as follows:

$$\kappa_q(\bar{\alpha}) = q^{1-\frac{1}{\bar{\alpha}}} < \frac{1}{2}\left(q^{1-\frac{1}{\alpha+\delta}} + q^{1-\frac{1}{\alpha-\delta}}\right)$$

So in practice, an estimated $\bar{\alpha}$ of around $3/2$, sometimes called the "half-cubic" exponent, would produce similar results as value of α much closer ro 1, as we used in the previous section. Simply $\kappa_q(\alpha)$ is convex, and dominated by the second order effect $\frac{\ln(q) q^{1-\frac{1}{\alpha+\delta}} (\ln(q) - 2(\alpha+\delta))}{(\alpha+\delta)^4}$, an effect that is exacerbated at lower values of α.

To show how unreliable the measures of inequality concentration from quantiles, consider that a standard error of 0.3 in the measurement of α causes $\kappa_q(\alpha)$ to rise by 0.25.

V. A Larger Total Sum is Accompanied by Increases in $\widehat{\kappa}_q$

There is a large dependence between the estimator $\widehat{\kappa}_q$ and the sum $S = \sum_{j=1}^{n} X_j$: conditional on an increase in $\widehat{\kappa}_q$ the expected sum is larger. Indeed, as shown in theorem 1, $\widehat{\kappa}_q$ and S are positively correlated.

For the case in which the random variables under concern are wealth, we observe as in Figure 3 such conditional increase; in other words, since the distribution is of the class of fat tails under consideration, the maximum is of the same order as the sum, additional wealth means more measured inequality. Under such dynamics, is quite absurd to assume that additional wealth will arise from the bottom or even the middle. (The same argument can be applied to wars, epidemics, size or companies, etc.)

VI. Conclusion and Proper Estimation of Concentration

Concentration can be high at the level of the generator, but in small units or subsections we will observe a lower κ_q. So examining times series, we can easily get a historical illusion of rise in, say, wealth concentration when it has been there all along at the level of the process; and an expansion in the size of the unit measured can be part of the explanation.[3]

[3]Accumulated wealth is typically thicker tailed than income, see [8].

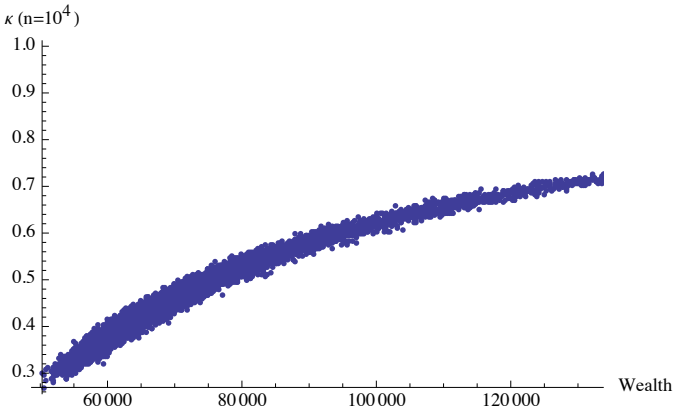


Fig. 3: Effect of additional wealth on $\hat{\kappa}$

Even the estimation of α can be biased in some domains where one does not see the entire picture: in the presence of uncertainty about the "true" α, it can be shown that, unlike other parameters, the one to use is not the probability-weighted exponents (the standard average) but rather the minimum across a section of exponents [6].

One must not perform analyses of year-on-year changes in $\widehat{\kappa}_q$ without adjustment. It did not escape our attention that some theories are built based on claims of such "increase" in inequality, as in [9], without taking into account the true nature of κ_q, and promulgating theories about the "variation" of inequality without reference to the stochasticity of the estimation – and the lack of consistency of κ_q across time and sub-units. What is worse, rejection of such theories also ignored the size effect, by countering with data of a different sample size, effectively making the dialogue on inequality uninformational statistically.[4]

The mistake appears to be commonly made in common inference about fat-tailed data in the literature. The very methodology of using concentration and changes in concentration is highly questionable. For instance, in the thesis by Steven Pinker [10] that the world is becoming less violent, we note a fallacious inference about the concentration of damage from wars from a $\widehat{\kappa}_q$ with minutely small population in relation to the fat-tailedness.[5] Owing to the fat-tailedness of war casualties and consequences of violent conflicts, an adjustment would rapidly invalidate such claims that violence from war has statistically experienced a decline.

A. Robust methods and use of exhaustive data

We often face argument of the type "the method of measuring concentration from quantile contributions $\hat{\kappa}$ is robust and based on a complete set of data". Robust methods, alas,

[4]*Financial Times*, May 23, 2014 "Piketty findings undercut by errors" by Chris Giles.

[5]Using Richardson's data, [10]: "(Wars) followed an 80:2 rule: almost eighty percent of the deaths were caused by *two percent* (his emph.) of the wars". So it appears that both Pinker and the literature cited for the quantitative properties of violent conflicts are using a flawed methodology, one that produces a severe bias, as the centile estimation has extremely large biases with fat-tailed wars. Furthermore claims about the mean become spurious at low exponents.

tend to fail with fat-tailed data, see [6]. But, in addition, the problem here is worse: even if such "robust" methods were deemed unbiased, a method of direct centile estimation is still linked to a static and specific population and does not aggregage. Accordingly, such techniques do not allow us to make statistical claims or scientific statements about the true properties which should necessarily carry out of sample.

Take an insurance (or, better, reinsurance) company. The "accounting" profits in a year in which there were few claims do not reflect on the "economic" status of the company and it is futile to make statements on the concentration of losses per insured event based on a single year sample. The "accounting" profits are not used to predict variations year-on-year, rather the exposure to tail (and other) events, analyses that take into account the stochastic nature of the performance. This difference between "accounting" (deterministic) and "economic" (stochastic) values matters for policy making, particularly under fat tails. The same with wars: we do not estimate the severity of a (future) risk based on past in-sample historical data.

B. How Should We Measure Concentration?

Practitioners of risk managers now tend to compute CVaR and other metrics, methods that are extrapolative and nonconcave, such as the information from the α exponent, taking the one closer to the lower bound of the range of exponents, as we saw in our extension to Theorem 2 and rederiving the corresponding κ, or, more rigorously, integrating the functions of α across the various possible states. Such methods of adjustment are less biased and do not get mixed up with problems of aggregation –they are similar to the "stochastic volatility" methods in mathematical finance that consist in adjustments to option prices by adding a "smile" to the standard deviation, in proportion to the variability of the parameter representing volatility and the errors in its measurement. Here it would be "stochastic alpha" or "stochastic tail exponent"[6] By extrapolative, we mean the built-in extension of the tail in the measurement by taking into account realizations outside the sample path that are in excess of the extrema observed.[7] [8]

ACKNOWLEDGMENT

The late Benoît Mandelbrot, Branko Milanovic, Dominique Guéguan, Felix Salmon, Bruno Dupire, the late Marc Yor, Albert Shiryaev, an anonymous referee, the staff at Luciano Restaurant in Brooklyn and Naya in Manhattan.

[6]Also note that, in addition to the centile estimation problem, some authors such as [11] when dealing with censored data, use Pareto interpolation for unsufficient information about the tails (based on tail parameter), filling-in the bracket with conditional average bracket contribution, which is not the same thing as using full power-law extension; such a method retains a significant bias.

[7]Even using a lognormal distribution, by fitting the scale parameter, works to some extent as a rise of the standard deviation extrapolates probability mass into the right tail.

[8]We also note that the theorems would also apply to Poisson jumps, but we focus on the powerlaw case in the application, as the methods for fitting Poisson jumps are interpolative and have proved to be easier to fit in-sample than out of sample, see [6].

REFERENCES

[1] B. Mandelbrot, "The pareto-levy law and the distribution of income," *International Economic Review*, vol. 1, no. 2, pp. 79–106, 1960.
[2] ——, "The stable paretian income distribution when the apparent exponent is near two," *International Economic Review*, vol. 4, no. 1, pp. 111–115, 1963.
[3] C. Dagum, "Inequality measures between income distributions with applications," *Econometrica*, vol. 48, no. 7, pp. 1791–1803, 1980.
[4] ——, *Income distribution models*. Wiley Online Library, 1983.
[5] S. Singh and G. Maddala, "A function for size distribution of incomes: reply," *Econometrica*, vol. 46, no. 2, 1978.
[6] N. N. Taleb, "Silent risk: Lectures on probability, vol 1," *Available at SSRN 2392310*, 2015.
[7] M. Falk *et al.*, "On testing the extreme value index via the pot-method," *The Annals of Statistics*, vol. 23, no. 6, pp. 2013–2035, 1995.
[8] X. Gabaix, "Power laws in economics and finance," National Bureau of Economic Research, Tech. Rep., 2008.
[9] T. Piketty, "Capital in the 21st century," 2014.
[10] S. Pinker, *The better angels of our nature: Why violence has declined*. Penguin, 2011.
[11] T. Piketty and E. Saez, "The evolution of top incomes: a historical and international perspective," National Bureau of Economic Research, Tech. Rep., 2006.

The Law of Large Numbers Under Fat Tails

Nassim Nicholas Taleb
Tandon School of Engineering, New York University and Real World Risk Institute, LLC.

I. Introduction

You observe data and get some confidence that the average is represented by the sample thanks to a standard metrified "n". Now what if the data were fat tailed? How much more do you need? What if the model were uncertain –we had uncertainty about the parameters or the probability distribution itself? Let us call "sample equivalence" the sample size that is needed to correspond to a Gaussian sample size of n.

It appears that 1) the statistical literature has been silent on the subject of sample equivalence –since the sample mean is not a good estimator under fat tailed distributions, 2) errors in the estimation of the mean can be several order of magnitudes higher than under corresponding thin tails, 3) many operators writing "scientific" papers aren't aware of it (which includes many statisticians), 4) model error compounds the issue.

We show that fitting tail exponents via ML methods have a small error in delivering the mean.

Main Technical Results In addition to the qualitative discussions about commonly made errors in violating the sample equivalence, the technical contribution is as follows:

- explicit extractions of partial expectations for alpha stable distributions
- the expression of how uncertainty about parameters (quantified in terms of parameter volatility) translates into a larger (or smaller) required n. In other words, the effect of model uncertainty, how the degree of model uncertainty worsens inference, in a quantifiable way.

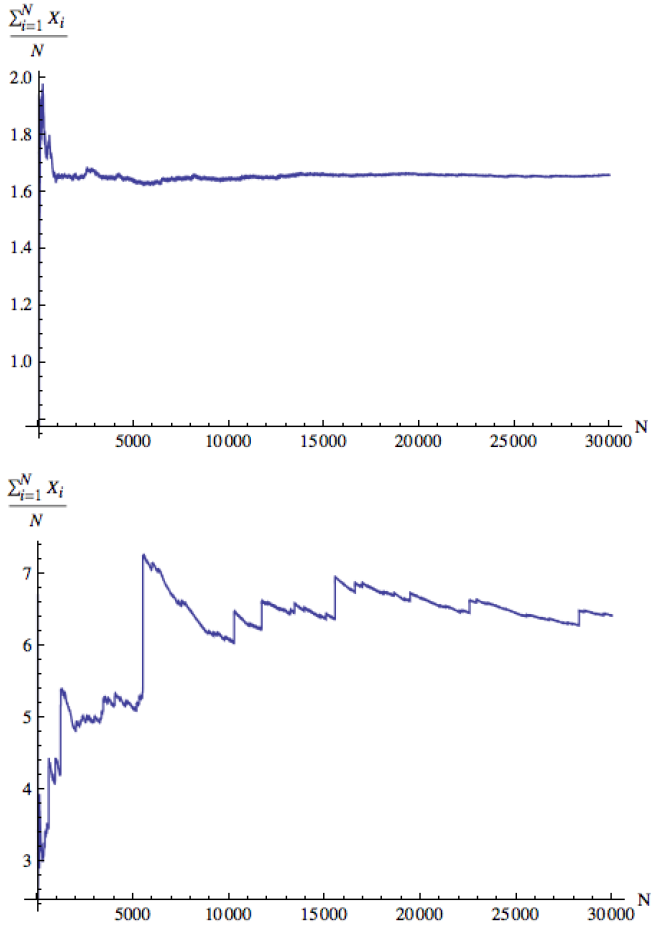


Fig. 1: How thin tails (Gaussian) and fat tails (1< α $\leq$2) converge to the mean.

II. Summary of the first result

The first discussion examines the issue of "sample equivalence" without any model uncertainty.

A. The problem

Let us summarize the standard convergence theorem. By the weak law of large numbers, a sum of random variables $X_1, \ldots, X_n$ with finite mean m, that is $\mathbb{E}(X) < \infty$, then $\frac{1}{n}\sum_{1\leq i\leq n} X_i$ converges to m in probability, as n $\to \infty$. Or, for any $\epsilon > 0$ $\lim_{n\to\infty} \mathbb{P}\big(|\overline{X}_n - m| > \epsilon\big) = 0$. In other words: the sample mean will end up converging to the true mean, should the latter exist.

But the result holds at infinity, while we live with finite n. There are several regimes of convergene.

- **Case 1a** when the variance and all other moments exist, and the data is i.i.d., there are two convergence effects at play, one, convergence to the Gaussian (by central limit), the second, the l.l.n., which accelerates the convergence. Some subcategories with higher kurtosis than the Gaussian, such as regime switching situations, or distributions entailing Poisson jumps or similar large deviations with small probability converge more slowly but these are special cases that we can ignore in this discussion since Case 2 is vastly more consequential in effect (it requires an extremely high kurtosis to slow down the central limit).
- **Case 1b** when the variance exists, but higher moments don't, the central limit theorem doesn't really work in practice (it is too slow for "real time") and the law of large numbers works more slowly than Case 1a, but works nevertheless. We consider this as "intermediate" case, more particularly with finite-variance power laws, those with the tail exponent ≥ 2 (or, more accurately, if the distribution is two-tailed, the lower of the left or right

tail exponent equal to or exceeding 2).

- **Case 2** when the mean exists, but the variance doesn't, the law of large numbers converges very, very slowly.

It is Case 2 that is the main object of this paper. More particularly cases where the lowest tail exponent $1 < \alpha \leq 2$. Of particular relevance is "80/20" where the $\alpha \approx 1.16$.

B. Discussion of the result about sample equivalence for fat tails

We assume that Case 1a converge to a Gaussian, hence approach the "Gaussian basin" which is the special case of stable distributions.

Table I shows the equivalence of number of summands between processes.

TABLE I: Corresponding n_α, or how many for equivalent α-stable distribution. The Gaussian case is the $\alpha = 2$. For the case with equivalent tails to the 80/20 one needs 10^{11} more data than the Gaussian.

α	n_α Symmetric	$n_\alpha^{\beta=\pm\frac{1}{2}}$ Skewed	$n_\alpha^{\beta=\pm 1}$ One-tailed
1	Fughedaboudit	-	-
$\frac{9}{8}$	6.09×10^{12}	2.8×10^{13}	1.86×10^{14}
$\frac{5}{4}$	574,634	895,952	1.88×10^{6}
$\frac{11}{8}$	5,027	6,002	8,632
$\frac{3}{2}$	567	613	737
$\frac{13}{8}$	165	171	186
$\frac{7}{4}$	75	77	79
$\frac{15}{8}$	44	44	44
2	30.	30	30

The "equivalence" is not straightforward.

Exposition of the problem

Let $X_{\alpha,1}, X_{\alpha,2}, \ldots, X_{\alpha,n_\alpha}$ be a sequence of i.i.d. powerlaw distributed variables with tail exponent $1 < \alpha \leq 2$ in at least one of the tails, that is, belonging to the class of distributions with at least one "power law" tail, that is:

$$\mathbb{P}(|X_\alpha| > |x|) \sim L(x)\,|x|^{-\alpha} \tag{1}$$

where $L : [x_0, \pm\infty) \to (0, \pm\infty)$ is a slowly varying function, defined as $\lim_{x\to\pm\infty} \frac{L(kx)}{L(x)} = 1$ for any $k > 0$.

Let $X_{g,1}, X_{g,2}, \ldots, X_{g,n}$ be a sequence of Gaussian variables with mean μ and scale σ. We are looking for values of n' corresponding to a given n_g:

$$n_{\min} = \inf\left\{ n_\alpha : \mathbb{E}\left(\left|\sum_{i=1}^{n_\alpha} \frac{X_{\alpha,i} - m_p}{n_\alpha}\right|\right) \leq \mathbb{E}\left(\left|\sum_{i=1}^{n_g} \frac{X_{g,i} - m_g}{n_g}\right|\right), n_\alpha > 0 \right\} \tag{2}$$

Instability of Mean Deviation and use of L^1 norm

And since we know that convergence for the Gaussian happens at speed $n_g^{\frac{1}{2}}$ (something we will redo using stable distributions), we can compare to convergence of other classes.

> The idea is to limit convergence to L^1 norm; we know clearly that there is no point using the L^2 norm, and even when (as in finite variance power laws, there is some convergence in L^2 (central limit), we ignore such situation for its difficulties in real time. As to the distribution of the maximum, that is, L^∞, fughedoubadit.

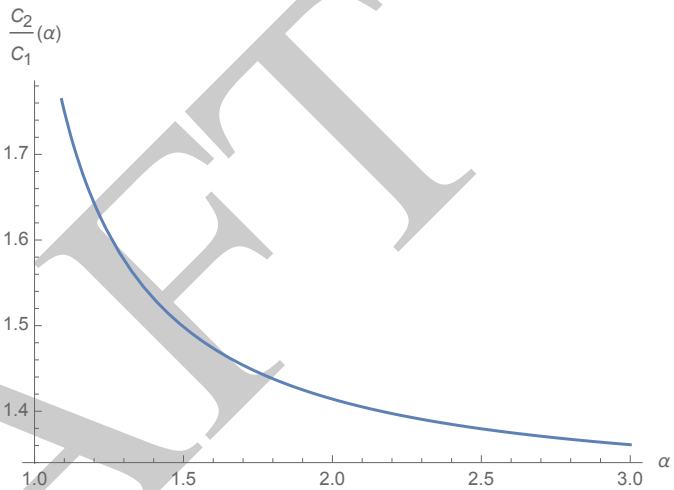


Fig. 2: The ratio of cumulants $\frac{C_2}{C_1}$ for a symmetric powerlaw, as a function of the tail exponent α.

We are expressing in Equation 2 the expected error (that is, a risk function) in $\mathbf{L}^1$ as mean absolute deviation from the observed average, to accommodate absence of variance –but assuming of course existence of first moment without which there is no point discussing averages.

Typically, in statistical inference, one uses standard deviations of the observations to establish the sufficiency of n. But in fat tailed data standard deviations do not exist, or, worse, when they exist, as in powerlaw with tail exponent > 3, they are extremely unstable, particularly in cases where kurtosis is infinite.

Using mean deviations of the samples (when these exist) doesn't accommodate the fact that fat tailed data hide properties. The "volatility of volatility", or the dispersion around the mean deviation increases nonlinearly as the tails get fatter. For instance, a stable distribution with tail exponent at $\frac{3}{2}$ matched to exactly the same mean deviation as the Gaussian will deliver measurements of mean deviation 1.4 times as unstable as the Gaussian.

Using mean absolute deviation for "volatility", and its mean deviation "volatility of volatility" expressed in the $\mathbf{L}^1$ norm, or C_1 and C_2 cumulant:

$$C_1 = \|.\|_1 = \mathbb{E}(|X - m|)$$

$$C_2 = \|(\|.\|_1)\|_1 = \mathbb{E}\left(|X - \mathbb{E}(|X - m|)|\right)$$

We can compare that matching mean deviations does not go very far matching cumulants.(see Appendix 1)

Further, a sum of Gaussian variables will have its extreme values distributed as a Gumbel while a sum of fat tailed will follow a Fréchet distribution *regardless of the the number of summands*. The difference is not trivial, as shown in figures , as in 10^6 realizations for an average with 100 summands, we can be expected observe maxima $> 4000\times$ the average while for a Gausthsian we can hardly encounter more than $> 5\times$.

III. Generalizing Mean Deviation as Partial Expectation

It is unfortunate that even if one matches mean deviations, the dispersion of the distributions of the mean deviations (and their skewness) would be such that a "tail" would remain markedly different in spite of a number of summands that allows the matching of the first order cumulant $\|.\|_1$. So we can match the special part of the distribution, the expectation $> K$ or $< K$, where K can be any arbitrary level.

Let $\Psi(t)$ be the characteristic function of the random variable. Let θ be the Heaviside theta function. Since $\text{sgn}(x) = 2\theta(x) - 1$

$$\Psi^{\theta,K}(t) = \int_{-\infty}^{\infty} e^{itx}\left(2\theta(x-K)-1\right)\,\mathrm{d}x = \frac{2ie^{iKt}}{t}$$

And define the partial expectation as $\mathbb{E}_K^+ := \int_K^\infty x\,\mathrm{d}F(x) = \mathbb{E}(X|_{X>K})\mathbb{P}(X>K)$. The special expectation becomes, by convoluting the Fourier transforms; where F is the distribution function for x:

$$\mathbb{E}_K^+ = -i\frac{\partial}{\partial t}\int_{-\infty}^{\infty}\Psi(t-u)\Psi^{\theta,K}(u)\mathrm{d}u|_{t=0} \tag{3}$$

Our method allows the computation of a conditional tail or "CVar" in the language of finance and insurance.

Note a similar approach using the Hilbert Transform for the absolute value of a Lévy stable r.v., see Hlusel, [1], Pinelis [2].

Mean deviation (under a symmetric distribution with mean μ, i.e. $\mathbb{P}(X>\mu) = \frac{1}{2}$) becomes a special case of equation 3, $\mathbb{E}(|X-\mu|) = \left(\int_\mu^\infty (x-\mu)\,\mathrm{d}F(x) - \int_{-\infty}^{\mu}(x-\mu)\,\mathrm{d}F(x)\right) = \mathbb{E}_\mu^+$.

IV. Class of Stable Distributions

Assume alpha-stable the class $\mathfrak{S}$ of probability distribution that is closed under convolution: $\mathbf{S}(\alpha,\beta,\mu,\sigma)$ represents the stable distribution with tail index $\alpha \in (0,2]$, symmetry parameter $\beta \in [0,1]$, location parameter $\mu \in \mathbb{R}$, and scale parameter $\sigma \in \mathbb{R}^+$. The Generalized Central Limit Theorem gives sequences a_n and b_n such that the distribution of the shifted and rescaled sum $Z_n = \left(\sum_i^n X_i - a_n\right)/b_n$ of n i.i.d. random variates X_i the distribution function of which $F_X(x)$ has asymptotes $1 - cx^{-\alpha}$ as $x \to +\infty$ and $d(-x)^{-\alpha}$ as $x \to -\infty$ weakly converges to the stable distribution

$$S(\wedge_{\alpha,2}, \mathbb{1}_{0<\alpha<2}\frac{c-d}{c+d}, 0, 1).$$

We note that the characteristic functions are real for all symmetric distributions. [We also note that the convergence is not clear across papers [3] but this doesn't apply to symmetric distributions.]

Note that the tail exponent α used in non stable cases is somewhat, but not fully, different for $\alpha = 2$, the Gaussian case where it ceases to be a powerlaw –the main difference is in the asymptotic interpretation. But for convention we retain the same symbol as it corresponds to tail exponent but use it differently in more general non-stable power law contexts.

The characteristic function $\Psi(t)$ of a variable X^α with scale σ will be, using the expression for $\alpha > 1$, See Zolotarev [4], Samorodnitsky and Taqqu [5]:

$$\Psi_\alpha = \exp\left(i\mu t - |t\sigma|^\alpha\left(1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(t)\right)\right)$$

which, for an n-summed variable (the equivalent of mixing with equal weights), becomes:

$$\Psi_\alpha(t) = \exp\left(i\mu nt - \left|n^{\frac{1}{\alpha}}t\sigma\right|^\alpha\left(1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(t)\right)\right)$$

A. Results

Let $X^\alpha \in \mathfrak{S}$, be the centered variable with a mean of zero, $X^\alpha = (Y^\alpha - \mu)$. We write $\mathbb{E}_K^+(\alpha,\beta,\mu,\sigma,K) := \mathbb{E}(X^\alpha|_{X^\alpha>K}\,\mathbb{P}(X^\alpha > K))$ under the stable distribution above. From Equation 3:

$$\begin{aligned}&\mathbb{E}_K^+(\alpha,\beta,\mu,\sigma,K)\\ &=\frac{1}{2\pi}\int_{-\infty}^{\infty}\alpha\sigma^\alpha|u|^{\alpha-2}\left(1\right.\\ &\quad\left.+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(u)\right)\exp\left(|u\sigma|^\alpha\left(-1\right.\right.\\ &\quad\left.\left.-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(u)\right)+iKu\right)\,du\end{aligned} \tag{4}$$

with explicit solution for $K = \mu = 0$:

$$\begin{aligned}\mathbb{E}_K^+(\alpha,\beta,0,\sigma,0) = -\sigma\frac{1}{\pi\alpha}\Gamma\left(-\frac{1}{\alpha}\right)\left(\left(1+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right.\\ \left.+\left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right).\end{aligned} \tag{5}$$

and semi-explicit generalized form for $K \neq \mu$:

$$\begin{aligned}&\mathbb{E}_K^+(\alpha,\beta,\mu,\sigma,K) \qquad (6)\\ &=\sigma\frac{\Gamma\left(\frac{\alpha-1}{\alpha}\right)\left(\left(1+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}+\left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right)}{2\pi}\\ &+\sum_{k=1}^{\infty}\frac{i^k(K-\mu)^k\Gamma\left(\frac{k+\alpha-1}{\alpha}\right)\left(\beta^2\tan^2\left(\frac{\pi\alpha}{2}\right)+1\right)^{\frac{1-k}{\alpha}}}{2\pi\sigma^{k-1}k!}\\ &\left((-1)^k\left(1+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{k-1}{\alpha}}+\left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{k-1}{\alpha}}\right)\end{aligned}$$

Our formulation in Equation 6 generalizes and simplifies the commonly used one from Wolfe [6] from which Hardin [7] got the explicit form, promoted in Samorodnitsky and Taqqu [5] and Zolotarev [4]:

$$\mathbb{E}(|X|) = \frac{1}{\pi}\sigma\left(2\Gamma\left(1-\frac{1}{\alpha}\right)\left(\beta^2\tan^2\left(\frac{\pi\alpha}{2}\right)+1\right)^{\frac{1}{2\alpha}} \cos\left(\frac{\tan^{-1}\left(\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}\right)\right) \quad (7)$$

Which allows us to prove the following statements:

1) Relative convergence: The general case with $\beta \neq 0$: for so and so, assuming so and so, (precisions) etc.,

$$n_\alpha^\beta = 2^{\frac{\alpha}{1-\alpha}}\pi^{\frac{\alpha}{2-2\alpha}}\left(\Gamma\left(\frac{\alpha-1}{\alpha}\right)\sqrt{n_g}\left(\left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{1}{\alpha}} + \left(1+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{1}{\alpha}}\right)\right)^{\frac{\alpha}{\alpha-1}} \quad (8)$$

with alternative expression:

$$n_\alpha^\beta = \pi^{\frac{\alpha}{2-2\alpha}}\left(\frac{\sec^2\left(\frac{\pi\alpha}{2}\right)^{-\frac{1}{2}/\alpha}\sec\left(\frac{\tan^{-1}\left(\tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}\right)}{\sqrt{n_g}\,\Gamma\left(\frac{\alpha-1}{\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} \quad (9)$$

Which in the symmetric case $\beta = 0$ reduces to:

$$n_\alpha = \pi^{\frac{\alpha}{2(1-\alpha)}}\left(\frac{1}{\sqrt{n_g}\,\Gamma\left(\frac{\alpha-1}{\alpha}\right)}\right)^{\frac{\alpha}{1-\alpha}} \quad (10)$$

2) Speed of convergence: $\forall k \in \mathbb{N}^+$ and $\alpha \in (1, 2]$

$$\mathbb{E}\left(\left|\sum_i^{kn_\alpha}\frac{X_i^\alpha - m_\alpha}{kn_\alpha}\right|\right)/\mathbb{E}\left(\left|\sum_i^{n_\alpha}\frac{X_i^\alpha - m_\alpha}{n_\alpha}\right|\right) = k^{\frac{1}{\alpha}-1} \quad (11)$$

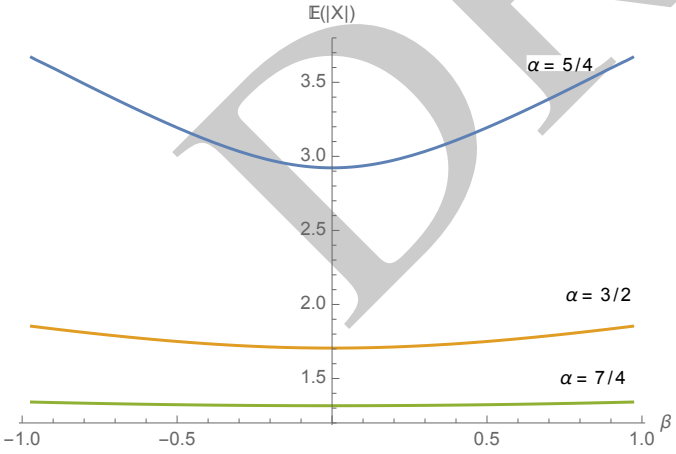


Fig. 3: Asymmetries and Mean Deviation.

Remark 1. *The ratio mean deviation of distributions in $\mathfrak{S}$ is homogeneous of degree $k^{\frac{1}{\alpha}-1}$. This is not the case for other classes "nonstable".*

Proof. (Sketch) From the characteristic function of the stable distribution. Other distributions need to converge to the basin $\mathfrak{S}$. □

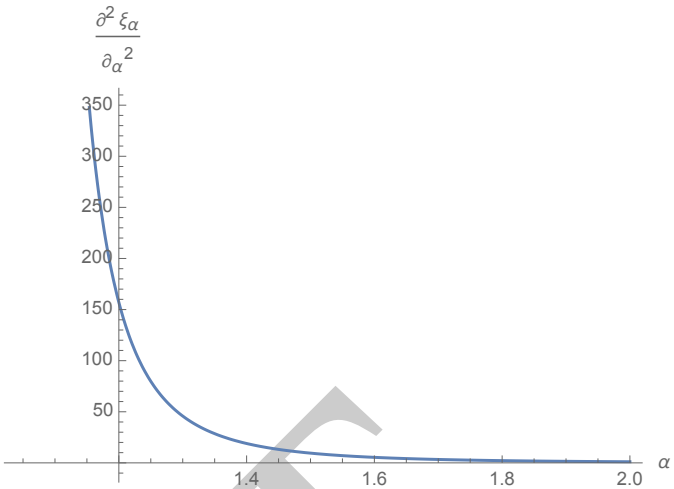


Fig. 4: Mixing distributions: the effect is pronounced at lower values of α, as tail uncertainty creates more fat-tailedness.

B. Stochastic Alpha or Mixed Samples

Define mixed population X_α and $\xi(X_\alpha)$ as the mean deviation of ...

Proposition 1. *For so and so*

$$\xi(X_{\bar{\alpha}}) \geq \sum_{i=1}^{m}\omega_i\xi(X_{\alpha_i})$$

where $\bar{\alpha} = \sum_{i=1}^m \omega_i\alpha_i$ and $\sum_{i=1}^m \omega_i = 1$.

Proof. A sketch for now: $\forall\alpha \in (1, 2)$, where γ is the Euler-Mascheroni constant ≈ 0.5772, $\psi^{(1)}$ the first derivative of the Poly Gamma function $\psi(x) = \Gamma'[x]/\Gamma[x]$, and H_n the n^{th} harmonic number:

$$\frac{\partial^2\xi}{\partial\alpha^2} = \frac{2\sigma\Gamma}{\pi\alpha^4}\left(\frac{\alpha-1}{\alpha}\right)n^{\frac{1}{\alpha}-1}\left(\psi^{(1)}\left(\frac{\alpha-1}{\alpha}\right) + \left(-H_{-\frac{1}{\alpha}}+\log(n)+\gamma\right)\left(2\alpha-H_{-\frac{1}{\alpha}}+\log(n)+\gamma\right)\right)$$

which is positive for values in the specified range, keeping $\alpha < 2$ as it would no longer converge to the Stable basin. □

Which is also negative with respect to *alpha* as can be seen in Figure 4. The implication is that one's sample underestimates the required "n". (Commentary).

V. Symmetric NonStable Distributions in the Subexponential Class

A. Symmetric Mixed Gaussians, Stochastic Mean

While mixing Gaussians the kurtosis rises, which makes it convenient to simulate fattailedness. But mixing means has the opposite effect, as if it were more "stabilizing". We can observe a similar effect of "thin-tailedness" as far as the n required to match the standard benchmark. The situation is the result of multimodality, noting that stable distributions are unimodal (Ibragimov and Chernin) [8] and infinitely divisible Wolfe [9]. For X_i Gaussian with mean μ, $\mathbb{E} = \mu\,\mathrm{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\mu^2}{2\sigma^2}}$, and keeping the average $\mu \pm \delta$

VIII. Alternative Methods for Mean

We saw that there are two ways to get the mean:

- The observed mean from data,
- The observed α from data, with corresponding distribution of the mean.

We will compare both –in fact there is a very large difference between the properties of both estimators.

Where $\mathcal{L}$ is the lognormal distribution, the idea is

$$\alpha \overset{d}{\sim} \mathcal{L}\left[\log(\alpha_0) - \frac{\sigma^2}{2}, \sigma\right]$$

For the most simplified Pareto distribution,

$$f(x) = \alpha L^{\alpha} x^{-\alpha-1}\,,\; x \in [L, \infty)$$

with expectation $\mathbb{E}(X) = \frac{\alpha L}{\alpha - 1}$. Since

$$f(\alpha) = \frac{e^{-\frac{\left(\log(\alpha) - \log(\alpha_0) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}}}{\sqrt{2\pi}\alpha\sigma}\,, \alpha \in (0, \infty)$$

we have $z(\alpha) : \mathbb{R}^+ \to \mathbb{R}\backslash[0, L); z \triangleq \frac{\alpha L}{\alpha - 1}$, with distribution:

$$g(z) = \frac{L \exp\left(-\frac{\left(-2\log(\alpha_0) + 2\log\left(\frac{z}{z-L}\right) + \sigma^2\right)^2}{8\sigma^2}\right)}{\sqrt{2\pi}\sigma z(z-L)}, z \in \mathbb{R}\backslash[0, L)$$

which we can verify as, interestingly $\int_{-\infty}^{0} g(z)\mathrm{d}z + \int_{L}^{\infty} g(z)\mathrm{d}z = 1$. Further, $\mathbb{P}(Z > 0) = \mathbb{P}(Z > L) = \frac{1}{2}\mathrm{erfc}\left(\frac{\sigma^2 - 2\log(\alpha_0)}{2\sqrt{2}\sigma}\right)$. The mean determined by the Hill estimator is unbiased since: we can show that

$$\lim_{\sigma \to 0} \frac{\int_L^{\infty} z\, g(z)\, \mathrm{d}z}{\int_L^{\infty} g(z)\, \mathrm{d}z} = L\frac{\alpha}{\alpha - 1} \tag{13}$$

The standard deviation of in sample α:

IX. Acknowledgement

Colman Humphrey,...

References

[1] M. Hlusek, "On distribution of absolute values," 2011.

[2] I. Pinelis, "Characteristic function of the positive part of a random variable and related results, with applications," *Statistics & Probability Letters*, vol. 106, pp. 281–286, 2015.

[3] V. V. Uchaikin and V. M. Zolotarev, *Chance and stability: stable distributions and their applications*. Walter de Gruyter, 1999.

[4] V. M. Zolotarev, *One-dimensional stable distributions*. American Mathematical Soc., 1986, vol. 65.

[5] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*. CRC Press, 1994, vol. 1.

[6] S. J. Wolfe, "On the local behavior of characteristic functions," *The Annals of Probability*, pp. 862–866, 1973.

[7] C. D. Hardin Jr, "Skewed stable variables and processes." DTIC Document, Tech. Rep., 1984.

[8] I. Ibragimov and K. Chernin, "On the unimodality of geometric stable laws," *Theory of Probability & Its Applications*, vol. 4, no. 4, pp. 417–419, 1959.

[9] S. J. Wolfe, "On the unimodality of infinitely divisible distribution functions," *Probability Theory and Related Fields*, vol. 45, no. 4, pp. 329–335, 1978.

[10] I. Zaliapin, Y. Y. Kagan, and F. P. Schoenberg, "Approximating the distribution of pareto sums," *Pure and Applied geophysics*, vol. 162, no. 6-7, pp. 1187–1228, 2005.