STATISTICAL CONSEQUENCES OF FAT TAILS

Real World Preasymptotics, Epistemology, and Applications

Nassim Nicholas Taleb

STEM
ACADEMIC PRESS

This format is based on André Miede's *ClassicThesis*, with adaptation from *Ars Classica*.

COAUTHORS [1]

Pasquale Cirillo (Chapters 10, 12, and 13)

Raphael Douady (Chapter 11)

Andrea Fontanari (Chapter 10)

Hélyette Geman (Chapter 22)

Donald Geman (Chapter 22)

Espen Haug (Chapter 20)

---

[1] Papers relied upon here are [33, 34, 35, 69, 76, 89, 111, 154, 155, 158, 159, 160, 162, 163, 164, 165, 171, 172]

CONTENTS

1 | PROLOGUE *,†

> The less you understand the world, the easier it is to make a decision.

Figure 1.1: *The problem is not awareness of "fat tails", but the lack of understanding of their consequences. Saying "it is fat tailed" implies much more than changing the name of the distribution, but a general overhaul of the statistical tools and types of decisions made.*

> The main idea behind the *Incerto* project is that while there is a lot of uncertainty and opacity about the world, and an incompleteness of information and understanding, there is little, if any, uncertainty about what actions should be taken based on such an incompleteness, in any given situation.

THIS BOOK consists in 1) published papers and 2) (uncensored) commentary, about classes of statistical distributions that deliver extreme events, and how we should deal with them for both statistical inference and decision making. Most "standard" statistics come from theorems designed for thin tails: they need to be adapted preasymptotically to fat tails, which is not trivial –or abandoned altogether.

So many times this author has been told *"of course we know this"* or the beastly *"nothing new"* about fat tails by a professor or practitioner who just produced an analysis using "variance", "GARCH", "kurtosis" , "Sharpe ratio", or "value at risk", or produced some "statistical significance" that is clearly not significant.

1
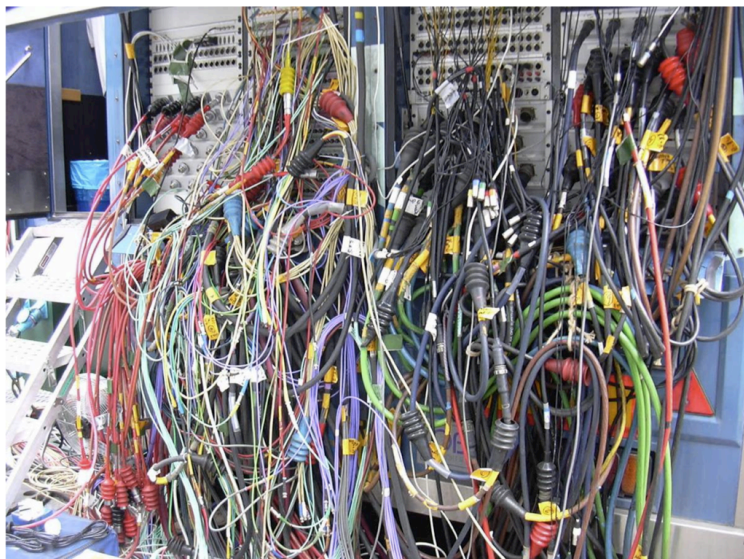
Figure 1.2: *Complication without insight: the clarity of mind of many professionals using statistics and data science without an understanding of the core concepts, what it is fundamentally about.*

Credit: Wikimedia.

More generally, this book draws on the author's multi-volume series, *Incerto* [157] and associated technical research program, which is about how to live in the real world, a world with a structure of uncertainty that is too complicated for us.

The *Incerto* tries to connects five different fields related to tail probabilities and extremes: mathematics, philosophy, social science, contract theory, decision theory, and the real world. If you wonder why contract theory, the answer is: option theory is based on the notion of contingent and probabilistic contracts designed to modify and share classes of exposures in the tails of the distribution; in a way option theory is mathematical contract theory. Decision theory is not about understanding the world, but getting out of trouble and ensuring survival. This point is the subject of the next volume of the *Technical Incerto*, with the temporary working title *Convexity, Risk, and Fragility*.

I presented this book and the main points at the Bloomberg Quant Conference in New York in September 2018. After the lecture, a prominent mathematical finance professor came to see me. "This is very typical Taleb", he said. "You show what's wrong but don't offer too many substitutes".

Clearly, in business or in anything subjected to the rigors of the real world, he would have been terminated. People who never had any skin in the game [167] cannot figure out the necessity of suspension of belief: *don't give a pilot a faulty metric, learn to only provide reliable information*. Nor can they get the outperformance of *via negativa* –Popperian science works by removal.The late David Freedman had tried unsuccessfully to tame vapid and misleading statistical modeling *outperformed* by "nothing".

But it is the case that the various chapters and papers here *do* offer solutions and alternatives, except that these aren't the most comfortable for some.
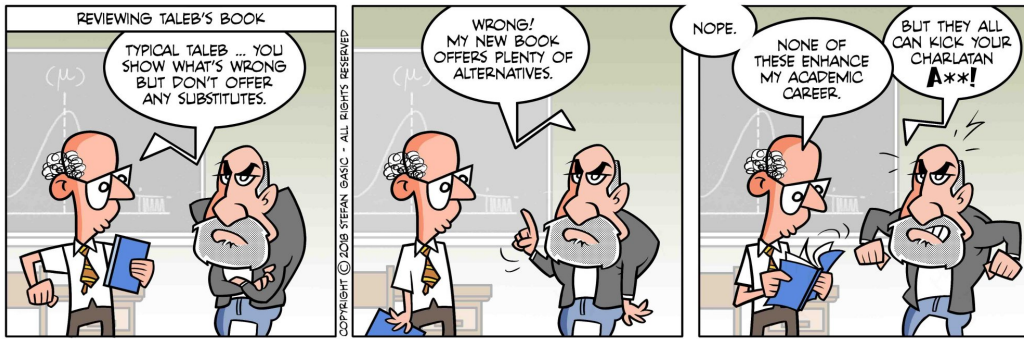
A WORD ON TERMINOLOGY

"Fat tails" is often used in many contexts. In ours, here, it is simply "higher kurtosis than the Gaussian" –to conform to the finance practitioner's lingo. For many it is meant to be a narrower definition, limited to "power laws" or "regular variations" –what we call "fat tails" may be "thick tails" for many. The next chapter will clarify.

ACKNOWLEDGMENTS

Part I

FAT TAILS AND THEIR EFFECTS, AN INTRODUCTION

2 | A NON–TECHNICAL OVERVIEW – THE DARWIN COLLEGE LECTURE[*,‡]

THIS chapter[1] presents a nontechnical yet comprehensive presentation of of the entire *statistical consequences of fat tails* project. It compresses the main ideas in one place. Mostly, it provides a list of more than a dozen consequences of fat tails on statistical inference.

2.1 ON THE DIFFERENCE BETWEEN THIN AND FAT TAILS

We begin with the notion of fat tails and how it relates to extremes using the two imaginary domains of Mediocristan (thin tails) and Extremistan (fat tails).

> - In Mediocristan, no observation can really change the statistical properties.
> - In Extremistan, the tails (the rare events) play a disproportionately large role in determining the properties.

Let us randomly select two people in Mediocristan with a (very unlikely) combined height of 4.1 meters a tail event. According to the Gaussian distribution (or, rather its one-tailed siblings), the most likely combination of the two heights is 2.05 metres and 2.05 metres.

Simply, the probability of exceeding 3 sigmas is 0.00135. The probability of exceeding 6 sigmas, twice as much, is 9.86×10^{-10}. The probability of two 3-sigma events occurring is 1.8×10^{-6}. Therefore the probability of two 3-sigma events occurring is considerably higher than the probability of one single 6-sigma event. This is using a class of distribution that is not fat-tailed. Figure 2.1 below shows that as we extend the ratio from the probability of two 3-sigma events divided by the probability of a 6-sigma event, to the probability of two 4-sigma events divided by the probability of an 8-sigma event, i.e., the further we go into the tail, we see that a large deviation can only occur via a combination (a sum) of a large number of intermediate deviations: the right side of Figure 2.1. In other words, for something bad to happen, it needs to come from a series of very unlikely events, not a single one. This is the logic of Mediocristan.

Let us now move to Extremistan and randomly select two people with combined wealth of $ 36 million. The most likely combination is not $18 million and $ 18 million. It is approximately $ 35,999,000 and $ 1,000.

This highlights the crisp distinction between the two domains; for the class of subexponential distributions, ruin is more likely to come from a single extreme event than from a series of bad episodes. This logic underpins classical risk theory as outlined by Lundberg early in the 20th Century [105] and formalized by Cramer[38], but forgotten by economists in recent times. This indicates that insurance can only work in Mediocristan; you should never write an uncapped insurance contract if there is a risk of catastrophe. The point is called the catastrophe principle.

As I said earlier, with fat tail distributions, extreme events away from the centre of the distribution play a very large role. Black Swans are not more frequent, they are more consequential. The fattest tail distribution has just one very large extreme deviation, rather than many departures form the norm. Figure 3.4 shows that if we take a distribution like the Gaussian and start fattening it, then the number of departures away from one standard deviation drops. The probability of an event staying within one standard deviation of the mean is 68 per cent. As the tails fatten, to mimic what happens in financial markets for example, the probability of an event staying within one standard deviation of the mean rises to between 75 and 95 per cent. So note that as we fatten the tails we get higher peaks, smaller shoulders, and a higher incidence of a very large deviation.
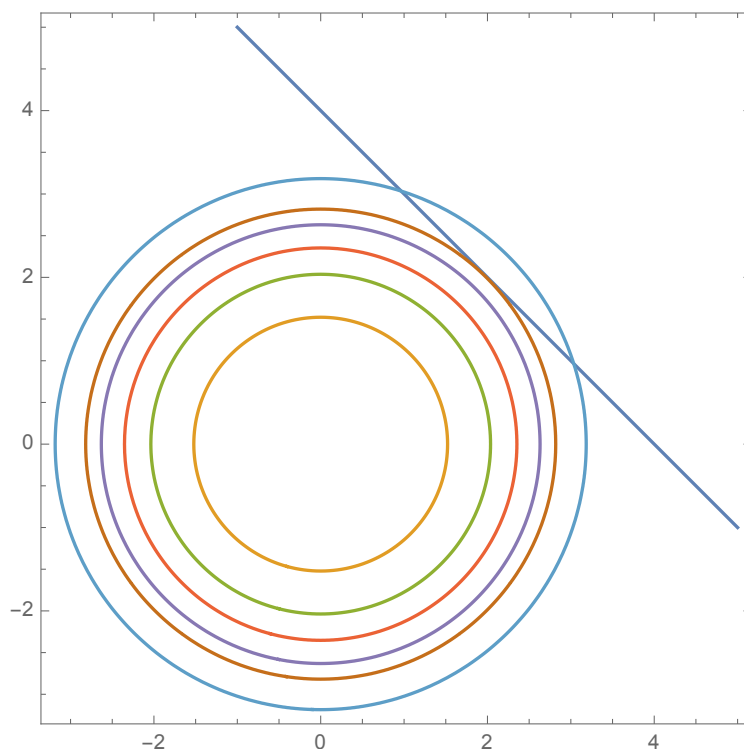
Figure 2.1: *Iso-densities for two independent Gaussian distributions. The line shows $x + y = 4.1$. Visibly the maximal probability is for $x = y = 2.05$.*
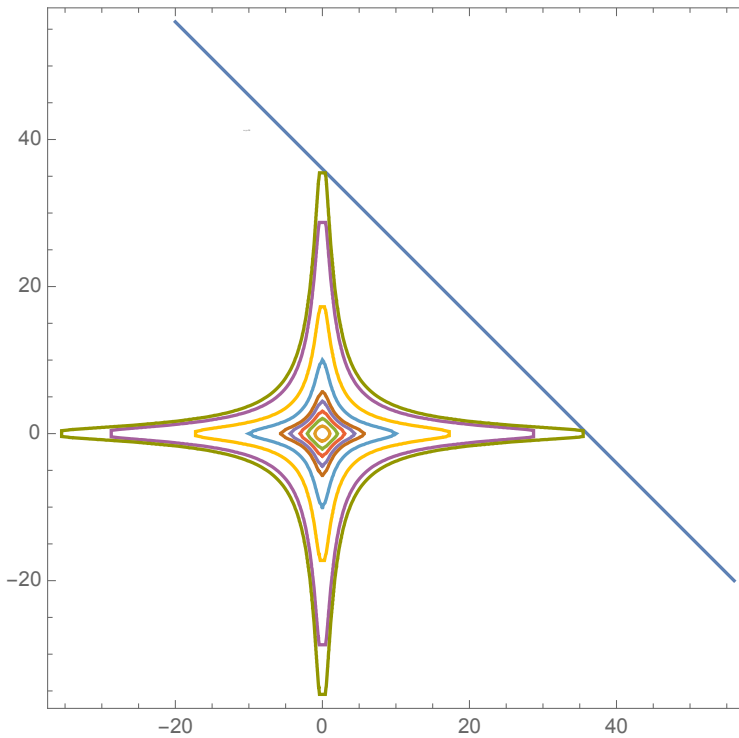
Figure 2.2: *Iso-densities for two independent Fat tailed distributions (in the power law class). The line shows $x + y = 36$. Visibly the maximal probability is for either $x = 36 - \epsilon$ or $y = 36 - \epsilon$, with ϵ going to 0 as the sum $x + y$ becomes larger, with the iso-densities looking more and more like a cross.*
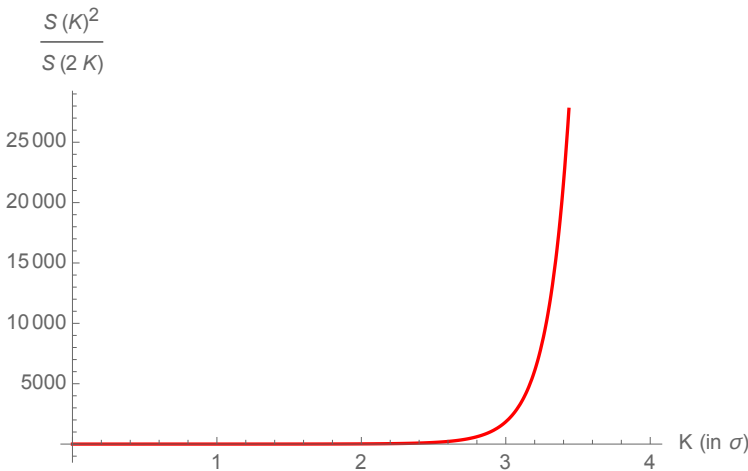
Figure 2.3: *Ratio of two occurrences of size K by one of 2K for a Gaussian distribution. The larger the K, that is, the more we are in the tails, the more likely the event is likely to come from 2 independent realizations of K (hence $P(K)^2$, and the less from a single event of magnitude 2K.* (Note: This is fudging for pedagogical simplicity. The correct approach would be to assume the most likely, say $S(K_0)$ deviations, times $S(K - K_0)$, but no worries since $S(K_0)$ is a constant.)

2.2 A (MORE ADVANCED) CATEGORIZATION AND ITS CONSEQUENCES

First there are entry level fat tails. This is any distribution with fatter tails than the Gaussian i.e. with more observations within one sigma and with kurtosis (a function of the fourth central moment) higher than three.

Second, there are subexponential distributions satisfying our thought experiment earlier. Unless they enter the class of power laws, these are not really fat tails because they do not have monstrous impacts from rare events. In other words, they have all the moments.

Level three, what is called by a variety of names, the power law, or slowly varying class, or "Pareto tails" class correspond to real fat tails.

High Variance Gaussian

$$\frac{1}{n}\sum_{i=1}^{n}x_i$$
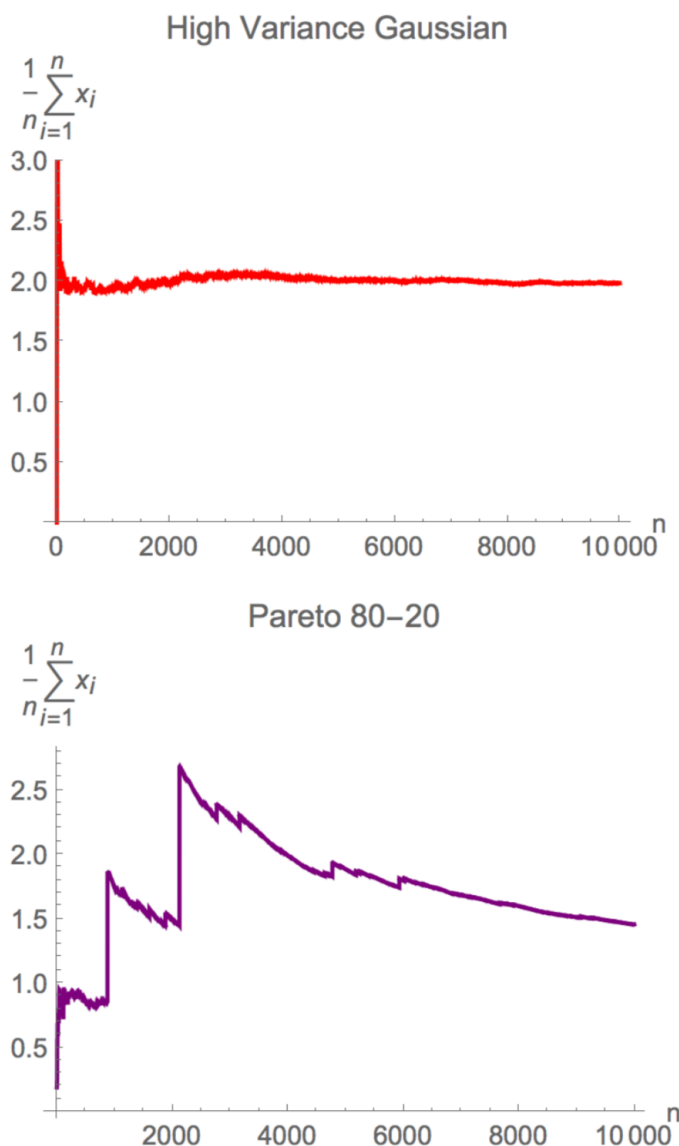
Figure 2.4: *The law of large numbers, that is how long it takes for the sample mean to stabilize, works much more slowly in Extremistan (here a Pareto distribution with 1.13 tail exponent , corresponding to the "Pareto 80-20". Both have the same mean absolute deviation. Note that the same applies to other forms of sampling, such as portfolio theory.*

Pareto 80–20

$$\frac{1}{n}\sum_{i=1}^{n}x_i$$

Working from the bottom left of Figure 2.5, we have the degenerate distribution where there is only one possible outcome i.e. no randomness and no variation. Then, above it, there is the Bernoulli distribution which has two possible outcomes. Then above it there are the two Gaussians. There is the natural Gaussian (with support on minus and plus infinity), and Gaussians that are reached by adding random walks (with compact support, sort of). These are completely different animals since one can deliver infinity and the other cannot (except asymptotically). Then above the Gaussians there is the subexponential class. Its members all have moments, but the subexponential class includes the lognormal, which is one of the strangest things on earth because sometimes it cheats and moves up to the top of the diagram. At low variance, it is thin-tailed, at high variance, it behaves like the very fat tailed.

Membership in the subexponential class satisfies the Cramer condition of possibility of insurance (losses are more likely to come from many events than a single one), as illus-
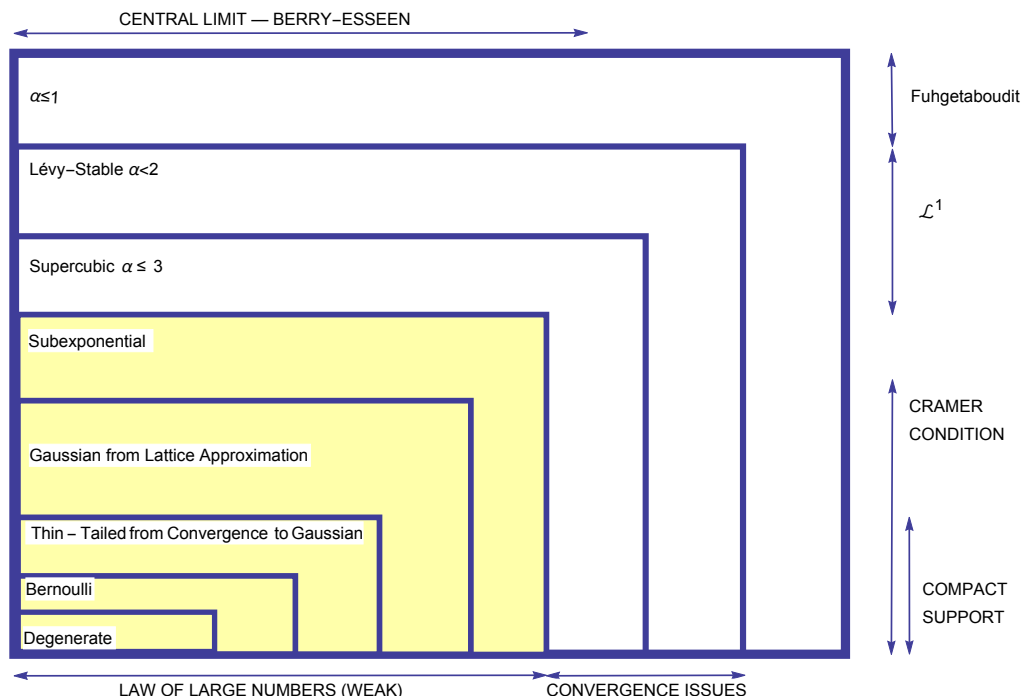
Figure 2.5: *The tableau of Fat tails, along the various classifications for convergence purposes (i.e., convergence to the law of large numbers, etc.) and gravity of inferential problems. Power Laws are in white, the rest in yellow. See Embrechts et al [60].*

trated in Figure 2.1. More technically it means that the expectation of the exponential of the random variable exists.[2]

Once we leave the yellow zone, where the law of large numbers largely works, then we encounter convergence problems. Here we have what are called power laws, such as Pareto laws. And then there is one called Supercubic, then there is Levy-Stable. From here there is no variance. Further up, there is no mean. Then there is a distribution right at the top, which I call the Fuhgetaboudit. If you see something in that category, you go home and you dont talk about it. In the category before last, below the top (using the parameter α, which indicates the "shape" of the tails, for $\alpha < 2$ but not $\alpha \leq 1$), there is no variance, but there is the mean absolute deviation as indicator of dispersion. And recall the Cramer condition: it applies up to the second Gaussian which means you can do insurance.

The traditional statisticians approach to fat tails has been to assume a different distribution but keep doing business as usual, using same metrics, tests, and statements of significance. But this is not how it really works and they fall into logical inconsistencies. Once we leave the yellow zone, for which statistical techniques were designed, things no longer work as planned. The next section presents a dozen issues, almost all terminal.

---

2 Technical point: Let X be a random variable. The Cramer condition: for all $r > 0$,

$$\mathbb{E}(e^{rX}) < +\infty.$$

The problem with overstandardized statistics

STATISTICAL ESTIMATION is based on two elements: the central limit theorem (which is assumed to work for "large" sums, thus making about everything conveniently normal) and that of the law of large numbers, which reduces the variance of the estimation as one increases the sample size. However, there are now caveats as we can see throughout this text. In Chapter x, we show a tableau of "what large means" for the central limit theorem: convergence can be very, very slow –it is distribution dependent. As shown by Bouchaud and Potters in [19] and Sornette in [149], the tails for some finite variance but infinite higher moments can converge to the Gaussian within $\pm\sqrt{n\log n}$, meaning the center of the distribution inside such band becomes Gaussian, but remote parts don't –and the remote parts determine so much of the properties. My paper (this Chapter 7) examines the LLN for various distributions.

Life happens in the preasymptotics.

Sadly, in the entry on estimators in the monumental *Encyclopedia of Statistical Science* [101], W. Hoeffding writes:

> "The exact distribution of a statistic is usually highly complicated and difficult to work with. Hence the need to approximate the exact distribution by a distribution of a simpler form whose properties are more transparent. The limit theorems of probability theory provide an important tool for such approximations. In particular, the classical central limit theorems state that the sum of a large number of independent random variables is approximately normally distributed under general conditions. In fact, the normal distribution plays a dominating role among the possible limit distributions. To quote from Gnedenko and Kolmogorov's text [[80], Chap. 5]:
>
>> *"Whereas for the convergence of distribution functions of sums of independent variables to the normal law only restrictions of a very general kind, apart from that of being infinitesimal (or asymptotically constant), have to be imposed on the summands, for the convergence to another limit law some very special properties are required of the summands".*
>
> Moreover, many statistics behave asymptotically like sums of independent random variables. All of this helps to explain the importance of the normal distribution as an asymptotic distribution."

Now what if we do not reach the normal distribution, as life happens before the asymptote? This is what this book is about.[a]

---

a The reader is invited to consult a "statistical estimation" entry in any textbook or online encyclopedia. Odds are that the notion of "what happens if we do not reach the asymptote" will never be discussed –as in the 9500 pages of the monumental *Encyclopedia of Statistics*. Further, ask a regular user of statistics about how much data one needs for such and such distributions, and don't be surprised at the answer. The problem is that people have too many prepackaged statistical tools in their heads, ones they never had to rederive themselves. The motto here is: "statistics is never standard".

2.3 THE MAIN CONSEQUENCES

Here are some consequences of moving out of the yellow zone:

1. **The law of large numbers, when it works, works too slowly in the real world**.
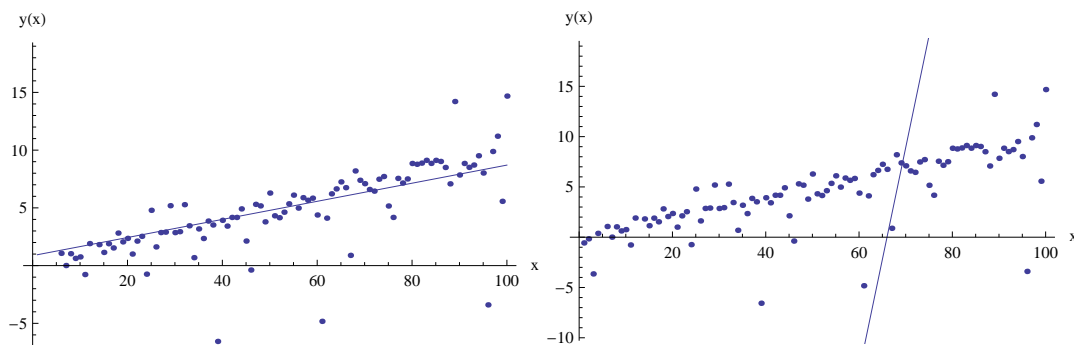
Figure 2.6: *In the presence of fat tails, we can fit markedly different regression lines to the same story (the Gauss-Markov theorem doesn't apply anymore). Left: a regular (naïve) regression. Right: a regression line that tries to accommodate the large deviation —a "hedge ratio" so to speak, one that protects the agent from a large deviation, but mistracks small ones. Missing the largest deviation can be fatal. Note that the sample doesn't include the critical observation, but it has been guessed using "shadow mean" methods.*

Figure 2.7: *Inequality measures such as the Gini coefficient require completely different methods of estimation under fat tails, as we will see in Part III. Science is hard.*

(This is more shocking than you think as it cancels most statistical estimators). See Figure 2.4 in this chapter for an illustration. The subject is treated in Chapter 7 and distributions are classified according to how fast they *preasymptotically* converge according to the law of large numbers.

2. **The mean of the distribution will not correspond to the sample mean, particularly if the distribution is skewed (or one-tailed).**

In fact, there is no fat tailed distribution in which the mean can be properly estimated directly from the sample mean, unless we have orders of magnitude more data than we do (people in finance still do not understand this). The point is discussed in the "shadow mean" chapters, such as Chapter 12 and Chapter 13. And we will introduce the notion of hidden properties are in 2.5.

3. **Standard deviations and variance are not useable.**

They fail out of sample –even when they exist; even when all moments exist. Discussed in Chapter 3.1.

4. **Beta, Sharpe Ratio and other common financial metrics are uninformative.**

(This is a simple consequence of the previous point). Either they require much more data, many more orders of magnitude, or some different model than the one being used, of which we are not yet aware. Further, stochastic correlations or covariances also represent a form of fat tails, discussed in Chapter 3.1 as well.

5. **Robust statistics is not robust at all**

The story of my life. The so-called "empirical distribution" is not empirical (as it misrepresents the expected payoffs in the tails). Discussed in Chapter 8.

6. **Linear least-square regression doesn't work (failure of the Gauss-Markov theorem)**.

See figure 2.6. Either we need a lot, a lot of data to minimize the squared deviations (in other words, the Gauss-Markov theorem applies, but not for our preasymptotic situation as the real world has no infinite date), or we can't because the second moment does not exist. In the latter case, if we minimize absolute deviations (MAD, mean absolute deviations), as we seen in 3.1, not only we may still be facing insufficient data but the deviation slope may not be unique.

7. **Maximum likelihood methods work for parameters of the shape of the distribution (good news)**.

Take a power law. We may estimate a parameter for its shape, the tail exponent (written α in this book), which, adding some other parameter (the scale) connects us back to its mean.

So we can produce more reliable (or at least less unreliable) plug-in estimators for, say, the tail exponent in *some* situations. But, of course, not all. Now what do we do when we do not have a reliable estimator? We do not expose ourselves to harm in the presence of fragility, but can still take decisions if we are bounded for maximum losses.

8. **The gap between disconfirmatory and confirmatory empiricism is wider than in situations covered by common statistics i.e. the difference between absence of evidence and evidence of absence becomes larger.**

From a controversy the author had with the cognitive linguist and science writer Steven Pinker: making pronouncement (and generating theories) from recent variations in data is not easily possible. Stating "violence has dropped" because the number of people killed in wars has dropped is not a scientific statement: a scientific claim distinguishes itself from an anecdote as it aims at affecting what happens out of sample, by focusing on statistically significance: non statistically significant statements are not the realm of science. However, saying violence has risen upon a single observation may be a rigorously scientific claim. The practice of reading into descriptive statistics may be acceptable under thin tails (as sample sizes do not have

to be large), but never so under fat tails, except, to repeat, in the presence of a large deviation.

9. **The principal component analysis is likely to produce false factors.**

   This point is a bit technical; it adapts the notion of sample insufficiency to large random vectors seen via the dimension reduction technique called principal component analysis (PCA) . The issue a higher dimensional version of our law of large number complications. The story is best explained in Figure 2.15, which shows the accentuation of what is called the "Wigner Effect", from insufficiency of data for the PCA. Also, to be technical, note that the Marchenko-Pastur distribution is not applicable in the absence of a finite fourth moment (or, has been shown in [15], for tail exponent in excess of 4 using our terminology here)[3].

10. **The method of moments (MoM) fails to work. Higher moments are uninformative or do not exist.**

    The same applies to the GMM, the generalized method of moment. This is a long story, but take for now that the estimation of a given distribution by moment matching fails if higher moments are not finite, so every sample delivers a different moment –as we will soon see with the 4th moment of the SP500.

11. **There is no such thing as a *typical* large deviation**

    Conditional on having a "large" move, the magnitude of such a move is not defined, especially under serious Fat Tails (the Power Law tails class). This is associated with the catastrophe principle we saw earlier. In the Gaussian world, the expectation of a movement, conditional that the movement exceeds 4 standard deviations, is about 4 standard deviations. For a Power Law it will be a multiple of that.

12. **The Gini coefficient ceases to be additive.**

    Methods of measuring sample data for Gini are interpolative –they in effect have the same problem we saw earlier with the sample mean underestimating or overestimating the true mean. Here, an additional complexity arises as the Gini becomes super-additive under fat tails. As the sampling space grows, the conventional Gini measurements give an illusion of large concentrations of wealth. (In other words, inequality in a continent, say Europe, can be higher than the average inequality of its members). The same applies to other measures of concentration such as the top 1% has x percent of the total wealth, etc. The derivations are in Chapters 10 and 11.

13. **Large Deviation Theory** (Varadan [185] , Dembo and Zeituni [45], etc.) fails to apply to fat tails. I mean, it really doesn't apply.

14. **Option risks are never mitigated by dynamic hedging**. This might be technical for nonfinance people but the entire basis of finance rests on the possibility and necessity of dynamic hedging, both of which will be shown to be erroneous in Chapters 18 19 and 20. The required exponential decline require the probability distribution to be outside the sub-exponential class. Again, we are talking about the Cramer condition.

15. **Forecasting in frequency space diverges from expected payoff**.

    This point is explored in the next section: the foolish notion of focus on frequency rather than expectation can carry a mild effect under thin tails; not under fat tails.

Let us discuss the major points.

---

3 To clear up the terminology: in this book, the tail exponent, commonly written α is the limit of quotient of the log of the survival function in excess of K over log K, which would be 1 for Cauchy. Some researchers use $\alpha - 1$ from the corresponding density function.

2.3.1 Forecasting

In *Fooled by Randomness* (2001/2005), the character is asked which was *more probable* that a given market would go higher or lower by the end of the month. Higher, he said, much more probable. But then it was revealed that he was making trades that benefit if that particular market *goes down*. This of course, appears to be paradoxical for the nonprobabilist but very ordinary for traders, particularly under nonstandard distributions (yes, the market is more likely to go up, but should it go down it will fall much much more). This illustrates the common confusion between a *forecast* and an exposure (a forecast is a binary outcome, an exposure has more nuanced results and depends on full distribution). This example shows one of the extremely elementary mistakes of talking about *probability presented* as single numbers not distributions of outcomes, but when we go deeper into the subject, many less obvious, or less known paradox-style problems occur. Simply, it is of the opinion of the author, that it is not rigorous to talk about "probability" as a final product, or even as a "foundation" of decisions.

In the real world one is not paid in probability, but in dollars (or in survival, etc.). The fatter the tails, the more one needs to worry about payoff space – the saying goes: "payoff swamps probability" (see box). One can be wrong very frequently if the cost is low, so long as one is convex to payoff (i.e. make large gains when one is right). Further, one can be forecasting with 99.99% accuracy and still go bust (in fact, more likely to go bust: funds with impeccable track records were those that went bust during the 2008-2009 rout [4]). A point that may be technical for those outside quantitative finance: it is the difference between a vanilla option and a corresponding binary of the same strike, as discussed in *Dynamic Hedging* [156]: counterintuitively, fat tailedness *lowers* the value of the binary and raise that of the vanilla. This is expressed by the author's adage: "I've never seem a rich forecaster." We will examine in depth in 3.3.1 where we show that fattening the tails cause the probability of events higher than 1 standard deviations to drop –but the consequences to rise (in term of contribution to moments, say effect on the mean or other metrics).

> **Remark 2.1**
> *Probabilistic forecast errors ("calibration") are in a different probability class from that true real-world P/L variations (or true payoffs).*
>
> *"Calibration", which is a measure of how accurate one's predictions, lies in probability space –between 0 and 1. Any measure of such calibration will necessarily be thin-tailed (and, if anything, extra-thin tailed since it is bounded) –whether the random variable under such prediction is fat tailed or not. On the other hand, payoffs in the real world can be fat tailed, hence the distribution of such "calibration" will follow the property of the random variable.*

We show full derivations and proofs in Appendix 9.

2.3.2 Ebola cannot be compared to falls from ladders

Let us illustrate one of the problem of thin-tailed thinking with a real world example. People quote so-called "empirical" data to tell us we are foolish to worry about ebola when only two Americans died of ebola in 2016. We are told that we should worry more about deaths from diabetes or people tangled in their bedsheets. Let us think about it in terms of tails. But, if we were to read in the newspaper that 2 billion people have died suddenly,

---

4 R. Douady, data from Risk Data about funds that collapsed in the 2008 crisis, personal communication

Payoff swamps probability in Extremistan: To see the main difference between Mediocristan and Extremistan, consider the *event* of a plane crash. A lot of people will lose their lives, something very sad, say between 100 and 400 people, so the event is counted as a bad episode, a single one. For forecasting and risk management, we work on minimizing such a probability to make it negligible.

Now, consider a type of plane crashes that will kill *all* the people who ever rode the plane, even all passengers who ever rode planes in the past. All. Is it the same type of event? The latter event is in Extremistan and, for these, we don't talk about probability but focus instead on the magnitude of the event.

- For the first type, management consists in reducing the probability –the frequency – of such events. Remember that we count events and aim at reducing their counts.

- For the second type, it consists in reducing the effect should such an event take place. We do not count events, we measure impact.

If you think the thought experiment is a bit weird, consider that the money center banks lost in 1982 more money than they ever made in their history, the Savings and Loans industry (now gone) did so in 1991, and the entire banking system lost every penny ever made in 2008-9. One can routinely witness people lose everything they earned cumulatively in a single market event. The same applies to many, many industries (e.g. automakers and airlines).

But banks are only about money; consider that for wars we cannot afford the naive focus on event frequency without taking into account the magnitude, as done by the science writer Steven Pinker in [134], discussed in Chapter 13. This is without even examining the ruin problems (and nonergodicity) in the end of this section. More technically, one needs to meet the Cramer condition of non-subexponentiality for a tally of events (taken at face value) for raw probability to have any meaning at all.

The plane analogy was proposed by the insightful Russ Robert during one of his econtalk podcasts with the author.
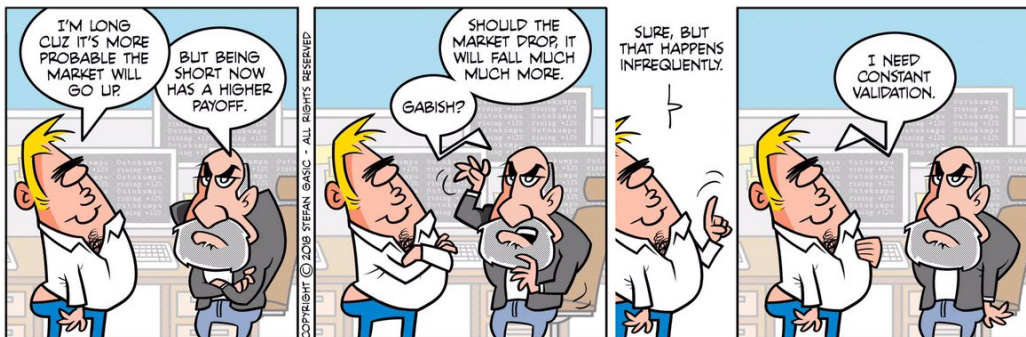
Figure 2.8: *Life is about payoffs, not forecasting, and the difference increases in Extremistan*

1

it is far more likely that they died of ebola than smoking or diabetes or tangled in their bedsheets? This is rule number one. "Thou shalt not compare a multiplicative fat-tailed process in Extremistan in the subexponential class to a thin-tailed process that has Chernoff bounds from Mediocristan". This is simply because of the catastrophe principle we saw

earlier, illustrated in Figure 2.1. It is naïve empiricism to compare these processes, to suggest that we worry too much about ebola and too little about diabetes. In fact it is the other way round. We worry too much about diabetes and too little about ebola and other multiplicative effects. This is an error of reasoning that comes from not understanding fat tails –sadly it is more and more common. What is worse, such errors of reasoning are promoted by **empirical psychology** which does not appear to be empirical.

2.3.3 The Law of Large Numbers

Let us now discuss the law of large numbers which is the basis of much of statistics. The law of large numbers tells us that as we add observations the mean becomes more stable, the rate being around $\sqrt{n}$. Figure 2.4 shows that it takes many more observations under a fat-tailed distribution (on the right hand side) for the mean to stabilize.

Table 2.1: *Corresponding n_α, or how many observations to get a drop in the error around the mean for an equivalent α-stable distribution (the measure is discussed in more details in Chapter 7). The Gaussian case is the $\alpha = 2$. For the case with equivalent tails to the 80/20 one needs at least 10^{11} more data than the Gaussian.*

α	n_α	$n_\alpha^{\beta=\pm\frac{1}{2}}$	$n_\alpha^{\beta=\pm 1}$
	Symmetric	Skewed	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.

One of the best known statistical phenomena is Paretos 80/20 e.g. twenty per cent of Italians own 80 per cent of the land. Table 2.1 shows that while it takes 30 observations in the Gaussian to stabilize the mean up to a given level, it takes 10^{11} observations in the Pareto to bring the sample error down by the same amount (assuming the mean exists).

Despite this being trivial to compute, few people compute it. You cannot make claims about the stability of the sample mean with a fat tailed distribution. There are other ways to do this, but not from observations on the sample mean.
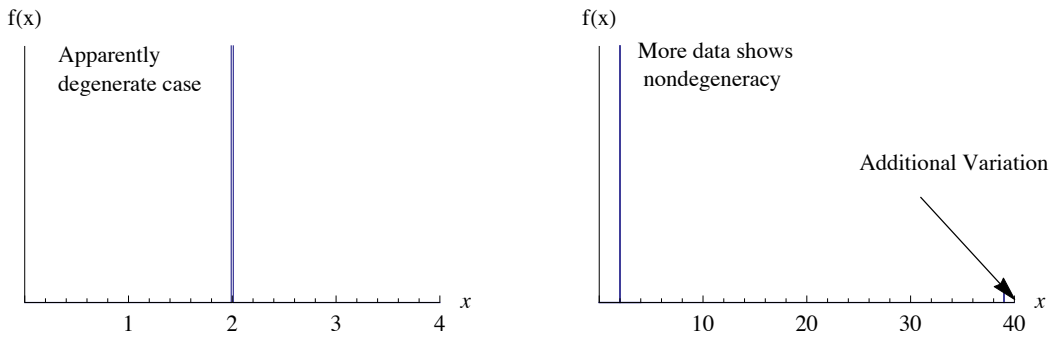
Figure 2.9: *The Masquerade Problem (or Central Asymmetry in Inference). To the left, a degenerate random variable taking seemingly constant values, with a histogram producing a Dirac stick. One cannot rule out nondegeneracy. But the right plot exhibits more than one realization. Here one can rule out degeneracy. This central asymmetry can be generalized and put some rigor into statements like "failure to reject" as the notion of what is rejected needs to be refined. We produce rules in Chapter* **??**.

2.3.4 Epistemology and Inferential Asymmetry

Let us now examine the epistemological consequences. Figure 2.9 illustrates the Masquerade Problem (or Central Asymmetry in Inference). On the left is a degenerate random variable taking seemingly constant values with a histogram producing a Dirac stick.

We have known at least since Sextus Empiricus that we cannot rule out degeneracy but there are situations in which we can rule out non-degeneracy. If I see a distribution that has no randomness, I cannot say it is not random. That is, we cannot say there are no Black Swans . Let us now add one observation. I can now see it is random, and I can rule out degeneracy. I can say it is not not random. On the right hand side we have seen a Black Swan , therefore the statement that there are no Black Swans is wrong. This is the negative empiricism that underpins Western science. As we gather information, we can rule things out. The distribution on the right can hide as the distribution on the left, but the distribution on the right cannot hide as the distribution on the left (check). This gives us a very easy way to deal with randomness. Figure 2.10 generalizes the problem to how we can eliminate distributions.

If we see a 20 sigma event, we can rule out that the distribution is thin-tailed. If we see no large deviation, we can not rule out that it is not fat tailed unless we understand the process very well. This is how we can rank distributions. If we reconsider Figure 2.5 we can start seeing deviation and ruling out progressively from the bottom. These are based on how they can deliver tail events. Ranking distributions becomes very simple because if someone tells you there is a ten-sigma event, it is much more likely that they have the wrong distribution than it is that you really have ten-sigma event. Likewise, as we saw, fat tailed distributions do not deliver a lot of deviation from the mean. But once in a while you get a big deviation. So we can now rule out what is not mediocristan. We can rule out where we are not we can rule out mediocristan. I can say this distribution is fat tailed by elimination. But I can not certify that it is thin tailed. This is the Black Swan problem.
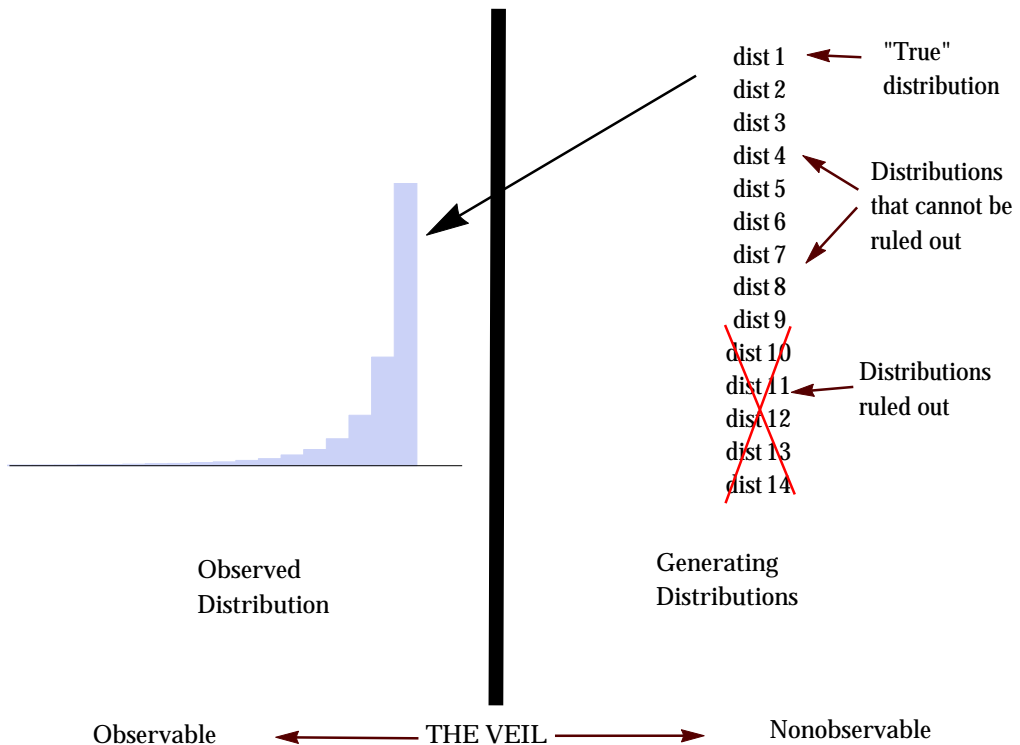
Figure 2.10: ***"The probabilistic veil"***. *Taleb and Pilpel [172] cover the point from an epistemological stand-point with the "veil" thought experiment by which an observer is supplied with data (generated by someone with "perfect statistical information", that is, producing it from a generator of time series). The observer, not knowing the generating process, and basing his information on data and data only, would have to come up with an estimate of the statistical properties (probabilities, mean, variance, value-at-risk, etc.). Clearly, the observer having incomplete information about the generator, and no reliable theory about what the data corresponds to, will always make mistakes, but these mistakes have a certain pattern. This is the central problem of risk management.*

2.4 PRIMER ON POWER LAWS (WITHOUT MATHEMATICS)

Let us now discuss the intuition behind the Pareto Law. It is simply defined as: say X is a random variable. For x sufficently large, the probability of exceeding $2x$ divided by the probability of exceeding x is no different from the probability of exceeding $4x$ divided by the probability of exceeding $2x$, and so forth. This property is called "scalability".[2]

So if we have a Pareto (or Pareto-style) distribution, the ratio of people with $ 16 million compared to $ 8 million is the same as the ratio of people with $ 2 million and $ 1 million. There is a constant inequality. This distribution has no characteristic scale which makes it very easy to understand. Although this distribution often has no mean and no standard deviation we can still understand it –in fact we can understand it much better than we do with more standard statistical distributions. But because it has no mean we have to ditch the statistical textbooks and do something more solid, more rigorous, even if it seems less mathematical.

A Pareto distribution has no higher moments: moments either do not exist or become statistically more and more unstable. So next we move on to a problem with economics and econometrics. In 2009 I took 55 years of data and looked at how much of the kurtosis
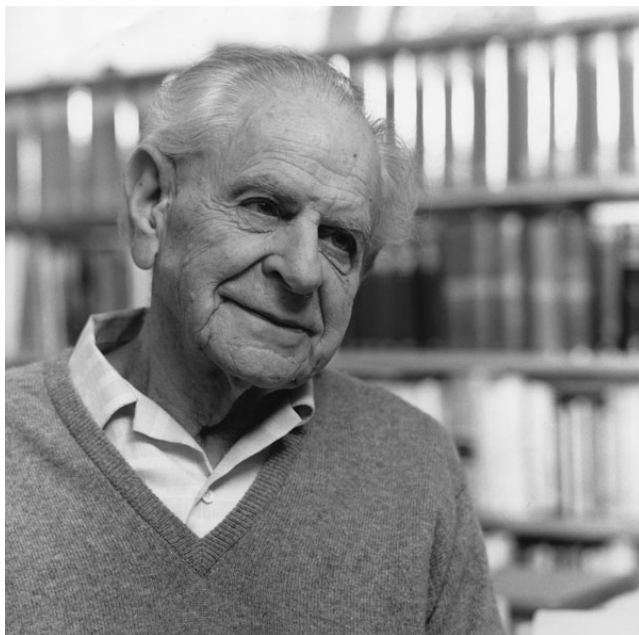
Figure 2.11: *Popper's solution of the problem of induction is in asymmetry: relying on confirmatory empiricism, that is focus on "ruling out" what fails to work, via negativa style. We extend this approach to statistical inference with the probabilistic veil by progressively ruling out entire classes of distributions.*

Scientific Rigor and Asymmetries by The Russian School of Probability

NE CAN BELIEVE in the rigor of mathematical statements about probability without falling into the trap of providing naive computations subjected to model error. There is a wonderful awareness of the asymmetry throughout the works of the Russian school of probability –and asymmetry here is the analog of Popper's idea in mathematical space.

Members across three generations: P.L. Chebyshev, A.A. Markov, A.M. Lyapunov, S.N. Bernshtein (ie. Bernstein), E.E. Slutskii, N.V. Smirnov, L.N. Bol'shev, V.I. Romanovskii, A.N. Kolmogorov, Yu.V. Linnik, and the new generation: V. Petrov, A.N. Nagaev, A. Shryayev, and a few more.

They had something rather potent in the history of scientific thought: they thought in inequalities, not equalities (most famous: Markov, Chebyshev, Bernstein, Lyapunov). They used bounds, not estimates. Even their central limit version was a matter of bounds, which we exploit later by seeing what takes place *outside the bounds*. They were world apart from the new generation of users who think in terms of precise probability –or worse, mechanistic social scientists. Their method accommodates skepticism, one-sided thinking: "A is $> x$, $AO(x)$ [Big-O: "of order" x], rather than $A = x$.

For those working on integrating the mathematical rigor in risk bearing they provide a great source. We always know one-side, not the other. We know the lowest value we are willing to pay for insurance, not necessarily the upper bound (or vice versa).[a]

---

a The way this connects asymmetry to robustness is as follows. Is robust what does not produce variability across perturbation of parameters of the probability distribution. If there is change, but with an asymmetry, i.e. a concave or convex response to such perturbations, the classification is fragility and antifragility, respectively, see [153].

(a function of the fourth moment) came from the largest observation –see Table 2.3. For a Gaussian the maximum contribution over the same time span should be around .008 $\pm$ .0028. For the S&P 500 it was about 80 per cent. This tells us that we dont know anything about kurtosis. Its sample error is huge; or it may not exist so the measurement is heavily

> **Principle 2.1** (Epistemology: the invisibility of the generator.)
>
> - *We do not observe probability distributions, just realizations.*
>
> - *A probability distribution cannot tell you if the realization belongs to it.*
>
> - *You need a meta-probability distribution to discuss tail events.*

Figure 2.12: *The Problem of Induction. The philosophical problem of enumerative induction, expressed in the question:*
"How many white swans do you need to count before ruling out the future occurrence of a black one?"
maps surprisingly perfectly to our problem of the working of the law of large numbers:
"How much data do you need before making a certain claim with an acceptable error rate?"
It turns out that the very nature of statistical inference reposes on a clear definition and quantitative measure of the inductive mechanism. It happens that, under fat tails, we need considerably more data; as we will see in Chapters 6 and 7 there is a way to gauge the relative speed of the inductive mechanism, even if ultimately the problem of induction cannot be perfectly solved.
The problem said of induction is generally misattributed to Hume, [158] .

sample dependent. If we dont know anything about the fourth moment, we know nothing about the stability of the second moment. It means we are not in a class of distribution that allows us to work with the variance, even if it exists. This is finance.

For silver futures, in 46 years 94 per cent of the kurtosis came from one observation. We cannot use standard statistical methods with financial data. GARCH (a method popular in academia) does not work because we are dealing with squares. The variance of the squares is analogous to the fourth moment. We do not know the variance. But we can work very easily with Pareto distributions. They give us less information, but nevertheless, it is more rigorous if the data are uncapped or if there are any open variables.

Table 2.3, for financial data, debunks all the college textbooks we are currently using. A lot of econometrics that deals with squares goes out of the window. This explains why economists cannot forecast what is going on they are using the wrong methods. It will work within the sample, but it will not work outside the sample. If we say that variance (or kurtosis) is infinite, we are not going to observe anything that is infinite within a sample.

Figure 2.13: *A Discourse to Show that Skeptical Philosophy is of Great Use in Science by François de La Mothe Le Vayer (1588-1672), apparently Bishop Huet's source. Every time I find a original thinker who figured out the skeptical solution to the Black Swan problem, it turns out that he may just be cribbing a predecessor –not maliciously, but we forget to dig to the roots. As we insist, "Hume's problem" has little to do with Hume, who carried the heavy multi-volume Dictionary of Pierre Bayle (his predecessors) across Europe. I thought it was Huet who was as one digs, new predecessors crop up*
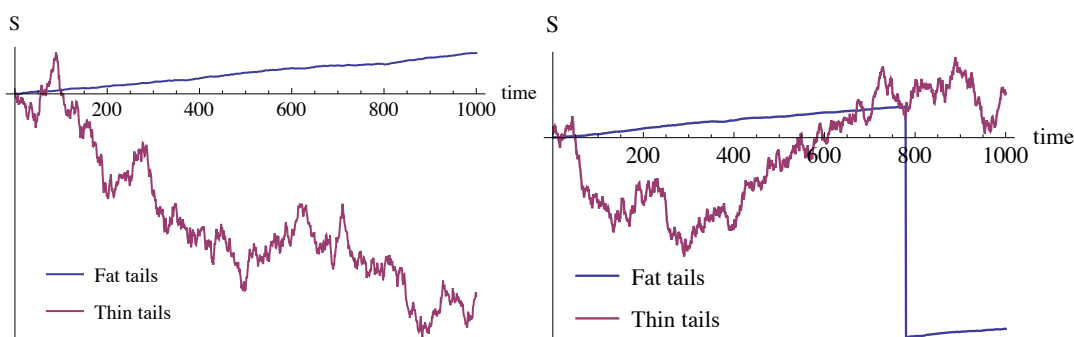
Figure 2.14: *It is not possible to "accept" thin tails, very easy to reject thintailedness. One distribution can produce jumps and quiet days will not help rule out their occurrence.*

Principal component analysis (Figure 2.15) is a dimension reduction method for big data and it works beautifully with thin tails (at least sometimes). But if there is not enough data there is an illusion of a structure. As we increase the data (the *n* variables), the structure becomes flat (something called in some circles the "Wigner effect" for random matrices, after Eugene Wigner —do not confuse with Wigner's discoveries about the dislocation of atoms under radiation). In the simulation, the data that has absolutely no structure: principal components (PCs) should be all equal (asymptotically, as data becomes large); but the small sample effect causes the ordered PCs to show a declining slope. We have zero correlation on the matrix. For a fat tailed distribution (the lower section), we need a lot more data for the spurious correlation to wash out i.e. dimension reduction does not work with fat tails.

Table 2.2: *An example of a power law*

Richer than 1 million	1 in 62.5
Richer than 2 million	1 in 250
Richer than 4 million	1 in 1,000
Richer than 8 million	1 in 4,000
Richer than 16 million	1 in 16,000
Richer than 32 million	1 in ?

Table 2.3: *Kurtosis from a single observation for financial data* $\dfrac{Max\left(X^4_{t-\Delta ti}\right)^n_{i=0}}{\sum^n_{i=0} X^4_{t-\Delta ti}}$

Security	Max Q	Years.
Silver	0.94	46.
SP500	0.79	56.
CrudeOil	0.79	26.
Short Sterling	0.75	17.
Heating Oil	0.74	31.
Nikkei	0.72	23.
FTSE	0.54	25.
JGB	0.48	24.
Eurodollar Depo 1M	0.31	19.
Sugar	0.3	48.
Yen	0.27	38.
Bovespa	0.27	16.
Eurodollar Depo 3M	0.25	28.
CT	0.25	48.
DAX	0.2	18.

2.5 WHERE ARE THE HIDDEN PROPERTIES?

The following summarizes everything that I wrote in *The Black Swan* (a message that somehow took more than a decade to go through without distortion. Distributions can be one-tailed (left or right) or two-tailed. If the distribution has a fat tail, it can be fat tailed one tail or it can be fat tailed two tails. And if is fat tailed one tail, it can be fat tailed left tail or fat tailed right tail.

See Figure 2.17 for the intuition: if it is fat tailed and we look at the sample mean, we observe fewer tail events. The common mistake is to think that we can naively derive the mean in the presence of one-tailed distributions. But there are unseen rare events and with time these will fill in. But by definition, they are low probability events. The trick is to estimate the distribution and then derive the mean. This is called plug-in estimation, see Table 2.4. It is not done by measuring the directly observable sample mean which is biased with fat-tailed distributions. This is why, outside a crisis, banks seem to make large profits. Then once in a while they lose everything and more and have to be bailed out by the taxpayer. The way we handle this is by differentiating the true mean (which I call "shadow") from the realized mean, as in the Tableau in Table 2.4.

We can also do that for the Gini coefficient to estimate the "shadow" one rather than the naively observed one.
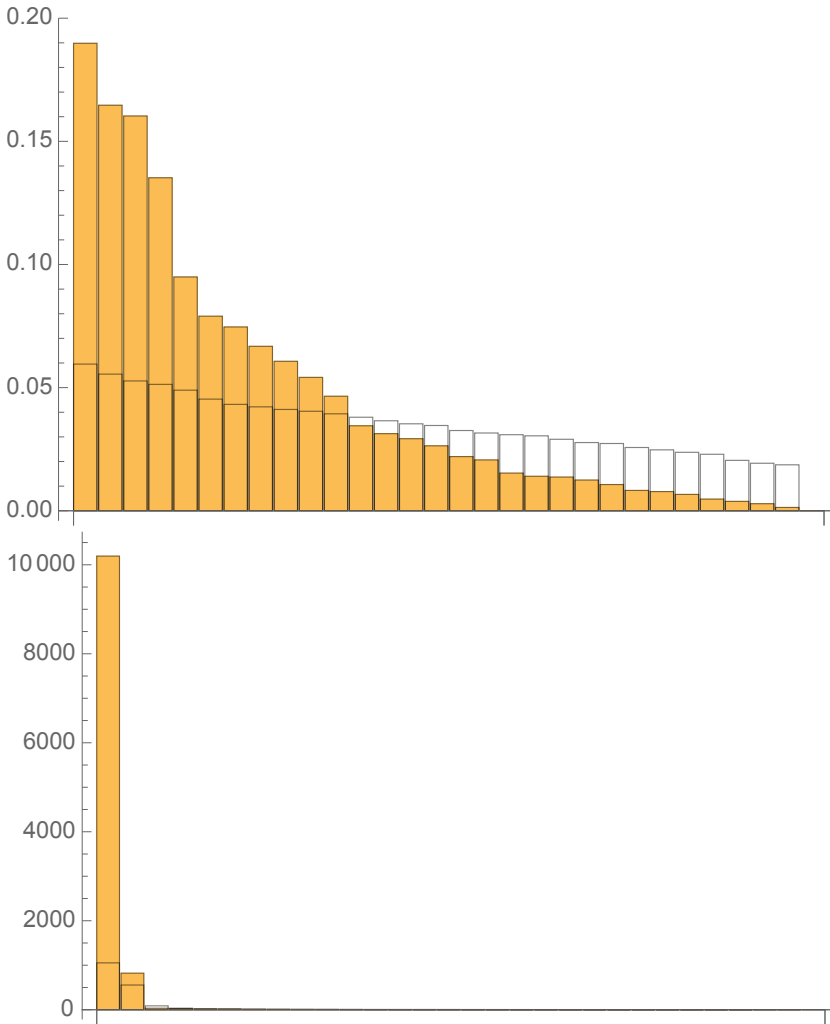
Figure 2.15: **Spurious PCAs Under Fat Tails:** *A Monte Carlo experiment that shows how spurious correlations and covariances are more acute under fat tails. Principal Components ranked by variance for 30 Gaussian uncorrelated variables, n=100 (above) and 1000 data points, and principal components ranked by variance for 30 Stable Distributed (with tail $\alpha = \frac{3}{2}$, symmetry $\beta = 1$, centrality $\mu = 0$, scale $\sigma = 1$) (below). Both are "uncorrelated" identically distributed variables.*

This is what we mean when we say that the "empirical" distribution is not "empirical". In other words: 1) there is a wedge between population and sample attributes and, 2) even exhaustive historical data can be seen as mere sampling from a broader phenomenon.

Once we have figured out the distribution, we can estimate the statistical mean. This works much better than directly measuring the sample mean. For a Pareto distribution, for instance, 98% of observations are below the mean. There is a bias in the observed mean. But once we know that we have a Pareto distribution, we should ignore the sample mean and look elsewhere.

Note that the field of Extreme Value Theory [84] [60] [85] focuses on tail properties, not the mean or statistical inference.

Figure 2.16: *A central asymmetry: the difference between absence of evidence and evidence of absence is compounded by fat tails. It requires a more elaborate understanding of random events —or a more naturalistic one. Courtesy Stefan Gasic.*
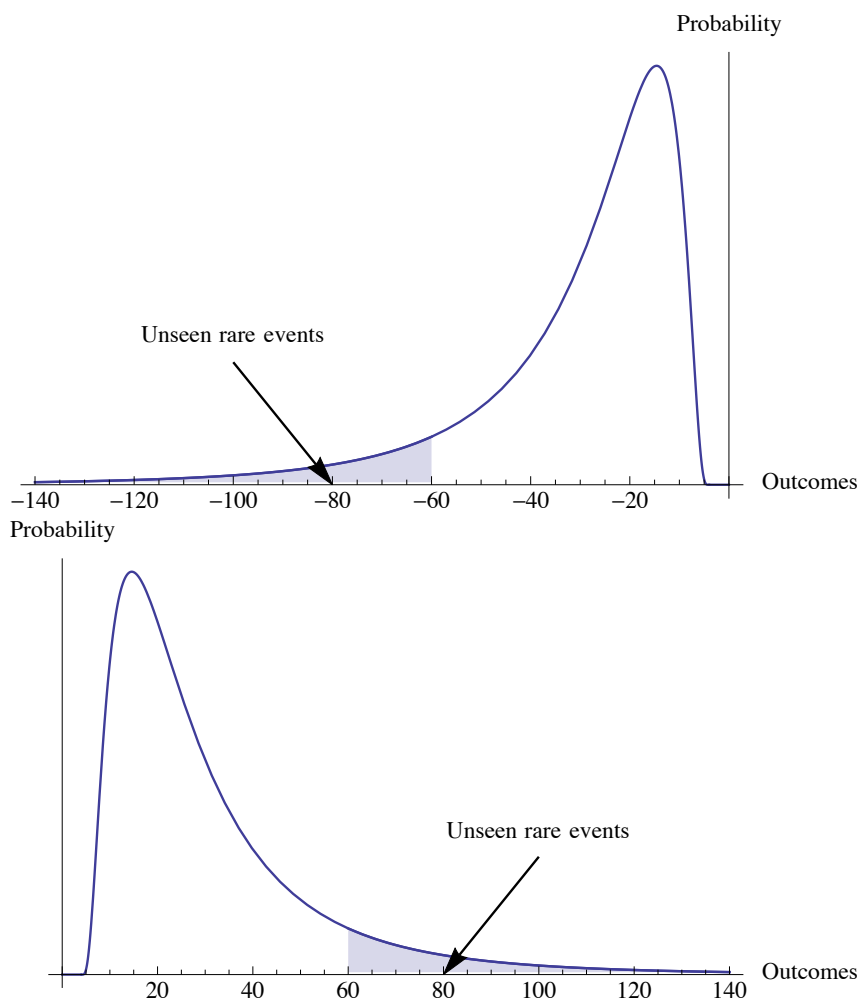
Figure 2.17: *Shadow Mean at work: Below: Inverse Turkey Problem The unseen rare event is positive. When you look at a positively skewed (antifragile) time series and make (nonparametric) inferences about the unseen, you miss the good stuff an underestimate the benefits. Above: The opposite problem. The filled area corresponds to what we do not tend to see in small samples, from insufficiency of data points. Interestingly the shaded area increases with model error.*

Table 2.4: *Shadow mean, sample mean and their ratio for different minimum thresholds. In bold the values for the 145k threshold. Rescaled data. From Cirillo and Taleb [34]*

Thresh.$\times 10^3$	Shadow$\times 10^7$	Sample$\times 10^7$	Ratio
50	1.9511	1.2753	1.5299
100	2.3709	1.5171	1.5628
145	**3.0735**	**1.7710**	**1.7354**
300	3.6766	2.2639	1.6240
500	4.7659	2.8776	1.6561
600	5.5573	3.2034	1.7348

2.6 BAYESIAN SCHMAYESIAN

In the absence of reliable information, Bayesian methods can be of little help. This author has faced since the publication of *The Black Swan* numerous questions concerning the use of something vaguely Bayesian to solve problems about the unknown under fat tails. Unless one cannot manufacture information beyond what's available, no technique, Bayesian nor Schmayesian can help. The key is that one needs a reliable prior, something not readily observable (see Diaconis and Friedman [48] for the difficulty for an agent in formulating a prior).

A problem is the speed of updating, as we will cover in Chapter 6, which is highly distribution dependent. The mistake in the rational expectation literature is to believe that two observers supplied with the same information would necessarily converge to the same view. Unfortunately, the conditions for that to happen in real time or to happen at all are quite specific.

One of course can use Bayesian methods (under adequate priors) for the estimation of parameters if 1) one has a clear idea about the range of values (say from universality classes or other stable basins) and 2) these parameters follow a tractable distribution with low variance such as, say, the tail exponent of a Pareto distribution (which is inverse-gamma distributed), [9].

> **Moral Hazard in Financial Education**: The most depressing experience I've had was when I taught a course on Fat Tails at the University of Massachussetts Amherst, at the business school, during my very brief stint there. One PhD student in finance told me bluntly that he liked my ideas but that a financial education career commanded "the highest salary in the land" (that is, among all other specialties in education). He preferred to use Markowitz methods as these were used by other professors, hence allowed him to get his papers published, and get a high paying job. I was disgusted, but predicted he would subsequently have a very successful career writing non-papers. He did.

2.7 x VS $f(x)$, EXPOSURES TO x CONFUSED WITH KNOWLEDGE ABOUT x

Take X a random or nonrandom variable, and $F(X)$ the exposure, payoff, the effect of X on you, the end bottom line. (X is often is higher dimensions but let's assume to simplify that it is a simple one-dimensional variable).

Practitioners and risk takers often observe the following disconnect: people (nonpractitioners) talking X (with the implication that practitioners should care about X in running their affairs) while practitioners think about $F(X)$, nothing but $F(X)$. And the straight confusion since Aristotle between X and $F(X)$ has been chronic as discussed in *Antifragile* [161] which is written around that theme. Sometimes people mention $F(X)$ as utility but miss the full payoff. And the confusion is at two level: one, simple confusion; second, in the decision-science literature, seeing the difference and not realizing that action on $F(X)$ is easier than action on X.

- The variable X can be unemployment in Senegal, $F_1(X)$ is the effect on the bottom line of the IMF, and $F_2(X)$ is the effect on your grandmother (which I assume is minimal).

- X can be a stock price, but you own an option on it, so $F(X)$ is your exposure an option value for X, or, even more complicated the utility of the exposure to the option value.

- X can be changes in wealth, $F(X)$ the convex-concave way it affects your well-being. One can see that $F(X)$ is vastly more stable or robust than X (it has thinner tails).

Convex vs. linear functions of a variable X Consider Fig. 2.19; confusing $F(X)$ (on the vertical) and X (the horizontal) is more and more significant when $F(X)$ is nonlinear. The more convex $F(X)$, the more the statistical and other properties of $F(X)$ will be divorced from those of X. For instance, the mean of $F(X)$ will be different from $F(\text{Mean of} X)$, by Jensen's inequality. But beyond Jensen's inequality, the difference in risks between the two will be more and more considerable. When it comes to probability, the more nonlinear F, the less the probabilities of X matters compared to that of F. Moral of the story: focus on F, which we can alter, rather than on the measurement of the elusive properties of X.

Probability Distribution of X *Probability Distribution of F(X)*
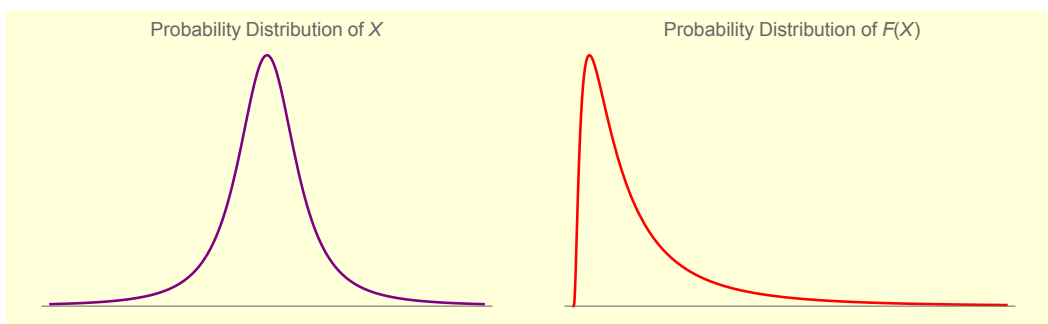
Figure 2.18: *The Conflation Problem X (random variable) and F(X) a function of it (or payoff). If F(X) is convex we don't need to know much about it –it becomes an academic problem. And it is safer to focus on transforming F(X) than X.*

Limitations of knowledge What is crucial, our limitations of knowledge apply to X not necessarily to $F(X)$. We have no control over X, some control over $F(X)$. In some cases a very, very large control over $F(X)$.

The danger with the treatment of the Black Swan problem is as follows: people focus on X ("predicting X"). My point is that, although we do not understand X, we can deal with it by working on F which we can understand, while others work on predicting X which we can't because small probabilities are incomputable, particularly in "fat tailed" domains. $F(x)$ is how the end result affects you.
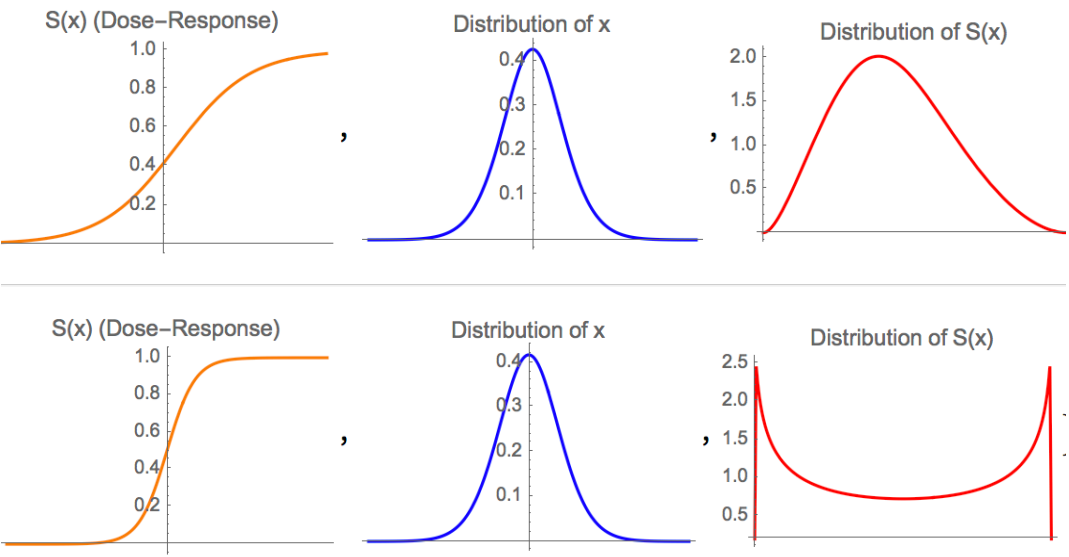
Figure 2.19: *The Conflation Problem: a convex-concave transformation of a fat tailed X produces a thin tailed distribution (above). A sigmoidal transformation (below) that is bounded on a distribution in* $(-\infty, \infty)$ *produces an ArcSine distribution, with compact support.*
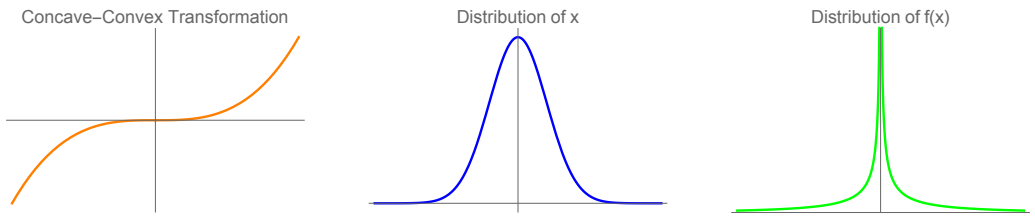
Figure 2.20: *A concave-convex transformation (of the style of a probit –an inverse CDF for the Gaussian– or of a logit) makes the tails of the distribution of* $f(x)$ *thicker*

The probability distribution of $F(X)$ is markedly different from that of X, particularly when $F(X)$ is nonlinear. We need a nonlinear transformation of the distribution of X to get $F(X)$. We had to wait until 1964 to start a discussion on "convex transformations of random variables", Van Zwet (1964)[184] –as the topic didn't seem important before.

Ubiquity of S curves F is almost always nonlinear (actually I know of no exception to nonlinearity), often "S curved", that is convex-concave (for an increasing function). See the longer discussion in F.

2.8 RUIN AND PATH DEPENDENCE

Let us finish with path dependence and time probability. Our greatgrandmothers did understand fat tails. These are not so scary; we figured out how to survive by making rational decisions based on deep statistical properties.

Path dependence is as follows. If I iron my shirts and then wash them, I get vastly different results compared to when I wash my shirts and then iron them. My first work,

Fragility and Antifragility   When $F(X)$ is concave (fragile), errors about X can translate into extreme negative values for $F(X)$. When $F(X)$ is convex, one is largely immune from severe negative variations. In situations of trial and error, or with an option, we do not need to understand X as much as our exposure to the risks. Simply the statistical properties of X are swamped by those of H. The point of *Antifragile* is that exposure is more important than the naive notion of "knowledge", that is, understanding X.

The more nonlinear F the less the probabilities of X matters in the probability distribution of the final package F.

Many people confuse the probabilites of X with those of F. I am serious: the *entire* literature reposes largely on this mistake. For Baal's sake, focus on F, not X.

Dynamic Hedging [156], was about how traders avoid the "absorbing barrier" since once you are bust, you can no longer continue: anything that will eventually go bust will lose *all* past profits.

The physicists Ole Peters and Murray Gell-Mann [126] shed new light on this point, and revolutionized decision theory showing that a key belief since the development of applied probability theory in economics was wrong. They pointed out that all economics textbooks make this mistake; the only exception are by information theorists such as Kelly and Thorp.

Let us explain ensemble probabilities.

Assume that 100 of us, randomly selected, go to a casino and gamble. If the 28th person is ruined, this has no impact on the 29th gambler. So we can compute the casinos return using the law of large numbers by taking the returns of the 100 people who gambled. If we do this two or three times, then we get a good estimate of what the casinos edge is. The problem comes when ensemble probability is applied to us as individuals. It does not work because if one of us goes to the casino and on day 28 is ruined, there is no day 29. This is why Cramer showed insurance could not work outside what he called the Cramer condition, which excludes possible ruin from single shocks. Likewise, no individual investor will achieve the alpha return on the market because no single investor has infinite pockets (or, as Ole Peters has observed, is running his life across branching parallel universes). We can only get the return on the market under strict conditions.

Time probability and ensemble probability are not the same. This only works if the risk takers has an allocation policy compatible with the Kelly criterion[97],[176] using logs. Peters wrote three papers on time probability (one with Murray Gell-Mann) and showed that a lot of paradoxes disappeared.

Let us see how we can work with these and what is wrong with the literature. If we visibly incur a tiny risk of ruin, but have a frequent exposure, it will go to probability one over time. If we ride a motorcycle we have a small risk of ruin, but if we ride that motorcycle a lot then we will reduce our life expectancy. The way to measure this is:

Behavioral finance so far makes conclusions from statics not dynamics, hence misses the picture. It applies trade-offs out of context and develops the consensus that people irrationally overestimate tail risk (hence need to be "nudged" into taking more of these exposures). But the catastrophic event is an absorbing barrier. No risky exposure can be analyzed in isolation: risks accumulate. If we ride a motorcycle, smoke, fly our own propeller plane, and join the mafia, these risks add up to a near-certain premature death. Tail risks are not a renewable resource.

Better be convex than right: In the fall of 2017, a firm went bust betting against volatility –they were predicting lower real market volatility (rather, variance) than "expected" by the market. *They were correct in the prediction, but went bust nevertheless.* They were just very concave in the payoff function. Recall that x is not $f(x)$ and that in the real world there are almost no linear $f(x)$.

The following example can show us how. Consider the following payoff in the figure below. The payoff function is $f(x) = 1 - x^2$ daily, meaning if x moves by up to 1 unit (say, standard deviation), there is a profit, losses beyond. This is a typical contract called "variance swap".

$f(x)=1-x^2$
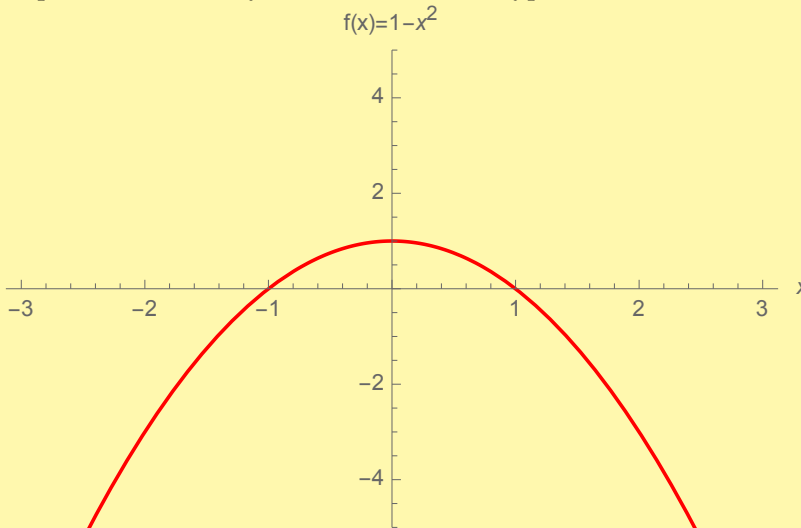
Now consider the following two types successions of deviations of x for 7 days (expressed in standard deviations).

Succession 1 (thin tails): $\{1,1,1,1,1,0,0\}$. Mean variation= 0.71. P/L= 2.

Succession 2 (fat tails): $\{0,0,0,0,0,0,5\}$. Mean variation= 0.71 (same). P/L=-18 (bust, really bust).

In both cases they forecast right, but the lumping of the volatility –the fatness of tails– made a huge difference.

This in a nutshell explains why, in the real world, "bad" forecasters can make great traders and decision makers and vice versa –something every operator knows but that the mathematically and practically unsophisticated "forecasting" literature, centuries behind practice, misses.

Every risk taker who survived understands this. Warren Buffett understands this. Goldman Sachs understands this. They do not want small risks, they want zero risk because that is the difference between the firm surviving and not surviving over twenty, thirty, one hundred years. This attitude to tail risk can explain that Goldman Sachs is 149 years old –it ran as partnership with unlimited liability for approximately the first 130 years, but was bailed out once in 2009, after it became a bank. This is not in the decision theory literature but we (people with skin in the game) practice it every day. We take a unit, look at how long a life we wish it to have and see by how much the life expectancy is reduced by *repeated* exposure.

> **Remark 2.2** (Psychology of decision making)
> *The psychological literature focuses on one-single episode exposures and narrowly defined cost-benefit analyses. Some analyses label people as paranoid for overestimating small risks, but don't get that if we had the smallest tolerance for collective tail risks, we would not have made it for the past several million years.*

Next let us consider layering, why systemic risks are in a different category from individual, idiosyncratic ones. Look at the (inverted) pyramid in Fig. 2.22: the worst-case scenario is not that an individual dies. It is worse if your family, friends and pets die. It is worse if you die and your arch enemy survives. They collectively have more life expectancy lost from a terminal tail event.

So there are layers. The biggest risk is that the entire ecosystem dies. The precautionary principle puts structure around the idea of risk for units expected to survive.

Ergodicity in this context means that your analysis for ensemble probability translates into time probability. If it doesn't, ignore ensemble probability altogether.

2.9 WHAT TO DO?

To summarize, we first need to make a distinction between mediocristan and Extremistan, two separate domains that about never overlap with one another. If we dont make that distinction, we dont have any valid analysis. Second, if we dont make the distinction
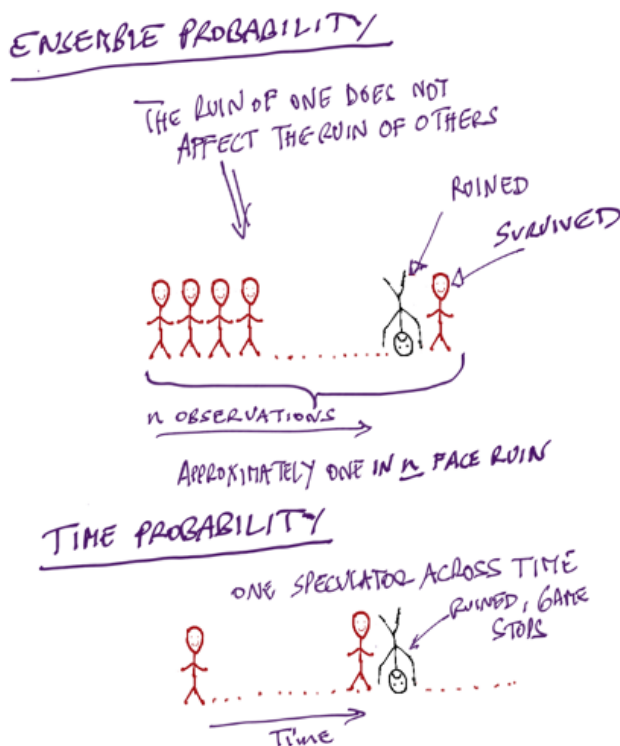
Figure 2.21: *Ensemble probability vs. time probability. The treatment by option traders is done via the absorbing barrier. I have traditionally treated this in Dynamic Hedging [156] and Antifragile[153] as the conflation between X (a random variable) and f(X) a function of said r.v., which may include an absorbing state.*
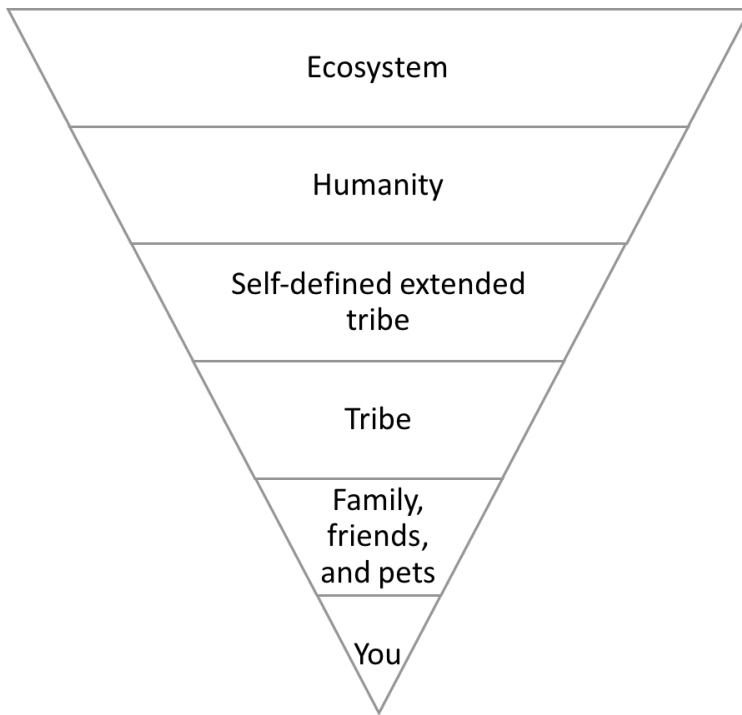
Figure 2.22: *A hierarchy for survival. Higher entities have a longer life expectancy, hence tail risk matters more for these. Lower entities such as you and I are renewable.*

Principle 2.2 (Repetition of exposures)
Focus only on the reduction of life expectancy of the unit assuming repeated exposure at a certain density or frequency.

between time probability (path dependent) and ensemble probability (path independent) we dont have a valid analysis.

The next phase of the *Incerto* project is to gain understanding of fragility, robustness, and, eventually, anti-fragility. Once we know something is fat-tailed, we can use heuristics to see how an exposure there reacts to random events: how much is a given unit harmed by them. It is vastly more effective to focus on being insulated from the harm of random events than try to figure them out in the required details (as we saw the inferential errors under fat tails are huge). So it is more solid, much wiser, more ethical, and more effective to focus on detection heuristics and policies rather than fabricate statistical properties.

The beautiful thing we discovered is that everything that is fragile has to present a concave exposure [153] similar –if not identical –to the payoff of a short option, that is, a negative exposure to volatility. It is nonlinear, necessarily. It has to have harm that accelerates with intensity, up to the point of breaking. If I jump 10m I am harmed more than 10 times than if I jump one metre. That is a necessary property of fragility. We just need to look at acceleration in the tails. We have built effective stress testing heuristics based on such an option-like property [169].

In the real world we want simple things that work [78]; we want to impress our accountant and not our peers. (My argument in the latest instalment of the *Incerto*, *Skin in the Game* is that systems judged by peers and not evolution rot from overcomplication). To survive we need to have clear techniques that map to our procedural intuitions.

The new focus is on how to detect and measure convexity and concavity. This is much, much simpler than probability.

NOTES

[1]"Gabish" rather than "capisce"? Gabish is the recreated pronunciation of Siculo-Galabrez (Calabrese); the "p" used to sound like a "b" and the "g" like a Semitic kof, a hard K, from Punic. Much like capicoli is "gabagool".

[2]More formally: let X be a random variable belonging to the class of distributions with a "power law" right tail:

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{2.1}$$

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 $k > 0$. We can apply the same to the negative domain.

3 | OVERVIEW OF FAT TAILS, PART I, THE UNIVARIATE CASE[†]

THIS CHAPTER is organized as follows. We look at three levels of fat-tails with more emphasis on the intuitions and heuristics than formal mathematical differences, which will be pointed out later in the discussions of limit theorems. The three levels are:

- Fat tails, entry level (sort of), i.e., finite moments
- Subexponential class
- Power Law class

Level one will be the longest as we will use it to build intuitions. While, mathematically, it is the least used (fat tails are usually associated with power laws and limit behavior); analytically and practically it is relied upon the most (we can get the immediate consequences of fat-tailedness with little effort, the equivalent of a functional derivative that provides a good grasp of local sensitivities). For instance, as a trader, the author was able to get most of the effect of fattailedness with a simple heuristic of averaging option prices across two volatilities.

3.1 LEVEL 1: FAT TAILS, BUT FINITE MOMENTS

In this section, we link fatness of tails to higher moments, but stay in the situation that no moment is infinite.

3.1.1 A Simple Heuristic to Create Mildly Fat Tails

Remark 3.1 (Fat Tails and Jensen's inequality)
For a Gaussian distribution (and, possibly, members of the location-scale family of distributions), tail probabilities are convex to the scale of the distribution, here the variance σ^2. This allows us to fatten the tails by "stochasticizing" the variance, checking the effect of Jensen's inequality on the total.

Heteroscedasticity is the general technical term often used in time series analysis to characterize a process with fluctuating scale. Our method "stochasticizes", that is perturbates the variance of the distribution under the constraint of conservation of the mean.
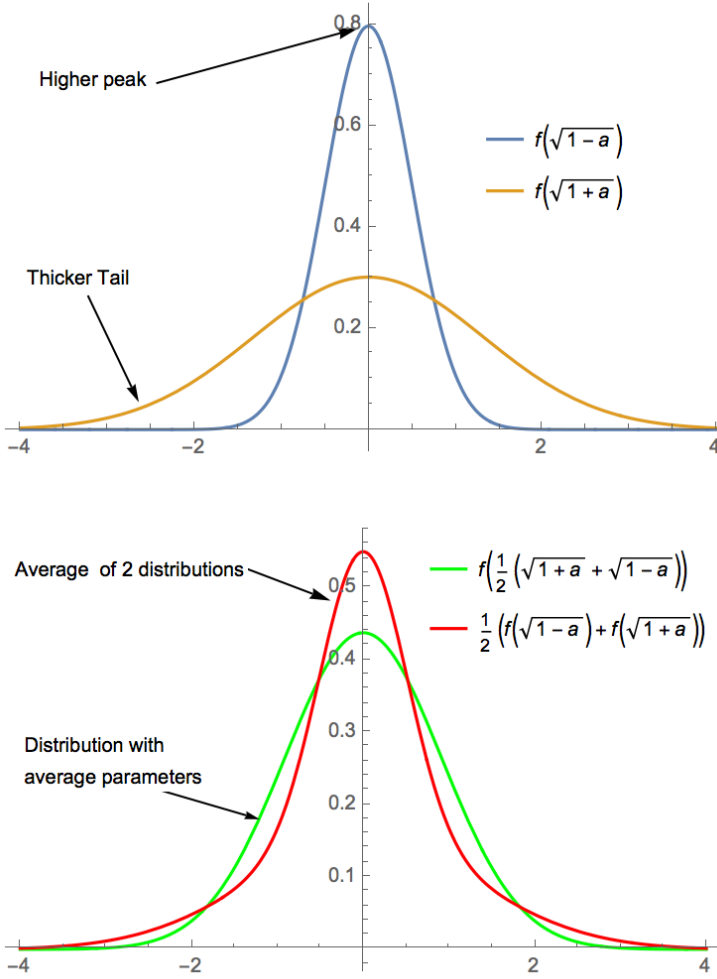
Figure 3.1: *How random volatility creates fatter tails owing to the convexity of some parts of the density to the scale of the distribution.*

But note that *any* fat tailed process, even a power law, can be *in sample* (that is finite number of observations necessarily discretized) described by a simple Gaussian process with changing variance, a regime switching process, or a combination of Gaussian plus a series of variable jumps (though not one where jumps are of equal size, see the summary in [118]).

This method will also allow us to answer the great question: "where do the tails start?" in 3.3.

Let $f(\sqrt{a}, x)$ be the density of the normal distribution (with mean 0) as a function of the variance for a given point x of the distribution. Let us compare $f\left(\frac{1}{2}\left(\sqrt{1-a} + \sqrt{a+1}, x\right)\right)$ to $\frac{1}{2}\left(f\left(\sqrt{1-a}, x\right) + f\left(\sqrt{a+1}, x\right)\right)$ by Jensen's inequality. We assume the average σ^2 constant, but the discussion works just as well if we just assumed σ constant —it is a long debate whether one should constraint on the mean of the variance or that of the standard deviation.

Since higher moments increase under fat tails, as compared to lower ones, it should be possible to simply increase fat tailedness while keeping lower moments (the first two or three) invariant.

3.1.2 A Variance-preserving heuristic

Keep $\mathbb{E}(X^2)$ constant and increase $\mathbb{E}(X^4)$, by "stochasticizing" the variance of the distribution, since $\mathbb{E}(X^4)$ is itself analog to the variance of $\mathbb{E}(X^2)$ measured across samples – $\mathbb{E}(X^4)$ is the noncentral equivalent of $\mathbb{E}\left((X^2 - \mathbb{E}(X^2))^2\right)$ so we will focus on the simpler version outside of situations where it matters. Further, we will do the "stochasticizing" in a more involved way in later sections of the chapter.

An effective heuristic to get some intuition about the effect of the fattening of tails consists in simulating a random variable set to be at mean 0, but with the following variance-preserving tail fattening trick: the random variable follows a distribution $N(0, \sigma\sqrt{1-a})$ with probability $p = \frac{1}{2}$ and $N\left(0, \sigma\sqrt{1+a}\right)$ with the remaining probability $\frac{1}{2}$, with $0 \leqslant a < 1$.

The characteristic function is

$$\phi(t, a) = \frac{1}{2} e^{-\frac{1}{2}(1+a)t^2\sigma^2}\left(1 + e^{at^2\sigma^2}\right) \tag{3.1}$$

Odd moments are nil. The second moment is preserved since

$$M(2) = (-i)^2 \partial^{t,2}\phi(t)|_0 = \sigma^2 \tag{3.2}$$

and the fourth moment

$$M(4) = (-i)^4 \partial^{t,4}\phi|_0 = 3\left(a^2 + 1\right)\sigma^4 \tag{3.3}$$

which puts the traditional kurtosis at $3\left(a^2 + 1\right)$ (assuming we do not remove 3 to compare to the Gaussian). This means we can get an "implied a from kurtosis. The value of a is roughly the mean deviation of the stochastic volatility parameter "volatility of volatility" or Vvol in a more fully parametrized form.

Limitations of the simple heuristic This heuristic, while useful for intuition building, is of limited powers as it can only raise kurtosis to twice that of a Gaussian, so it should be used only to getting some intuition about the effects of the convexity. Section 3.1.3 will present a more involved technique.

> **Remark 3.2** (Peaks)
> *As Figure 3.4 shows: fat tails manifests themselves with higher peaks, a concentration of observations around the center of the distribution.*

3.1.3 Fattening of Tails With Skewed Variance

We can improve on the fat-tail heuristic in 3.1.1, (which limited the kurtosis to twice the Gaussian) as follows. We Switch between Gaussians with variance:

$$\begin{cases} \sigma^2(1+a), & \text{with probability } p \\ \sigma^2(1+b), & \text{with probability } 1-p \end{cases} \tag{3.4}$$

with $p \in [0,1)$, both a, b $\in$ (-1,1) and b= $-a\frac{p}{1-p}$, giving a characteristic function:

$$\phi(t,a) = p\, e^{-\frac{1}{2}(a+1)\sigma^2 t^2} - (p-1)\, e^{-\frac{\sigma^2 t^2 (ap+p-1)}{2(p-1)}}$$

with Kurtosis $\frac{3((1-a^2)p-1)}{p-1}$ thus allowing polarized states and high kurtosis, all variance preserving, conditioned on, when a > (<)0, a < (>)$\frac{1-p}{p}$.

Thus with, say, p = 1/1000, and the corresponding maximum possible a = 999, kurtosis can reach as high a level as 3000.

This heuristic approximates quite well the effect on probabilities of a lognormal weighting for the characteristic function

$$\phi(t,V) = \int_0^\infty \frac{e^{-\frac{t^2 v}{2} - \frac{\left(\log(v) - v0 + \frac{Vv^2}{2}\right)^2}{2Vv^2}}}{\sqrt{2\pi v V v}}\, dv \tag{3.5}$$

where v is the variance and Vv is the second order variance, often called volatility of volatility. Thanks to integration by parts we can use the Fourier transform to obtain all varieties of payoffs (see Gatheral [73]). But the absence of a closed-form distribution can be remedied as follows, with the use of distributions for the variance that are analytically more tractable.
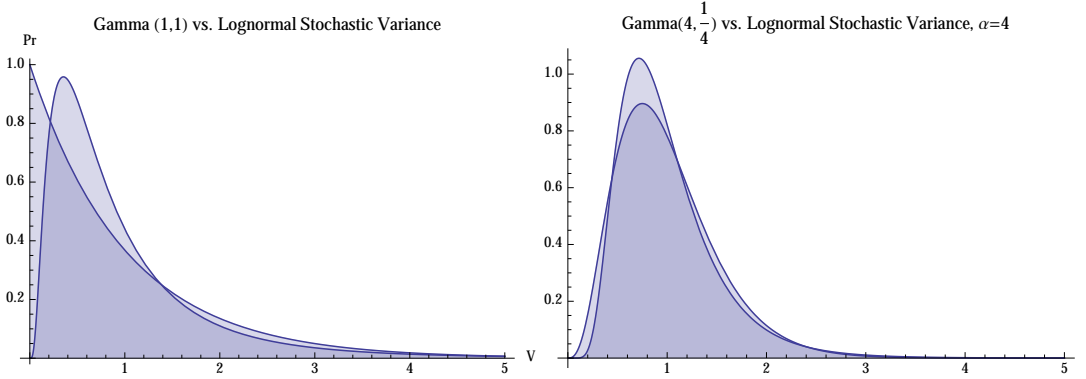
Figure 3.2: *Stochastic Variance: Gamma distribution and Lognormal of same mean and variance.*

Gamma Variance   The gamma distribution applied to the variance of a Gaussian is is a useful shortcut for a full distribution of the variance, which allows us to go beyond the narrow scope of heuristics [25]. It is easier to manipulate analytically than the Lognormal.

Assume that the variance of the Gaussian follows a gamma distribution.

$$\Gamma_a(v) = \frac{v^{a-1}\left(\frac{V}{a}\right)^{-a} e^{-\frac{av}{V}}}{\Gamma(a)}$$

with mean V and standard deviation $\frac{V}{\sqrt{a}}$. Figure 3.2 shows the matching to a lognormal with same first two moments where we calibrate the lognormal to mean $\frac{1}{2}\log\left(\frac{aV^3}{aV+1}\right)$ and

Gaussian With Gamma Variance
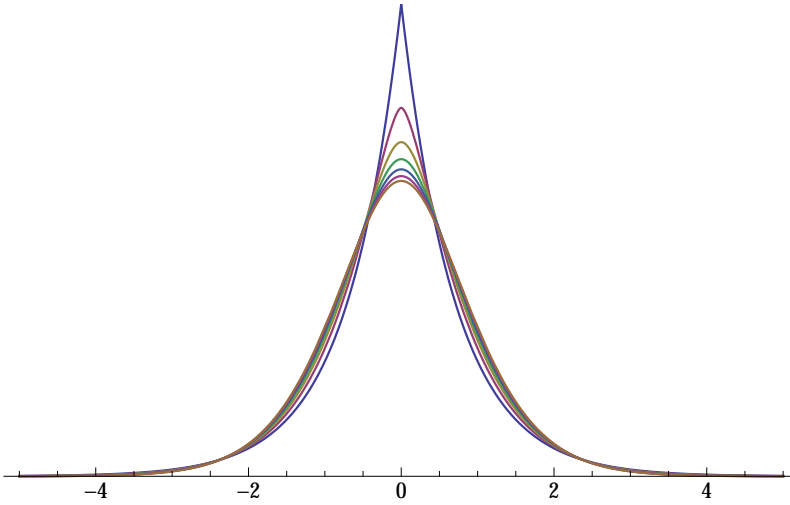
Figure 3.3: *Stochastic Variance using Gamma distribution by perturbating α in equation 3.7.*

standard deviation $\sqrt{-\log\left(\frac{aV}{aV+1}\right)}$. The final distribution becomes (once again, assuming the same mean as a fixed volatility situation:

$$f_{a,V}(x) = \int_0^\infty \frac{e^{-\frac{(x-\mu)^2}{2v}}}{\sqrt{2\pi}\sqrt{v}} \Gamma_a(v) \mathrm{d}v \tag{3.6}$$

allora:

$$f_{\alpha,V}(x) = \frac{2^{\frac{3}{4}-\frac{a}{2}} a^{\frac{a}{2}+\frac{1}{4}} V^{-\frac{a}{2}-\frac{1}{4}} |x-\mu|^{a-\frac{1}{2}} K_{a-\frac{1}{2}}\left(\frac{\sqrt{2}\sqrt{a}|x-\mu|}{\sqrt{V}}\right)}{\sqrt{\pi}\Gamma(a)}. \tag{3.7}$$

where $K_n(z)$ is the Bessel K function, which satisfies the differential equation $-y\left(n^2+z^2\right) + z^2 y'' + zy' = 0$.

Let us now get deeper into the different forms of stochastic volatility.

3.2 DOES STOCHASTIC VOLATILITY GENERATE POWER LAWS?

Assume the base distribution is normally distributed, $\mathcal{N}(\mu,\sigma)$. Now there are different ways to make σ, the scale, stochastic. Note that we need sigma to follow some one-tailed distribution.

- We can make σ^2 (or, possibly σ) follow a Lognormal distribution. It does not yield closed form solutions, but we can get the moments and verify it is not a power law.
- We can make σ^2 (or σ) follow a gamma distribution. It does yield closed form solutions, as we saw in the example above, in Eq. 3.7.
- We can make $\frac{1}{\sigma^2}$ –the precision parameter –follow a gamma distribution.
- We can make $\frac{1}{\sigma^2}$ follow a lognormal distribution.

Let X be any random variable with pdf $f(.)$ in the location-scale family, and λ any random variable with pdf $g(.)$; X and λ are assumed to be independent. Since by standard theorem, the moments of order p for the product and the ratio $\frac{X}{\lambda}$ are:

$$\mathbb{E}\left((XY)^p\right) = \mathbb{E}\left(X^p\right)\mathbb{E}\left(Y^p\right)$$

and

$$\mathbb{E}\left(\left(\frac{X}{\lambda}\right)^p\right) = \mathbb{E}\left(\left(\frac{1}{\lambda}\right)^p\right)\mathbb{E}\left(X^p\right).$$

(via the Mellin transform).

Note that as proprety of location-scale family, $\frac{1}{\lambda}f_{\frac{x}{\lambda}}(\frac{x}{\lambda}) = f_x(\frac{x}{\lambda})$ so, for instance, if $x \sim \mathcal{N}(0,1)$ (that is, normally distributed), then $\frac{x}{\sigma} \sim \mathcal{N}(0,\sigma)$.

3.3 THE BODY, THE SHOULDERS, AND THE TAILS

We assume tails start at the level of convexity of the segment of the probability distribution to the scale of the distribution –in other words, affected by the stochastic volatility effect.
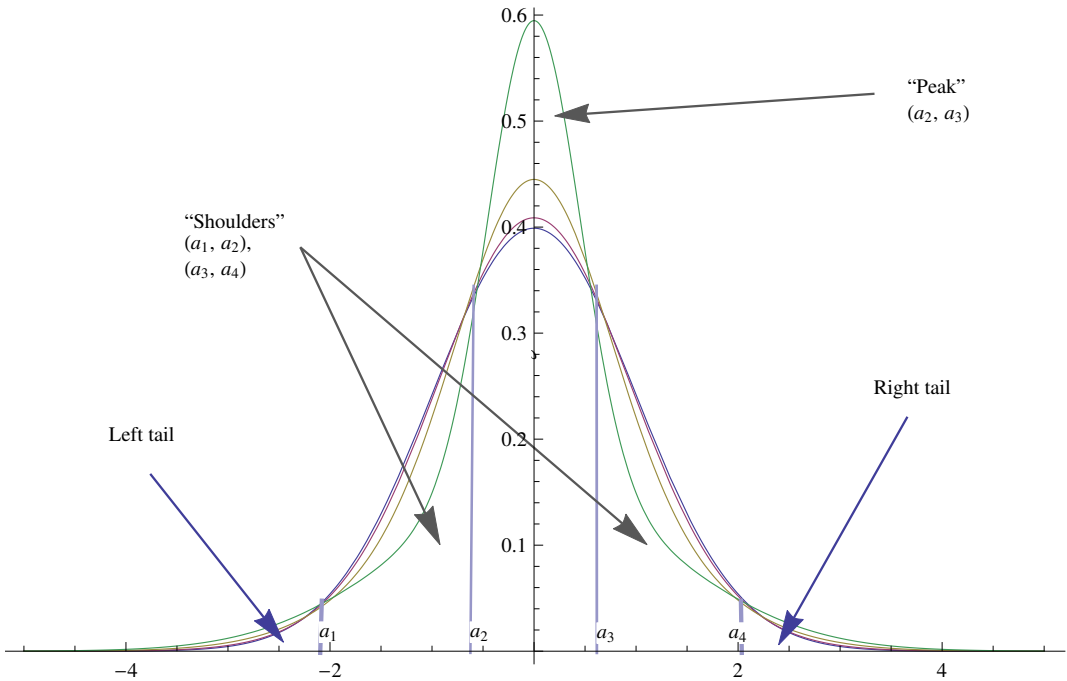
Figure 3.4: *Where do the tails start? Fatter and fatter fails through perturbation of the scale parameter σ for a Gaussian, made more stochastic (instead of being fixed). Some parts of the probability distribution gain in density, others lose. Intermediate events are less likely, tails events and moderate deviations are more likely. We can spot the crossovers a_1 through a_4. The "tails" proper start at a_4 on the right and a_1 on the left.*

Table 3.1: *Transformations for Stochastic Vol. We can see from the density of the transformations $\frac{1}{x}$ or $\frac{1}{\sqrt{x}}$ if we have a power law on hand. $\mathcal{LN}$, $\mathcal{N}$, $\mathcal{G}$ and $\mathcal{P}$ are the Lognormal, Normal, Gamma, and Pareto distributions, respectively.*

distr	$p(x)$	$p\left(\frac{1}{x}\right)$	$p\left(\frac{1}{\sqrt{x}}\right)$
$\mathcal{LN}(m,s)$	$\dfrac{e^{-\frac{(m-\log(x))^2}{2s^2}}}{\sqrt{2\pi}sx}$	$\dfrac{e^{-\frac{(m+\log(x))^2}{2s^2}}}{\sqrt{2\pi}sx}$	$\dfrac{\sqrt{\frac{2}{\pi}}e^{-\frac{(m+2\log(x))^2}{2s^2}}}{sx}$
$\mathcal{N}(m,s)$	$\dfrac{e^{-\frac{(m-x)^2}{2s^2}}}{\sqrt{2\pi}s}$	$\dfrac{e^{-\frac{\left(m-\frac{1}{x}\right)^2}{2s^2}}}{\sqrt{2\pi}sx^2}$	$\dfrac{\sqrt{\frac{2}{\pi}}e^{-\frac{\left(m-\frac{1}{x^2}\right)^2}{2s^2}}}{sx^3}$
$\mathcal{G}(a,b)$	$\dfrac{b^{-a}x^{a-1}e^{-\frac{x}{b}}}{\Gamma(a)}$	$\dfrac{b^{-a}x^{-a-1}e^{-\frac{1}{bx}}}{\Gamma(a)}$	$\dfrac{2b^{-a}x^{-2a-1}e^{-\frac{1}{bx^2}}}{\Gamma(a)}$
$\mathcal{P}(1,\alpha)$	$\alpha x^{-\alpha-1}$	$\alpha x^{\alpha-1}$	$2\alpha x^{2\alpha-1}$

Table 3.2: *The p-moments of possible distributions for variance*

distr	$\mathbb{E}\left(X^p\right)$	$\mathbb{E}\left((\frac{1}{X})^p\right)$	$\mathbb{E}\left((\frac{1}{\sqrt{X}})^p\right)$
$\mathcal{LN}(m,s)$	$e^{mp+\frac{p^2s^2}{2}}$	$e^{\frac{1}{2}p\left(ps^2-2m\right)}$	$e^{\frac{1}{8}p\left(ps^2-4m\right)}$
$\mathcal{G}(a,b)$	$b^p(a)_p$	$\dfrac{(-1)^pb^{-p}}{(1-a)_p}$, $p < a$	fughedaboudit
$\mathcal{P}(1,\alpha)$	$\dfrac{\alpha}{\alpha-p}$, $p < \alpha$	$\dfrac{\alpha}{\alpha+p}$	$\dfrac{2\alpha}{2\alpha+p}$

The Black Swan Problem: As we saw, it is not merely that events in the tails of the distributions matter, happen, play a large role, etc. The point is that these events play the major role *and* their probabilities are not computable, not reliable for any effective use. The implication is that Black Swans do not necessarily come from fat tails; le problem can result from an incomplete assessment of tail events.

3.3.1 The Crossovers and Tunnel Effect.

Notice in Figure 3.4 a series of crossover zones, invariant to a. Distributions called "bell shape" have a convex-concave-convex shape (or quasi-concave shape).

Let X be a random variable with distribution with PDF $p(x)$ from a general class of all unimodal one-parameter continuous pdfs p_σ with support $\mathcal{D} \subseteq \mathbb{R}$ and scale parameter σ. Let $p(.)$ be quasi-concave on the domain, but neither convex nor concave. The density function $p(x)$ satisfies: $p(x) \geq p(x+\epsilon)$ for all $\epsilon > 0$, and $x > x^*$ and $p(x) \geq p(x-\epsilon)$ for all $x < x^*$ with $\{x^* : p(x^*) = \max_x p(x)\}$. The class of quasiconcave functions is defined as follows: for all x and y in the domain and $\omega \in [0,1]$,

$$p\left(\omega x + (1-\omega)y\right) \geq \min\left(p(x), p(y)\right).$$

A- If the variable is "two-tailed", that is, its domain of support $\mathcal{D}=(-\infty,\infty)$, and where $p^\delta(x) \triangleq \frac{p(x,\sigma+\delta)+p(x,\sigma-\delta)}{2}$,

1. There exist a "high peak" inner tunnel, $A_T = (a_2, a_3)$ for which the δ-perturbed σ of the probability distribution $p^\delta(x) \geq p(x)$ if $x \in (a_2, a_3)$

2. There exists outer tunnels, the "tails", for which $p^\delta(x) \geq p(x)$ if $x \in (-\infty, a_1)$ or $x \in (a_4, \infty)$

3. There exist intermediate tunnels, the "shoulders", where $p^\delta(x) \leq p(x)$ if $x \in (a_1, a_2)$ or $x \in (a_3, a_4)$

Let $A = \{a_i\}$ the set of solutions $\left\{ x : \frac{\partial^2 p(x)}{\partial \sigma^2}\big|_a = 0 \right\}$.

For the Gaussian (μ, σ), the solutions obtained by setting the second derivative with respect to σ to 0 are:

$$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(2\sigma^4 - 5\sigma^2(x-\mu)^2 + (x-\mu)^4\right)}{\sqrt{2\pi}\sigma^7} = 0,$$

which produces the following crossovers:

$$\{a_1, a_2, a_3, a_4\} = \left\{ \mu - \sqrt{\frac{1}{2}\left(5 + \sqrt{17}\right)}\sigma, \mu - \sqrt{\frac{1}{2}\left(5 - \sqrt{17}\right)}\sigma, \right.$$

$$\left. \mu + \sqrt{\frac{1}{2}\left(5 - \sqrt{17}\right)}\sigma, \mu + \sqrt{\frac{1}{2}\left(5 + \sqrt{17}\right)}\sigma \right\} \tag{3.8}$$

In figure 3.4, the crossovers for the intervals are numerically $\{-2.13\sigma, -.66\sigma, .66\sigma, 2.13\sigma\}$.

As to a symmetric power law(as we will see further down), the Student T Distribution with scale s and tail exponent α:

$$p(x) \triangleq \frac{\left(\frac{\alpha}{\alpha + \frac{x^2}{s^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}s B\left(\frac{\alpha}{2}, \frac{1}{2}\right)}$$

$$\{a_1, a_2, a_3, a_4\} = \left\{ -\frac{\sqrt{\frac{5\alpha - \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, \frac{\sqrt{\frac{5\alpha - \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, \right.$$

$$\left. -\frac{\sqrt{\frac{5\alpha + \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, \frac{\sqrt{\frac{5\alpha + \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}} \right\}$$

where $B(.)$ is the Beta function $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 dt t^{a-1}(1-t)^{b-1}$.

When the Student is "cubic", that is, $\alpha = 3$:

$$\{a_1, a_2, a_3, a_4\} = \left\{ -\sqrt{4 - \sqrt{13}}s, -\sqrt{4 + \sqrt{13}}s, \sqrt{4 - \sqrt{13}}s, \sqrt{4 + \sqrt{13}}s \right\}$$

We can verify that when $\alpha \to \infty$, the crossovers become those of a Gaussian. For instance, for a_1:

$$\lim_{\alpha \to \infty} -\frac{\sqrt{\frac{5\alpha - \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}} = -\sqrt{\frac{1}{2}\left(5 - \sqrt{17}\right)}s$$

> **In Summary, Where Does the Tail Start?**
>
> For a general class of symmetric distributions with power laws, the tail starts at:
> $\pm\frac{\sqrt{\frac{5\alpha+\sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}$, with α infinite in the stochastic volatility Gaussian case and s the
> standard deviation. The "tail" is located between around 2 and 3 standard deviations.
> This flows from our definition: which part of the distribution is convex to errors in
> the estimation of the scale.
> But in practice, because historical measurements of STD will be biased lower because
> of small sample effects (as we repeat fat tails accentuate small sample effects), the
> deviations will be > 2-3 STDs.
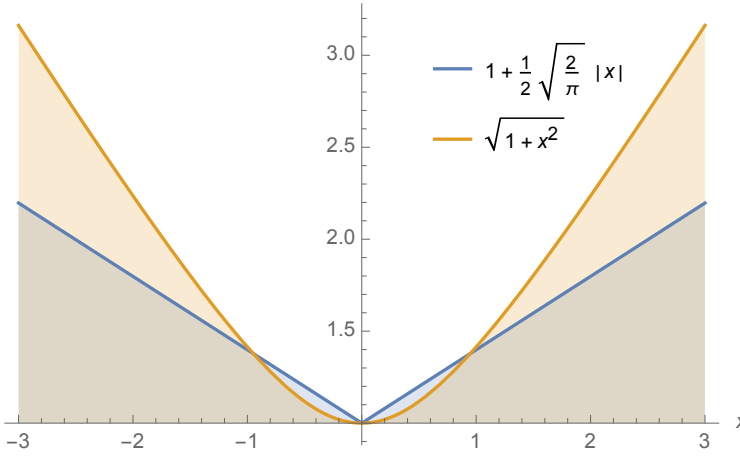
Figure 3.5: *We compare the behavior of $\sqrt{K+x^2}$ and $K+|x|$. The difference between the two weighting functions increases for large values of the random variable x, which explains the divergence of the two (and, more generally, higher moments) under fat tails.*

B- For some one-tailed distribution that have a "bell shape" of convex-concave-convex shape, under some conditions, the same 4 crossover points hold. The Lognormal is a special case.

$$\{a_1, a_2, a_3, a_4\} = \left\{ e^{\frac{1}{2}\left(2\mu-\sqrt{2}\sqrt{5\sigma^2-\sqrt{17}\sigma^2}\right)}, \right.$$
$$\left. e^{\frac{1}{2}\left(2\mu-\sqrt{2}\sqrt{\sqrt{17}\sigma^2+5\sigma^2}\right)}, e^{\frac{1}{2}\left(2\mu+\sqrt{2}\sqrt{5\sigma^2-\sqrt{17}\sigma^2}\right)}, e^{\frac{1}{2}\left(2\mu+\sqrt{2}\sqrt{\sqrt{17}\sigma^2+5\sigma^2}\right)} \right\}$$

Stochastic Parameters The problem of elliptical distributions is that they do not map the return of securities, owing to the absence of a single variance at any point in time, see Bouchaud and Chicheportiche (2010) [30]. When the scales of the distributions of the individuals move but not in tandem, the distribution ceases to be elliptical. Figure 4.2 shows the effect of applying the equivalent of stochastic volatility methods: the more annoying stochastic correlation. Instead of perturbating the correlation matrix Σ as a unit as in section 4.1, we perturbate the correlations with surprising effect.

3.4 FAT TAILS, MEAN DEVIATION AND THE RISING NORMS

Next we discuss the beastly use of standard deviation and its interpretation.

3.4.1 The common errors

We start by looking at standard deviation and variance as the properties of higher moments. Now, What is standard deviation? It appears that the same confusion about fat tails has polluted our understanding of standard deviation.

> The difference between standard deviation (assuming a mean of 0 to simplify) $\sigma = \sqrt{\frac{1}{n}\sum x_i^2}$ and mean absolute deviation $MAD = \frac{1}{n}\sum |x_i|$ increases under fat tails, as one can see in Figure 3.5 . This can provide a conceptual approach to the notion.

Dan Goldstein and the author [82] put the following question to investment professionals and graduate students in financial engineering –people who work with risk and deviations all day long.

> A stock (or a fund) has an average return of 0%. It moves on average 1% a day in absolute value; the average up move is 1% and the average down move is 1%. It does not mean that all up moves are 1% –some are .6%, others 1.45%, and so forth.
>
> Assume that we live in the Gaussian world in which the returns (or daily percentage moves) can be safely modeled using a Normal Distribution. Assume that a year has 256 business days. What is its standard deviation of returns (that is, of the percentage moves), the sigma that is used for volatility in financial applications?
>
> What is the daily standard deviation?
>
> What is the yearly standard deviation?

As the reader can see, the question described mean deviation. And the answers were overwhelmingly wrong. For the daily question, almost all answered 1%. Yet a Gaussian random variable that has a daily percentage move in absolute terms of 1% has a standard deviation that is higher than that, about 1.25%. It should be up to 1.7% in empirical distributions. The most common answer for the yearly question was about 16%, which is about 80% of what would be the true answer. The professionals were scaling daily volatility to yearly volatility by multiplying by $\sqrt{256}$ which is correct provided one had the correct daily volatility.
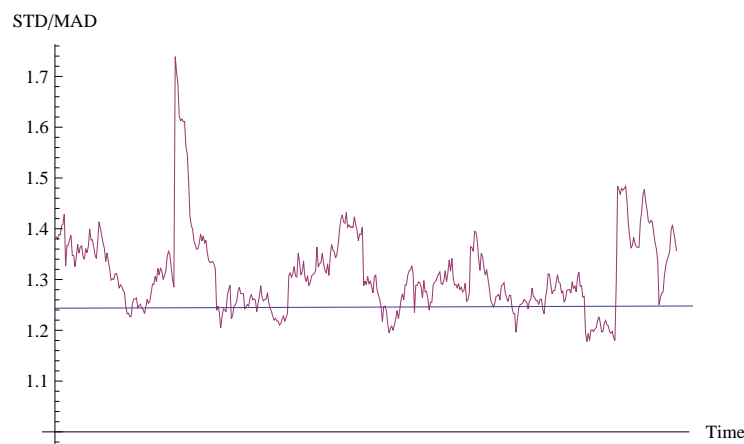
Figure 3.6: *The Ratio STD/MAD for the daily returns of the SP500 over the past 47 years, seen with a monthly rolling window. We can see 1.25 as approximately the value for Gaussian deviations, as the cut point for fat tailedness.*

So subjects tended to provide MAD as their intuition for STD. When professionals involved in financial markets and continuously exposed to notions of volatilityă talk about standard deviation, they use the wrong measure, mean absolute deviation (MAD) instead of standard deviation (STD), causing an average underestimation of between 20 and 40%. In some markets it can be up to 90%.

Further, responders rarely seemed to immediately understand the error whenăit was pointed out to them. However when asked to present the equation for standard deviation they effectively expressed it as the mean root mean square deviation. Some were puzzled as they were not aware of the existence of MAD.

Why this is relevant: Here you have decision-makers walking around talking about "volatility" and not quite knowing what it means. We note some clips in the financial press to that effect in which the journalist, while attempting to explain the "VIX", i.e., volatility index, makes the same mistake. Even the website of the department of commerce misdefined volatility.

Further, there is an underestimation as MAD is by Jensen's inequality lower (or equal) than STD.

How the ratio rises   For a Gaussian the ratio ~ 1.25, and it rises from there with fat tails.

Example: Take an extremely fat tailed distribution with $n=10^6$, observations are all -1 except for a single one of 10^6,

$$X = \left\{ -1, -1, ..., -1, 10^6 \right\}.$$

The mean absolute deviation, MAD $(X) = 2$. The standard deviation STD $(X)=1000$. The ratio standard deviation over mean deviation is 500.

3.4.2   Some Analytics

The ratio for thin tails   As a useful heuristic, consider the ratio h

$$h = \frac{\sqrt{\mathbb{E}\left(X^2\right)}}{\mathbb{E}(|X|)}$$

where $\mathbb{E}$ is the expectation operator (under the probability measure of concern and X is a centered variable such $\mathbb{E}(x) = 0$); the ratio increases with the fat tailedness of the distribution; (The general case corresponds to $\frac{\left(\mathbb{E}_p(x^p)\right)^{\frac{1}{p}}}{\mathbb{E}_n(|x|)}$, $p > 1$, under the condition that the distribution has finite moments up to n, and the special case here $n = 2$).[1]

Simply, x^n is a weighting operator that assigns a weight, x^{n-1}, which is large for large values of X, and small for smaller values.

The effect is due to the convexity differential between both functions, $|X|$ is piecewise linear and loses the convexity effect except for a zone around the origin.

---

1 The word "infinite" moment is a big ambiguous, it is better to present the problem as "undefined" moment in the sense that it depends on the sample, and does not replicate outside. Say, for a two-tailed distribution (i.e. with support on the real line), the designation"infinite" variance might apply for the fourth moment, but not to the third.

Mean Deviation vs Standard Deviation, more technical Why the [REDACTED] did statistical science pick STD over Mean Deviation? Here is the story, with analytical derivations not seemingly available in the literature. In Huber [90]:

> There had been a dispute between Eddington and Fisher, around 1920, about the relative merits of dn (mean deviation) and Sn (standard deviation). Fisher then pointed out that for **exactly normal** observations, Sn is 12% more efficient than dn, and this seemed to settle the matter. (My emphasis)

Let us rederive and see what Fisher meant.

Let n be the number of summands:

$$\text{Asymptotic Relative Efficiency (ARE)} = \lim_{n\to\infty} \left(\frac{\mathbb{V}(Std)}{\mathbb{E}(Std)^2} \Big/ \frac{\mathbb{V}(Mad)}{\mathbb{E}(Mad)^2} \right)$$

Assume we are certain that X_i, the components of sample follow a Gaussian distribution, normalized to mean=0 and a standard deviation of 1.

Relative Standard Deviation Error The characteristic function $\Psi_1(t)$ of the distribution of x^2: $\Psi_1(t) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2}+itx^2}}{\sqrt{2\pi}} dx = \frac{1}{\sqrt{1-2it}}$. With the squared deviation $z = x^2$, f, the pdf for n summands becomes:

$$f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itz) \left(\frac{1}{\sqrt{1-2it}} \right)^n dt = \frac{2^{-\frac{n}{2}} e^{-\frac{z}{2}} z^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, z > 0.$$

Now take $y = \sqrt{z}$, $f_Y(y) = \frac{2^{1-\frac{n}{2}} e^{-\frac{z^2}{2}} z^{n-1}}{\Gamma\left(\frac{n}{2}\right)}$, $z > 0$, which corresponds to the Chi Distribution with n degrees of freedom. Integrating to get the variance: $\mathbb{V}_{std}(n) = n - \frac{2\Gamma\left(\frac{n+1}{2}\right)^2}{\Gamma\left(\frac{n}{2}\right)^2}$. And, with the mean equalling $\frac{\sqrt{2}\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$, we get $\frac{\mathbb{V}(Std)}{\mathbb{E}(Std)^2} = \frac{n\Gamma\left(\frac{n}{2}\right)^2}{2\Gamma\left(\frac{n+1}{2}\right)^2} - 1$.

Relative Mean Deviation Error Characteristic function again for $|x|$ is that of a folded Normal distribution, but let us redo it:

$\Psi_2(t) = \int_0^{\infty} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}+itx} = e^{-\frac{t^2}{2}} \left(1 + i\operatorname{erfi}\left(\frac{t}{\sqrt{2}}\right)\right)$, where erfi is the imaginary error function $erf(iz)/i$.

The first moment: $M_1 = -i\frac{\partial}{\partial t^1} \left(e^{-\frac{t^2}{2n^2}} \left(1 + i\operatorname{erfi}\left(\frac{t}{\sqrt{2n}}\right)\right) \right)^n \Big|_{t=0} = \sqrt{\frac{2}{\pi}}$.

The second moment, $M_2 = (-i)^2 \frac{\partial^2}{\partial t^2} \left(e^{-\frac{t^2}{2n^2}} \left(1 + i\operatorname{erfi}\left(\frac{t}{\sqrt{2n}}\right)\right) \right)^n \Big|_{t=0} = \frac{2n+\pi-2}{\pi n}$. Hence,

$\frac{\mathbb{V}(Mad)}{\mathbb{E}(Mad)^2} = \frac{M_2 - M_1^2}{M_1^2} = \frac{\pi-2}{2n}$.

Finalmente, the Asymptotic Relative Efficiency For a Gaussian

$$\text{ARE} = \lim_{n\to\infty} \frac{n\left(\frac{n\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n+1}{2}\right)^2} - 2\right)}{\pi - 2} = \frac{1}{\pi - 2} \approx .875$$

which means that the standard deviation is 12.5% more "efficient" than the mean deviation *conditional on the data being Gaussian* and these blokes bought the argument. Except that the slightest contamination blows up the ratio. We will show later why Norm ℓ^2 is not appropriate for about anything; but for now let us get a glimpse on how fragile the STD is.

3.4.3 Effect of Fatter Tails on the "efficiency" of STD vs MD

Consider a standard mixing model for volatility with an occasional jump with a probability p. We switch between Gaussians (keeping the mean constant and central at 0) with:

$$\mathbb{V}(x) = \begin{cases} \sigma^2(1+a) & \text{with probability } p \\ \sigma^2 & \text{with probability } (1-p) \end{cases}$$

For ease, a simple Monte Carlo simulation would do. Using $p = .01$ and $n = 1000$... Figure 3.7 shows how a=2 causes degradation. A minute presence of outliers makes MAD more "efficient" than STD. Small "outliers" of 5 standard deviations cause MAD to be five times more efficient.
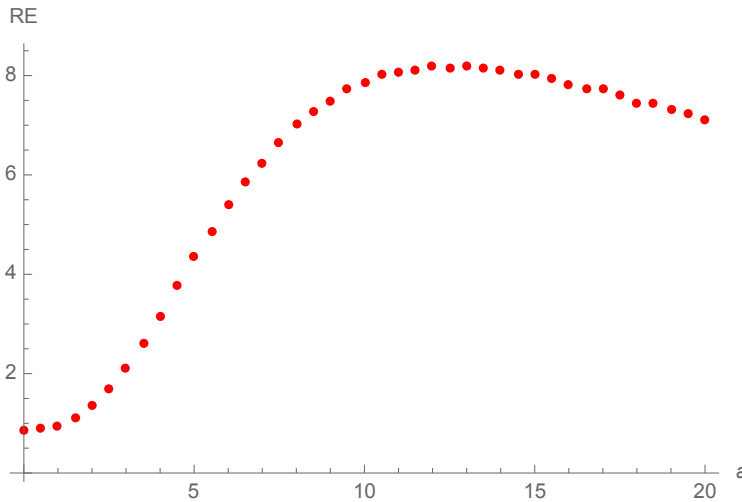
Figure 3.7: *A simulation of the Relative Efficiency ratio of Standard deviation over Mean deviation when injecting a jump size $\sqrt{(1+a)} \times \sigma$, as a multiple of σ the standard deviation.*

3.4.4 Moments and The Power Mean Inequality

recheck derivations/proofs

Let $X \triangleq (x_i)_{i=1}^n$,

$$\|X\|_p \triangleq \left(\frac{\sum_{i=1}^n |x_i|^p}{n} \right)^{1/p}$$

For any $0 < p < q$ the following inequality holds:

$$\sqrt[p]{\sum_{i=1}^n w_i |x_i|^p} \leq \sqrt[q]{\sum_{i=1}^n w_i |x_i|^q} \tag{3.9}$$

where the positive weights w_i sum to unity.

Proof. The proof for positive p and q is as follows: Define the following function: $f : R^+ \to R^+$; $f(x) = x^{\frac{q}{p}}$. f is a power function, so it does have a second derivative:

$$f''(x) = \left(\frac{q}{p}\right)\left(\frac{q}{p} - 1\right) x^{\frac{q}{p} - 2}$$

which is strictly positive within the domain of f, since $q > p - f$ is convex. Hence, by Jensen's inequality : $f\left(\sum_{i=1}^n w_i x_i^p\right) \leq \sum_{i=1}^n w_i f(x_i^p)$, so $\sqrt[q]{\sum_{i=1}^n w_i x_i^p} \leq \sum_{i=1}^n w_i x_i^q$ after raising both side to the power of $1/q$ (an increasing function, since $1/q$ is positive) we get the inequality. □

What is critical for our exercise and the study of the effects of fat tails is that, for a given norm, dispersion of results increases values. For example, take a flat distribution, $X = \{1, 1\}$. $\|X\|_1 = \|X\|_2 = \ldots = \|X\|_n = 1$. Perturbating while preserving $\|X\|_1$, $X = \left\{\frac{1}{2}, \frac{3}{2}\right\}$ produces rising higher norms:

$$\{\|X\|_n\}_{n=1}^5 = \left\{1, \frac{\sqrt{5}}{2}, \frac{\sqrt[3]{7}}{2^{2/3}}, \frac{\sqrt[4]{41}}{2}, \frac{\sqrt[5]{61}}{2^{4/5}}\right\}. \tag{3.10}$$

Trying again, with a wider spread, we get even higher values of the norms, $X = \left\{\frac{1}{4}, \frac{7}{4}\right\}$,

$$\{\|X\|_n\}_{n=1}^5 = \left\{1, \frac{5}{4}, \frac{\sqrt[3]{\frac{43}{2}}}{2}, \frac{\sqrt[4]{1201}}{4}, \frac{\sqrt[5]{2101}}{2 \times 2^{3/5}}\right\}. \tag{3.11}$$

So we can see it becomes rapidly explosive.

One property quite useful with power laws with infinite moment:

$$\|X\|_\infty = \sup\left(\frac{1}{n}|x_i|\right)_{i=1}^n \tag{3.12}$$

Gaussian Case For a Gaussian, where $x \sim N(0, \sigma)$, as we assume the mean is 0 without loss of generality,

Let $\mathbb{E}(X)$ be the expectation operator for X,

$$\frac{\mathbb{E}(X)^{1/p}}{\mathbb{E}(|X|)} = \frac{\pi^{\frac{p-1}{2p}} \left(2^{\frac{p}{2}-1}((-1)^p + 1)\Gamma\left(\frac{p+1}{2}\right)\right)^{\frac{1}{p}}}{\sqrt{2}}$$

or, alternatively

$$\frac{\mathbb{E}(X^p)}{\mathbb{E}(|X|)} = 2^{\frac{1}{2}(p-3)}(1 + (-1)^p)\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2} - \frac{p}{2}} \Gamma\left(\frac{p+1}{2}\right) \tag{3.13}$$

where $\Gamma(z)$ is the Euler gamma function; $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. For odd moments, the ratio is 0. For even moments:

$$\frac{\mathbb{E}\left(X^2\right)}{\mathbb{E}\left(|X|\right)} = \sqrt{\frac{\pi}{2}}\,\sigma$$

hence

$$\frac{\sqrt{\mathbb{E}\left(X^2\right)}}{\mathbb{E}\left(|X|\right)} = \frac{\text{Standard Deviation}}{\text{Mean Absolute Deviation}} = \sqrt{\frac{\pi}{2}}$$

As to the fourth moment, it equals $3\sqrt{\frac{\pi}{2}}\sigma^3$.

For a Power Law distribution with tail exponent $\alpha=3$, say a Student T

$$\frac{\sqrt{\mathbb{E}\left(X^2\right)}}{\mathbb{E}\left(|X|\right)} = \frac{\text{Standard Deviation}}{\text{Mean Absolute Deviation}} = \frac{\pi}{2}$$

We will return to other metrics and definitions of fat tails with Power Law distributions when the moments are said to be "infinite", that is, do not exist. Our heuristic of using the ratio of moments to mean deviation works only in sample, not outside.

"Infinite" moments   Infinite moments, say infinite variance, always manifest themselves as computable numbers in observed sample, yielding an estimator M, simply because the sample is finite. A distribution, say, Cauchy, with undefined means will always deliver a measurable mean in finite samples; but different samples will deliver completely different means. Figures **??** and **??** illustrate the "drifting" effect of M a with increasing information.

3.4.5   Comment: Why We Should Retire Standard Deviation

The notion of standard deviation has confused hordes of scientists; it is time to retire it from common use and replace it with the more effective one of mean deviation. Standard deviation, STD, should be left to mathematicians, physicists and mathematical statisticians deriving limit theorems. There is no scientific reason to use it in statistical investigations in the age of the computer, as it does more harm than good-particularly with the growing class of people in social science mechanistically applying statistical tools to scientific problems.

Say someone just asked you to measure the "average daily variations" for the temperature of your town (or for the stock price of a company, or the blood pressure of your uncle) over the past five days. The five changes are: (-23, 7, -3, 20, -1). How do you do it?

Do you take every observation: square it, average the total, then take the square root? Or do you remove the sign and calculate the average? For there are serious differences between the two methods. The first produces an average of 15.7, the second 10.8. The first is technically called the root mean square deviation. The second is the mean absolute deviation, MAD. It corresponds to "real life" much better than the first-and to reality. In fact, whenever people make decisions after being supplied with the standard deviation number, they act as if it were the expected mean deviation.

It is all due to a historical accident: in 1893, the great Karl Pearson introduced the term "standard deviation" for what had been known as "root mean square error". The confusion started then: people thought it meant mean deviation. The idea stuck: every time a

newspaper has attempted to clarify the concept of market "volatility", it defined it verbally as mean deviation yet produced the numerical measure of the (higher) standard deviation.

But it is not just journalists who fall for the mistake: I recall seeing official documents from the department of commerce and the Federal Reserve partaking of the conflation, even regulators in statements on market volatility. What is worse, Goldstein and I found that a high number of data scientists (many with PhDs) also get confused in real life.

It all comes from bad terminology for something non-intuitive. By a psychological phenomenon called attribute substitution, some people mistake MAD for STD because the former is easier to come to mind – this is "Lindy" as it is well known by cheaters and illusionists.

1) MAD is more accurate in sample measurements, and less volatile than STD since it is a natural weight whereas standard deviation uses the observation itself as its own weight, imparting large weights to large observations, thus overweighing tail events.

2) We often use STD in equations but really end up reconverting it within the process into MAD (say in finance, for option pricing). In the Gaussian world, STD is about 1.25 time MAD, that is, $\sqrt{\frac{\pi}{2}}$. But we adjust with stochastic volatility where STD is often as high as 1.6 times MAD.

3) Many statistical phenomena and processes have "infinite variance" (such as the popular Pareto 80/20 rule) but have finite, and sometimes very well behaved, mean deviations. Whenever the mean exists, MAD exists. The reverse (infinite MAD and finite STD) is never true.

4) Many economists have dismissed "infinite variance" models thinking these meant "infinite mean deviation". Sad, but true. When the great Benoit Mandelbrot proposed his infinite variance models fifty years ago, economists freaked out because of the conflation.

It is sad that such a minor point can lead to so much confusion: our scientific tools are way too far ahead of our casual intuitions, which starts to be a problem with science. So I close with a statement by Sir Ronald A. Fisher: 'The statistician cannot evade the responsibility for understanding the process he applies or recommends.'

Note   The usual theory is that if random variables $X_1, \ldots, X_n$ are independent, then

$$\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n).$$

by the linearity of the variance. But then it assumes that one cannot use another metric then by simple transformation make it additive[2]. As we will see, for the Gaussian md$(X) = \sqrt{\frac{2}{\pi}}\sigma$ —for the Student T with 3 degrees of freedom, the factor is $\frac{2}{\pi}$, etc.

3.5   LEVEL 2: SUBEXPONENTIALITY

Table 3.3: *Ranking distributions*

	Class	Description
$\mathcal{D}_1$	True Thin Tails	Compact support (e.g. : Bernouilli, Binomial)
$\mathcal{D}_2$	Thin tails	Gaussian reached organically through summation of true thin tails, by Central Limit; compact support except at the limit $n \to \infty$
$\mathcal{D}_{3a}$	Conventional Thin tails	Gaussian approximation of a natural phenomenon
$\mathcal{D}_{3b}$	Starter Fat Tails	Higher kurtosis than the Gaussian but rapid convergence to Gaussian under summation
$\mathcal{D}_5$	Subexponential	(e.g. lognormal)
$\mathcal{D}_6$	Supercubic α	Cramer conditions do not hold for $t > 3, \int e^{-tx} d(Fx) = \infty$
$\mathcal{D}_7$	Infinite Variance	Levy Stable $\alpha < 2$, $\int e^{-tx} dF(x) = \infty$
$\mathcal{D}_8$	Undefined First Moment	Fuhgetaboutdit

3.5.1 Revisiting the Rankings

Probability distributions range between extreme thin-tailed (Bernoulli) and extreme fat tailed. Among the categories of distributions that are often distinguished due to the convergence properties of moments are:

1. Having a support that is compact but not degenerate

2. Subgaussian

3. Subexponential

4. Power Law with exponent greater than 2

5. Power Law with exponent less than or equal to 2. In particular, Power Law distributions have a finite mean only if the exponent is greater than 1, and have a finite variance only if the exponent exceeds 2.

6. Power Law with exponent less than 1.

Our interest is in distinguishing between cases where tail events dominate impacts, as a formal definition of the boundary between the categories of distributions to be considered as mediocristan and Extremistan.

Centrally, a subexponential distribution is the cutoff between "thin" and "fat" tails. It is defined as follows.

The mathematics is crisp: the exceedance probability or survival function needs to be exponential in one not the other. Where is the border?

---

2 For instance option pricing in the Black Scholes formula is done using variance, but the price maps directly to MAD; an at-the-money straddle is just a conditional mean deviation. So we translate MAD into standard deviation, then back to MAD

The natural boundary between mediocristan and Extremistan occurs at the subexponential class which has the following property:

Let $\mathbf{X} = (X_i)_{1 \leq i \leq n}$ be a sequence of independent and identically distributed random variables with support in $(\mathbb{R}^+)$, with cumulative distribution function F. The subexponential class of distributions is defined by [174],[136].

$$\lim_{x \to +\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2 \tag{3.14}$$

where $F^{*2} = F' * F$ is the cumulative distribution of $X_1 + X_2$, the sum of two independent copies of X. This implies that the probability that the sum $X_1 + X_2$ exceeds a value x is twice the probability that either one separately exceeds x. Thus, every time the sum exceeds x, for large enough values of x, the value of the sum is due to either one or the other exceeding x—the maximum over the two variables—and the other of them contributes negligibly.

More generally, it can be shown that the sum of n variables is dominated by the maximum of the values over those variables in the same way. Formally, the following two properties are equivalent to the subexponential condition [31],[62]. For a given $n \geq 2$, let $S_n = \Sigma_{i=1}^{n} x_i$ and $M_n = \max_{1 \leq i \leq n} x_i$

a) $\lim_{x \to \infty} \frac{P(S_n > x)}{P(X > x)} = n,$

b) $\lim_{x \to \infty} \frac{P(S_n > x)}{P(M_n > x)} = 1.$

Thus the sum S_n has the same magnitude as the largest sample M_n, which is another way of saying that tails play the most important role.

Intuitively, tail events in subexponential distributions should decline more slowly than an exponential distribution for which large tail events should be irrelevant. Indeed, one can show that subexponential distributions have no exponential moments:

$$\int_0^{\infty} \mathbf{e}^{\epsilon x} \, dF(x) = +\infty \tag{3.15}$$

for all values of ϵ greater than zero. However, the converse isn't true, since distributions can have no exponential moments, yet not satisfy the subexponential condition.

We note that if we choose to indicate deviations as negative values of the variable x, the same result holds by symmetry for extreme negative values, replacing $x \to +\infty$ with $x \to -\infty$. For two-tailed variables, we can separately consider positive and negative domains.

3.5.2 What is a probability distribution?

The best way to figure out a probability distribution is to... invent one. In fact in the next section, 3.5.3, we will build one that is the exact borderline between thin and fat tails *by construction*.

Let s be the survival function. We have $s : \mathbb{R} \to [0, 1]$ that satisfies

$$\lim_{x \to +\infty} \frac{s(x)^n}{s(nx)} = 1, \tag{3.16}$$

and

$$\lim_{x \to +\infty} s(x) = 0$$

$$\lim_{x \to -\infty} s(x) = 1$$

Note : another property of the demarcation is the absence of Lucretius problem from *The Black Swan*:

$$\lim_{x \to +\infty} \mathbb{E}(x - K | x > K) \quad \begin{aligned} &= \kappa K, \ K, \kappa > 0 \text{ for fat tails} \\ &= \kappa \text{ for borderline subexponential} \\ &= 0 \text{ for thin tails} \end{aligned} \tag{3.17}$$

3.5.3 Let us invent a distribution

Find functions $f : \mathbb{R} \to [0, 1]$ that satisfy:

$$\lim_{x \to +\infty} \frac{f(x)^2}{f(x^*)f(2x)} = 1,$$

$$f'(x) \leq 0 \ \ \forall x,$$

and

$$\lim_{x \to +\infty} f(x) = 0.$$

$$\lim_{x \to -\infty} f(x) = 1.$$

(when the variable is time, this describes the condition for the *Lindy Effect*<mark>explain the Lindy effect</mark> to take place). Let us assume a candidate function a sigmoid, using the hyperbolic tangent $F^\kappa(x) := \frac{1}{2} \tanh\left(\frac{\kappa x}{\pi}\right) + \frac{1}{2}$, with $\kappa \in (0, \infty)$. We use this as a kernel distribution (we mix later to modify the kurtosis).

For $\kappa > 0$

$$\lim x \to +\infty \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{\kappa x}{\pi}\right) = 1$$

$$\lim x \to -\infty \frac{1}{2} + \frac{1}{2} \mathrm{Tanh}\left[\frac{\kappa x}{\pi}\right]$$

and

$$f(x) = D(\frac{1}{2} \tanh\left(\frac{\kappa x}{\pi}\right) + \frac{1}{2}, x) = \frac{\kappa \mathrm{Sech}\left[\frac{x\kappa}{\pi}\right]^2}{2\pi} \tag{3.18}$$

All functions

$$f_a(x) = \frac{1}{2}\left(1 - \tanh(ax)\right), a > 0$$

solve our requirements. Since a is just a rescaling, we need only show this for $a = 1$. Let $f(x) = f_1(x)$. The derivative is $f'(x) = \frac{1}{2}\left(\tanh(x)^2 - 1\right) \leq 0$ and

$$\frac{f(x)^2}{f(2x)} = \frac{1}{2} \frac{2 \cosh(x)^2 - 1}{\cosh(x)^2} = 1 - \frac{1}{2 \cosh(x)^2} < 1$$

with a limit 1 for $x \to \infty$. The limits of $f(x)$ for $x \to \pm\infty$ are trivial.
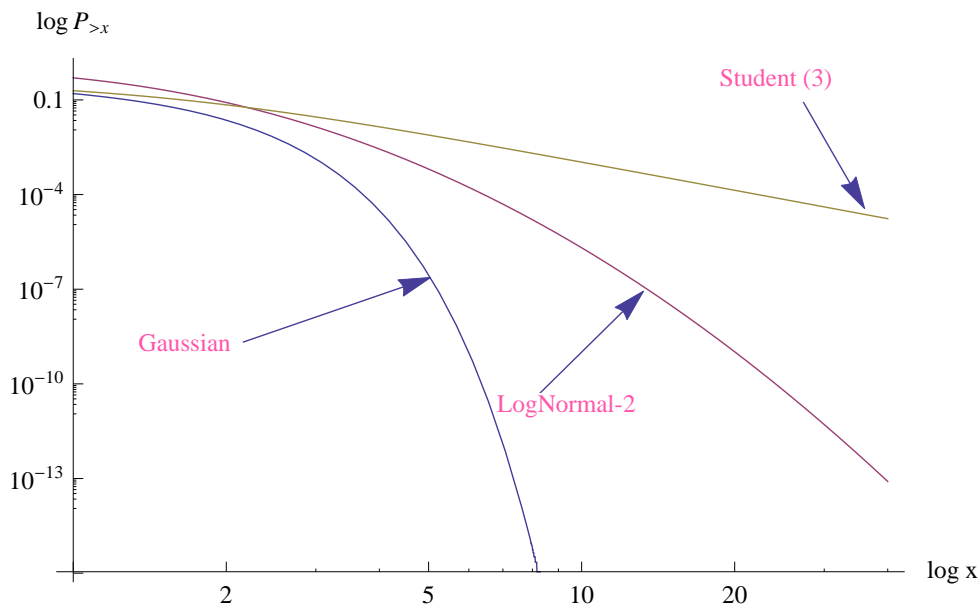
Figure 3.8: *Three Types of Distributions. As we hit the tails, the Student remains scalable while the Standard Lognormal shows an intermediate position before eventually ending up getting an infinite slope on a log-log plot. But beware the lognormal as it may have some surprises (Chapter 7)*
.

3.6 LEVEL 3: SCALABILITY AND POWER LAWS

Now we get into the serious business.

Why power laws? There are a lot of theories on why things should be power laws, as sort of exceptions to the way things work probabilistically. But it seems that the opposite idea is never presented: power should can be the norm, and the Gaussian a special case as we will see in Chapter x (effectively the topic of *Antifragile* and Vol 2 of the *Technical Incerto*), of concave-convex responses (sort of dampening of fragility and antifragility, bringing robustness, hence thinning tails).

3.6.1 Scalable and Nonscalable, A Deeper View of Fat Tails

So far for the discussion on fat tails we stayed in the finite moments case. For a certain class of distributions, those with finite moments, $\frac{P_{X>nK}}{P_{X>K}}$ depends on n and K. For a scale-free distribution, with K "in the tails", that is, large enough, $\frac{P_{X>nK}}{P_{X>K}}$ depends on n not K. These latter distributions lack in characteristic scale and will end up having a Paretian tail, i.e., for x large enough, $P_{X>x} = Cx^{-\alpha}$ where α is the tail and C is a scaling constant.

Note: We can see from the scaling difference between the Student and the Pareto the conventional definition of a Power Law tailed distribution is expressed more formally as $\mathbb{P}(X > x) = L(x)x^{-\alpha}$ where $L(x)$ is a "slow varying function", which satisfies the following:

$$\lim_{x \to \infty} \frac{L(t\,x)}{L(x)} = 1$$

for all constants $t > 0$.

Table 3.4: *Scalability, comparing slowly varying functions/powerlaws to other distributions*

k	$\mathbb{P}(X > k)^{-1}$ (Gaussian)	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\,k)}$ (Gaussian)	$\mathbb{P}(X > k)^{-1}$ Student(3)	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\,k)}$ Student (3)	$\mathbb{P}(X > k)^{-1}$ Pareto(2)	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\,k)}$ Pareto (2)
2	44	720	14.4	4.9	8	4
4	31600.	5.1×10^{10}	71.4	6.8	64	4
6	1.01×10^{9}	5.5×10^{23}	216	7.4	216	4
8	1.61×10^{15}	9×10^{41}	491	7.6	512	4
10	1.31×10^{23}	9×10^{65}	940	7.7	1000	4
12	5.63×10^{32}	fughedaboudit	1610	7.8	1730	4
14	1.28×10^{44}	fughedaboudit	2530	7.8	2740	4
16	1.57×10^{57}	fughedaboudit	3770	7.9	4100	4
18	1.03×10^{72}	fughedaboudit	5350	7.9	5830	4
20	3.63×10^{88}	fughedaboudit	7320	7.9	8000	4

For x large enough, $\frac{\log P_{>x}}{\log x}$ converges to a constant, namely the tail exponent $-\alpha$. A scalable should produce the slope α in the tails on a log-log plot, as $x \to \infty$. Compare to the Gaussian (with STD σ and mean μ), by taking the PDF this time instead of the exceedance probability $\log\left(f(x)\right) = \frac{(x-\mu)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \approx -\frac{1}{2\sigma^2}x^2$ which goes to $-\infty$ faster than $-\log(x)$ for $\pm x \to \infty$.

So far this gives us the intuition of the difference between classes of distributions. Only scalable have "true" fat tails, as others turn into a Gaussian under summation. And the tail exponent is asymptotic; we may never get there and what we may see is an intermediate version of it. The figure above drew from Platonic off-the-shelf distributions; in reality processes are vastly more messy, with switches between exponents.

3.6.2   Grey Swans

Why do we use Student T to simulate symmetric power laws?   For convenience, only for convenience. It is not that we *believe* that the generating process is Student T. Simply, the center of the distribution does not matter much for the properties involved in certain classes of decision making.

The lower the exponent, the less the center plays a role. The higher the exponent, the more the student T resembles the Gaussian, and the more justified its use will be accordingly.

More advanced methods involving the use of Levy laws may help in the event of asymmetry, but the use of two different Pareto distributions with two different exponents, one for the left tail and the other for the right one would do the job (without unnecessary complications).

Estimation issues   Note that there are many methods to estimate the tail exponent α from data, what is called a "calibration. However, we will see, the tail exponent is rather
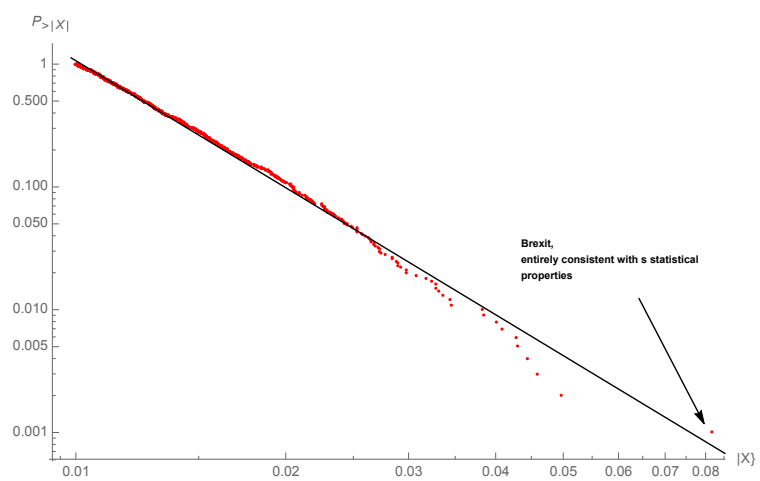
Figure 3.9: *The Grey Swan of Brexit when seen using a power law.*
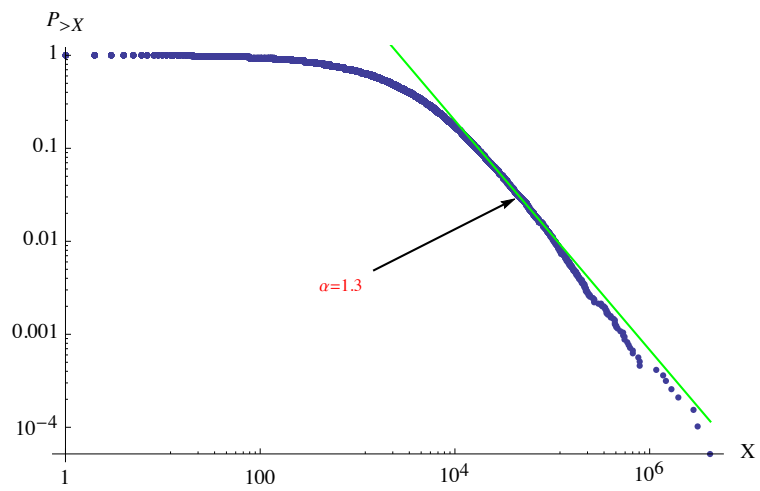
Figure 3.10: *Book Sales: the near tail can be robust for estimation of sales from rank and vice versa –it works well and shows robustness so long as one doesn't compute general expectations or higher non-truncated moments.*
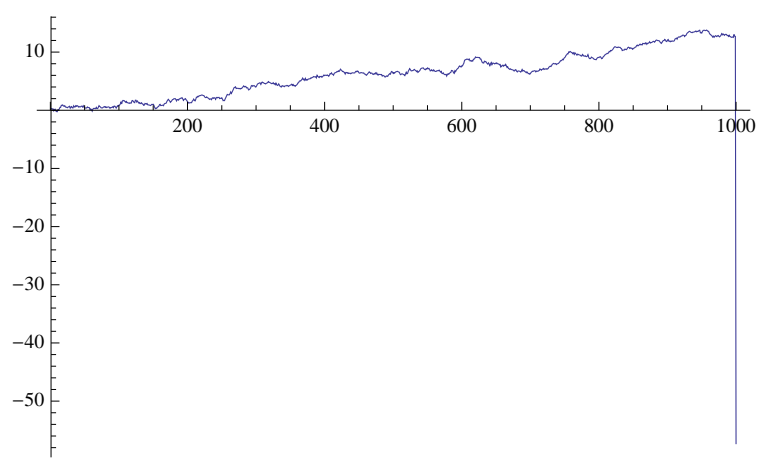
Figure 3.11: *The Turkey Problem, where nothing in the past properties seems to indicate the possibility of the jump.*

hard to guess, and its calibration marred with errors, owing to the insufficiency of data in the tails. In general, the data will show thinner tail than it should.

We will return to the issue in Chapter **??**.

3.7 BELL SHAPED VS NON BELL SHAPED POWER LAWS

The slowly varying function effect, a case study   The fatter the tails, the less the "body" matters for the moments (which become infinite, eventually). But for power laws with thinner tails, the zone that is not power law (the slowly moving part) plays a role –"slowly varying" is more formally defined in 15.2.2 and 3.6.1. This section will show how apparently equal distributions can have different shapes.

Let us compare a double Pareto distribution with the following PDF:

$$
f_P(x) = \begin{cases} \alpha(1+x)^{-\alpha-1} & x \ge 0 \\[2ex] \alpha(1-x)^{-\alpha-1} & x < 0 \end{cases}
$$

to a Student T with same centrality parameter 0, scale parameter s and PDF $f_S(x) = \frac{\alpha^{\alpha/2}\left(\alpha+\frac{x^2}{s^2}\right)^{\frac{1}{2}(-\alpha-1)}}{sB\left(\frac{\alpha}{2},\frac{1}{2}\right)}$ where $B(.)$ is the Euler beta function, $B(a,b) = \frac{(\Gamma(a))(\Gamma(b))}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt$.

We have two ways to compare distributions.

- Equalizing by tail ratio: setting $\lim_{x\to\infty} \frac{f_p(x)}{f_s(x)} = 1$ to get the same tail ratio, we get the equivalent "tail" distribution with $s = \left(\alpha^{1-\frac{\alpha}{2}}B\left(\frac{\alpha}{2},\frac{1}{2}\right)\right)^{1/\alpha}$.

- Equalizing by standard deviations (when finite): we have, with $\alpha > 2$, $\mathbb{E}(X_P^2) = \frac{2}{\alpha^2-3\alpha+2}$ and $\mathbb{E}(X_S^2) = \frac{\alpha\left(\alpha^{1-\frac{\alpha}{2}}B\left(\frac{\alpha}{2},\frac{1}{2}\right)\right)^{2/\alpha}}{\alpha-2}$.

   So we could set $\sqrt{\mathbb{E}(X_P^2)} = \sqrt{k}\sqrt{\mathbb{E}(X_S^2)}$ $k \to \left. \frac{2\alpha^{-2/\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)^{-2/\alpha}}{\alpha-1}\right\}$.

Finally, we have the comparison "bell shape" semi-concave vs the angular double-convex one as seen in Fig. 3.12.
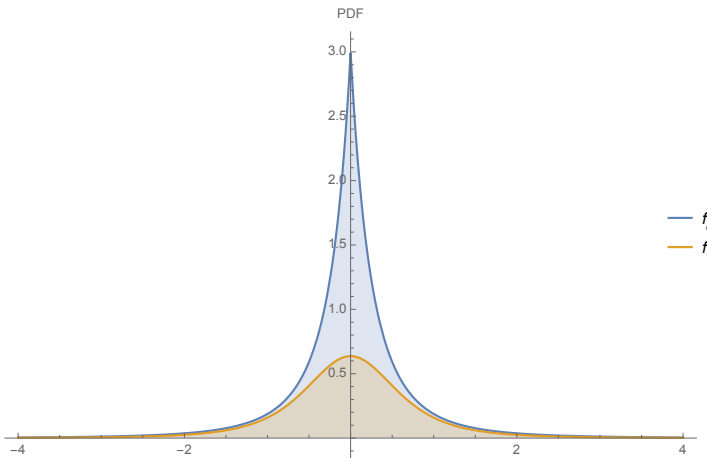
Figure 3.12: *Comparing two symmetric power laws of same exponent, one with a brief slowly varying function, the other with an extended one. All moments eventually become the same in spite of the central differences in their shape for small deviations.*

4 | OVERVIEW OF FAT TAILS, PART 2 (HIGHER DIMENSIONS)†

THIS DISCUSSION is about as simplified as possible handling of higher dimensions. We will look at 1) the simple effect of fat-tailedness for multiple random variables, 2) Ellipticality and distributions, 3) random matrices and the associated distribution of eigenvalues, 4) How we can look at covariance and correlations when moments don't exist (say, as in the Cauchy case).
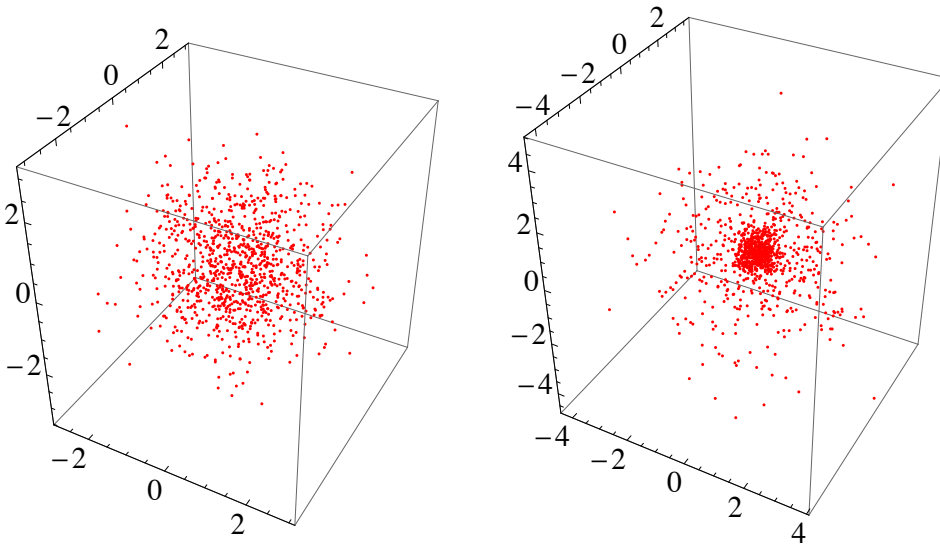
Figure 4.1: *Fat tails in higher dimensions: For a 3 dimentional vector, thin tails (left) and fat tails (right) of the same variance. In place of a bell curve with higher peak (the "tunnel") of the univariate case, we see an increased density of points towards the center.*

4.1 FAT TAILS IN HIGHER DIMENSION, FINITE MOMENTS

Let $\overrightarrow{X} = (X_1, X_2, \ldots, X_m)$ be a $p \times 1$ random vector with the variables assumed to be drawn from a multivariate Gaussian. Consider the joint probability distribution $f(x_1, \ldots, x_m)$.

We denote the m-variate multivariate Normal distribution by $\mathcal{N}(0, \Sigma)$, with mean vector $\vec{\mu}$, variance-covariance matrix Σ, and joint pdf,

$$f\left(\vec{x}\right) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x} - \vec{\mu}\right)\right) \tag{4.1}$$

where $\vec{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, and Σ is a symmetric, positive definite ($m \times m$) matrix.

We can apply the same simplied variance preserving heuristic as in 3.1.1 to fatten the tails:

$$f_a\left(\vec{x}\right) = \frac{1}{2}(2\pi)^{-m/2}|\Sigma_1|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \Sigma_1^{-1}\left(\vec{x} - \vec{\mu}\right)\right)$$
$$+ \frac{1}{2}(2\pi)^{-m/2}|\Sigma_2|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \Sigma_2^{-1}\left(\vec{x} - \vec{\mu}\right)\right) \tag{4.2}$$

where a is a scalar that determines the intensity of stochastic volatility, $\Sigma_1 = \Sigma(1 + a)$ and $\Sigma_2 = \Sigma(1 - a)$.[1]
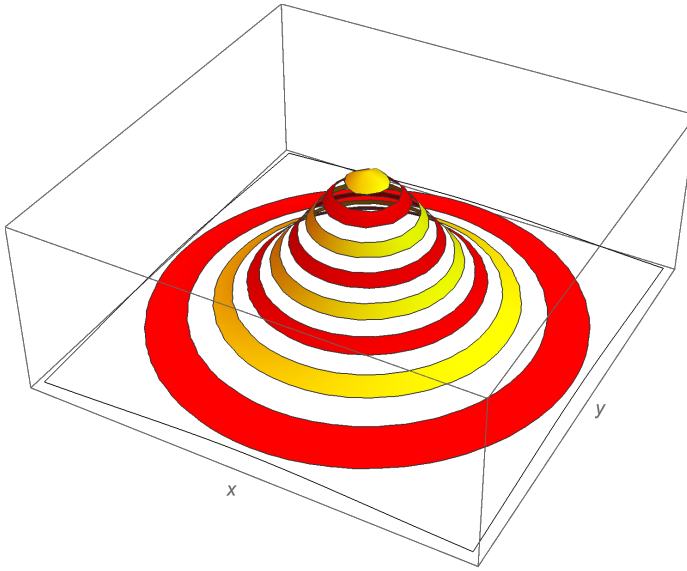
Figure 4.2: *Elliptically Contoured Joint Returns of Powerlaw (Student T)*

Notice in Figure 4.1, as with the one-dimensional case, a concentration in the middle part of the distribution

4.2 JOINT FAT–TAILEDNESS AND ELLIPTICALITY OF DISTRIBUTIONS

There is another aspect, beyond our earlier definition(s) of fat-tailedness, once we increase the dimensionality into random vectors:

What is an Elliptical Distribution? From the definition in [66], **X**, a $p \times 1$ random vector is said to have an elliptical (or elliptical contoured) distribution with location parameters

---

1 We can simplify by assuming as we did in the single dimension case, without any loss of generality, that $\vec{\mu} = (0, \ldots, 0)$.
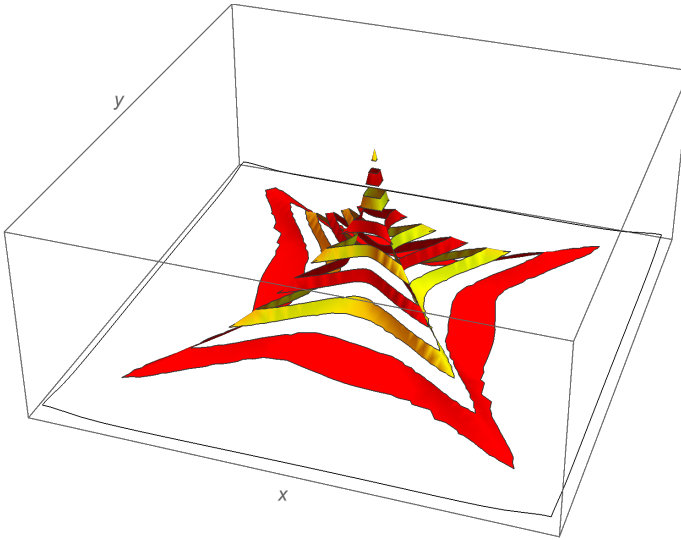
Figure 4.3: *NonElliptical Joint Returns, from stochastic correlations*
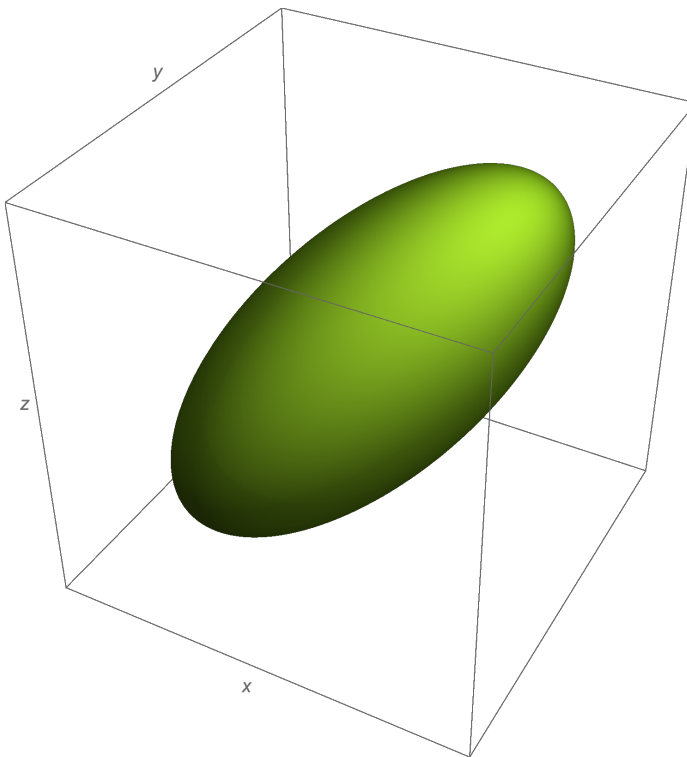
Figure 4.4: *Elliptically Contoured Joint Returns for for a multivariate distribution (x, y, z) solving to the same density.*

μ, a non-negative matrix Σ and some scalar function Ψ if its characteristic functionis of the form $exp(it'\mu)\Psi(t\Sigma t')$.

Intuitively an elliptical distribution should show an ellipse for iso-density plots when represented in 2-D (for a bivariate) and 3-D (for a trivariate) as in Figures 4.2 and 4.4; a noneliptical would violate the shape as in Figures 4.3 and 4.5.

The main property of the class of elliptical distribution is that it is closed under linear transformation. This leads to attractive properties in the building of portfolios, and in the results of portfolio theory (in fact one cannot have portfolio theory without ellitipcality of distributions). Under ellipticality, all portfolios can be characterized completely by their
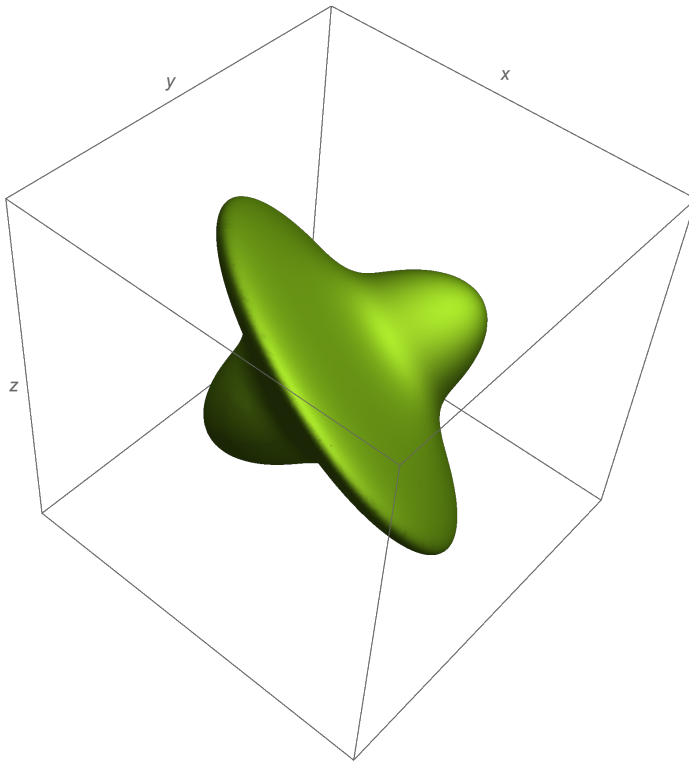
Figure 4.5: *NonElliptical Joint Returns, from stochastic correlations, for a multivariate distribution (x, y, z), solving to the same density.*
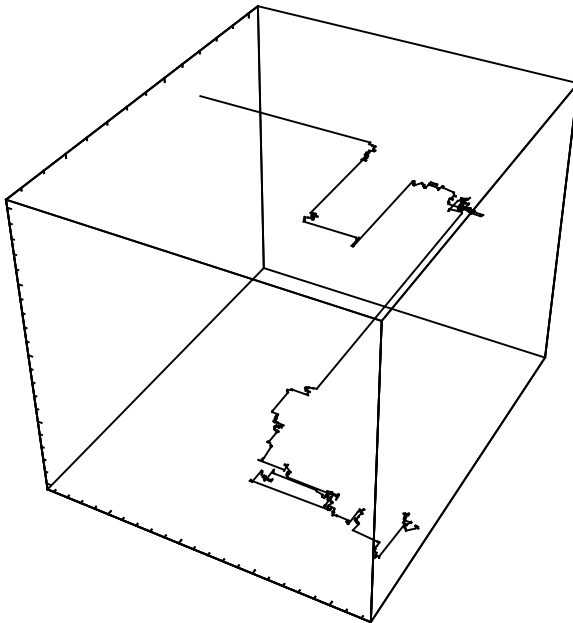
Figure 4.6: **History moves by jumps**: *A fat tailed historical process, in which events are distributed according to a power law that corresponds to the "80/20", with $\alpha \simeq 1.13$, the equivalent of a 3-D Brownian motion.*

location and scale and any two portfolios with identical location and scale (in return space) have identical distributions returns.

Note that (ironically) Lévy-Stable distributions are elliptical –but only in the way they are defined.
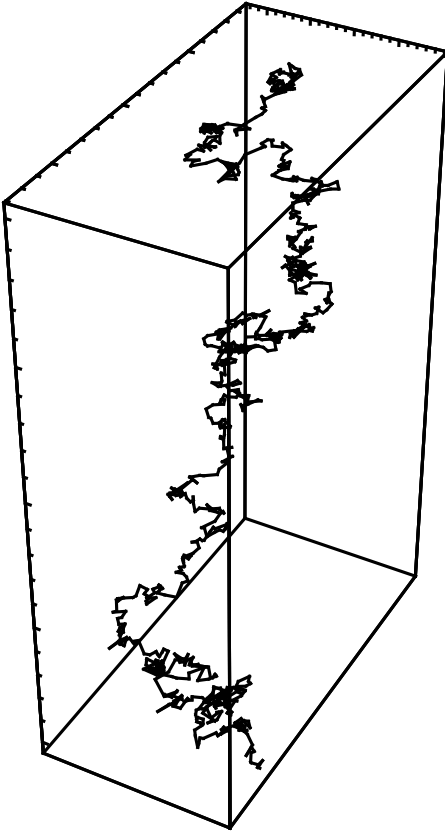
Figure 4.7: *What the proponents of "great modera-tion" or "long peace" have in mind: history as a thin-tailed process.*

4.2.1 Ellipticality and Independence for Fat Tails

Take the product of two Cauchy densities for x and y (what we used in Fig. 2.1):

$$f(x)f(y) = \frac{1}{\pi^2 \left(x^2 + 1\right)\left(y^2 + 1\right)} \tag{4.3}$$

which, patently, as we saw in Chapter 2 (with the example of the two randomly selected persons with a total net worth of \$36 million), is not elliptical. Compare to the joint distribution $f_\rho(x, y)$:

$$f_\rho(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}\left(y\left(\frac{y}{1-\rho^2} - \frac{\rho x}{1-\rho^2}\right) + x\left(\frac{x}{1-\rho^2} - \frac{\rho y}{1-\rho^2}\right) + 1\right)^{3/2}}, \tag{4.4}$$

and setting $\rho = 0$ to get no correlation,

$$f_0(x, y) = \frac{1}{2\pi\left(x^2 + y^2 + 1\right)^{3/2}} \tag{4.5}$$

which is elliptical. This shows that absence of correlation is not independence as:
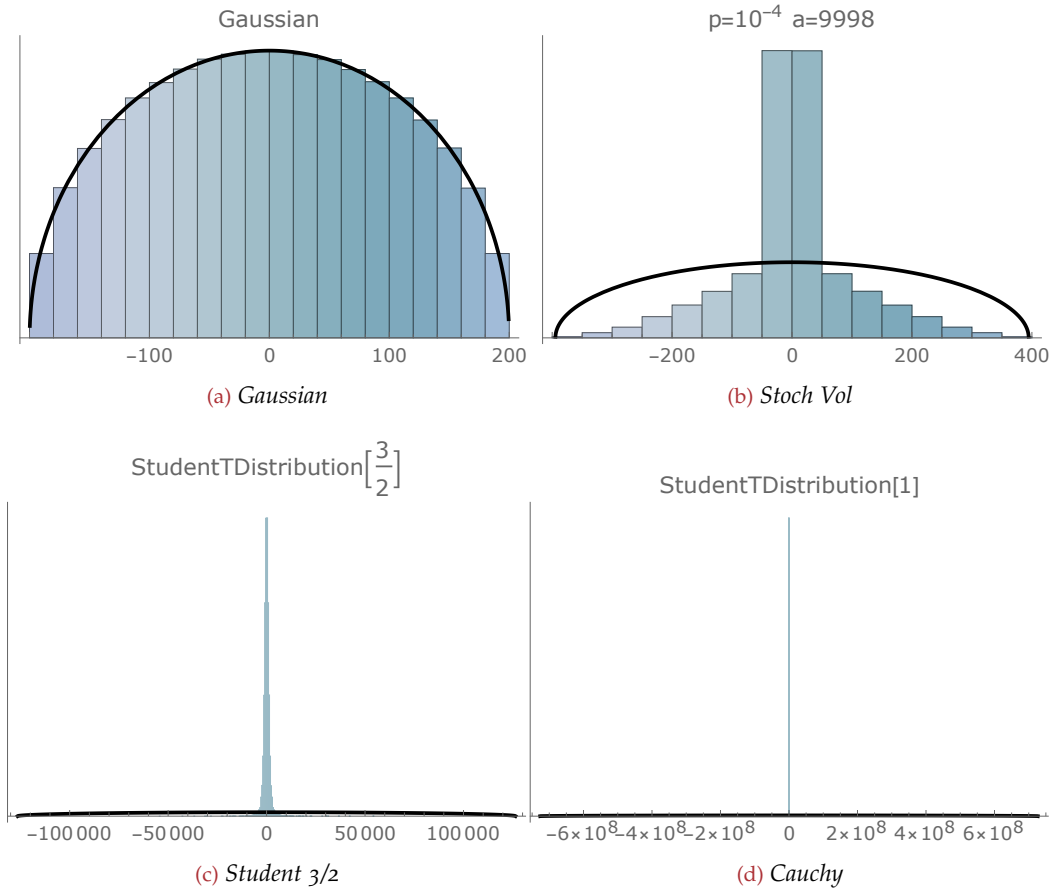
Figure 4.8: *The various shapes of the distribution of the eigenvalues for random matrices, which in the Gaussian case follow the Wigner semi-circle distribution. The Cauchy case corresponds to the Student parametrized to have 1 degrees of freedom*

Independence between two variables X and Y is defined by:

$$\frac{f(x,y)}{f(x)} = 1,$$

regardless of the correlation coefficient. In the class of elliptical distributions, the bivariate Gaussian with coefficient 0 is both independent and uncorrelated.

The reason the multivariate stable distribution with correlation coefficient set to 0 is not independent is the following.

A random vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is said to have the multivariate stable distribution if every linear combination of its components $Y = a_1 X_1 + \cdots + a_k X_k$ has a stable distribution. That is, for any constant vector $\mathbf{a} \in \mathbb{R}^k$, the random variable $\mathbf{Y} = a^T \mathbf{X}$ should have a univariate stable distribution. And to have a linear combination remain within the same class requires ellipticality. Hence *by construction*, $f_0(x,y)$ is not necessarily equal to $f(x)f(y)$. Consider the Cauchy case that has an explicit density function. The denominator of the product of densities includes an additional term, $x^2 y^2$, which pushes the iso-densities in one direction or another.

4.3 FAT TAILS AND RANDOM MATRICES, A RAPID INTERLUDE

The eigenvalues of matrices themselves have an analog to Gaussian convergence: the semi-circle distribution.

Let $\mathbf{M}$ be a (n, n) symmetric matrix. We have the eivenvalues λ_i, $1 \leq i, \leq n$ such that $\mathbf{M}.\mathbf{V}_i = \lambda_i \mathbf{V}_i$ where $\mathbf{V}_i$ is the ith eigenvector.

The Wigner semicircle distribution with support $[-R, R]$ has for PDF f presenting a semicircle of radius R centered at $(0, 0)$ and then suitably normalized :

$$f(\lambda) = \frac{2}{\pi R^2} \sqrt{R^2 - \lambda^2} \text{ for } -R \leq \lambda \leq R. \tag{4.6}$$

This distribution arises as the limiting distribution of eigenvalues of (n, n) symmetric matrices with finite moments as the size n of the matrix approaches infinity.

We will tour the "fat-tailedness" of the random matrix in what follows as well as the convergence.

This is the equivalent of fat tails for matrices. Consider for now that the 4th moment reaching Gaussian levels (i.e. 3) for an univariate situation is equivalent to the eigenvalues reaching Wigner's semicircle.

4.4 MULTIVARIATE SCALE

Let $\mathbf{X}$ be a $(p \times 1)$ vector following a multivariate Student T distribution, $\mathbf{X} \sim \mathcal{S}_t (\mathbf{M}, \mathbf{\Sigma}, \alpha)$, where $\mathbf{\Sigma}$ is a $(p \times p)$ matrix, $\mathbf{M}$ a p length vector and α a Paretian tail exponent with PDF

$$f(\mathbf{X}) = \left(\frac{(\mathbf{X} - \mathbf{M}).\mathbf{\Sigma}^{-1}.(\mathbf{X} - \mathbf{M})}{\nu} + 1 \right)^{-\frac{1}{2}(\nu + p)}. \tag{4.7}$$

In the most simplified case, with $p = 2$, $M = (0, 0)$, and $\mathbf{\Sigma} = = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

$$f(x_1, x_2) = \frac{\nu \sqrt{1 - \rho^2} \left(\frac{-\nu\rho^2 + \nu - 2\rho x_1 x_2 + x_1^2 + x_2^2}{\nu - \nu\rho^2} \right)^{-\frac{\nu}{2} - 1}}{2\pi (\nu - \nu\rho^2)}. \tag{4.8}$$

4.5 CORRELATION AND UNDEFINED VARIANCE

insert section on Student/Cauchy, also show proof of finiteness of correlation

4.5.1 Finiteness of Correlation

Question: Why it is that a fat tailed distribution in the power law class $\mathfrak{P}$ with infinite or undefined mean (and higher moments) would have, in higher dimensions, undefined (or infinite) covariance but finite correlation?

Consider a distribution with support in $(-\infty, \infty)$. It has no moments: $\mathbb{E}(X)$ is indeterminate, $\mathbb{E}(X^2) = \infty$, no covariance, $\mathbb{E}(XY)$ is indeterminate. But the (noncentral) correlation for n variables is bounded by -1 and 1.

$$r \triangleq \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}, \quad n = 2, 3, \dots$$

By the subexponentiality property, we have $\mathbb{P}(X_1 + \dots + X_n) > x \sim \mathbb{P}(\max(X_1, \dots X_n) > x)$ as $x \to \infty$. We note that the power law class is included in the subexponential class $\mathfrak{S}$.

Order the variables in absolute values such that $|x_1| \leq |x_2| \leq \dots \leq |x_n|$

Let $\kappa_1 = \sum_{i=1}^{n-1} x_i y_i$, $\kappa_2 = \sum_{i=1}^{n-1} x_i^2$, and $\kappa_3 = \sum_{i=1}^{n-1} y_i^2$.

$$\lim_{x_n \to \infty} \frac{x_n y_n + \kappa_1}{\sqrt{x_n^2 + \kappa_2} \sqrt{y_n^2 + \kappa_3}} = \frac{y_n}{\sqrt{\kappa_3 + y_n^2}},$$

$$\lim_{y_n \to \infty} \frac{x_n y_n + \kappa_1}{\sqrt{x_n^2 + \kappa_2} \sqrt{y_n^2 + \kappa_3}} = \frac{x_n}{\sqrt{\kappa_2 + x_n^2}}$$

$$\lim_{x_n \to +\infty, y_n \to +\infty} = 1$$

$$\lim_{x_n \to +\infty, y_n \to -\infty} = -1$$

and

$$\lim_{x_n \to -\infty, y_n \to +\infty} = -1$$

for all values of $n \geq 2$.

4.5.2 Example of finiteness of Correlation
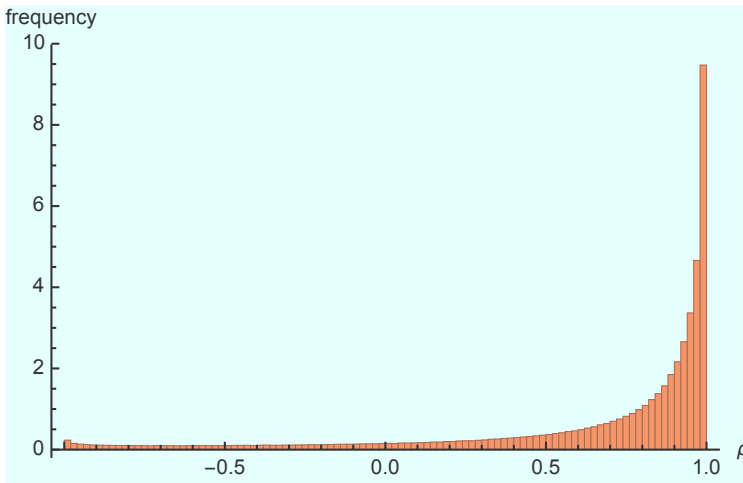
Figure 4.9: *Sample distribution of correlation for a sample of 10^3. The correlation exists for a bivariate T distribution (exponent $\frac{2}{3}$, correlation $\frac{3}{4}$) but... not useable*

FURTHER READING

For subexponentiality: Pitman [136], Embrechts and Goldie (1982) [61]Embrechts (1979 which seems to be close to his doctoral thesis), [62], Chistyakov (1964) [31], Goldie (1978) [81], Pitman [136], Teugels [174], and, more general, [63].

5 | THE EMPIRICAL DISTRIBUTION IS NOT EMPIRICAL

Crash Beliefs From Investor Surveys

William N. Goetzmann

Yale School of Management, Yale University

Dasol Kim

Weatherhead School of Management, Case Western Reserve University

Robert J. Shiller

Yale University

Abstract: Historical data suggest that the base rate for a severe, single-day stock market crash is relatively low. Surveys of individual and institutional investors, conducted regularly over a 26 year period in the United States, show that they assess the probability to be much higher. We examine the factors that influence investor responses and test the role of media influence. We find evidence consistent with an availability bias. Recent market declines and adverse market events made salient by the financial press are associated with higher subjective crash probabilities. Non-market-related, rare disasters are also associated with higher subjective crash probabilities.

Figure 5.1: *The base rate fallacy, revisited —or, rather in the other direction. The "base rate" is an empirical evaluation that bases itself on the worst past observations, an error identified in [158] as the fallacy identified by the Roman poet Lucrecius in* De rerum natura *of thinking the tallest future mountain equals the tallest on has previously seen. Quoted without permission after warning the author.*

HERE is a prevalent confusion about the nonparametric empirical distribution based on the following powerful property: as n grows, the errors around the empirical histogram for cumulative frequencies are Gaussian *regardless of the base distribution*, even if the true distribution is fat-tailed (assuming infinite support). For the CDF (or survival functions) are both uniform on $[0, 1]$, and, further, by the Donsker theorem, the sequence $\sqrt{n}\left(F_n(x) - F(x)\right)$ (F_n is the observed CDF or survival function for n summands, F the true CDF or survival function) converges in distribution to a Normal Distribution with mean 0 and variance $F(x)\left(1 - F(x)\right)$ (one may find even stronger forms of convergence via the Glivenko– Cantelli theorem).

Owing to this remarkable property, one may mistakenly assume that the effect of tails of the distribution converge in the same manner independently of the distribution. Further, and what contributes to the confusion, the variance, $F(x)\left(1 - F(x)\right)$ for both empirical CDF and survival function, drops at the extremes.

69

Figure 5.2: *The high watermark: the level of flooding in Paris in 1910 as a maxima. Clearly one has to consider that such record will be topped some day in the future and proper risk management consists in "how much" more than such a level one should seek protection.*

In truth, and that is a property of extremes, the error effectively increases in the tails if one multiplies by the deviation that corresponds to the probability.

Let χ_n be the difference between the empirical and the distributional conditional mean, defined as:

$$\chi_n = \sum_{i=1}^{n} x_i \mathbb{1}_{x_i \geq K} - \int_K^\infty x dF(x)$$

$$= K(F_n(K) - F(K)) + \sum_i^{\frac{x_{max}}{\delta}} \left(\bar{F}_n(K + (i+1)\delta) - \bar{F}_n(K + i\delta K) - \int_{K+i\delta}^{K+(i+1)\delta} d\bar{F}(K) \right) - \int_{x_{max}}^\infty dF(x),$$

$$(5.1)$$

where $x_{max} = \bar{F}_n^{-1}(0)$, that is where the distribution is truncated. χ_n recovers the dispersion of the distribution of x which remains fat tailed. Another way to see it is that for fat tailed variables, probabilities are more stable than their realizations and, more generally, the lowest moment will always disproportionately be the most stable one.

Biases of the empirical method under Fat Tails

We note that, owing of the convergence to the Gaussian, by Donsker's theorem:

$$\chi_n = \int_{x_{max}}^{\infty} dF(x) + O\left(\frac{F(x)\,(1 - F(x))}{\sqrt{n}}\right) \tag{5.2}$$

so, for sufficiently large (but not too large) n,

$$\chi_n \approx \int_{x_{max}}^{\infty} dF(x) \tag{5.3}$$

yet, under a Paretian regime, x_{max} is distributed according to a Fréchet, as we will see in Section TK.

Theorem 5.1

For an empirical distribution with a sample size n, the underestimation of the conditional tail expectation χ_n for a Paretian with scale L and tail index α is:

$$\varphi(\chi, n) = \left(\frac{\alpha - 1}{\alpha}\right)^{\frac{1}{\alpha - 1}} nL^{\frac{\alpha^2}{1 - \alpha} + 1} \chi^{\frac{1}{\alpha - 1}} \left(\exp\left(\left(\frac{\alpha - 1}{\alpha}\right)^{\frac{\alpha}{\alpha - 1}} n\left(-L^{\frac{\alpha^2}{1 - \alpha} + 1}\right)\chi^{\frac{\alpha}{\alpha - 1}}\right)\right) \tag{5.4}$$

and its expectation

$$\mathbb{E}(\chi_n) = \Gamma\left(\frac{\alpha - 1}{\alpha}\right) L^{\alpha + \frac{1}{\alpha} - 1} n^{\frac{1}{\alpha} - 1}$$

Proof. The maximum of n variables is in the MDA (Maximum domain of attraction) of Fréchet with scale $\beta = (Ln)^{1/\alpha}$. We have the conditional expectation $> \chi$: $\mathbb{E}(x)|_{x > \chi}\mathbb{P}(x > \chi) = \frac{\alpha L^\alpha \chi^{1 - \alpha}}{\alpha - 1}$. Randomizing χ and doing a probability transformation we get the density $\varphi(.)$. □

A | ECONOMETRICS IMAGINES FUNCTIONS IN L^2 SPACE†

THERE IS SOMETHING WRONG with econometrics, as almost all papers don't replicate in the real world. Two reliability tests in Chapter 8, one about parametric methods the other about robust statistics, show that there is something rotten in econometric methods, fundamentally wrong, and that the methods are not dependable enough to be of use in anything remotely related to risky decisions. Practitioners keep spinning inconsistent *ad hoc* statements to explain failures.

This is a brief nontechnical exposition from the results in [159].

With economic variables one single observation in 10,000, that is, one single day in 40 years, can explain the bulk of the "kurtosis", the finite-moment standard measure of "fat tails", that is, both a measure how much the distribution under consideration departs from the standard Gaussian, or the role of remote events in determining the total properties. For the U.S. stock market, a single day, the crash of 1987, determined 80% of the kurtosis for the period between 1952 and 2008. The same problem is found with interest and exchange rates, commodities, and other variables. Redoing the study at different periods with different variables shows a total instability to the kurtosis. The problem is not just that the data had "fat tails", something people knew but sort of wanted to forget; it was that we would never be able to determine "how fat" the tails were within standard methods. Never.

The implication is that those tools used in economics that are *based on squaring variables* (more technically, the $\mathcal{L}^2$ norm), such as standard deviation, variance, correlation, regression, the kind of stuff you find in textbooks, are not valid *scientifically* (except in some rare cases where the variable is bounded). The so-called "p values" you find in studies have no meaning with economic and financial variables. Even the more sophisticated techniques of stochastic calculus used in mathematical finance do not work in economics except in selected pockets.

A.0.1 Performance of Standard Parametric Risk Estimators

The implication is that those tools used in economics that are **based on squaring variables** (more technically, the Euclidian, or ℓ^2 norm), such as standard deviation, variance, correlation, regression, the kind of stuff you find in textbooks, are not valid *scientifically*(except in some rare cases where the variable is bounded). The so-called "p values" you find in studies have no meaning with economic and financial variables. Even the more sophisticated

Table A.1: *Maximum contribution to the fourth moment from a single daily observation*

Security	Max Q	Years.
Silver	0.94	46.
SP500	0.79	56.
CrudeOil	0.79	26.
Short Sterling	0.75	17.
Heating Oil	0.74	31.
Nikkei	0.72	23.
FTSE	0.54	25.
JGB	0.48	24.
Eurodollar Depo 1M	0.31	19.
Sugar #11	0.3	48.
Yen	0.27	38.
Bovespa	0.27	16.
Eurodollar Depo 3M	0.25	28.
CT	0.25	48.
DAX	0.2	18.

techniques of stochastic calculus used in mathematical finance do not work in economics except in selected pockets.

The results of most papers in economics based on these standard statistical methods are thus not expected to replicate, and they effectively don't. Further, these tools invite foolish risk taking. Neither do alternative techniques yield reliable measures of rare events, except that we can tell if a remote event is underpriced, without assigning an exact value.

From [159]), using log returns, $X_t := \log\left(\frac{P(t)}{P(t-i\Delta t)}\right)$. Consider the n-sample maximum quartic observation $\mathrm{Max}(X_{t-i\Delta t}^4)_{i=0}^n$. Let $Q(n)$ be the contribution of the maximum quartic variations over n samples and frequency Δt.

$$Q(n) := \frac{\mathrm{Max}\left(X_{t-i\Delta t}^4\right)_{i=0}^n}{\sum_{i=0}^n X_{t-i\Delta t}^4}$$

Note that for our purposes, where we use central or noncentral kurtosis makes no difference –results are nearly identical.

For a Gaussian (i.e., the distribution of the square of a Chi-square distributed variable) show $Q\left(10^4\right)$ the maximum contribution should be around .008 $\pm$ .0028. Visibly we can see that the observed distribution of the 4th moment has the property

$$P\left(X > \max(x_i^4)_{i \le 2 \le n}\right) \approx P\left(X > \sum_{i=1}^n x_i^4\right)$$

Recall that, naively, the fourth moment expresses the stability of the second moment. And the second moment expresses the stability of the measure across samples.

Note that taking the snapshot at a different period would show extremes coming from other variables while these variables showing high maximma for the kurtosis, would drop, a mere result of the instability of the measure across series and time.

Description of the dataset All tradable macro markets data available as of August 2008, with "tradable" meaning actual closing prices corresponding to transactions (stemming from markets not bureaucratic evaluations, includes interest rates, currencies, equity indices).

Share of Max Quartic
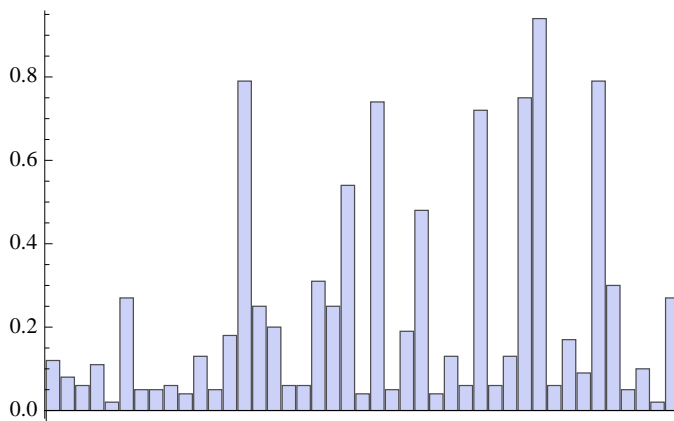
Figure A.1: *Max quartic across securities in Table A.1.*

EuroDepo 3M: Annual Kurt 1981−2008
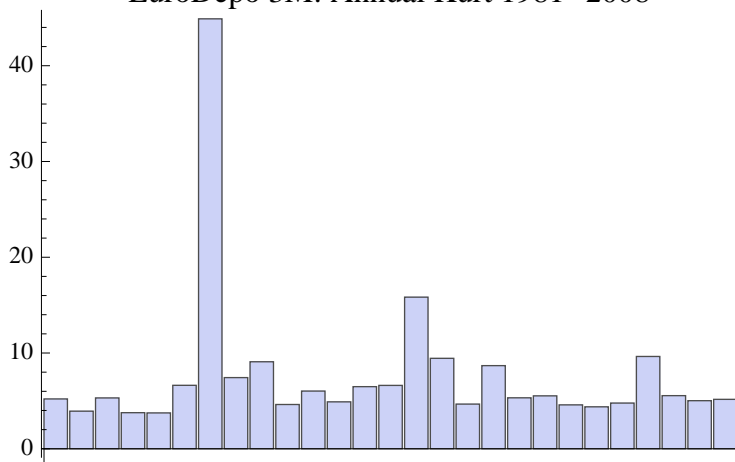
Figure A.2: *Kurtosis across nonoverlapping periods for Eurodeposits.*

Monthly Vol
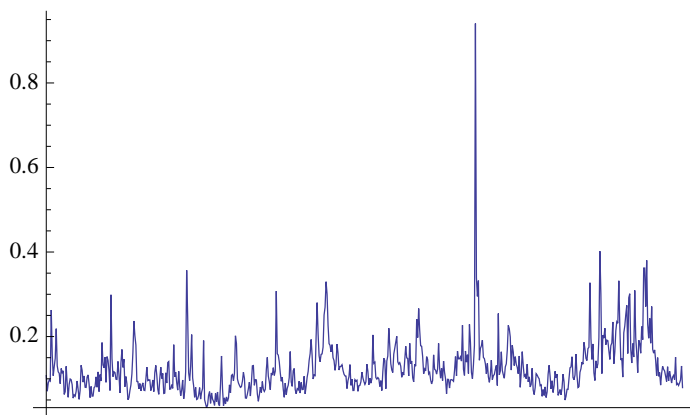
Figure A.3: *Monthly delivered volatility in the SP500 (as measured by standard deviations). The only structure it seems to have comes from the fact that it is bounded at 0. This is standard.*

Vol of Vol
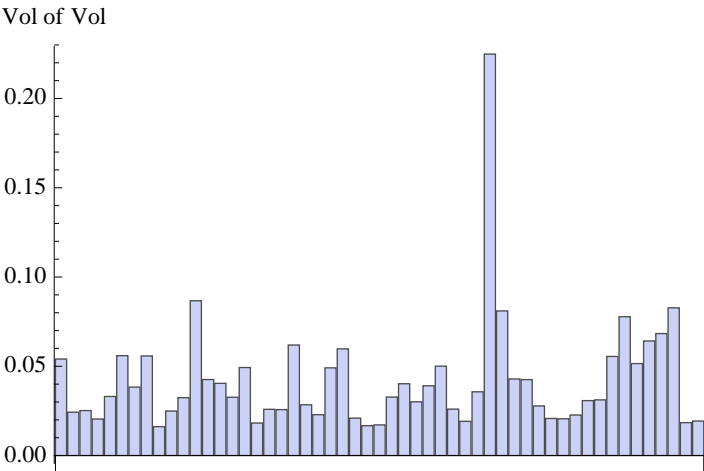
Figure A.4: *Montly volatility of volatility from the same dataset in Table A.1, predictably unstable.*

A.0.2 Performance of Standard NonParametric Risk Estimators

Does the past resemble the future in the tails? The following tests are nonparametric, that is entirely based on empirical probability distributions.
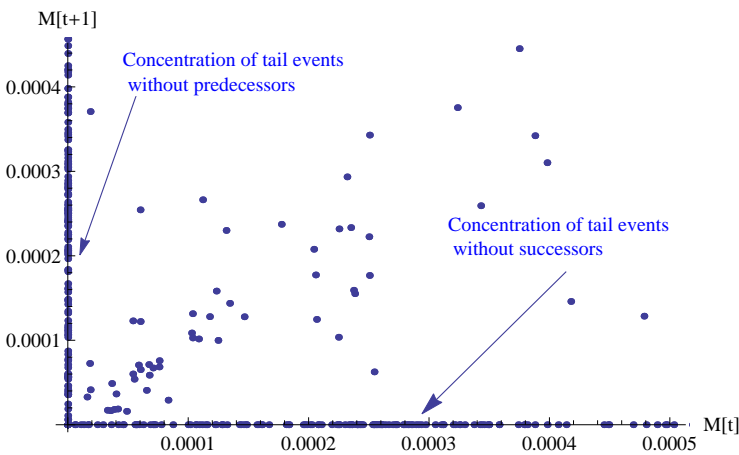
Figure A.5: *Comparing one absolute deviation M[t] and the subsequent one M[t+1] over a certain threshold (here 4% in stocks); illustrated how large deviations have no (or few) predecessors, and no (or few) successors– over the past 50 years of data.*
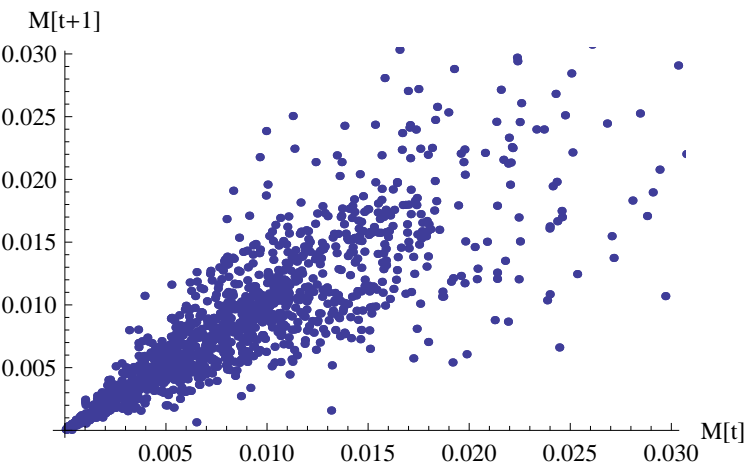
Figure A.6: *The "regular" is predictive of the regular, that is mean deviation. Comparing one absolute deviation M[t] and the subsequent one M[t+1] for macroeconomic data.*

So far we stayed in dimension 1. When we look at higher dimensional properties, such as covariance matrices, things get worse. We will return to the point with the treatment of model error in mean-variance optimization.

When x_t are now in $\mathbb{R}^N$, the problems of sensitivity to changes in the covariance matrix makes the estimator M extremely unstable. Tail events for a vector are vastly more difficult to calibrate, and increase in dimensions.
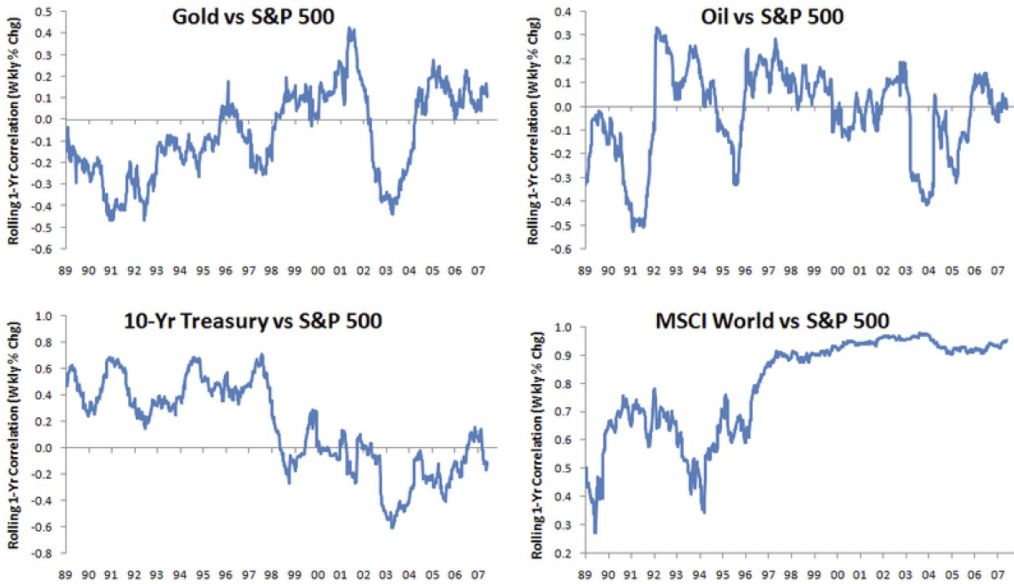
Figure A.7: *Correlations are also problematic, which flows from the instability of single variances and the effect of multiplication of the values of random variables.*

The Responses so far by members of the economics/econometrics establishment   No answer as to why they still use STD, regressions, GARCH , value-at-risk and similar methods.

Peso problem   Benoit Mandelbrot used to insist that one can fit anything with Poisson jumps.

Many researchers invoke "outliers" or "peso problem"[1] as acknowledging fat tails (or the role of the tails for the distribution), yet ignore them analytically (outside of Poisson models that are not possible to calibrate except after the fact: conventional Poisson jumps are thin-tailed). Our approach here is exactly the opposite: do not push outliers under the rug, rather build everything around them. In other words, just like the FAA and the FDA who deal with safety by focusing on catastrophe avoidance, we will throw away the ordinary under the rug and retain extremes as the sole sound approach to risk management. And this extends beyond safety since much of the analytics and policies that can be destroyed by tail events are inapplicable.

Peso problem confusion about the Black Swan problem   :

"(...) "Black Swans" (Taleb, 2007). These cultural icons refer to disasters that occur so infrequently that they are virtually impossible to analyze using standard statistical

---

1 The peso problem is a discovery of an outlier in money supply, became a name for outliers and unexplained behavior in econometrics.

inference. However, we find this perspective less than helpful because it suggests a state of hopeless ignorance in which we resign ourselves to being buffeted and battered by the unknowable."

Andrew Lo, who obviously did not bother to read the book he was citing.

Lack of skin in the game.   Indeed one wonders why econometric methods keep being used while being wrong, so shockingly wrong, how "University" researchers (adults) can partake of such acts of artistry. Basically these capture the ordinary and mask higher order effects. Since blowups are not frequent, these events do not show in data and the researcher looks smart most of the time while being fundamentally wrong. At the source, researchers, "quant" risk manager, and academic economist do not have skin in the game so they are not hurt by wrong risk measures: other people are hurt by them. And the artistry should continue perpetually so long as people are allowed to harm others with impunity. (More in Taleb and Sandis [173], Taleb [167]).

B | SPECIAL CASES OF FAT TAILS
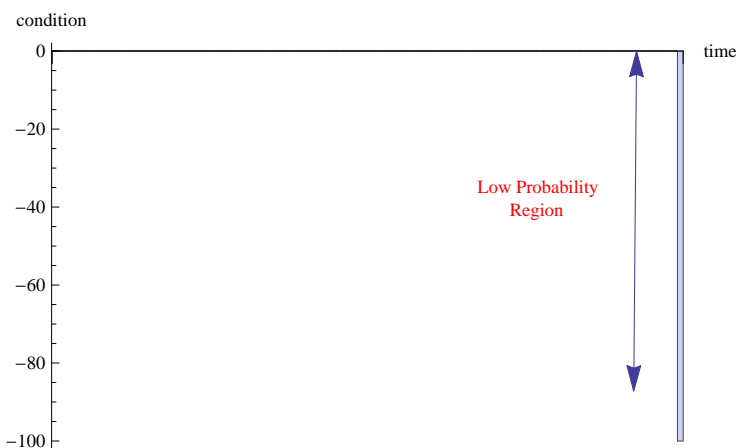
Figure B.1: *A coffee cup is less likely to incur "small" than large harm. It shatters, hence is exposed to (almost) everything or nothing. The same type of payoff is prevalent in markets with, say, (reval)devaluations, where small movements beyond a barrier are less likely than larger ones.*

For monomodal distributions, fat tails are the norm: one can look at tens of thousands of time series of the socio-economic variables without encountering a single episode of "platykurtic" distributions. But for multimodal distributions, some surprises can occur.
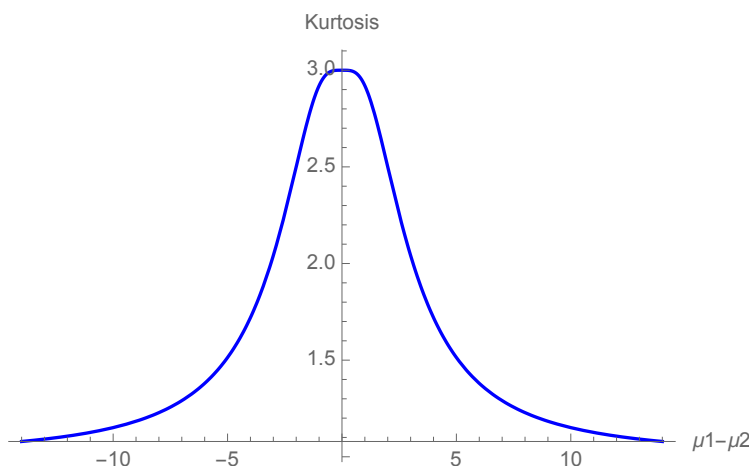
Figure B.2: *Negative (relative) kurtosis and bimodality (3 is the Gaussian).*

B.1 MULTIMODALITY AND FAT TAILS, OR THE WAR AND PEACE MODEL

We noted earlier in 3.1.1 that stochasticizing, ever so mildly, variances, the distribution gains in fat tailedness (as expressed by kurtosis). But we maintained the same mean.

But should we stochasticize the mean as well (while preserving the initial average), and separate the potential outcomes wide enough, so that we get many modes, the "kurtosis" (as measured by the fourth moment) would drop. And if we associate different variances with different means, we get a variety of "regimes", each with its set of probabilities.

Either the very meaning of "fat tails" loses its significance under multimodality, or takes on a new one where the "middle", around the expectation ceases to matter.[5, 106].

Now, there are plenty of situations in real life in which we are confronted to many possible regimes, or states. Assuming finite moments for all states, consider the following structure: s_1 a calm regime, with expected mean m_1 and standard deviation σ_1, s_2 a violent regime, with expected mean m_2 and standard deviation σ_2, or more such states. Each state has its probability p_i.

Now take the simple case of a Gaussian with switching means and variance: with probability $\frac{1}{2}$, $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and with probability $\frac{1}{2}$, $X \sim \mathcal{N}(\mu_2, \sigma_2)$. The kurtosis will be

$$Kurtosis = 3 - \frac{2\left((\mu_1 - \mu_2)^4 - 6\left(\sigma_1^2 - \sigma_2^2\right)^2\right)}{\left((\mu_1 - \mu_2)^2 + 2\left(\sigma_1^2 + \sigma_2^2\right)\right)^2} \tag{B.1}$$

As we see the kurtosis is a function of $d = \mu_1 - \mu_2$. For situations where $\sigma_1 = \sigma_2$, $\mu_1 \neq \mu_2$, the kurtosis will be below that of the regular Gaussian, and our measure will naturally be negative. In fact for the kurtosis to remain at 3,

$$|d| = \sqrt[4]{6}\sqrt{\max(\sigma_1, \sigma_2)^2 - \min(\sigma_1, \sigma_2)^2},$$

the stochasticity of the mean offsets the stochasticity of volatility.

Assume, to simplify a one-period model, as if one was standing in front of a discrete slice of history, looking forward at outcomes. (Adding complications (transition matrices between different regimes) doesn't change the main result.)

The characteristic function $\phi(t)$ for the mixed distribution becomes:

$$\phi(t) = \sum_{i=1}^{N} p_i e^{-\frac{1}{2}t^2\sigma_i^2 + itm_i}$$

For $N = 2$, the moments simplify to the following:

$$M_1 = p_1 m_1 + (1 - p_1) m_2$$

$$M_2 = p_1\left(m_1^2 + \sigma_1^2\right) + (1 - p_1)\left(m_2^2 + \sigma_2^2\right)$$

$$M_3 = p_1 m_1^3 + (1 - p_1) m_2\left(m_2^2 + 3\sigma_2^2\right) + 3m_1 p_1 \sigma_1^2$$

$$M_4 = p_1\left(6m_1^2\sigma_1^2 + m_1^4 + 3\sigma_1^4\right) + (1 - p_1)\left(6m_2^2\sigma_2^2 + m_2^4 + 3\sigma_2^4\right)$$

Let us consider the different varieties, all characterized by the condition $p_1 < (1 - p_1)$, $m_1 < m_2$, preferably $m_1 < 0$ and $m_2 > 0$, and, at the core, the central property: $\sigma_1 > \sigma_2$.

Variety 1: War and Peace.   Calm period with positive mean and very low volatility, turmoil with negative mean and extremely low volatility.
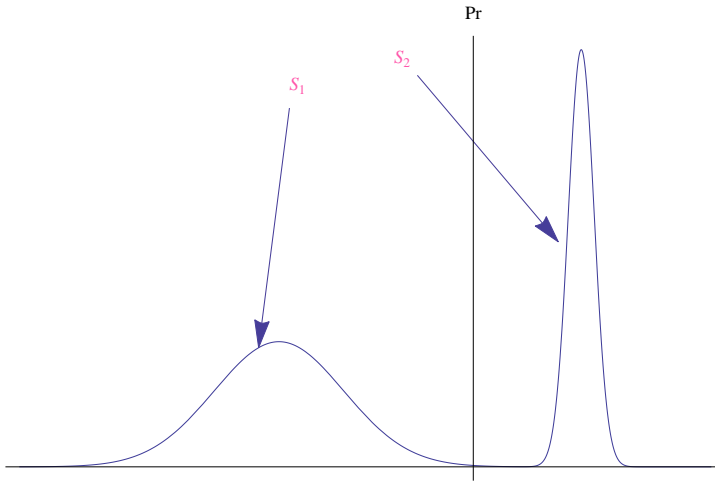
Figure B.3: *The War and peace model. Kurtosis K=1.7, much lower than the Gaussian.*

Variety 2: Conditional deterministic state   Take a bond B, paying interest r at the end of a single period. At termination, there is a high probability of getting $B(1 + r)$, a possibility of defaut. Getting exactly B is very unlikely. Think that there are no intermediary steps between war and peace: these are separable and discrete states. Bonds don't just default "a little bit". Note the divergence, the probability of the realization being at or close to the mean is about nil. Typically, $p(\mathbb{E}(x))$ the probabilitity densities of the expectation are smaller than at the different means of regimes, so $\mathbb{P}(x = \mathbb{E}(x)) < \mathbb{P}(x = m_1)$ and $< \mathbb{P}(x = m_2)$, but in the extreme case (bonds), $\mathbb{P}(x = \mathbb{E}(x))$ becomes increasingly small. The tail event is the realization around the mean.
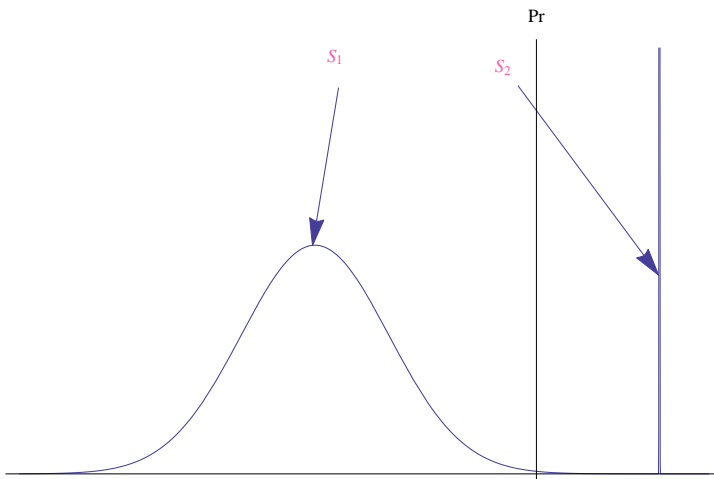
Figure B.4: *The Bond payoff/Currency peg model. Absence of volatility stuck at the peg, deterministic payoff in regime 2, mayhem in regime 1. Here the kurtosis K=2.5. Note that the coffee cup is a special case of both regimes 1 and 2 being degenerate .*

The same idea applies to currency pegs, as devaluations cannot be "mild", with all-or-nothing type of volatility and low density in the "valley" between the two distinct regimes.
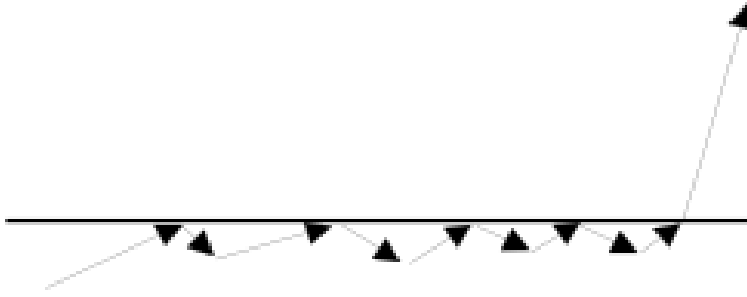
Figure B.5: *Pressure on the peg which may give a Dirac PDF in the "no devaluation" regime (or, equivalently, low volatility). It is typical for finance imbeciles to mistake regime S_2 for low volatility.*

With option payoffs, this bimodality has the effect of raising the value of at-the-money options and lowering that of the out-of-the-money ones, causing the exact opposite of the so-called "volatility smile".

Note the coffee cup has no state between broken and healthy. And the state of being broken can be considered to be an absorbing state (using Markov chains for transition probabilities), since broken cups do not end up fixing themselves.

Nor are coffee cups likely to be "slightly broken", as we see in figure B.1.

A brief list of other situations where bimodality is encountered:

1. Currency pegs
2. Mergers
3. Professional choices and outcomes
4. Conflicts: interpersonal, general, martial, any situation in which there is no intermediary between harmonious relations and hostility.
5. Conditional cascades

B.2 TRANSITION PROBABILITES: WHAT CAN BREAK WILL BREAK

So far we looked at a single period model, which is the realistic way since new information may change the bimodality going into the future: we have clarity over one-step but not more. But let us go through an exercise that will give us an idea about fragility. Assuming the structure of the model stays the same, we can look at the longer term behavior under transition of states. Let P be the matrix of transition probabilitites, where $p_{i,j}$ is the transition from state i to state j over Δt, (that is, where S(t) is the regime prevailing over period t, $P\left(S(t + \Delta t) = s_j \middle| S(t) = s_j\right)$)

$$P = \begin{pmatrix} p_{1,1} & p_{2,1} \\ p_{1,2} & p_{2,2} \end{pmatrix}$$

After n periods, that is, n steps,

$$P^n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}$$

Where

$$a_n = \frac{(p_{1,1} - 1)(p_{1,1} + p_{2,2} - 1)^n + p_{2,2} - 1}{p_{1,1} + p_{2,2} - 2}$$

$$b_n = \frac{(1 - p_{1,1})((p_{1,1} + p_{2,2} - 1)^n - 1)}{p_{1,1} + p_{2,2} - 2}$$

$$c_n = \frac{(1 - p_{2,2})((p_{1,1} + p_{2,2} - 1)^n - 1)}{p_{1,1} + p_{2,2} - 2}$$

$$d_n = \frac{(p_{2,2} - 1)(p_{1,1} + p_{2,2} - 1)^n + p_{1,1} - 1}{p_{1,1} + p_{2,2} - 2}$$

The extreme case to consider is the one with the absorbing state, where $p_{1,1} = 1$, hence (replacing $p_{i,\neq i | i = 1,2} = 1 - p_{i,i}$).

$$P^n = \begin{pmatrix} 1 & 0 \\ 1 - p_{2,2}^N & p_{2,2}^N \end{pmatrix}$$

and the "ergodic" probabilities:

$$\lim_{n \to \infty} P^n = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

The implication is that the absorbing state regime 1 S(1) will end up dominating with probability 1: what can break and is irreversible will eventually break.

With the "ergodic" matrix,

$$\lim_{n \to \infty} P^n = \pi . \mathbf{1}^{\mathsf{T}}$$

where $\mathbb{1}^{\mathsf{T}}$ is the transpose of unitary vector $\{1,1\}$, π the matrix of eigenvectors.

The eigenvalues become $\lambda = \begin{pmatrix} 1 \\ p_{1,1} + p_{2,2} - 1 \end{pmatrix}$ and associated eigenvectors $\pi = \begin{pmatrix} 1 & 1 \\ \frac{1 - p_{1,1}}{1 - p_{2,2}} & 1 \end{pmatrix}$.

C | PSEUDO-STOCHASTIC VOLATILITY: A CASE STUDY

<mark>remove and merge</mark>
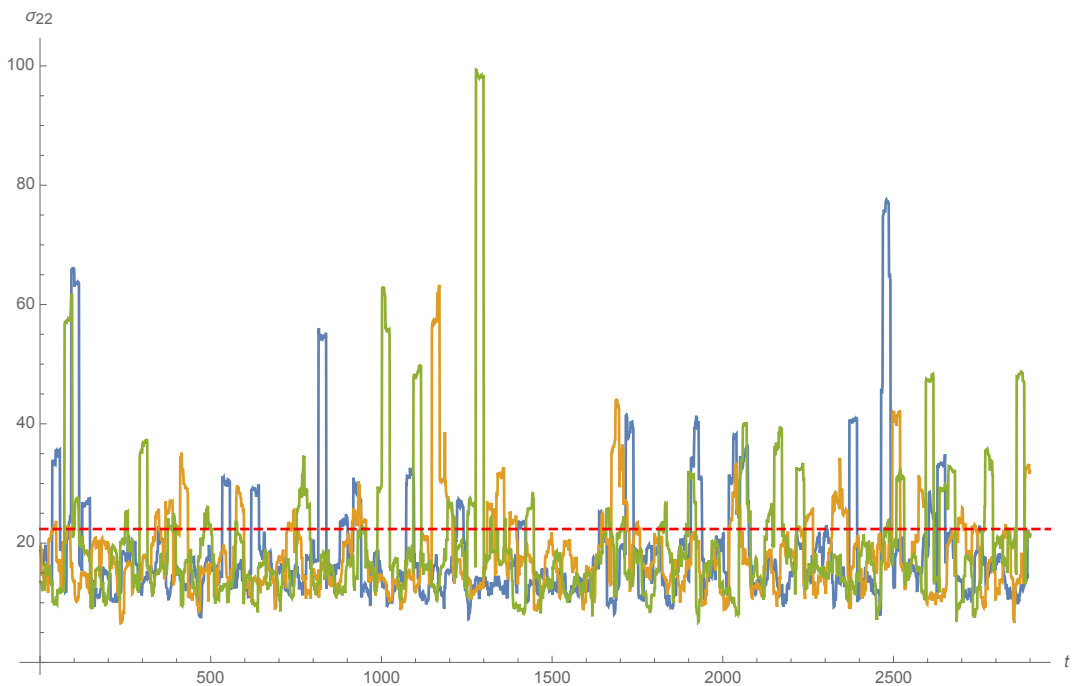
Figure C.1: *Running 22-day (i.e., corresponding to monthly) realized volatility (standard deviation) for a Student T distributed returns sampled daily. It gives the impression of stochastic volatility when in fact the scale of the distribution is constant.*

Fig. C.1 shows the volatility of returns of a market that greatly resemble ones should one use a standard simple stochastic volatility process. By stochastic volatility we assume the variance is distributed randomly (although the expressions are usually about the standard deviation, modeling is off the variance—note that the two have different expectations).

Let X be the returns with mean 0 and scale σ, with PDF $\varphi(.)$:

$$\varphi(x) = \frac{\left(\frac{\alpha}{\alpha + \frac{x^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\,\sigma B\left(\frac{\alpha}{2}, \frac{1}{2}\right)}, x \in (-\infty, \infty).$$

Transforming to get $Y = X^2$ (to get the distribution of the variance), ψ, the PDF for Y becomes,

$$\psi(y) = \frac{\left(\frac{\alpha\sigma^2}{\alpha\sigma^2+y}\right)^{\frac{\alpha+1}{2}}}{\sigma B\left(\frac{\alpha}{2},\frac{1}{2}\right)\sqrt{\alpha y}}, y \in (-\infty, \infty),$$

which we can see transforms into a power law with asymptotic tail exponent $\frac{\alpha}{2}$. The characteristic function $\omega\chi_y = \mathbb{E}(\exp(i\omega Y))$ can be written as

$$\omega\chi_y = \frac{\pi\sqrt{\alpha}\sigma\sqrt{\frac{1}{\alpha\sigma^2}}((\pi\alpha)\csc)\left(\frac{\sqrt{\pi}\,{}_1\tilde{F}_1\left(\frac{1}{2};1-\frac{\alpha}{2};-i\alpha\sigma^2\omega\right)}{\Gamma\left(\frac{\alpha+1}{2}\right)} - \left(\frac{1}{\alpha\sigma^2}\right)^{-\frac{\alpha}{2}}(-i\omega)^{\alpha/2}\,{}_1\tilde{F}_1\left(\frac{\alpha+1}{2};\frac{\alpha+2}{2};-i\alpha\sigma^2\omega\right)\right)}{2B\left(\frac{\alpha}{2},\frac{1}{2}\right)}$$

From which we get the mean deviation of the variance.

α	Mean Deviation
$\frac{5}{2}$	$\dfrac{\sqrt[4]{\frac{5}{3}}2^{3/4}\left(2\,{}_2F_1\left(\frac{1}{4},\frac{7}{4};\frac{5}{4};-\frac{5}{6}\right)+3\left(\frac{6}{11}\right)^{3/4}\right)\sigma^2\Gamma\left(\frac{7}{4}\right)}{\sqrt{\pi}\Gamma\left(\frac{5}{4}\right)}$
3	$\dfrac{6\sigma^2}{\pi}$
$\frac{7}{2}$	$\dfrac{5\,7^{3/4}\left(7\,{}_2F_1\left(\frac{3}{4},\frac{9}{4};\frac{7}{4};-\frac{7}{6}\right)-3\,{}_2F_1\left(\frac{7}{4},\frac{9}{4};\frac{11}{4};-\frac{7}{6}\right)\right)\sigma^2\Gamma\left(\frac{5}{4}\right)}{6\,6^{3/4}\sqrt{\pi}\Gamma\left(\frac{7}{4}\right)}$
4	$\dfrac{1}{7}\left(3\sqrt{21}-7\right)\sigma^2$
$\frac{9}{2}$	$\dfrac{3\sqrt[4]{\frac{3}{2}}\left(6\left(\frac{2}{5}\right)^{3/4}-6\,{}_2F_1\left(\frac{5}{4},\frac{11}{4};\frac{9}{4};-\frac{3}{2}\right)\right)\sigma^2\Gamma\left(\frac{11}{4}\right)}{5\sqrt{\pi}\Gamma\left(\frac{9}{4}\right)}$
5	$\dfrac{\sigma^2\left(7\sqrt{15}-16\tan^{-1}\left(\sqrt{\frac{5}{3}}\right)\right)}{6\pi}$

D | CASE STUDY: HOW THE MYOPIC LOSS AVERSION IS MISSPECIFIED

We fatten tails of the distribution with stochasticity of, say, the scale parameter, and can see what happens to some results in the literature that seem absurd at face value, and in fact are absurd under more rigorous use of probabilistic analyses.

Myopic loss aversion
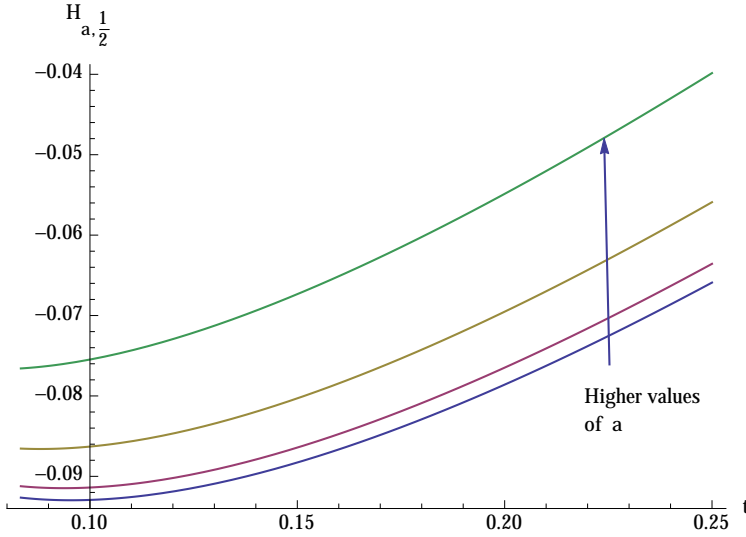
Figure D.1: *The effect of $H_{a,p}(t)$ "utility" or prospect theory of under second order effect on variance. Here $\sigma = 1, \mu = 1$ and t variable.*

Take the prospect theory valuation w function for x changes in wealth.

$$w_{\lambda,\alpha}(x) = x^{\alpha}\, \mathbb{1}_{x \geq 0} - \lambda(-x^{\alpha})\, \mathbb{1}_{x < 0}$$

Where $\phi_{\mu t, \sigma\sqrt{t}}(x)$ is the Normal Distribution density with corresponding mean and standard deviation (scaled by t)

The expected "utility" (in the prospect sense):

$$H_0(t) = \int_{-\infty}^{\infty} w_{\lambda,\alpha}(x)\phi_{\mu t, \sigma\sqrt{t}}(x)\, \mathrm{d}x \tag{D.1}$$

87
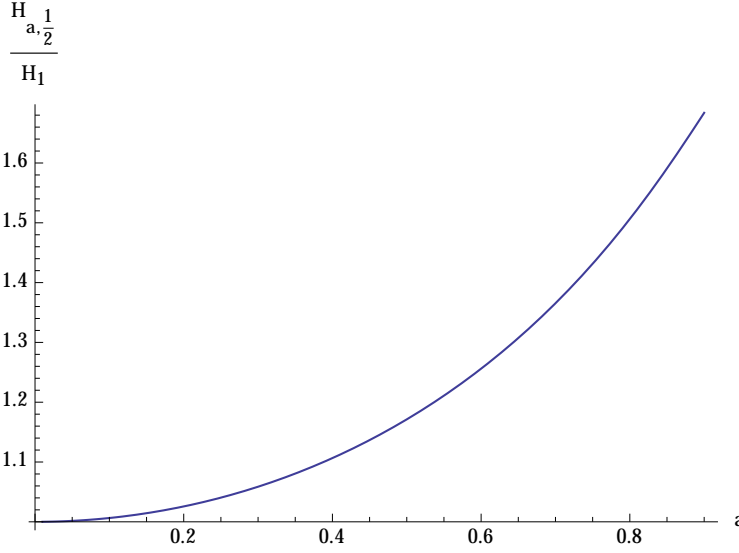
Figure D.2: The ratio $\frac{H_{a,\frac{1}{2}}(t)}{H_0}$ or the degradation of "utility" under second order effects.

$$= \frac{1}{\sqrt{\pi}} 2^{\frac{\alpha}{2}-2} \left(\frac{1}{\sigma^2 t}\right)^{-\frac{\alpha}{2}} \left(\Gamma\left(\frac{\alpha+1}{2}\right) \left(\sigma^\alpha t^{\alpha/2} \left(\frac{1}{\sigma^2 t}\right)^{\alpha/2}\right.\right.$$

$$\left.- \lambda\sigma\sqrt{t}\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(-\frac{\alpha}{2};\frac{1}{2};-\frac{t\mu^2}{2\sigma^2}\right) \qquad \text{(D.2)}$$

$$+ \frac{1}{\sqrt{2}\sigma}\mu\Gamma\left(\frac{\alpha}{2}+1\right) \left(\sigma^{\alpha+1} t^{\frac{\alpha}{2}+1} \left(\frac{1}{\sigma^2 t}\right)^{\frac{\alpha+1}{2}} + \sigma^\alpha t^{\frac{\alpha+1}{2}} \left(\frac{1}{\sigma^2 t}\right)^{\alpha/2}\right.$$

$$\left.\left.+ 2\lambda\sigma t\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(\frac{1-\alpha}{2};\frac{3}{2};-\frac{t\mu^2}{2\sigma^2}\right)\right)$$

We can see from D.2 that the more frequent sampling of the performance translates into worse utility. So what Benartzi and Thaler did was try to find the sampling period "myopia" that translates into the sampling frequency that causes the "premium" —the error being that they missed second order effects.

Now under variations of σ with stochatic effects, heuristically captured, the story changes: what if there is a very small probability that the variance gets multiplied by a large number, with the total variance remaining the same? The key here is that we are not even changing the variance at all: we are only shifting the distribution to the tails. We are here generously assuming that by the law of large numbers it was established that the "equity premium puzzle" was true and that stocks *really* outperformed bonds.

So we switch between two states, $(1+a)\sigma^2$ w.p. p and $(1-a)$ w.p. $(1-p)$.

Rewriting D.1

$$H_{a,p}(t) = \int_{-\infty}^{\infty} w_{\lambda,\alpha}(x) \left(p\,\phi_{\mu t,\sqrt{1+a}\,\sigma\sqrt{t}}(x) + (1-p)\,\phi_{\mu t,\sqrt{1-a}\,\sigma\sqrt{t}}(x)\right) \mathrm{d}x \qquad \text{(D.3)}$$

Result Conclusively, as can be seen in figures D.1 and D.2, second order effects cancel the statements made from "myopic" loss aversion. This doesn't mean that myopia doesn't have effects, rather that it cannot explain the "equity premium", not from the outside (i.e.

the distribution might have different returns, but from the inside, owing to the structure of the Kahneman-Tversky value function $v(x)$.

Comment We used the $(1 + a)$ heuristic largely for illustrative reasons; we could use a full distribution for σ^2 with similar results. For instance the gamma distribution with density $f(v) = \frac{v^{\gamma-1} e^{-\frac{\alpha v}{V}} \left(\frac{V}{\alpha}\right)^{-\gamma}}{\Gamma(\gamma)}$ with expectation V matching the variance used in the "equity premium" theory.

Rewriting D.3 under that form,

$$\int_{-\infty}^{\infty} \int_{0}^{\infty} w_{\lambda,\alpha}(x) \phi_{\mu t, \sqrt{v t}}(x) f(v) \, dv \, dx$$

Which has a closed form solution (though a bit lengthy for here).

True problem with Benartzi and Thaler Of course the problem has to do with fat tails and the convergence under LLN, which we treat separately.

Time preference under model error

Another example of the effect of the randomization of a parameter.

This author once watched with a great deal of horror one Laibson [103] at a conference in Columbia University present the idea that having one massage today to two tomorrow, but reversing in a year from now is irrational and we need to remedy it with some policy. (For a review of time discounting and intertemporal preferences, see [70], as economists temps to impart what seems to be a varying "discount rate" in a simplified model).[1]

Intuitively, what if I introduce the probability that the person offering the massage is full of balloney? It would clearly make me both prefer immediacy at almost any cost and conditionally on his being around at a future date, reverse the preference. This is what we will model next.

First, time discounting has to have a geometric form, so preference doesn't become negative: linear discounting of the form Ct, where C is a constant ant t is time into the future is ruled out: we need something like C^t or, to extract the rate, $(1 + k)^t$ which can be mathematically further simplified into an exponential, by taking it to the continuous time limit. Exponential discounting has the form e^{-kt}. Effectively, such a discounting method using a shallow model prevents "time inconsistency", so with $\delta < t$:

$$\lim_{t \to \infty} \frac{e^{-kt}}{e^{-k(t-\delta)}} = e^{-k\delta}$$

Now add another layer of stochasticity: the discount parameter, for which we use the symbol λ, is now stochastic.

So we now can only treat $H(t)$ as

$$H(t) = \int e^{-\lambda t} \phi(\lambda) \, d\lambda$$

---

[1] It came to my attention that [67] Farmer and Geanakoplos have applied a similar approach to Hyperbolic discounting

It is easy to prove the general case that under symmetric stochasticization of intensity $\Delta\lambda$ (that is, with probabilities $\frac{1}{2}$ around the center of the distribution) using the same technique we did in 3.1.1:

$$H'(t, \Delta\lambda) = \frac{1}{2}\left(e^{-(\lambda-\Delta\lambda)t} + e^{-(\lambda+\Delta\lambda)t}\right)$$

$$\frac{H'(t, \Delta\lambda)}{H'(t, 0)} = \frac{1}{2}e^{\lambda t}\left(e^{(-\Delta\lambda-\lambda)t} + e^{(\Delta\lambda-\lambda)t}\right) = \cosh(\Delta\lambda t)$$

Where cosh is the cosine hyperbolic function — which will converge to a certain value where intertemporal preferences are flat in the future.

Example: Gamma Distribution   Under the gamma distribution with support in $\mathbb{R}^+$, with parameters α and β, $\phi(\lambda) = \frac{\beta^{-\alpha}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)}$

we get:

$$H(t, \alpha, \beta) = \int_0^\infty e^{-\lambda t}\frac{\left(\beta^{-\alpha}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}\right)}{\Gamma(\alpha)}\, d\lambda = \beta^{-\alpha}\left(\frac{1}{\beta}+t\right)^{-\alpha}$$

so

$$\lim_{t\to\infty}\frac{H(t, \alpha, \beta)}{H(t-\delta, \alpha, \beta)} = 1$$

Meaning that preferences become flat in the future no matter how steep they are in the present, which explains the drop in discount rate in the economics literature.

Further, fudging the distribution and normalizing it, when

$$\phi(\lambda) = \frac{e^{-\frac{\lambda}{k}}}{k},$$

we get the *normatively obtained* so-called hyperbolic discounting:

$$H(t) = \frac{1}{1+kt}$$

which turns out to not be the empirical pathology that nerdy researchers have claimed it to be.

E | THE LARGE DEVIATION PRINCIPLE, IN BRIEF

ᴇᴛ ᴜs ᴅɪsᴄᴜss the Cramer bound with a rapid exposition of the surrounding literature. The idea behind the tall vs. rich outliers in 2.1 is that under some conditions, the tail probabilities decay exponentially. Such a property that is central in risk management –as we mentioned earlier, the catastrophe principle explains that for diversification to be effective, such exponential decay is necessary.

The large deviation principle helps us understand such a tail behavior.It also helps us figure out why things do not blow-up under thin-tailedness –but, more significantly, why they could under fat tails, or where the Cramèr condition is not satisfied.

Let M_N be the mean of a sequence of realizations (identically distributed) of N random variables. For large N, consider the tail probability:

$$\mathbb{P}(M_N > x) \approx e^{-NI(x)},$$

where $I(.)$ is the Cramer (or rate) function (Varadhan [185], Denbo and Zeitouni [45]). If we know the distribution of X, then, by Legendre transformation, $I(x) = \sup_{\theta>0} (\theta x - \lambda(\theta))$, where $\lambda(\theta) = \log \mathbb{E}\left(e^{\theta(X)}\right)$ is the cumulant generating function.

The behavior of the function $\theta(x)$ informs us on the contribution of a single event to the overall payoff. (It connects us to the *Cramer condition* which requires existence of exponential moments).

A special case for Bernoulli variables is the Chernoff Bound, which provides tight bounds for such a class of discrete variables with compact support.

SIMPLE CASE: CHERNOFF BOUND

A binary payoff is subjected to very tight bounds. Let $(X_i)_{1<i\leq n}$ be a sequence of independent Bernouilli trials taking values in $\{0,1\}$, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Consider the sum $S_n = \sum_{1<i\leq n} X_i$. with expectation $\mathbb{E}(S_n) = np = \mu$. Taking δ as a "distance from the mean", the Chernoff bounds gives:
For any $\delta > 0$

$$\mathbb{P}\left(S \geq (1+\delta)\mu\right) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

and for $0 < \delta \leq 1$

$$\mathbb{P}\left(S \geq (1+\delta)\mu\right) \leq 2e^{-\frac{\mu\delta^2}{3}}$$

Let us compute the probability of coin flips n of having 50% higher than the true mean, with p= $\frac{1}{2}$ and $\mu = \frac{n}{2}$: $\mathbb{P}\left(S \geq \left(\frac{3}{2}\right)\frac{n}{2}\right) \leq 2e^{-\frac{\mu\delta^2}{3}} = e^{-n/24}$, which for $n = 1000$ happens every 1 in 1.24×10^{18}.

Proof   The Markov bound gives: $\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$, but allows us to substitute X with a positive function $g(x)$, hence $\mathbb{P}(g(x) \geq g(c)) \leq \frac{\mathbb{E}(g(X))}{g(c)}$. We will use this property in what follows, with $g(X) = e^{\omega X}$.

Now consider $(1+\delta)$, with $\delta > 0$, as a "distance from the mean", hence, with $\omega > 0$,

$$\mathbb{P}\left(S_n \geq (1+\delta)\mu\right) = \mathbb{P}\left(e^{\omega S_n} \geq e^{\omega(1+\delta)\mu}\right) \leq e^{-\omega(1+\delta)\mu}\mathbb{E}(e^{\omega S_n}) \tag{E.1}$$

Now $E(e^{\omega S_n}) = \mathbb{E}(e^{\omega \Sigma(X_i)}) = \mathbb{E}(e^{\omega X_i})^n$, by independence of the stopping time, becomes $\left(\mathbb{E}(e^{\omega X})\right)^n$.

We have $\mathbb{E}(e^{\omega X}) = 1 - p + pe^{\omega}$. Since $1 + x \leq e^x$,

$$\mathbb{E}(e^{\omega S_n}) \leq e^{\mu(e^{\omega a}-1)}$$

Substituting in E.1, we get:

$$\mathbb{P}\left(e^{\omega S_n} \geq e^{\omega(1+\delta)\mu}\right) \leq e^{-\omega(1+\delta)\mu}e^{\mu(e^{\omega}-1)} \tag{E.2}$$

We tighten the bounds by playing with values of ω that minimize the right side. $\omega^* = \left\{\omega : \frac{\partial e^{\mu(e^{\omega}-1)-(\delta+1)\mu\omega}}{\partial \omega} = 0\right\}$ yields $\omega^* = \log(1 + \delta)$.

Which recovers the bound: $e^{\delta\mu}(\delta + 1)^{(-\delta-1)\mu}$.

F | MACHINE LEARNING CONSIDERATIONS

E HAVE learned from option trading that you can express any one-dimensional function as a weighted linear combination of call or put options –smoothed by adding time value to the option. An option becomes a building block. A payoff constructed via option is more precisely as follows $S = \sum_i^n \omega_i\, C(K_i, t_i)$, $i = 1, 2, \ldots, n$, where C is the call price (or, rather, valuation), ω is a weight K is the strike price, and t the time to expiration of the option. A European call C delivers $max(S - K, 0)$ at expiration t. [1]

Neural networks and nonlinear regression, the predecessors of machine learning, on the other hand, focused on the Heaviside step function, again smoothed to produce a sigmoid type "S" curve. A collection of different sigmoids would fit *in sample*.

f(x)

x
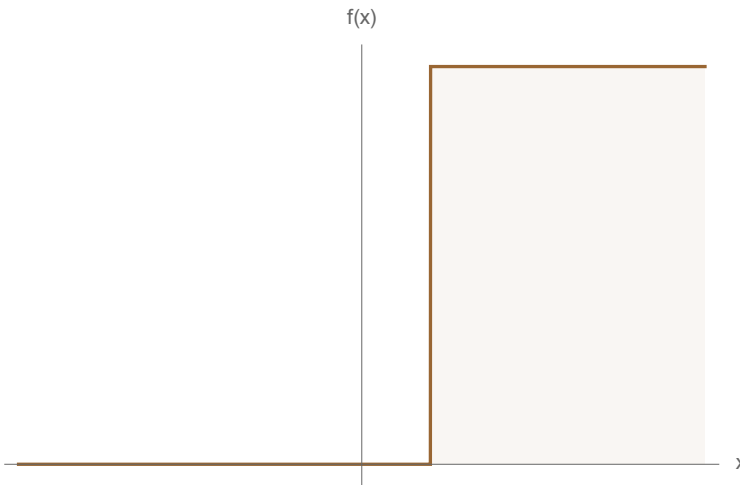
Figure F.1: *The heaviside θ function: note that it is the payoff of the "binary option" and can be decomposed as* $\lim_{\Delta K \to 0} \frac{C(K) - C(K + \Delta K)}{\delta K}$.

So this discussion is about ...fattailedness and how the different building blocks can accommodate them. Statistical machine learning switched to "ReLu" or "ramp" functions that act exactly like call options rather than an aggregation of "S" curves. Researchers then discovered that it allows better handling of out of sample tail events (since there are by definition no unexpected tail events in sample) owing to the latter's extrapolation properties.

What is a sigmoid? Consider a payoff function as shown in F.7 that can be expressed with formula $S : (-\infty, \infty) \to (0, 1)$, $S(x) = \frac{1}{2} \tanh\left(\frac{\kappa x}{\pi}\right) + \frac{1}{2}$, or, more precisely, a three parameter function $S_i : (-\infty, \infty) \to (0, a_1)$ $S_i(x) = \frac{a_i}{e^{(c_i - b_i x)} + 1}$. It can also be the cumulative normal

---

[1] This appears to be an independent discovery by traders of the universal approximation theorem, initially for sigmoid functions, which are discussed further down (Cybenko [39]).

distribution, $\mathcal{N}(\mu, \sigma)$ where σ controls the smoothness (it then becomes the Heaviside of Fig. F.7 at the limit of $\sigma \to 0$). The (bounded) sigmoid is the smoothing using parameters of the Heaviside function.
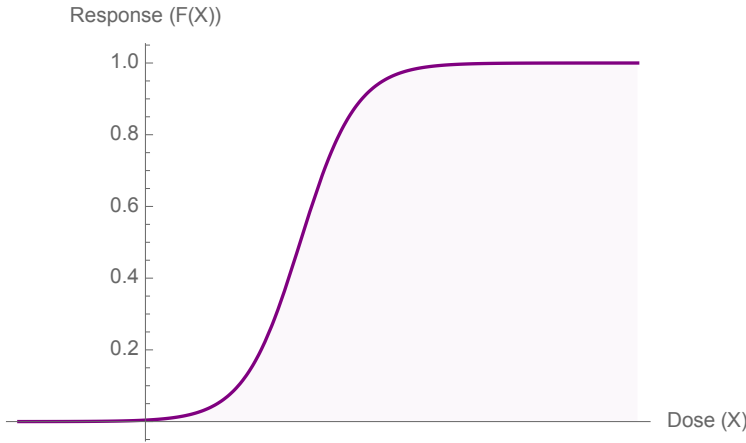
Figure F.2: *The sigmoid function ; note that it is bounded to both the left and right sides owing to saturation: it looks like a smoothed Heaviside θ.*

We can build composite "S" functions with n summands $\chi^n(x) = \sum_i^n \omega_i S_i(x)$ as in F.3. But:

Remark F.1
For $\chi^n(x) \in [0, \infty) \vee [-\infty, 0) \vee (-\infty, \infty)$, we must have $n \to \infty$.

We need an infinity of summands for an unbounded function. So wherever the "empirical distribution" will be maxed, the last observation will match the flat part of the sig. For the definition of an empirical distribution see 5.
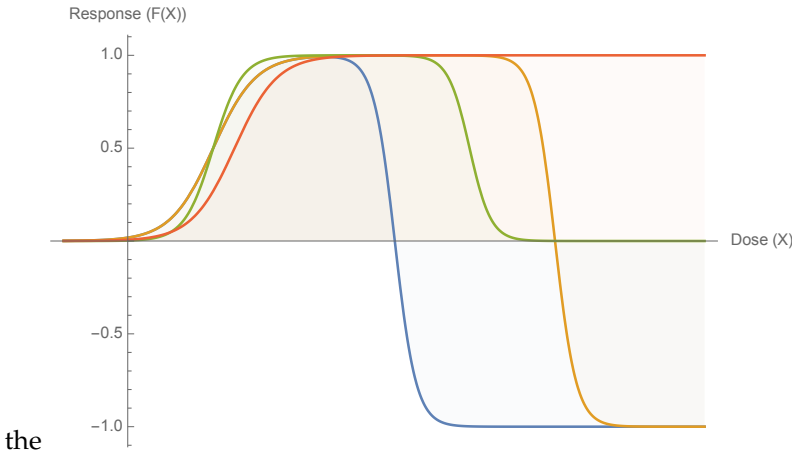
Figure F.3: *A sum of sigmoids will always be bounded, so one needs an infinite sum to replicate an "open" payoff, one that is not subjected to saturation.*

the

Now let us consider option payoffs. Fig.F.4 shows the payoff of a regular option at expiration –the definition of which which matches a Rectifier Linear Unit (ReLu) in machine learning. Now Fig. F.5 shows the following function: consider a function $\rho : (-\infty, \infty) \to [k, \infty)$, with $K \in \mathbb{R}$:

$$\rho(x, K, p) = k + \frac{\log\left(e^{p(x-K)} + 1\right)}{p} \tag{F.1}$$

We can sum the function as $\sum_i = 1^n \rho(x, K_i, p_i)$ to fit a nonlinear function, which in fact replicates what we did with call options –the parameters p_i allow to smooth time value.
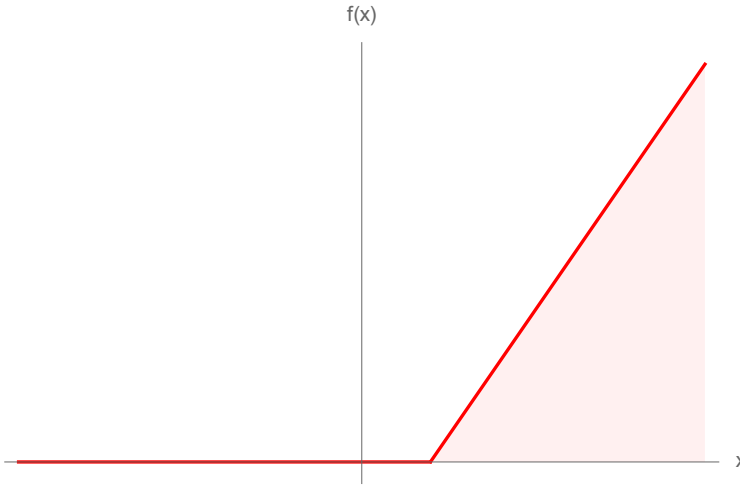
Figure F.4: *An option payoff at expiration, open on the right.*
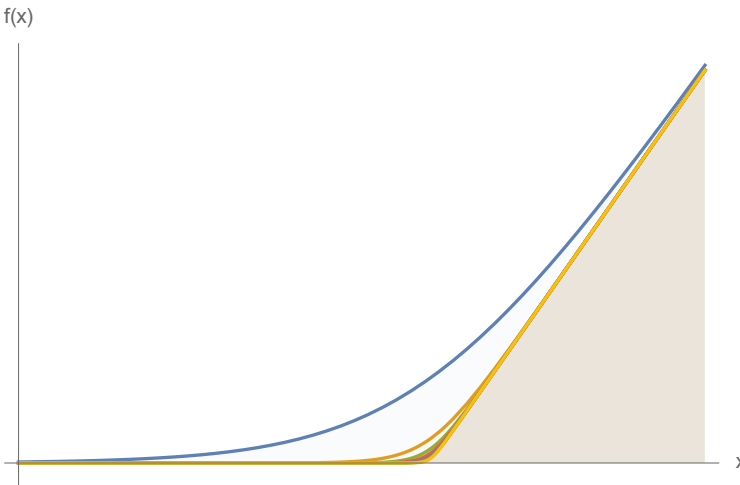
Figure F.5: *ρ function, from Eq. F.1 , with $k = 0$. We calibrate and smooth the payoff with different values of p.*
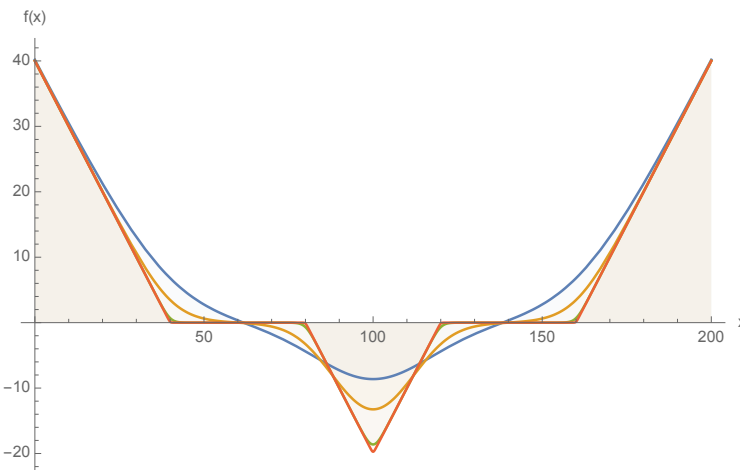
Figure F.6: *A butterfly (built via a sum of options/ReLu, not sigmoids), with open tails on both sides and flipping first and second derivatives. This example is particularly potent as it has no verbalistic correspondence but can be understood by option traders and machine learning.*

F.0.1 Calibration via angles

From figure F.6 we can see that, in the equation, $S = \sum_i^n \omega_i \, C(K_i, t_i)$, the ω_i corresponds to the arc tangent of the angle made –if positive (as illustrated in figure F.7), or the negative of the arctan of the supplementary angle.
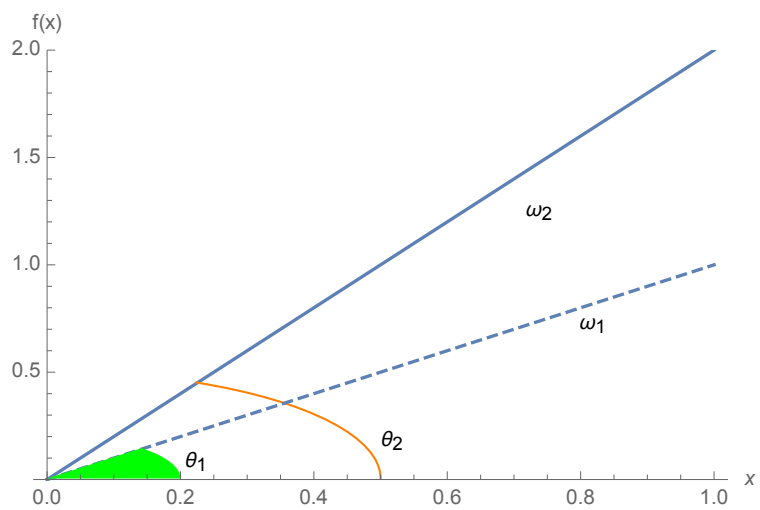
Figure F.7: *How* $\omega = \arctan\theta$. *By fitting angles we can translate a nonlinear function into its option summation.*

Summary

We can express all nonlinear univariate functions using a weighted sum of call options of different strikes, which in machine learning applications maps to the tails better than a sum of sigmoids (themselves a net of a long and a short options of neighboring strikes). We can get the weights implicitly using the angles of the functions relative to Cartesian coordinates.

Part II

THE LAW OF MEDIUM NUMBERS

6 | LIMIT DISTRIBUTIONS, A CONSOLIDATION*,†

N THIS expository chapter we proceed to consolidate the literature on limit distributions seen from our purpose, with some shortcuts where indicated. First we show the intuition behind the Central Limit Theorem and illustrate how it varies preasymptotically across distributions. Then we discuss the law of large numbers as applied to higher moments. A more formal approach is presented in the next chapter.

6.1 CENTRAL LIMIT IN ACTION
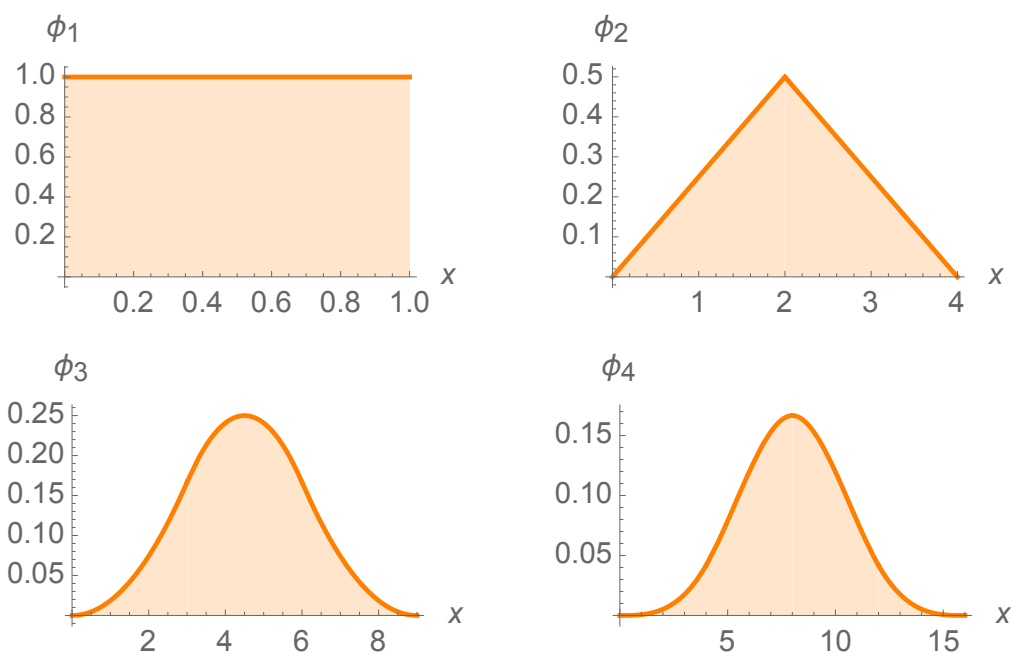
Figure 6.1: *The fastest CLT: the Uniform becomes Gaussian in a few steps. We have, successively, 1, 2, 3, and 4 summands. With 3 summands we see a well formed bell shape.*

The simplified version of the generalized central limit theorem (GCLT) (formulated by Paul Lévy) is as follows:
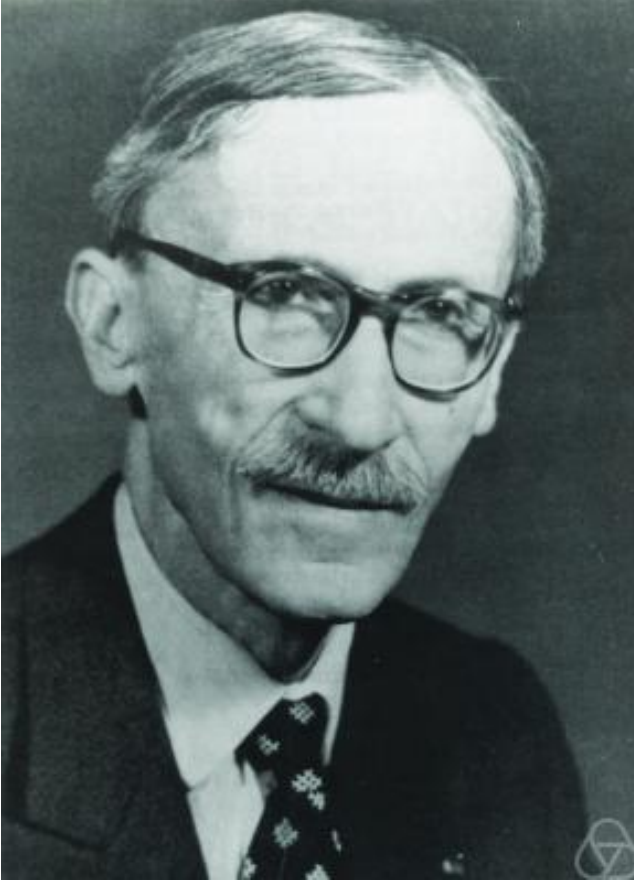
Figure 6.2: *Paul Lévy, 1886-1971, formulated the generalized central limit theorem.*

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. Consider their sum S_n. We have

$$\frac{Sn - a_n}{b_n} \xrightarrow{D} X_s, \tag{6.1}$$

where X_s follows a stable distribution $\mathcal{S}$, a_n and b_n are norming constants, and $\xrightarrow{D}$ denotes convergence in distribution (the distribution of X as $n \to \infty$). The properties of $\mathcal{S}$ will be more properly defined and explored in the next chapter. Take it for now that a random variable X_s follows a stable (or α-stable) distribution, symbolically $X_s \sim S(\alpha_s, \beta, \mu, \sigma)$, if its characteristic function $\chi(t) = \mathbb{E}(e^{itX_s})$ is of the form:

$$\chi(t) = e^{\left(i\mu t - |t\sigma|_s^\alpha \left(1 - i\beta \tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(t)\right)\right)} \text{ when } \alpha_s \neq 1. \tag{6.2}$$

The constraints are $-1 \geq \beta \geq 1$ and $0 < \alpha_s \leq 2$.[1]

The designation stable distribution implies that the distribution (or class) is stable under summation: you sum up random variables following any the various distributions that are members of the class $\mathfrak{S}$ explained next chapter (actually the same distribution with different parametrizations of the characteristic function), and you stay within the same distribution. Intuitively, $\chi(t)^n$ is the same form as $\chi(t)$, with $\mu \to n\mu$, and $\sigma \to n^{\frac{1}{\alpha}}\sigma$. The well known distributions in the class (or some people call it a "basin") are: the Gaussian,

---

[1] We will try to use $\alpha_s \in (0, 2]$ to denote the exponent of the limiting and Platonic stable distribution and $\alpha_p \in (0, \infty)$ the corresponding Paretian (preasymptotic) equivalen but only in situations where there could be some ambiguity. Plain α should be understood in context.

the Cauchy and the Lévy with $\alpha = 2, 1$, and $\frac{1}{2}$, respectively. Other distributions have no closed form density.[2]

We note that if X has a finite variance, X_s will be Gaussian. But note that X_s is a limiting construct as $n \to \infty$ and there are many, many complication with "how fast" we get there. Let us consider 4 cases that illustrate both the idea of CLT and the speed of it.

6.1.1 Fast convergence: the uniform dist.

Consider a uniform distribution –the simplest of all. If its support is in $[0, 1]$, it will simply have a density of $\phi(x_1) = 1$ for $0 \le x_1 \le 1$ and integrates to 1. Now add another variable, x_2, identically distributed and independent. The sum $x_1 + x_2$ immediately changed in shape! Look at $\phi_2(.)$, the density of the sum in Fig. 6.1. It is now a triangle. Add one variable and now consider the density ϕ_3 of the distribution of $X_1 + X_2 + X_3$. It is already almost bell shaped, with $n = 3$ summands.

The uniform sum distribution

$$\phi_n(x) = \sum_{k=0}^{n} (-1)^k \begin{pmatrix} n \\ k \end{pmatrix} \left(\frac{x - L}{H - L} - k \right)^{n-1} \text{sgn} \left(\frac{x - L}{H - L} - k \right) \text{ for } nL \le x \le nH$$

6.1.2 Semi-slow convergence: the exponential

Let us consider a sum of exponential random variables.

We have for initial density

$$\phi_1(x) = \lambda e^{-\lambda x}, x \ge 0$$

and for n summands[3]

$$\phi_n(x) = \left(\frac{1}{\lambda} \right)^{-n} \frac{x^{n-1} e^{-\lambda x}}{\Gamma(n)}$$

We have, replacing x by n/λ (and later in the illustrations in Fig. 6.3 $\lambda = 1$),

$$\frac{\left(\frac{1}{\lambda} \right)^{-n} x^{n-1} e^{\lambda(-x)}}{\Gamma(n)} \xrightarrow[n \to \infty]{} \frac{\lambda e^{-\frac{\lambda^2 \left(x - \frac{n}{\lambda} \right)^2}{2n}}}{\sqrt{2\pi}\sqrt{n}},$$

which is the density of the normal distribution with mean $\frac{n}{\lambda}$ and variance $\frac{n}{\lambda^2}$.

We can see how we get more slowly to the Gaussian, as shown in Fig. 6.3, mostly on account of its skewness. Getting to the Gaussian requires symmetry.

---

2 Actually, there are ways to use special functions; for instance one discovered accidentally by the author: for the Stable $\mathcal{S}$ with standard parameters $\alpha = \frac{3}{2}, \beta = 1, \mu = 0, \sigma = 1$, $PDF(x) =$
$$-\frac{\sqrt[3]{2}e^{\frac{x^3}{27}} \left(\sqrt[3]{3}x \text{Ai} \left(\frac{x^2}{3\,2^{2/3}\sqrt[3]{3}} \right) + 3\sqrt[3]{2}\text{Ai}' \left(\frac{x^2}{3\,2^{2/3}\sqrt[3]{3}} \right) \right)}{3\,3^{2/3}}$$ used further down in the example on the limit distribution for Pareto sums.

3 We derive the density of sums either by convolving, easy in this case, or as we will see with the Pareto, via characteristic functions.
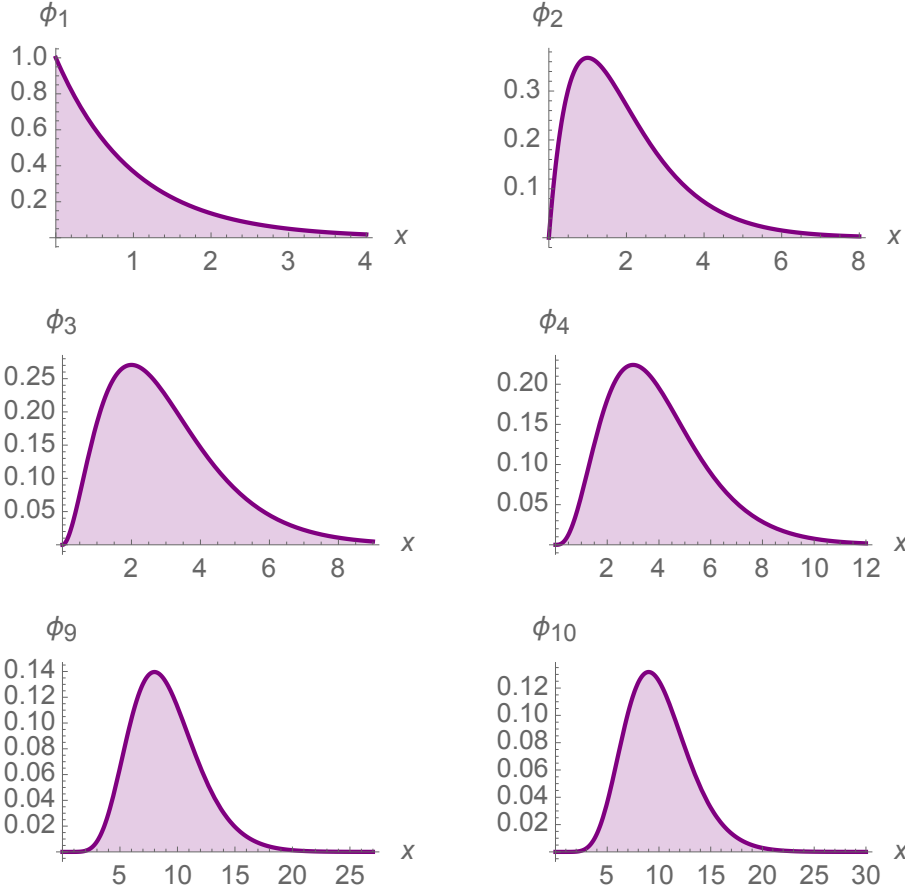
Figure 6.3: *The exponential distribution,φ indexed by the number of summands. Slower but good enough.*

6.1.3 The slow Pareto

Consider the simplest Pareto distribution on $[1, \infty)$:

$$\phi_1(x) = 2x^{-3}$$

and inverting the characteristic function,

$$\phi_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)(2E_3(-it))^n \, dt, \, x \geq n$$

Where $E_{(.)}(.)$ is the exponential integral $E_n(z) = \int_1^{\infty} \frac{dt e^{t(-z)}}{t^n}$. Clearly, the integration is done numerically (so far nobody has managed to pull out the distribution of a Pareto sum). It can be exponentially slow (up to 24 hours for $n = 50$ vs. 45 seconds for $n = 2$), so we have used Monte Carlo simulations for Figs. 6.1.1.

Recall from Eq. 6.1 that the convergence requires norming constants a_n and b_n. From Uchaikin and Zolotarev [182], we have (narrowing the situation for $1 < \alpha_p \leq 2$):

$$P(X > x) = cx^{-\alpha_p}$$
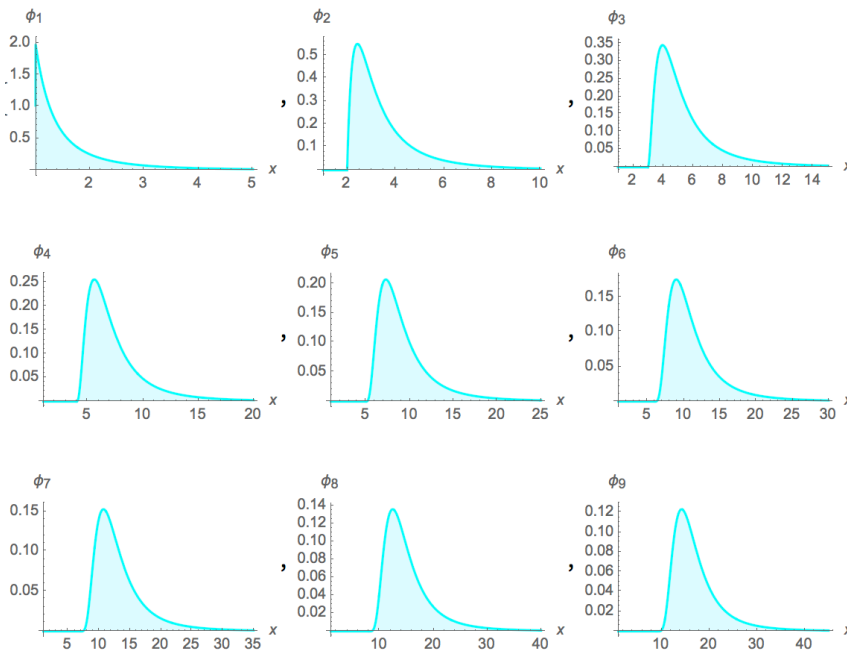
Figure 6.4: *The Pareto distribution. Doesn't want to lose its skewness, although in this case it should converge to the Gaussian... eventually.*

as $x \to \infty$ (assume here that c is a constant we will present more formally the "slowly varying function" in the next chapter, and

$$P(X < x) = d|x|^{-\alpha_p}$$

as $x \to \infty$. The norming constants become $a_n = n\ \mathbb{E}(X)$ for $\alpha_p > 1$ (for other cases, consult [182] as these are not likely to occur in practice), and

$$b_n = \begin{cases} \pi n^{1/\alpha_p} \left(2\sin\left(\frac{\pi \alpha_p}{2}\right) \Gamma(\alpha_p) \right)^{-\frac{1}{\alpha_p}} (c+d)^{1/\alpha_p} & \text{for } 1 < \alpha_p < 2 \\ \sqrt{c+d}\sqrt{n\log(n)} & \text{for } \alpha_p = 2 \end{cases} \tag{6.3}$$

And the symmetry parameter $\beta = \frac{c-d}{c+d}$. Clearly, the situation where the Paretian parameter α_p is greater than 2 leads to the Gaussian.

6.1.4 The half-cubic Pareto and its basin of convergence

Of interest is the case of $\alpha = \frac{3}{2}$. Unlike the situations where as in Fig. 6.1.1, the distribution ends up slowly being symmetric. But, as we will cover in the next chapter, it is erroneous to conflate its properties with those of a stable. It is, in a sense, more fat-tailed.
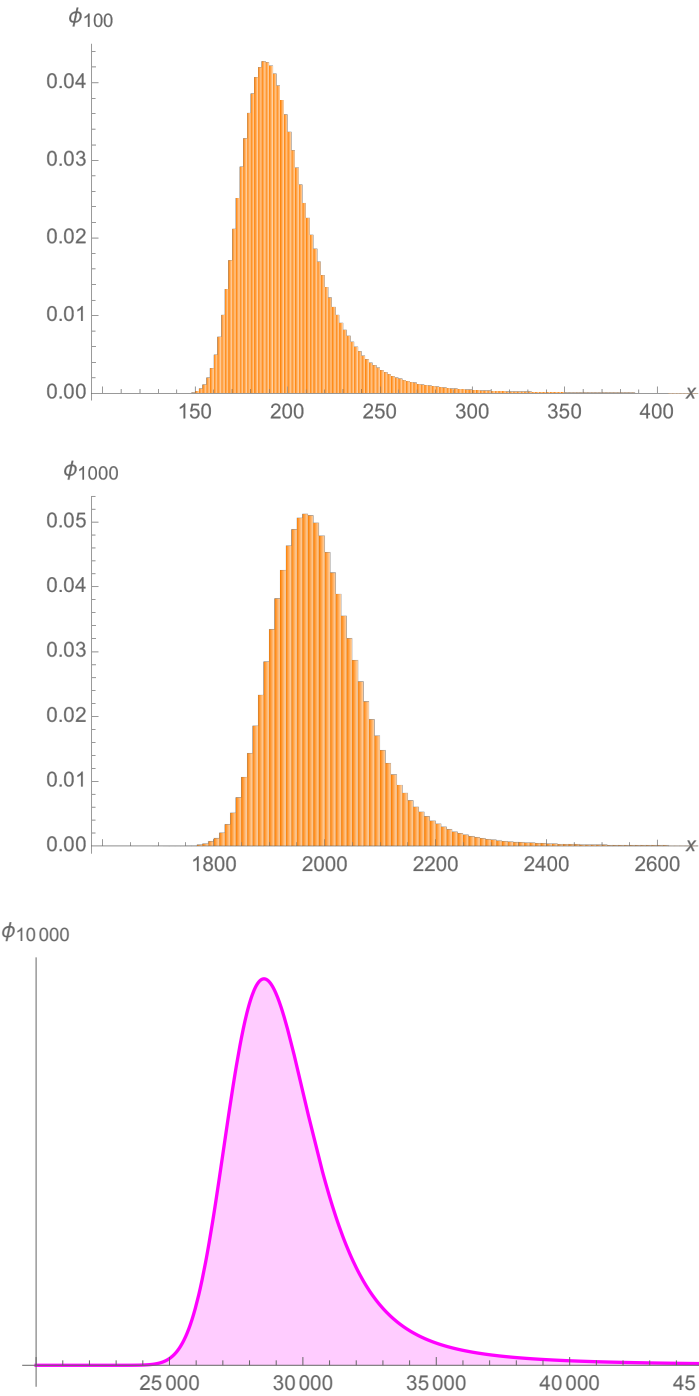
Figure 6.5: *The Pareto distribution, ϕ_{100} and ϕ_{1000}, not much improvement towards Gaussianity, but an alpha = 2 will eventually get you there if you are patient and have a long, very long, life.*

Figure 6.6: *The half-cubic Pareto distribution never becomes symmetric. Here $n = 10^4$*

6.2 CUMULANTS AND CONVERGENCE

Since the Gaussian (as a basin of convergence) has skewness of 0 and (raw) kurtosis of 3, we can heuristically examine the convergence of these moments to establish the speed of the workings under CLT.

Definition 6.1 (Excess p-cumulants)
Let $\chi(\omega)$ be characteristic functionof a given distribution, n the number of summands (for inde-
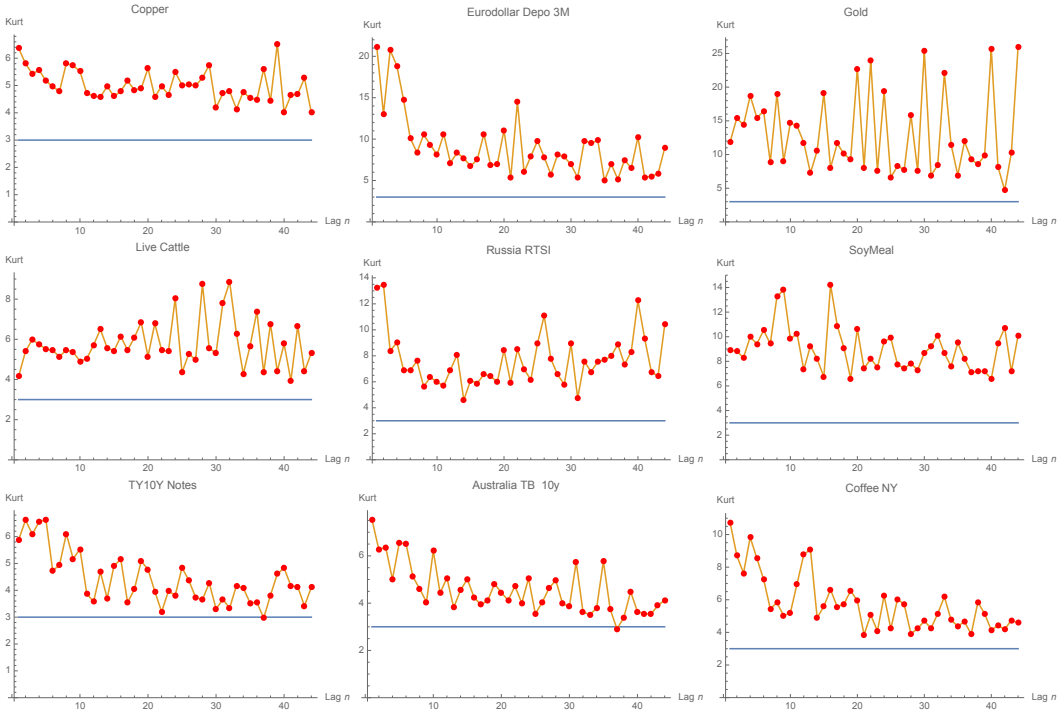
Figure 6.7: *Behavior of 4th moment under aggregation for a few financial securities deemed to converge to the Gaussian but in fact do not converge (backup data for [159]). There is no conceivable way to claim convergence to Gaussian for data sampled at a lower frequency.*

pendent random variables), p the order of the moment. We define the ratio of cumulants for the corresponding pth moment:

$$K_k^p \triangleq \frac{(-i)^p \partial^p \log(\chi(\omega)^n)}{\left(-\partial^2 \log(\chi(\omega)^n)\right)^2}$$

K(n) is a metric of excess pth moment over that of a Gaussian, p > 2; in other words, $K_n^4 = 0$ denotes Gaussianity for n independent summands.

Remark 6.1

We note that

$$\lim_{n \to \infty} K_N^p = 0$$

for all probability distributions outside the Power Law class.

We also note that $\lim_{p \to \infty} K_n^p$ is finite for the thin-tailed class. In other words, we face a clear-cut basin of converging vs. diverging moments.

For distributions outside the Power Law basin, $\forall p \in \mathbb{N}_{>2}$, K_n^p decays at a rate N^{p-2}.

A sketch of the proof can be done using the stable distribution as the limiting basin and the nonderivability at order p greater than its tail index, using Eq. 7.4.

Table 6.1 for N-summed p cumulants. It describes $K(n)$

We would expect a drop at a rate $\frac{1}{N^2}$ for stochastic volatility (gamma variance wlog). However, figure 8.9 shows the drop does not take place at any such speed. Visibly we are not in the basin. As seen in [159] there is an absence of convergence of kurtosis under summation across economic variables.

Table 6.1: *Table of Normalized Cumulants For Thin Tailed Distributions Speed of Convergence for N Independent Summands*

-

Distr.	Poisson (λ)	Expon. (λ)	Gamma (a,b)	Symmetric 2-state vol (σ_1, σ_2)	Γ-Variance (a, b)
$\kappa(2)$	1	1	1	1	1
$K(3)$	$\frac{1}{n\lambda}$	$\frac{2\lambda}{n}$	$\frac{2}{a\,b\,n}$	0	0
$K(4)$	$\frac{1}{n\lambda^2}$	$\frac{3!\lambda^2}{n}$	$\frac{3!}{a^2\,b^2\,n}$	$\frac{3(1-p)p}{n} \times \frac{\left(\sigma_1^2 - \sigma_2^2\right)^2}{\left(p\sigma_1^2 - (p-1)\sigma_2^2\right)^3}$	$\frac{3b}{n}$

6.3 THE LAW OF LARGE NUMBERS
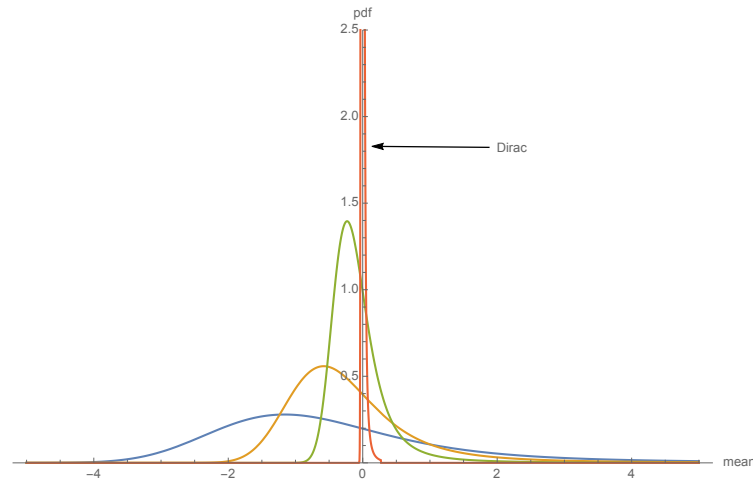
Figure 6.8: *The law of large numbers show a tightening distribution around the mean leading to degeneracy converging to a Dirac stick at the exact mean.*

By the weak law of large numbers, a sum of i.i.d. random variables $X_1, \ldots, X_n$ with finite mean m, that is $\mathbb{E}(X) < +\infty$, then $\frac{1}{n}\sum X_i$ converges to m in probability, as n $\to$ +∞. Or, for any $\epsilon > 0$ $\lim_{n\to+\infty} \mathbb{P}\left(|\overline{X}_n - m| > \epsilon\right) = 0$.

By standard results, we can observe the law of large numbers at work for the stable distribution, illustrated in Fig. 6.8:

$$\lim_{n\to+\infty} \chi\left(\frac{t}{n}\right)^n = e^{i\mu t}, \, 1 < \alpha_s \leq 2 \tag{6.4}$$

which is the characteristic functionof a Dirac delta at μ, a degenerate distribution, since the Fourier transform (here parametrized to be the inverse of the characteristic function) is:

$$\frac{1}{\sqrt{2\pi}} \mathcal{F}_t\left[e^{i\mu t}\right](x) = \delta(\mu + x). \tag{6.5}$$

Further, we can observe the "real-time" operation for all $1 < n < +\infty$ in the following way.

6.4 THE LAW OF LARGE NUMBERS FOR HIGHER MOMENTS

Table 6.2: *Fourth noncentral moment at daily, 10-day, and 66-day windows for the random variables*

	K(1)	K(10)	K(66)	Max Quartic	Years
Australian Dollar/USD	6.3	3.8	2.9	0.12	22.
Australia TB 10y	7.5	6.2	3.5	0.08	25.
Australia TB 3y	7.5	5.4	4.2	0.06	21.
BeanOil	5.5	7.0	4.9	0.11	47.
Bonds 30Y	5.6	4.7	3.9	0.02	32.
Bovespa	24.9	5.0	2.3	0.27	16.
British Pound/USD	6.9	7.4	5.3	0.05	38.
CAC40	6.5	4.7	3.6	0.05	20.
Canadian Dollar	7.4	4.1	3.9	0.06	38.
Cocoa NY	4.9	4.0	5.2	0.04	47.
Coffee NY	10.7	5.2	5.3	0.13	37.
Copper	6.4	5.5	4.5	0.05	48.
Corn	9.4	8.0	5.0	0.18	49.
Crude Oil	29.0	4.7	5.1	0.79	26.
CT	7.8	4.8	3.7	0.25	48.
DAX	8.0	6.5	3.7	0.20	18.
Euro Bund	4.9	3.2	3.3	0.06	18.
Euro Currency/DEM previously	5.5	3.8	2.8	0.06	38.
Eurodollar Depo 1M	41.5	28.0	6.0	0.31	19.
Eurodollar Depo 3M	21.1	8.1	7.0	0.25	28.
FTSE	15.2	27.4	6.5	0.54	25.
Gold	11.9	14.5	16.6	0.04	35.
Heating Oil	20.0	4.1	4.4	0.74	31.
Hogs	4.5	4.6	4.8	0.05	43.
Jakarta Stock Index	40.5	6.2	4.2	0.19	16.
Japanese Gov Bonds	17.2	16.9	4.3	0.48	24.
Live Cattle	4.2	4.9	5.6	0.04	44.
Nasdaq Index	11.4	9.3	5.0	0.13	21.
Natural Gas	6.0	3.9	3.8	0.06	19.
Nikkei	52.6	4.0	2.9	0.72	23.
Notes 5Y	5.1	3.2	2.5	0.06	21.
Russia RTSI	13.3	6.0	7.3	0.13	17.
Short Sterling	851.8	93.0	3.0	0.75	17.
Silver	160.3	22.6	10.2	0.94	46.
Smallcap	6.1	5.7	6.8	0.06	17.
SoyBeans	7.1	8.8	6.7	0.17	47.
SoyMeal	8.9	9.8	8.5	0.09	48.
Sp500	38.2	7.7	5.1	0.79	56.
Sugar #11	9.4	6.4	3.8	0.30	48.
SwissFranc	5.1	3.8	2.6	0.05	38.
TY10Y Notes	5.9	5.5	4.9	0.10	27.
Wheat	5.6	6.0	6.9	0.02	49.
Yen/USD	9.7	6.1	2.5	0.27	38.
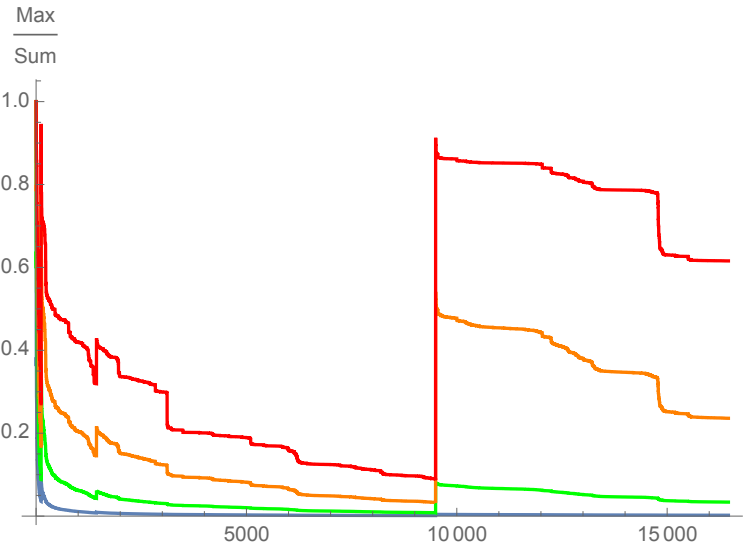
Figure 6.9: *Cumulative moments $p = 1, 2, 3, 4$.*
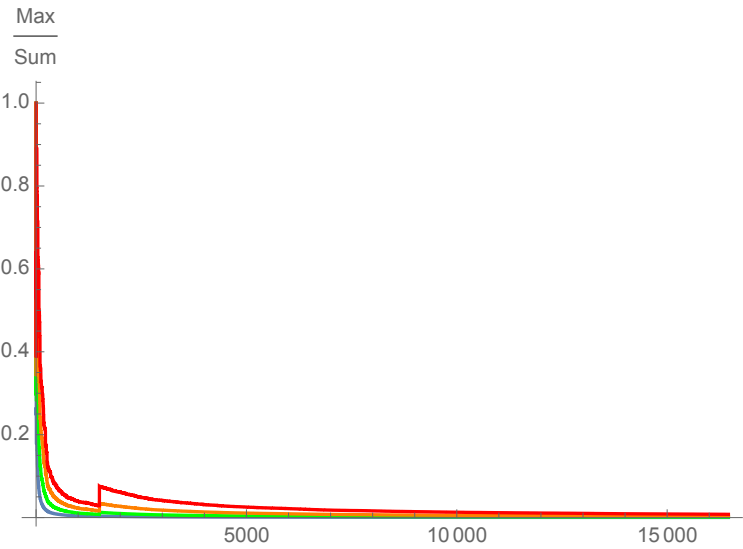
Figure 6.10: *Gaussian Control.*
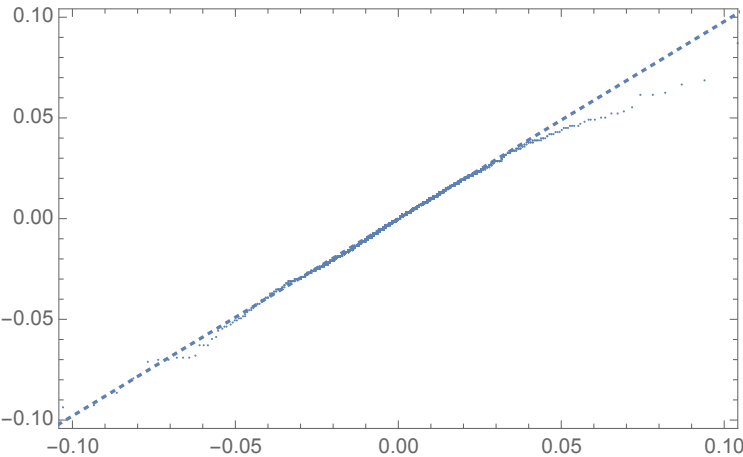
Figure 6.11: *QQ Plot of Student T: left tail fits, not the right tail.*

6.5 MEAN DEVIATION FOR A STABLE DISTRIBUTIONS

Let us prepare a result for the next chapter using the norm L^1 for situations of finite mean but infinite variance.[4] Clearly we have no way to measure the compression of the distribution around the mean within the norm L^2.

The error of a sum in the norm L^1 is as follows. Since $\text{sgn}(x) = 2\theta(x) - 1$:

$$\chi^{\text{sgn}(x)}(t) = \frac{2i}{t} \tag{6.6}$$

Let $\chi^d(.)$ be the characteristic function of any nondegenerate distribution. Convoluting $\chi^{\text{sgn}(x)} * (\chi^d)^n$, we obtain the characteristic functionfor the positive variations for n independent summands

$$\chi^m = \int_{-\infty}^{\infty} \chi^{\text{sgn}(x)}(t) \chi^d(u-t)^n \, dt.$$

In our case of mean absolute deviation being twice that of the positive values of X:

$$\chi(|S_n|) = (2i) \int_{-\infty}^{\infty} \frac{\chi(t-u)^n}{t} \, du$$

which is the Hilbert transform of χ when $\int$ is taken in the p.v. sense (Pinelis, 2015)[133]. In our situation, given that all independents summands are copies from the same distribution, we can replace the product $\chi(t)^n$ with $\chi_s(t)$ which is the same characteristic functionwith $\sigma_s = n^{1/\alpha}\sigma$, β remains the same:

$$\mathbb{E}(|X|) = 2i \frac{\partial}{\partial u} \text{p.v.} \int_{-\infty}^{\infty} \frac{\chi_s(t-u)}{t} \, dt|_{t=0} \tag{6.7}$$

Now, [133] the Hilbert transform,

$$= \frac{2}{\pi i} \int_0^{\infty-} \chi_s(u+t) - \chi_s(u-t) \, dt$$

which can be rewritten as

$$= -i \frac{\partial}{\partial u} \left(1 + \chi_s(u) + \frac{1}{\pi i} \int_0^{\infty-} \chi_s(u+t) - \chi_s(u-t) - \chi_s(t) + \chi_s(-t) \frac{dt}{t} \right). \tag{6.8}$$

Deriving first inside the integral and using a change of variable, $z = \log(t)$,

$$\mathbb{E}|X||_{(\tilde{\alpha},\beta,\sigma_s,0)} =$$
$$\int_{-\infty}^{\infty} 2i\alpha e^{-(\sigma_s e^z)^\alpha - z} \left(\sigma_s e^z \right)^\alpha \left(\beta \tan \left(\frac{\pi\alpha}{2} \right) \sin \left(\beta \tan \left(\frac{\pi\alpha}{2} \right) (\sigma_s e^z)^\alpha \right) \right.$$
$$\left. + \cos \left(\beta \tan \left(\frac{\pi\alpha}{2} \right) (\sigma_s e^z)^\alpha \right) \right) \, dz$$

which then integrates nicely to:

$$\mathbb{E}|X||_{(\tilde{\alpha},\beta,\sigma_s,0)} = \frac{\sigma_s}{2\pi} \Gamma \left(\frac{\alpha-1}{\alpha} \right) \left(\left(1 + i\beta \tan \left(\frac{\pi\alpha}{2} \right) \right)^{1/\alpha} + \left(1 - i\beta \tan \left(\frac{\pi\alpha}{2} \right) \right)^{1/\alpha} \right). \tag{6.9}$$

---

4 We say, again by convention, *infinite* for the situation where the random variable, say X^2 (or the variance of any random variable), is one-tailed –bounded on one side– and undefined in situations where the variable is two-tailed, e.g. the infamous Cauchy.

7

HOW MUCH DATA DO YOU NEED? AN OPERATIONAL METRIC FOR FAT-TAILEDNESS‡

I̶N THIS (RESEARCH) CHAPTER we discuss the laws of medium number-sLaw of medium numbers. We present an operational metric for univariate unimodal probability distributions with finite first moment, in $[0, 1]$ where 0 is maximally thin-tailed (Gaussian) and 1 is maximally fat-tailed. It is based on "how much data one needs to make meaningful statements about a given dataset?"

Applications: Among others, it

- helps assess the sample size n needed for statistical significance outside the Gaussian,

- helps measure the speed of convergence to the Gaussian (or stable basin),

- allows practical comparisons across classes of fat-tailed distributions,

- allows the assessment of the number of securities needed in portfolio construction to achieve a certain level of stability from diversification,

- helps understand some inconsistent attributes of the lognormal, pending on the parametrization of its variance.

The literature is rich for what concerns asymptotic behavior, but there is a large void for finite values of n, those needed for operational purposes.

Background [1] : Conventional measures of fat-tailedness, namely 1) the tail index for the Power Law class, and 2) Kurtosis for finite moment distributions fail to apply to some distributions, and do not allow comparisons across classes and parametrization, that is between power laws outside the Levy-Stable basin, or power laws to distributions in other classes, or power laws for different number of summands. How can one compare a sum of 100 Student T distributed random variables with 3 degrees of freedom to one in a Levy-Stable or a Lognormal class? How can one compare a sum of 100 Student T with 3 degrees of freedom to a single Student T with 2 degrees of freedom?

We propose an operational and heuristic metric that allows us to compare n-summed independent variables under all distributions with finite first moment. The method is

---

based on the rate of convergence of the law of large numbers for finite sums, n-summands specifically.

We get either explicit expressions or simulation results and bounds for the lognormal, exponential, Pareto, and the Student T distributions in their various calibrations –in addition to the general Pearson classes.
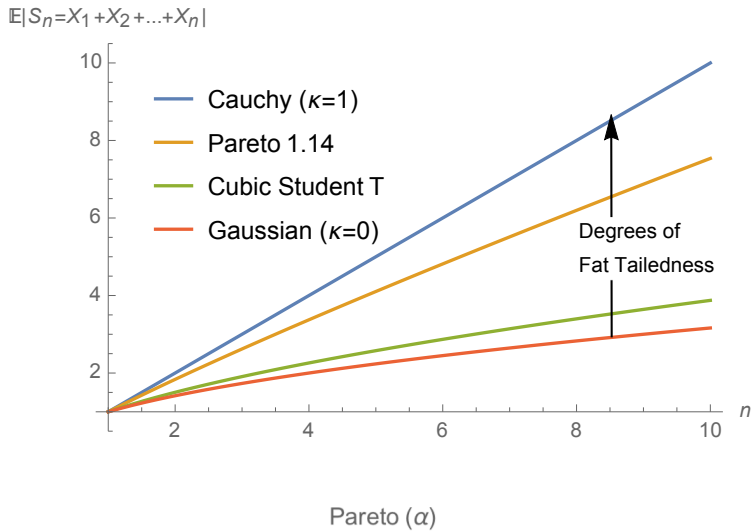
$\mathbb{E}|S_n = X_1 + X_2 + ... + X_n|$

Figure 7.1: *The intuition of what κ is measuring: how the mean deviation of the sum of identical copies of a r.v. $S_n = X_1 + X_2 + ... X_n$ grows as the sample increases and how we can compare preasymptotically distributions from different classes.*
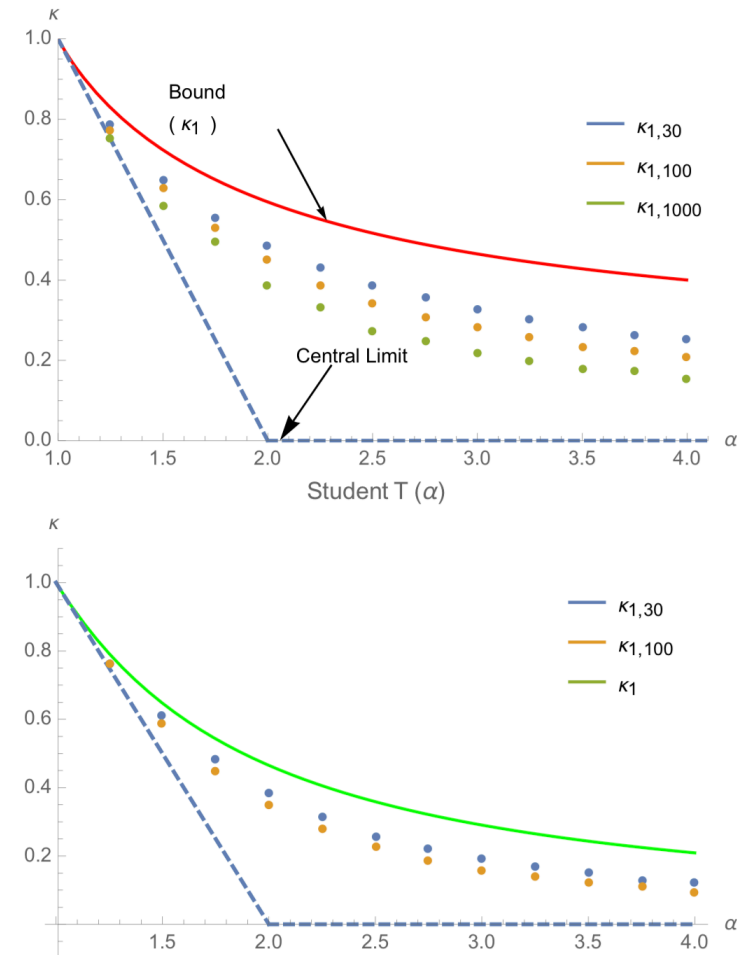
Figure 7.2: *Watching the effect of the Generalized Central Limit Theorem: Pareto and Student T Distribution, in the $\mathfrak{P}$ class, with α exponent, κ converge to $2 - (\mathbb{1}_{\alpha<2}\alpha + \mathbb{1}_{\alpha\geq2}2)$, or the Stable $\mathfrak{S}$ class. We observe how slow the convergence, even after 1000 summands. This discounts Mandelbrot's assertion that an infinite variance Pareto can be subsumed into a stable distribution.*

7.1 INTRODUCTION AND DEFINITIONS

How can one compare a Pareto distribution with tail α = 2.1 that is, with finite variance, to a Gaussian? Asymptotically, these distributions in the regular variation class with finite second moment, under summation, become Gaussian, but pre-asymptotically, we have no standard way of comparing them given that metrics that depend on higher moments, such as kurtosis, cannot be of help. Nor can we easily compare an infinite variance Pareto distribution to its limiting α-Stable distribution (when both have the same tail index or tail exponent). Likewise, how can one compare the "fat-tailedness" of, say a Student T with 3 degrees of freedom to that of a Levy-Stable with tail exponent of 1.95? Both distributions have a finite mean; of the two, only the first has a finite variance but, for a small number of summands, behaves more "fat-tailed" according to some operational criteria.

Criterion for "fat-tailedness"   There are various ways to "define" Fat Tails and rank distributions according to each definition. In the narrow class of distributions having all moments finite, it is the kurtosis, which allows simple comparisons and measure departures from the Gaussian, which is used as a norm. For the Power Law class, it can be the tail exponent . One can also use extremal values, taking the probability of exceeding a maximum value, adjusted by the scale (as practiced in extreme value theory). For operational uses, practitioners' fat-tailedness is a degree of concentration, such as "how much of the statistical properties will be attributable to a single observation?", or, appropriately adjusted by the scale (or the mean dispersion), "how much is the total wealth of a country in the hands of the richest individual?"

  Here we use the following criterion for our purpose, which maps to the measure of concentration in the past paragraph: "How much will additional data (under such a probability distribution) help increase the stability of the observed mean". The purpose is not entirely statistical: it can equally mean: "How much will adding an additional security into my portfolio allocation (i.e., keeping the total constant) increase its stability?"

  Our metric differs from the asymptotic measures (particularly ones used in extreme value theory) in the fact that it is fundamentally preasymptotic.

  Real life, and real world realizations, are outside the asymptote.

What does the metric do?   The metric we propose, κ does the following:

- Allows comparison of n-summed variables of different distributions for a given number of summands , or same distribution for different n, and assess the preasymptotic properties of a given distributions.
- Provides a measure of the distance from the limiting distribution, namely the Lévy α-Stable basin (of which the Gaussian is a special case).
- For statistical inference, allows assessing the "speed" of the law of large numbers, expressed in change of the mean absolute error around the average thanks to the increase of sample size n.
- Allows comparative assessment of the "fat-tailedness" of two different univariate distributions, when both have finite first moment.
- Allows us to know ahead of time how many runs we need for a Monte Carlo simulation.

The state of statistical inference   The last point, the "speed", appears to have been ignored (see earlier comments in Chapter 2 about the 9,400 pages of the *Encyclopedia of*

Statistical Science [101]). It is very rare to find a discussion about how long it takes to reach the asymptote, or how to deal with n summands that are large but perhaps not sufficiently so for the so-called "normal approximation".

To repeat our motto, "statistics is never standard". This metric aims at showing *how standard is standard*, and measure the exact departure *from the standard* from the standpoint of statistical significance.

7.2 THE METRIC

Table 7.1: *Kappa for 2 summands, κ_1.*

Distribution	κ_1
Student T (α)	$2 - \dfrac{2\log(2)}{2\log\left(\dfrac{2^{2-\alpha}\Gamma\left(\alpha-\frac{1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)^2}\right)+\log(\pi)}$
Exponential/Gamma	$2 - \dfrac{\log(2)}{2\log(2)-1} \approx .21$
Pareto (α)	$2 - \dfrac{\log(2)}{\log\left((\alpha-1)^{2-\alpha}\alpha^{\alpha-1}\int_0^{\frac{2}{\alpha-1}} -2\alpha^2(y+2)^{-2\alpha-1}\left(\frac{2}{\alpha-1}-y\right)\left(B_{\frac{1}{y+2}}(-\alpha,1-\alpha)-B_{\frac{y+1}{y+2}}(-\alpha,1-\alpha)\right)dy\right)}$ [a]
Normal (μ,σ) with switching variance $\sigma^2 a$ w.p p. [b]	$2 - \dfrac{\log(2)}{\log\left(\dfrac{\sqrt{2}\left(\sqrt{\frac{ap}{p-1}+\sigma^2}+p\left(-2\sqrt{\frac{ap}{p-1}+\sigma^2}+p\left(\sqrt{\frac{ap}{p-1}+\sigma^2}-\sqrt{2a\left(\frac{1}{p-1}+2\right)+4\sigma^2}+\sqrt{a+\sigma^2}\right)+\sqrt{2a\left(\frac{1}{p-1}+2\right)+4\sigma^2}\right)\right)}{p\sqrt{a+\sigma^2}-(p-1)\sqrt{\frac{ap}{p-1}+\sigma^2}}\right)}$
Lognormal (μ,σ)	$\approx 2 - \dfrac{\log(2)}{\log\left(\dfrac{2\,\mathrm{erf}\left(\dfrac{\sqrt{\log\left(\frac{1}{2}\left(e^{\sigma^2}+1\right)\right)}}{2\sqrt{2}}\right)}{\mathrm{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)}\right)}.$

a $B.(.,.)$ is the incomplete Beta function: $B_z(a,b) = \int_0^z t^{a-1}(1-t)^{b-1}dt$; erf(.) is the error function erf(z) = $\frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$.

b See comments and derivations in the appendix for switching both variance and mean as it can produce negative values for kappa.

Definition 7.1 (the κ metric)
Let $X_1,\ldots,X_n$ be i.i.d. random variables with finite mean, that is $\mathbb{E}(X) < +\infty$. Let $S_n = X_1 + X_2 + \ldots + X_n$ be a partial sum. Let $\mathbb{M}(n) = \mathbb{E}(|S_n - \mathbb{E}(S_n)|)$ be the expected mean absolute deviation from the mean for n summands. Define the "rate" of convergence for n additional summands starting with n_0:

Table 7.2: *Summary of main results*

Distribution	κ_n
Exponential/Gamma	Explicit
Lognormal (μ, σ)	No explicit κ_n but explicit lower and higher bounds (low or high σ or n). Approximated with Pearson IV for σ in between.
Pareto (α) (Constant)	Explicit for κ_2 (lower bound for all α).
Student T(α) (slowly varying function)	Explicit for κ_1 , $\alpha = 3$.
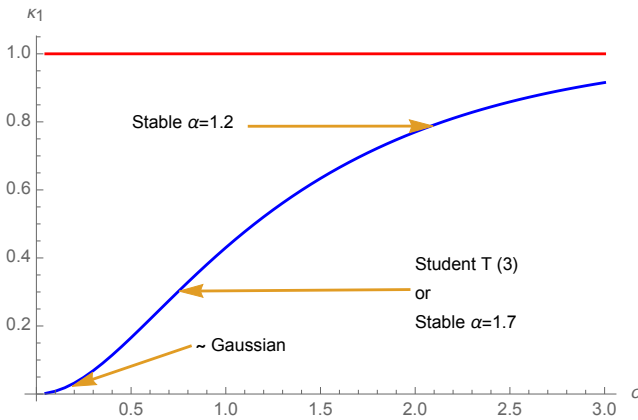
Figure 7.3: *The lognormal distribution behaves like a Gaussian for low values of σ, but becomes rapidly equivalent to a power law. This illustrates why, operationally, the debate on whether the distribution of wealth was lognormal (Gibrat) or Pareto (Zipf) doesn't carry much operational significance.*

$$\kappa_{n_0,n} = \min \left\{ \kappa_{n_0,n} : \frac{\mathbb{M}(n)}{\mathbb{M}(n_0)} = \left(\frac{n}{n_0} \right)^{\frac{1}{2-\kappa_{n_0,n}}} , n_0 = 1, 2, ... \right\},$$

$n > n_0 \geq 1$, *hence*

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\log \left(\frac{\mathbb{M}(n)}{\mathbb{M}(n_0)} \right)}. \tag{7.1}$$

Further, for the baseline values $n = n_0 + 1$, we use the shorthand κ_{n_0}.

We can also decompose $\kappa(n_0, n)$ in term of "local" intermediate ones similar to "local" interest rates, under the constraint.

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\sum_{i=0}^{n} \frac{\log(i+1) - \log(i)}{2 - \kappa(i, i+1)}}. \tag{7.2}$$

Table 7.3: *Comparing Pareto to Student T (Same tail exponent α)*

α	Pareto κ_1	Pareto $\kappa_{1,30}$	Pareto $\kappa_{1,100}$	Student κ_1	Student $\kappa_{1,30}$	Student $\kappa_{1,100}$
1.25	0.829	0.787	0.771	0.792	0.765	0.756
1.5	0.724	0.65	0.631	0.647	0.609	0.587
1.75	0.65	0.556	0.53	0.543	0.483	0.451
2.	0.594	0.484	0.449	0.465	0.387	0.352
2.25	0.551	0.431	0.388	0.406	0.316	0.282
2.5	0.517	0.386	0.341	0.359	0.256	0.227
2.75	0.488	0.356	0.307	0.321	0.224	0.189
3.	0.465	0.3246	0.281	0.29	0.191	0.159
3.25	0.445	0.305	0.258	0.265	0.167	0.138
3.5	0.428	0.284	0.235	0.243	0.149	0.121
3.75	0.413	0.263	0.222	0.225	0.13	0.10
4.	0.4	0.2532	0.211	0.209	0.126	0.093

Use of Mean Deviation   Note that we use for measure of dispersion around the mean the mean absolute deviation, to stay in norm L^1 in the absence of finite variance –actually, even in the presence of finite variance, under Power Law regimes, distributions deliver an unstable and uninformative second moment. Mean deviation proves far more robust there. (Mean absolute deviation can be shown to be more "efficient" except in the narrow case of kurtosis equals 3 (the Gaussian), see a longer discussion in [168]; for other advantages, see [127].)

7.3 STABLE BASIN OF CONVERGENCE AS BENCHMARK

Definition 7.2 (the class $\mathfrak{P}$)
The $\mathfrak{P}$ class of power laws (regular variation) is defined for r.v. X as follows:

$$\mathfrak{P} = \left\{ X : \mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \right\} \tag{7.3}$$

where $\sim$ means that the limit of the ratio or rhs to lhs goes to 1 as $x \to \infty$. $L : [x_{\min}, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x \to +\infty} \frac{L(kx)}{L(x)} = 1$ for any $k > 0$. The constant $\alpha > 0$.

Next we define the domain of attraction of the sum of identically distributed variables, in our case with identical parameters.

Definition 7.3
(stable $\mathfrak{S}$ class) A random variable X follows a stable (or α-stable) distribution, symbolically $X \sim S(\tilde{\alpha}, \beta, \mu, \sigma)$, if its characteristic function $\chi(t) = \mathbb{E}(e^{itX})$ is of the form:

$$\chi(t) = \begin{cases} e^{\left(i\mu t - |t\sigma|^{\tilde{\alpha}} \left(1 - i\beta \tan\left(\frac{\pi\tilde{\alpha}}{2}\right) sgn(t)\right)\right)} & \tilde{\alpha} \neq 1 \\[2em] e^{it\left(\frac{2\beta\sigma \log(\sigma)}{\pi} + \mu\right) - |t\sigma|\left(1 + \frac{2i\beta sgn(t) \log(|t\sigma|)}{\pi}\right)} & \tilde{\alpha} = 1 \end{cases}, \tag{7.4}$$

Next, we define the corresponding stable $\tilde{\alpha}$:

$$\tilde{\alpha} \triangleq \begin{cases} \alpha \; \mathbb{1}_{\alpha<2} + 2 \; \mathbb{1}_{\alpha\geq 2} & \text{if X is in } \mathfrak{P} \\ 2 & \text{otherwise.} \end{cases} \tag{7.5}$$

Further discussions of the class $\mathfrak{S}$ are as follows.

7.3.1 Equivalence for stable distributions

For all n_0 and $n \geq 1$ in the Stable $\mathfrak{S}$ class with $\tilde{\alpha} \geq 1$:

$$\kappa_{(n_0,n)} = 2 - \tilde{\alpha},$$

simply from the property that

$$\mathbb{M}(n) = n^{\frac{1}{\alpha}} \mathbb{M}(1) \tag{7.6}$$

This, simply shows that $\kappa_{n_0,n} = 0$ for the Gaussian.

The problem of the preasymptotics for n summands reduces to:

- What is the property of the distribution for $n_0 = 1$ (or starting from a standard, off-the shelf distribution)?
- What is the property of the distribution for n_0 summands?
- How does $\kappa_n \to 2 - \tilde{\alpha}$ and at what rate?

7.3.2 Practical significance for sample sufficiency

> **Confidence intervals**: As a simple heuristic, the higher κ, the more disproportionally insufficient the confidence interval. Any value of κ above .15 effectively indicates a high degree of unreliability of the "normal approximation". One can immediately doubt the results of numerous research papers in fat-tailed domains.

Computations of the sort done Table 7.2 for instance allows us to compare various distributions under various parametriazation. (comparing various Pareto distributions to symmetric Student T and, of course the Gaussian which has a flat kappa of 0)

As we mentioned in the introduction, required sample size for statistical inference is driven by n, the number of summands. Yet the law of large numbers is often invoked in erroneous conditions; we need a rigorous sample size metric.

Many papers, when discussing financial matters, say [72] use finite variance as a binary classification for fat tailedness: power laws with a tail exponent greater than 2 are therefore classified as part of the "Gaussian basin", hence allowing the use of variance and other such metrics for financial applications. A much more natural boundary is finiteness of expectation for financial applications [160]. Our metric can thus be useful as follows:

Let $X_{g,1}, X_{g,2}, \ldots, X_{g,n_g}$ be a sequence of Gaussian variables with mean μ and scale σ. Let $X_{v,1}, X_{v,2}, \ldots, X_{v,n_v}$ be a sequence of some other variables scaled to be of the same $\mathbb{M}(1)$, namely $\mathbb{M}^v(1) = \mathbb{M}^g(1) = \sqrt{\frac{2}{\pi}}\sigma$. We would be looking for values of n_v corresponding to a given n_g.

> κ_n is indicative of both the rate of convergence under the law of large number, and for $\kappa_n \to 0$, for rate of convergence of summands to the Gaussian under the central limit, as illustrated in Figure 7.2.

$$n_{\min} = \inf \left\{ n_v : \mathbb{E}\left(\left| \sum_{i=1}^{n_v} \frac{X_{v,i} - m_p}{n_v} \right| \right) \leq \mathbb{E}\left(\left| \sum_{i=1}^{n_g} \frac{X_{g,i} - m_g}{n_g} \right| \right), n_v > 0 \right\} \qquad (7.7)$$

which can be computed using $\kappa_n = 0$ for the Gaussian and backing our from κ_n for the target distribution with the simple approximation:

$$n_v = n_g^{-\frac{1}{\kappa_{1,n_g}-1}} \approx n_g^{-\frac{1}{\kappa_1-1}}, n_g > 1 \qquad (7.8)$$

The approximation is owed to the slowness of convergence. So for example, a Student T with 3 degrees of freedom ($\alpha = 3$) requires 120 observations to get the same drop in variance from averaging (hence confidence level) as the Gaussian with 30, that is 4 times as much. The one-tailed Pareto with the same tail exponent $\alpha = 3$ requires 543 observations to match a Gaussian sample of 30, 4.5 times more than the Student, which shows 1) finiteness of variance is not an indication of fat tailedness (in our statistical sense), 2) neither are tail exponent s good indicators 3) how the symmetric Student and the Pareto distribution are not equivalent because of the "bell-shapedness" of the Student (from the slowly varying function) that dampens variations in the center of the distribution.

We can also elicit quite counterintuitive results. From Eq. 7.8, the "Pareto 80/20" in the popular mind, which maps to a tail exponent around $\alpha \approx 1.14$, requires $> 10^9$ more observations than the Gaussian.

7.4 TECHNICAL CONSEQUENCES

7.4.1 Some oddities with asymmetric distributions

The stable distribution, when skewed, has the same κ index as a symmetric one (in other words, κ is invariant to the β parameter in Eq. 7.4, which conserves under summation). But a one-tailed simple Pareto distribution is fatter tailed (for our purpose here) than an equivalent symmetric one.

This is relevant because the stable is never really observed in practice and used as some limiting mathematical object, while the Pareto is more commonly seen. The point is not well grasped in the literature. Consider the following use of the substitution of a stable for a Pareto. In Uchaikin and Zolotarev [182]:

> Mandelbrot called attention to the fact that the use of the extremal stable distributions (corresponding to $\beta = 1$) to describe empirical principles was preferable to the use of the Zipf-Pareto distributions for a number of reasons. It can be seen from many publications, both theoretical and applied, that Mandelbrot's ideas receive more and more wide recognition of experts. In this way, the hope arises to confirm empirically established principles in the framework of mathematical models and, at the same time, to clear up the mechanism of the formation of these principles.

These are not the same animals, even for large number of summands.

7.4.2 Rate of convergence of a student T distribution to the Gaussian Basin

We show in the appendix –thanks to the explicit derivation of κ for the sum of students with $\alpha = 3$, the "cubic" commonly noticed in finance –that the rate of convergence of κ to 0 under summation is $\frac{1}{\log(n)}$. This (and the semi-closed form for the density of an n-summed cubic Student) complements the result in Bouchaud and Potters [20] (see also [149]), which is as follows. Their approach is to separate the "Gaussian zone" where the density is approximated by that of a Gaussian, and a "Power Law zone" in the tails which retains the original distribution with Power Law decline. The "crossover" between the two moves right and left of the center at a rate of $\sqrt{n \log(n)}$ standard deviations) which is excruciatingly slow. Indeed, one can note that more summands fall at the center of the distribution, and fewer outside of it, hence the speed of convergence according to the central limit theorem will differ according to whether the density concerns the center or the tails.

Further investigations would concern the convergence of the Pareto to a Levy-Stable, which so far we only got numerically.

7.4.3 The lognormal is neither thin nor fat tailed

Naively, as we can see in Figure 7.2, at low values of the parameter σ, the lognormal behaves like a Gaussian, and, at high σ, it appears to have the behavior of a Cauchy of sorts (a one-tailed Cauchy, rather a stable distribution with $\alpha = 1$, $\beta = 1$), as κ gets closer and closer to 1. This gives us an idea about some aspects of the debates as to whether some variable is Pareto or lognormally distributed, such as, say, the debates about wealth [109], [40], [41]. Indeed, such debates can be irrelevant to the real world. As P. Cirillo [32] observed, many cases of Paretianity are effectively lognormal situations with high variance; the practical statistical consequences, however, are smaller than imagined.

7.4.4 Can kappa be negative?

Just as kurtosis for a mixed Gaussian (i.e., with stochastic mean, rather than stochastic volatility) can dip below 3 (or become "negative" when one uses the convention of measuring kurtosis as excess over the Gaussian by adding 3 to the measure), the kappa metric can become negative when kurtosis is "negative". These situations require bimodality (i.e., a switching process between means under fixed variance, with modes far apart in terms of standard deviation). They do not appear to occur with unimodal distributions.

Details and derivations are presented in the appendix.

7.5 CONCLUSION AND CONSEQUENCES

To summarize, while the limit theorems (the law of large numbers and the central limit) are concerned with the behavior as $n \to +\infty$, we are interested in finite and exact n both small and large.

We may draw a few operational consequences:
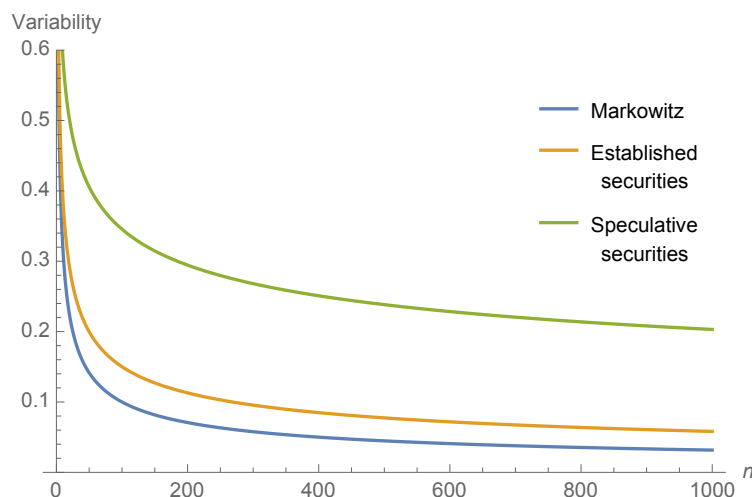
Figure 7.4: *In short, why the 1/n heuristic works: it takes many, many more securities to get the same risk reduction as via portfolio allocation according to Markowitz . We assume to simplify that the securities are independent, which they are not, something that compounds the effect.*

7.5.1 Portfolio pseudo-stabilization

Our method can also naturally and immediately apply to portfolio construction and the effect of diversification since adding a security to a portfolio has the same "stabilizing" effect as adding an additional observation for the purpose of statistical significance. "How much data do you need?" translates into "How many securities do you need?". Clearly, the Markowicz allocation method in modern finance[112] (which seems to not be used by Markowitz himself for his own portfolio [120]) applies only for κ near 0; people use convex heuristics, otherwise they will underestimate tail risks and "blow up" the way the famed portfolio-theory oriented hedge fund Long Term Management did in 1998 [167] [176].)

We mentioned earlier that a Pareto distribution close to the "80/20" requires up to 10^9 more observations than a Gaussian; consider that the risk of a portfolio under such a distribution would be underestimated by at least 8 orders of magnitudes if one uses modern portfolio criteria. Following such a reasoning, one simply needs broader portfolios.

It has also been noted that there is practically no financial security that is not fatter tailed than the Gaussian, from the simple criterion of kurtosis [159], meaning Markowitz portfolio allocation is *never* the best solution. It happens that agents wisely apply a noisy approximation to the $\frac{1}{n}$ heuristic which has been classified as one of those biases by behavioral scientists but has in fact been debunked as false (a false bias is one in which, while the observed phenomenon is there, it does not constitute a "bias" in the bad sense of the word; rather it is the researcher who is mistaken owing to using the wrong tools instead of the decision-maker). This tendency to "overdiversify" has been deemed a departure from optimal investment behavior by Benartzi and Thaler [13], explained in [12] "when faced with n options, divide assets evenly across the options. We have dubbed this heuristic the "1/n rule."" However, broadening one's diversification is effectively as least as optimal as standard allocation(see critique by Windcliff and Boyle [188] and [46]). In short, an equally weighted portfolio outperforms the SP500 across a broad range range of metrics. But even the latter two papers didn't conceive of the full effect and properties of fat tails, which we can see here with some precision. Fig. 7.5 shows the effect for securities compared to Markowitz.

This false bias is one in many examples of policy makers "nudging" people into the wrong rationality [167] and driving them to increase their portfolio risk many folds.

A few more comments on financial portfolio risks. The SP500 has a κ of around .2, but one needs to take into account that it is itself a basket of $n = 500$ securities, albeit unweighted and consisting of correlated members, overweighing stable stocks. Single stocks have kappas between .3 and .7, meaning a policy of "overdiversification" is a must.

Likewise the metric gives us some guidance in the treatment of data for forecasting, by establishing sample sufficiency, to state such matters as how many years of data do we need before stating whether climate conditions "have changed", see [108].

7.5.2 Other aspects of statistical inference

So far we considered only univariate distributions. For higher dimensions, a potential area of investigation is an equivalent approach to the multivariate distribution of extreme fat tailed variables, the sampling of which is not captured by the Marchenko-Pastur (or Wishhart) distributions. As in our situation, adding variables doesn't easily remove noise from random matrices.

7.5.3 Final comment

As we keep saying, "statistics is never standard"; however there are heuristics methods to figure out where and by how much we depart from the standard.

APPENDIX

We show here some derivations

7.5.4 Cubic Student T (Gaussian Basin)

The Student T with 3 degrees of freedom is of special interest in the literature owing to its prevalence in finance [72]. It is often mistakenly approximated to be Gaussian owing to the finiteness of its variance. Asymptotically, we end up with a Gaussian, but this doesn't tell us anything about the rate of convergence. Mandelbrot and Taleb [111] remarks that the cubic acts more like a powerlaw in the distribution of the extremes, which we will elaborate here thanks to an explicit PDF for the sum.

Let X be a random variable distributed with density $p(x)$:

$$p(x) = \frac{6\sqrt{3}}{\pi \left(x^2 + 3\right)^2} \; , x \in (-\infty, \infty) \tag{7.9}$$

Proposition 7.1
Let Y be a sum of $X_1, \ldots, X_n$, n identical copies of X. Let $\mathbb{M}(n)$ be the mean absolute deviation from the mean for n summands. The "rate" of convergence $\kappa_{1,n} = \left\{ \kappa : \frac{\mathbb{M}(n)}{\mathbb{M}(1)} = n^{\frac{1}{2-\kappa}} \right\}$ is:

$$\kappa_{1,n} = 2 - \frac{\log(n)}{\log\left(e^n n^{-n} \Gamma(n+1, n) - 1\right)} \tag{7.10}$$

where $\Gamma(.,.)$ is the incomplete gamma function $\Gamma(a, z) = \int_z^\infty dt\, t^{a-1} e^{-t}$.

Since the mean deviation $\mathbb{M}(n)$:

$$\mathbb{M}(n) = \begin{cases} \frac{2\sqrt{3}}{\pi} & \text{for } n = 1 \\ \frac{2\sqrt{3}}{\pi} \left(e^n n^{-n} \Gamma(n+1,n) - 1\right) & \text{for } n > 1 \end{cases} \tag{7.11}$$

The derivations are as follows. For the pdf and the MAD we followed different routes. We have the characteristic function for n summands:

$$\varphi(\omega) = (1 + \sqrt{3}|\omega|)^n \, e^{-n\sqrt{3}|\omega|}$$

The pdf of Y is given by:

$$p(y) = \frac{1}{\pi} \int_0^\infty (1 + \sqrt{3}\,\omega)^n \, e^{-n\sqrt{3}\omega} \cos(\omega y) \, d\omega$$

After arduous integration we get the result in 7.11. Further, since the following result does not appear to be found in the literature, we have a side useful result: the PDF of Y can be written as

$$p(y) = \frac{e^{n - \frac{iy}{\sqrt{3}}} \left(e^{\frac{2iy}{\sqrt{3}}} E_{-n}\left(n + \frac{iy}{\sqrt{3}}\right) + E_{-n}\left(n - \frac{iy}{\sqrt{3}}\right)\right)}{2\sqrt{3}\pi} \tag{7.12}$$

where $E_{(.)}(.)$ is the exponential integral $E_n z = \int_1^\infty \frac{e^{t(-z)}}{t^n} dt$.

Note the following identities (from the updating of Abramowitz and Stegun) [51]

$$n^{-n-1}\Gamma(n+1,n) = E_{-n}(n) = e^{-n} \frac{(n-1)!}{n^n} \sum_{m=0}^n \frac{n^m}{m!}$$

As to the asymptotics, we have the following result (proposed by Michail Loulakis): Re-expressing Eq. 7.11:

$$\mathbb{M}(n) = \frac{2\sqrt{3}n!}{\pi n^n} \sum_{m=0}^{n-1} \frac{n^m}{m!}$$

Further,

$$e^{-n} \sum_{m=0}^{n-1} \frac{n^m}{m!} = \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right)$$

(From the behavior of the sum of Poisson variables as they converge to a Gaussian by the central limit theorem: $e^{-n} \sum_{m=0}^{n-1} \frac{n^m}{m!} = \mathbb{P}(X_n < n)$ where X_n is a Poisson random variable with parameter n. Since the sum of n independent Poisson random variables with parameter 1 is Poisson with parameter n, the Central Limit Theorem says the probability distribution of $Z_n = (X_n - n)/\sqrt{n}$ approaches a standard normal distribution. Thus $\mathbb{P}(X_n < n) = \mathbb{P}(Z_n < 0) \to 1/2$ as $n \to \infty$.[2] For another approach, see [121] for proof that $1 + \frac{n}{1!} + \frac{n^2}{2!} + \cdots + \frac{n^{n-1}}{(n-1)!} \sim \frac{e^n}{2}$.)

Using the property that $\lim_{n \to \infty} \frac{n!\exp(n)}{n^n \sqrt{n}} = \sqrt{2\pi}$, we get the following exact asymptotics:

$$\lim_{n \to \infty} \log(n)\kappa_{1,n} = \frac{\pi^2}{4}$$

---

2 Robert Israel on Math Stack Exchange

thus κ goes to 0 (i.e, the average becomes Gaussian) at speed $\frac{1}{\log(n)}$, which is excruciatingly slow. In other words, even with 10^6 summands, the behavior cannot be summarized as that of a Gaussian, an intuition often expressed by B. Mandelbrot [111].

7.5.5 Lognormal Sums

From the behavior of its cumulants for n summands, we can observe that a sum behaves likes a Gaussian when σ is low, and as a lognormal when σ is high –and in both cases we know explicitly κ_n.

The lognormal (parametrized with μ and σ) doesn't have an explicit characteristic function. But we can get cumulants K_i of all orders i by recursion and for our case of summed identical copies of r.v. X_i, $K_i^n = K_i(\sum_n X_i) = nK_i(X_1)$.

Cumulants:

$$K_1^n = ne^{\mu + \frac{\sigma^2}{2}}$$
$$K_2^n = n\left(e^{\sigma^2} - 1\right)e^{2\mu + \sigma^2}$$
$$K_3^n = n\left(e^{\sigma^2} - 1\right)^2\left(e^{\sigma^2} + 2\right)e^{3\mu + \frac{3\sigma^2}{2}}$$
$$K_4^n = \dots$$

Which allow us to compute: *Skewness* $= \dfrac{\sqrt{e^{\sigma^2} - 1}\left(e^{\sigma^2} + 2\right)e^{\frac{1}{2}\left(2\mu + \sigma^2\right) - \mu - \frac{\sigma^2}{2}}}{\sqrt{n}}$ and *Kurtosis* $= 3 +$
$\dfrac{e^{2\sigma^2}\left(e^{\sigma^2}\left(e^{\sigma^2} + 2\right) + 3\right) - 6}{n}$

We can immediately prove from the cumulants/moments that:

$$\lim_{n \to +\infty} \kappa_{1,n} = 0, \lim_{\sigma \to 0} \kappa_{1,n} = 0$$

and our bound on κ becomes explicit:

Let $\kappa_{1,n}^*$ be the situation under which the sums of lognormal conserve the lognormal density, with the same first two moments. We have

$$0 \le \kappa_{1,n}^* \le 1,$$

$$\kappa_{1,n}^* = 2 - \frac{\log(n)}{\log\left(\dfrac{n\,\mathrm{erf}\left(\dfrac{\sqrt{\log\left(\frac{n + e^{\sigma^2} - 1}{n}\right)}}{2\sqrt{2}}\right)}{\mathrm{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)}\right)}$$

Heuristic attempt Among other heuristic approaches, we can see in two steps how 1) under high values of σ, $\kappa_{1,n} \to \kappa_{1,n}^*$, since the law of large numbers slows down, and 2) $\kappa_{1,n}^* \overset{\sigma \to \infty}{\to} 1$.

Loulakis' Proof   Proving the upper bound, that for high variance $\kappa_{1,n}$ approaches 1 has been shown formally my Michail Loulakis[3] which we summarize as follows. We start with the identify $\mathbb{E}\left(|X - m|\right) = 2\int_m^\infty (x - m)f(x)dx = 2\int_m^\infty \bar{F}_X(t)dt$, where $f(.)$ is the density, m is the mean, and $\bar{F}_X(.)$ is the survival function. Further, $\mathbb{M}(n) = 2\int_{nm}^\infty \bar{F}(x)dx$. Assume $\mu = \frac{1}{2}\sigma^2$, or $X = \exp\left(\sigma Z - \frac{\sigma^2}{2}\right)$ where Z is a standard normal variate. Let S_n be the sum $X_1 + \ldots + X_n$; we get $\mathbb{M}(n) = 2\int_n^\infty \mathbb{P}(S_n > t)dt$. Using the property of subexponentiality ([136]), $\mathbb{P}(S_n > t) \geq \mathbb{P}(\max_{0<i\leq n}(X_i) > t) \geq n\mathbb{P}(X_1 > t) - \binom{n}{2}\mathbb{P}(X_1 > t)^2$. Now $\mathbb{P}(X_1 > t) \overset{\sigma\to\infty}{\to} 1$ and the second term to 0 (using Hölder's inequality).

Skipping steps, we get $\liminf_{\sigma\to\infty} \frac{\mathbb{M}(n)}{\mathbb{M}(1)} \geq n$, while at the same time we need to satisfy the bound $\frac{\mathbb{M}(n)}{\mathbb{M}(1)} \leq n$. So for $\sigma \to \infty$, $\frac{\mathbb{M}(n)}{\mathbb{M}(1)} = n$, hence $\kappa_{1,n} \overset{\sigma\to\infty}{\to} 1$.

Pearson Family approach for computation   For computational purposes, for the σ parameter not too large (below $\approx .3$, we can use the Pearson family for computational convenience –although the lognormal does not belong to the Pearson class (the normal does, but we are close enough for computation). Intuitively, at low *sigma*, the first four moments can be sufficient because of the absence of large deviations; not at higher *sigma* for which conserving the lognormal would be the right method.

The use of Pearson class is practiced in some fields such as information/communication theory, where there is a rich literature: for summation of lognormal variates see Nie and Chen, [122], and for Pearson IV, [29], [47].

The Pearson family is defined for an appropriately scaled density f satisfying the following differential equation.

$$f'(x) = -\frac{(a_0 + a_1 x)}{b_0 + b_1 x + b_2 x^2}f(x) \tag{7.13}$$

We note that our parametrization of a_0, b_2, etc. determine the distribution within the Pearson class –which appears to be the Pearson IV. Finally we get an expression of mean deviation as a function of n, σ, and μ.

Let m be the mean. Diaconis et al [49] from an old trick by De Moivre, Suzuki [152] show that we can get explicit mean absolute deviation. Using, again, the identity $\mathbb{E}(|X - m|) = 2\int_m^\infty (x - m)f(x)dx$ and integrating by parts,

$$\mathbb{E}(|X - m|) = \frac{2\left(b_0 + b_1 m + b_2 m^2\right)}{a_1 - 2b_2}f(m) \tag{7.14}$$

---

3 Review of this paper; Loulakis proposed a formal proof in place of the heuristic derivation.

We use cumulants of the n-summed lognormal to match the parameters. Setting $a_1 = 1$, and $m = \frac{b_1 - a_0}{1 - 2b_2}$, we get

$$
\begin{cases}
a_0 = \dfrac{e^{\mu + \frac{\sigma^2}{2}} \left(-12n^2 + (3-10n)e^{4\sigma^2} + 6(n-1)e^{\sigma^2} + 12(n-1)e^{2\sigma^2} - (8n+1)e^{3\sigma^2} + 3e^{5\sigma^2} + e^{6\sigma^2} + 12 \right)}{2 \left(6(n-1) + e^{2\sigma^2} \left(e^{\sigma^2} \left(5e^{\sigma^2} + 4 \right) - 3 \right) \right)} \\[4mm]
b_2 = \dfrac{e^{2\sigma^2} \left(e^{\sigma^2} - 1 \right) \left(2e^{\sigma^2} + 3 \right)}{2 \left(6(n-1) + e^{2\sigma^2} \left(e^{\sigma^2} \left(5e^{\sigma^2} + 4 \right) - 3 \right) \right)} \\[4mm]
b_1 = \dfrac{\left(e^{\sigma^2} - 1 \right) e^{\mu + \frac{\sigma^2}{2}} \left(e^{\sigma^2} \left(e^{\sigma^2} \left(e^{\sigma^2} \left(-4n + e^{\sigma^2} \left(e^{\sigma^2} + 4 \right) + 7 \right) - 6n + 6 \right) + 6(n-1) \right) + 12(n-1) \right)}{2 \left(6(n-1) + e^{2\sigma^2} \left(e^{\sigma^2} \left(5e^{\sigma^2} + 4 \right) - 3 \right) \right)} \\[4mm]
b_0 = -\dfrac{n \left(e^{\sigma^2} - 1 \right) e^{2 \left(\mu + \sigma^2 \right)} \left(e^{\sigma^2} \left(-2(n-1)e^{\sigma^2} - 3n + e^{3\sigma^2} + 3 \right) + 6(n-1) \right)}{2 \left(6(n-1) + e^{2\sigma^2} \left(e^{\sigma^2} \left(5e^{\sigma^2} + 4 \right) - 3 \right) \right)}
\end{cases}
$$

Polynomial expansions   Other methods, such as Gram-Charlier expansions, such as Schleher [146], Beaulieu,[10], proved less helpful to obtain κ_n. At high values of σ, the approximations become unstable as we include higher order Lhermite polynomials. See review in Dufresne [52] and [53].

7.5.6 Exponential

The exponential is the "entry level" fat tails, just at the border.

$$
f(x) = \quad \lambda e^{-\lambda x}, \quad x \geq 0.
$$

By convolution the sum $Z = X_1, X_2, \ldots X_n$ we get, by recursion, since $f(y) = \int_0^y f(x) f(y - x)\, dx = \lambda^2 y e^{-\lambda y}$:

$$
f_n(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!} \tag{7.15}
$$

which is the gamma distribution; we get the mean deviation for n summands:

$$
\mathbb{M}(n) = \frac{2e^{-n} n^n}{\lambda \Gamma(n)}, \tag{7.16}
$$

hence:

$$
\kappa_{1,n} = 2 - \frac{\log(n)}{n \log(n) - n - \log(\Gamma(n)) + 1} \tag{7.17}
$$

We can see the asymptotic behavior is equally slow (similar to the student) although the exponential distribution is sitting at the cusp of subexponentiality:

$$
\lim_{n \to \infty} \log(n) \kappa_{1,n} = 4 - 2 \log(2\pi)
$$
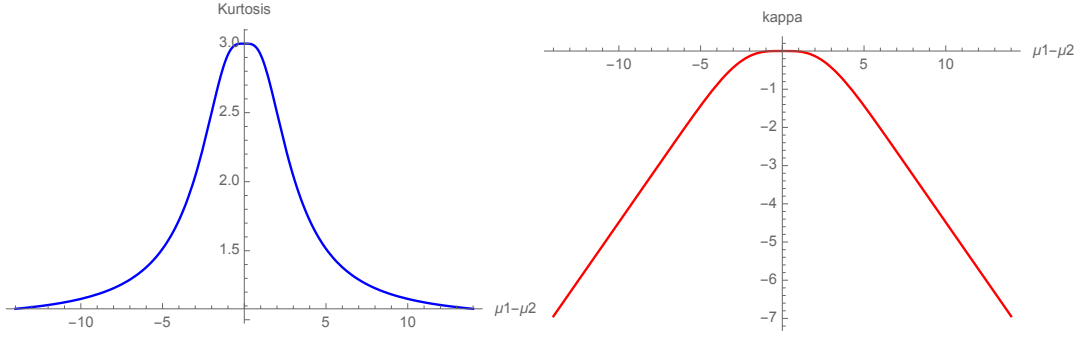
Figure 7.5: *Negative kurtosis from B.2 and corresponding kappa.*

7.5.7 Negative kappa, negative kurtosis

Consider the simple case of a Gaussian with switching means and variance: with probability $\frac{1}{2}$, $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and with probability $\frac{1}{2}$, $X \sim \mathcal{N}(\mu_2, \sigma_2)$.

These situations with thinner tails than the Gaussian are encountered with bimodal situations where μ_1 and μ_2 are separated; the effect becomes acute when they are separated by several standard deviations. Let d= $\mu_1 - \mu_2$ and $\sigma = \sigma_1 = \sigma_2$ (to achieve minimum kurtosis),

$$\kappa_1 = \frac{\log(4)}{\log(\pi) - 2\log\left(\frac{\sqrt{\pi}de^{\frac{d^2}{4\sigma^2}}\operatorname{erf}\left(\frac{d}{2\sigma}\right)+2\sqrt{\sigma^2}e^{\frac{d^2}{4\sigma^2}}+2\sigma}{de^{\frac{d^2}{4\sigma^2}}\operatorname{erf}\left(\frac{d}{2\sqrt{2}\sigma}\right)+2\sqrt{\frac{2}{\pi}}\sigma e^{\frac{d^2}{8\sigma^2}}}\right)} + 2 \tag{7.18}$$

which we see is negative for wide values of $\mu_1 - \mu_2$.

8 | DIAGNOSTIC TOOLS FOR FAT TAILS. WITH APPLICATION TO THE SP500[†]

IN THIS (RESEARCH) CHAPTER we show some reasoning errors in the literature on the "overestimation" of tail risks in general and in the stock market and other economic random variables in particular. We isolate the three various methods to study tail risks and show the statistical invalidity of the first two under the power law/slow variation class.

We propose a battery of tests to assess if one fails to reject Paretianity compared to other ad hoc adjustments such as stochastic volatility /Poisson. Applying to the SP500, we show that the process cannot be reliably estimated outside the slow variation class, with or possibly finite variance –and more so for the tails.

Analyses in $L2$ such as GARCH , conditional variance, or stochastic volatility are methodologically (and practically) invalid.

We also present the notion of ergodicity of portfolio in the context of tail pricing.

We show how conclusions about "overpricing" of tail events tend to underestimate the tail and "deep tail" (catastrophes) by up to 70 times.

8.1 INTRODUCTION

The problem If you use the wrong distribution (or method)... then all the consequences may be wrong, and your approach is most certainly unscientific. Changing from thin-tailed to fat-tailed is not just *changing the color of the dress*. The finance and economic idiots hold the message "we know it is fat tailed" but then fail to grasp the consequences on many things such as the slowness of the law of large numbers and the failure of sample means to be sufficient statistic (as well as the ergodicity effect, among others).

Here we just focus on the inadequate estimation of tail risks.

Some practical consequences of the problem For the U.S. stock market indices, while the first method is deemed to be ludicrous, using the second method leads to an underestimation of the payoff in the tails of between 5 and 70 times.

The organization of the discussion We contrast the three possible approaches to tail risks: Methods 1, 2, and 3. We establish that Power Law distributions can only be estimated by Method 3. We show the battery of tests for failure to reject if a given data belong to the

$$\frac{\int_K^\infty \bar{\phi}_s \, dx}{\int_K^\infty \bar{\phi}_e \, (x) \, dx}$$
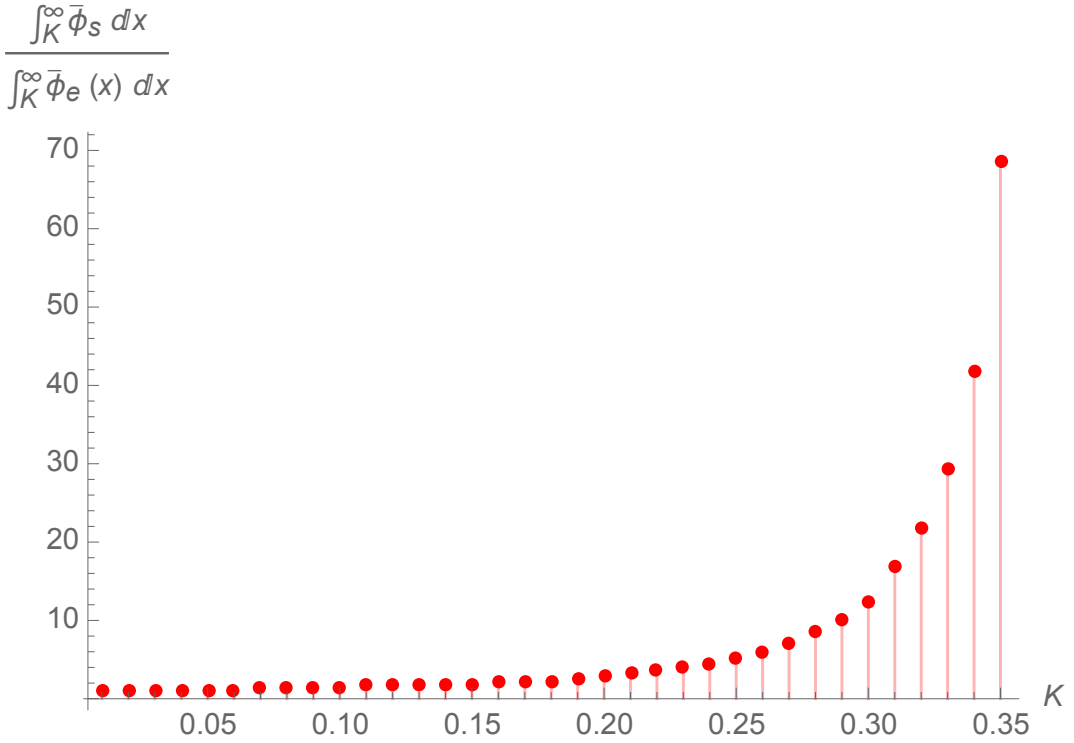
Figure 8.1: *This figure represents the extent of the underestimation. It shows the relative value of tail CVar-style measure compared to that from the empirical distribution. The deep tail is underestimated up to 70 times by current methods, even those deemed "empirical".*

Power Law basin (based on criteria of stability of moments and convergence). Method 3 is extrapolative and extends the tails beyond in-sample extrema.

Next, we show the tail risk underestimation. Finally we connect the point to the regular argument of ergodicity of probability.

8.2 METHODS 1 THROUGH 3

There are three general statistical methods to capture tail risk while dealing with independent data.

Method 1 (Naive Parametric)

Method 1 is the "naive" parametric method based on thin-tails distributions, where the maximum likelihood of the conditional mean represents the true mean, for all subsegments of the distributions. The conditional "tail mean" $\mu_K \triangleq \mathbb{E}(X|_{X>K})$ (or $\mu_{<K} \triangleq \mathbb{E}(-X|_{-X>K})$ is derived parametrically to match the *sample* mean $m(n) = \frac{1}{n} \sum_{i \leq n} x_i$ and the *sample* second moment of a parametric distribution, $m_2(n) = \frac{1}{n} \sum_{i \leq n} x_i^2$ thanks to the working of the Central Limit Theorem (in its Lindeberg formulation).

Note: in the rest of the paper we will express $\mathbb{E}(-X|_{-X>K})$ "the negative tail" in positive numbers except when otherwise mentioned.
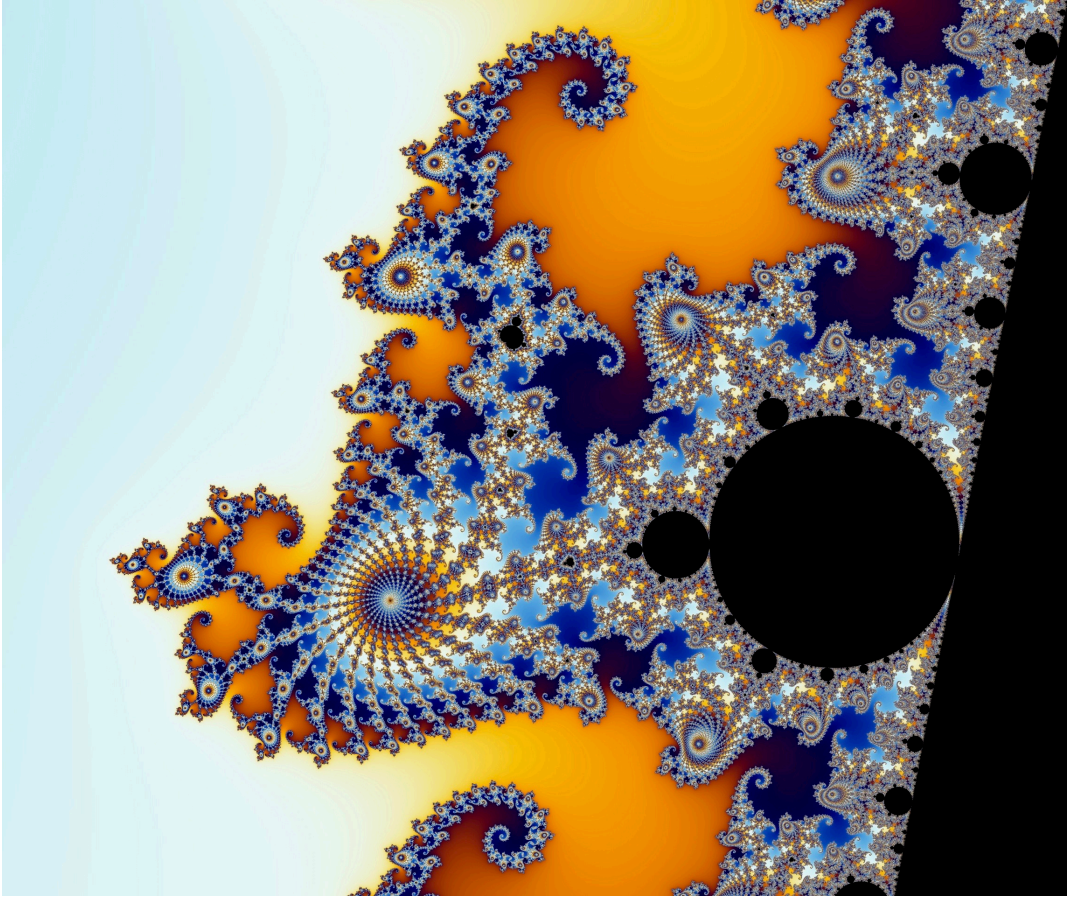
Figure 8.2: *The great Benoit Mandelbrot linked fractal geometry to statistical distributions via self-affinity at all scales. When asked to explain his work, he said: "rugosité", meaning"roughness" –it took him fifty years to realize that was his specialty. (Seahorse Created by Wolfgang Beyer, Wikipedia Commons.)*

The limit for $n \to \infty$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right). \tag{8.1}$$

where $\xrightarrow{D}$ indicates convergence in distribution. Critically, Method 1 assumes that n is sufficiently large to assume 8.1, and that the same applies to higher moments.

Definition 8.1 (Method 1)
Method 1 assumes that the data allows the matching of mean and variance to that of a standard parametric distribution and, when n is sufficiently large, to that of a Gaussian, and estimate conditional tail means μ_K from that of the parametric distribution.

We note that Method 1 is the standard approach in finance.

Remark 8.1
Method 1 relies on both the central limit theorem (CLT) and the law of large numbers (LLN) .

In other words, Method 1 assumes that the maximum likelihood estimation of the mean and variance are derived from the sample first two moments. The way CLT and LLN mix inextricably is one of the problems we faced in Chapter x.

Remark 8.2
Method 1 assumes that the mean and variance are sufficient statistics for conditional tail means μ_K and $\mu_{<K}$.

Method 2 (Nonparametric)

Method 2 is the nonparametric method based on the empirical distribution, which *appear to* adjust for fat tails. Empirical distribution, or survival function $\bar{F}(t)$: Let $X_1, X_2, \ldots X_n$ be independent, identically distributed real random variables with the common cumulative distribution function $F(t)$.

$$\bar{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \geq t},$$

where $\mathbb{1}_A$ is the indicator function.

Definition 8.2
Method 2 is where the conditional mean $\mu > K \triangleq \mathbb{E}(X|x > K)$ is assumed to be represented by the sample conditional mean

$$m_K \triangleq \frac{\sum_{i=1}^{n} x_i \mathbb{1}_{x_i \geq K}}{\sum_{i=1}^{n} \mathbb{1}_{x_i \geq K}}. \tag{8.2}$$

and the survival probabilities are represented by $\bar{F}_n(K)$.

We note that the empirical distributions are necessarily censured on the interval $[x_{min}, x_{max}]$. On the other hand, owing to the finiteness of moments inherent in Method 1, the latter extrapolate very little outside such a range.

We next discuss the properties of the extremes of Medhod 2.[REPEAT]

The so-called "empirical distribution" is not quite empirical There is a prevalent confusion about the nonparametric empirical distribution based on the following powerful property: as n grows, it converges to the Gaussian *regardless of the base distribution*, even if fat-tailed (assuming infinite support). For the CDF (or survival functions) are both uniform on $[0, 1]$, and, further, by the Donsker theorem, the sequence $\sqrt{n}\,(F_n(x) - F(x))$ converges in distribution to a Normal Distribution with mean 0 and variance $F(x)\,(1 - F(x))$ (one may find even stronger forms of convergence via the Glivenko– Cantelli theorem). Owing to this remarkable property, one may mistakenly assume that the tails of the distribution converge in the same manner independently of the distribution. Further, and what contributes to the confusion, given the variance, $F(x)\,(1 - F(x))$ for both empirical CDF and survival function, drops at the extremes.

In truth, and that is a property of extremes, the error effectively increases in the tails if one multiplies by the divergence. Let χ_n be the difference between the empirical and the distributional conditional mean, defined as:

$$\chi_n = \sum_{i=1}^{n} x_i \mathbb{1}_{x_i \geq K} - \int_{K}^{\infty} x\, dF(x)$$

$$= K(F_n(K) - F(K)) + \sum_{i}^{\frac{x_{max}}{\delta}} \left(\bar{F}_n(K + (i+1)\delta) - \bar{F}_n(K + i\delta K) - \int_{K+i\delta}^{K+(i+1)\delta} d\bar{F}(K) \right) - \int_{max}^{\infty} dF(x), \tag{8.3}$$

where $x_{max} = \bar{F}_n^{-1}(0)$, that is where the distribution is truncated. χ_n recovers the dispersion of the distribution of x which remains fat tailed. Another way to see it is that for fat tailed

variables, probabilities are more stable than their realizations and, more generally, the lowest moment will always disproportionately be the most stable one.

> **Biases of the empirical method under Fat Tails**
>
> We note that, owing of the convergence to the Gaussian, by Donsker's theorem:
>
> $$\chi_n = \int_{x_{max}}^{\infty} dF(x) + O\left(\frac{F(x)\,(1 - F(x))}{\sqrt{n}}\right) \tag{8.4}$$
>
> so, for sufficiently large (but not too large) n,
>
> $$\chi_n \approx \int_{x_{max}}^{\infty} dF(x) \tag{8.5}$$
>
> yet, under a Paretian regime, x_{max} is distributed according to a Fréchet, as we will see in Section TK.

Theorem 8.1

For an empirical distribution with a sample size n, the underestimation of the conditional tail expectation χ_n for a Paretian with scale L and tail index α is:

$$\varphi(\chi, n) = \left(\frac{\alpha - 1}{\alpha}\right)^{\frac{1}{\alpha-1}} nL^{\frac{\alpha^2}{1-\alpha}+1} \chi^{\frac{1}{\alpha-1}} \left(\exp\left(\left(\frac{\alpha-1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} n\left(-L^{\frac{\alpha^2}{1-\alpha}+1}\right) \chi^{\frac{\alpha}{\alpha-1}}\right)\right) \tag{8.6}$$

and its expectation

$$\mathbb{E}(\chi_n) = \Gamma\left(\frac{\alpha-1}{\alpha}\right) L^{\alpha+\frac{1}{\alpha}-1} n^{\frac{1}{\alpha}-1}$$

Proof. The maximum of n variables is in the MDA (Maximum domain of attraction) of Fréchet with scale $\beta = (Ln)^{1/\alpha}$. We have the conditional expectation $> \chi$: $\mathbb{E}(x)|_{x>\chi} \mathbb{P}(x > \chi) = \frac{\alpha L^\alpha \chi^{1-\alpha}}{\alpha-1}$. Randomizing χ and doing a probability transformation we get the density $\varphi(.)$. $\qquad \square$

Remark 8.3

Method 2 relies on the law of large numbers without the central limit theorem.

Method 3 (Nonrestricted Maximum Likelihood parametric)

Method 3 is a more general Maximum Likelihood parametric approach that finds a theoretical distribution and a parameter that has converged according to the law of large numbers. This allows to extend the tails and extrapolate outside the sample based on statistical properties, not just past maxima.

Definition 8.3

Method 3 fits those parameters of the distribution that can satisfy the law of large numbers for sample size n. More specifically, it does not hold that the LLN holds for the mean n given n observations, even if $\mu > K < \infty$.

Equivalently, Method 3 does not accept the sample conditional mean m_K as statistical mean for all K in the support of the distribution.

Method 3 finds the parameter for which the law of large numbers holds.

The latter approach is necessary for distribution in the Power Law basin, which is characterized by the slowness of the mean. [34], [33].

Definition 8.4 (Power Law Class $\mathfrak{P}$)
The r.v. $X \in \mathbb{R}$ belongs to $\mathfrak{P}$, the class of slowly varying functions (a.k.a. Paretian tail or power law-tailed) if its survival function (for the variable taken in absolute value) decays asymptotically at a fixed exponent α, or α', that is

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{8.7}$$

(right tail) or

$$\mathbb{P}(-X > x) \sim L(x)\, x^{-\alpha'} \tag{8.8}$$

(left tail)
where $\alpha, \alpha' > 0$ and $L : (0, \infty) \to (0, \infty)$ is a slowly varying function, defined as $\lim_{x \to \infty} \frac{L(kx)}{L(x)} = 1$ for all $k > 0$.

The happy result is that the parameter α obeys an inverse gamma distribution that converges rapidly to a Gaussian and does not require a large n to get a good estimate. This is illustrated in Figure 8.3, where we can see the difference in fit.

8.3 THE LAW OF LARGE NUMBERS UNDER PARETIANITY
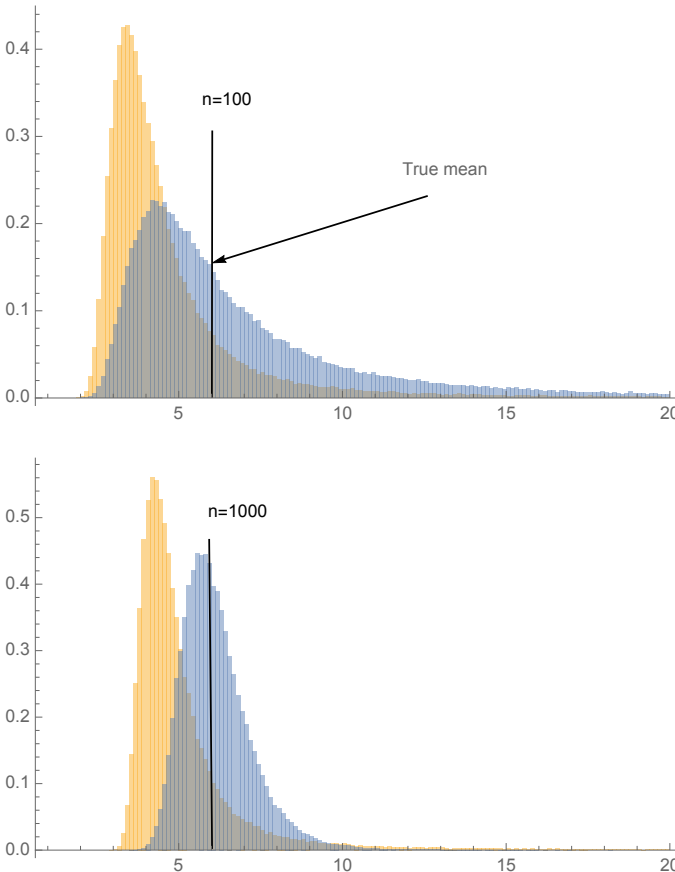
Figure 8.3: *Monte Carlo Simulation (10^5) of a comparison of sample mean (Methods 1 and 2) vs maximum likelihood mean estimations (Method 3) for a Pareto Distribution with $\alpha = 1.2$ (yellow and blue respectively), for $n = 100, 1000$. We can see how the MLE tracks the distribution more reliably. We can also observe the bias as Methods 1 and 2 underestimate the sample mean in the presence of skewness in the data. We need 10^7 more data in order to get the same error rate.*

> Why the sample mean, if it converges asymptotically for a sample size n, as $n \to \infty$, does not necessarily behave reliably for finite n, even larges values; this is a common mistake in the literature (as researchers use mechanistically asymptotic properties without verifying pre-asymptotic attributes).

To illustrate the slowness of the LLN in the slowly varying function class: for a distribution of the type "Pareto 80-20", that is with an exponent $\alpha \approx 1.14$, it takes $> 10^{11}$ more data than the Gaussian for *sample mean* $\to$ *mean*; accordingly neither Method 1 nor Method 2 are statistically reliable.

Simply, on the surface, the law of large number can be phrased as follows: For $\mathbb{E}(X) = \mu < \infty$ (meaning finite first moment but no restriction on higher moments),

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu \qquad \text{when } n \to \infty$$

That is to say that for any positive number ϵ:

$$\lim_{n \to \infty} \Pr\left(|\overline{X}_n - \mu| > \varepsilon \right) = 0.$$

The confusion arises as to "speed". To what extent the n-sized sample first moment $\frac{1}{n} \sum_{i=1}^{n} X_i$ is representative of μ? Well, it depends on α.

First, we need to look at "errors" about the mean in mean deviation terms (since we put no restriction on higher moments than 1).

Definition 8.5
Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean, that is $\mathbb{E}(X) < +\infty$. Let $S_n = X_1 + X_2 + \ldots + X_n$ be a partial sum. Let $MD(n) = \mathbb{E}(|S_n - \mathbb{E}(S_n)|)$ be the expected mean absolute deviation from the mean for n summands. Let $\kappa_{n_0,n}$ be the "speed" of convergence for additional n summands given that already we have n_0 summands:

For a thin-tailed variable, errors about the mean are, for n summands, simply $n^{-2}\sigma\frac{\sqrt{2}}{\pi}$. For a symmetric Stable Distribution (about which, see x) with otherwise same summands (and adapting to fit a scale σ):

$$MD_{\alpha,\beta}(n) = \frac{1}{2\pi}\sqrt{2}\sigma\Gamma\left(\frac{\alpha-1}{\alpha}\right) n^{\frac{1}{\alpha}-1}\left(\left(1+i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha} + \left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right) \quad (8.9)$$

which becomes simple when $\beta = 0$ (Samorodnitsky and Takku,[145], Uchaikin and Zolotarev [182]):

$$MD_\alpha(n) = \frac{\sqrt{2}\sigma\Gamma\left(\frac{\alpha-1}{\alpha}\right) n^{\frac{1}{\alpha}-1}}{\pi} \quad (8.10)$$

and defining n_2 the number of summands required to make mean deviation equal to that of a Gaussian with n_1 summands. $\{n_2 : MD_\alpha(n_2) = MD_G(n_1)\}$ where "G" is the Gaussian of the same scale

$$n_{2,\beta=1} = 2^{\frac{\alpha}{1-\alpha}} \pi^{\frac{\alpha}{2-2\alpha}}\left(\sqrt{n_1}\left(\left(1+i\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha} + \left(1-i\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right)\Gamma\left(\frac{\alpha-1}{\alpha}\right)\right)^{\frac{\alpha}{\alpha-1}}$$

$$(8.11)$$

which simplifies for a symmetric distribution:

$$n_2 = \pi^{\frac{\alpha}{2-2\alpha}} \left(\frac{1}{\sqrt{n_1}\Gamma\left(\frac{\alpha-1}{\alpha}\right)} \right)^{\frac{1}{\frac{1}{\alpha}-1}} \tag{8.12}$$

We can thus verify that for $\alpha = 1.14$, $n_{2,\beta=1} > 10^{11} \times n_1$. Again, Figure 8.3 is representative as applying to $\alpha = 1.2$, we get 10^7.

Secondly, there is a problem with finite variance power laws: for $n < 10^7$, even $\alpha > 2$ becomes meaningless as convergence to the Gaussian basin is very slow. See paper on Kappa-n, still incomplete.

8.4 DISTRIBUTION OF THE TAIL EXPONENT

> The distribution of α, the tail exponent of a Paretian distribution is thin-tailed and converges rapidly to the Gaussian.

Consider wlog the standard Pareto distribution for a random variable X with pdf:

$$\phi_X(x) = \alpha L^\alpha x^{-\alpha-1}, x > L \tag{8.13}$$

Assume $L = 1$ by scaling.

The likelihood function is $\mathcal{L} = \prod_{i=1}^n \alpha x_i^{-\alpha-1}$. Maximizing the Log of the likelihood function (assuming we set the minimum value) $\log(\mathcal{L}) = n(\log(\alpha) + \alpha \log(L)) - (\alpha+1) \sum_{i=1}^n \log(x_i)$ yields: $\hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(x_i)}$. Now consider $l = -\frac{\sum_{i=1}^n \log X_i}{n}$. Using the characteristic function to get the distribution of the average logarithm yield:

$$\psi(t)^n = \left(\int_1^\infty f(x) \exp\left(\frac{it \log(x)}{n} \right) dx \right)^n = \left(\frac{\alpha n}{\alpha n - it} \right)^n$$

which is the characteristic function of the gamma distribution $(n, \frac{1}{\alpha n})$. A standard result is that $\hat{\alpha}' \triangleq \frac{1}{l}$ will follow the inverse gamma distribution with density:

$$\phi_{\hat{\alpha}}(a) = \frac{e^{-\frac{\alpha n}{\hat{\alpha}}} \left(\frac{\alpha n}{\hat{\alpha}} \right)^n}{\hat{\alpha}\Gamma(n)}, a > 0$$

.

Debiasing   Since $\mathbb{E}(\hat{\alpha}) = \frac{n}{n-1}\alpha$ we elect another –unbiased– random variable $\hat{\alpha}' = \frac{n-1}{n}\hat{\alpha}$ which, after scaling, will have for distribution $\phi_{\hat{\alpha}'}(a) = \frac{e^{\frac{\alpha-\alpha n}{a}} \left(\frac{\alpha(n-1)}{a} \right)^{n+1}}{\alpha\Gamma(n+1)}$.

Truncating for $\alpha > 1$ Given that values of $\alpha \le 1$ lead to absence of mean we restrict the distribution to values greater than $1 + \epsilon$, $\epsilon > 0$. Our sampling now applies to lower-truncated values of the estimator, those strictly greater than 1, with a cut point $\epsilon > 0$, that

is, $\sum \frac{n-1}{\log(x_i)} > 1 + \epsilon$, or $\mathbb{E}(\hat{\alpha}|_{\hat{\alpha}>1+\epsilon})$: $\phi_{\hat{\alpha}'}(a) = \frac{\phi_{\hat{\alpha}'}(a)}{\int_{1+\epsilon}^{\infty} \phi_{\hat{\alpha}'}(a)\, da}$, hence the distribution of the values of the exponent conditional of it being greater than 1 becomes:

$$\phi_{\hat{\alpha}''}(a) = \frac{e^{\frac{\alpha n^2}{a-an}} \left(\frac{\alpha n^2}{a(n-1)}\right)^n}{a \left(\Gamma(n) - \Gamma\left(n, \frac{n^2 \alpha}{(n-1)(\epsilon+1)}\right)\right)}, \ a \geq 1 + \epsilon \tag{8.14}$$

8.5 DEPENDENCE AND ASYMMETRIES

8.5.1 Records and Extrema

The Gumbel record methods is as follows (Embrechts et al [60]). Let $X_1, X_2, \ldots$ be a discrete time series, with a maximum at period $t \geq 2$, $M_t = \max(X_1, X_2, \ldots, X_t)$, we have the record counter $N_{1,t}$ for n data points.

$$N_{1,t} = 1 + \sum_{k=2}^{t} \mathbb{1}_{X_t > M_{t-1}} \tag{8.15}$$

Regardless of the underlying distribution, the expectation $\mathbb{E}(N_t)$ is the Harmonic Number H_t, and the variance $H_t - H_t^{(2)}$, where $H_t = \sum_{i=1}^{t} \frac{1}{i^r}$. We note that the harmonic number is concave and very slow in growth, logarithmic, as it can be approximated with $log(n) + \gamma$, where γ is the Euler Mascheroni constant. The approximation is such that $\frac{1}{2(n+1)} \leq H_t - log(t)t - \gamma \leq \frac{1}{2n}$ (*Wolfram Mathworld* [187]).
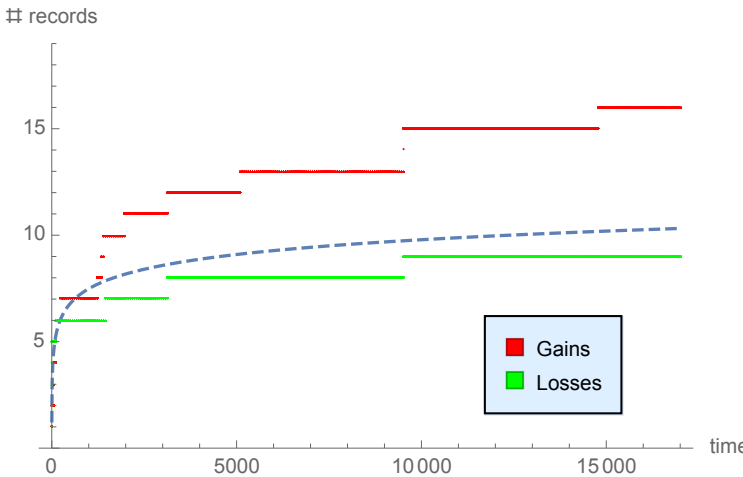
Figure 8.4: *The record test shows independence for extremes of negative returns, dependence for positive ones. The number of records for independent observations grows over time at the harmonic number $H(t)$ (dashed line), $\approx$ logarithmic but here appears to grow > 2.5 standard deviations faster for positive returns, hence we cannot assume independence for extremal gains. The test does not make assertions about dependence outside extremes.*

Remark 8.4
The Gumbel test of independence above is sufficient condition for the convergence of extreme negative values of the log-returns of the SP500 to the Maximum Domain of Attraction (MDA) of the extreme value distribution.

Entire series We reshuffled the SP500 (i.e. bootstrapped without replacement, using a sample size equal to the original ≈ 17000 points, with 10^3 repeats) and ran records across all of them. As shown in Fig. 8.6 and 8.5, the mean was 10.4 (approximated by the harmonic number, with a corresponding standard deviation.) The survival function S(.) of

$N_{1.7\times10^4} = 16$, $S(16) = \frac{1}{40}$ which allows us to consider the independence of positive extrema implausible.

On the other hand the negative extrema (9 counts) show realizations close to what is expected (10.3), diverting by $\frac{1}{2}$ a s.t.d. from expected, enough to justify a failure to reject independence.

Subrecords If instead of taking the data as one block over the entire period, we broke the period into sub-periods, we get (because of the concavity of the measure and Jensen's inequality), $N_{t_1+\delta,t_1+\Delta+\delta}$, we obtain T/δ observations. We took $\Delta = 10^3$ and $\delta = 10^2$, thus getting 170 subperiods for the $T \approx 17 \times 10^3$ days. The picture as shown in Fig. 8.7 cannot reject independence for both positive and reject observations.

It appears that the total is

Conclusion We can at least use EVT for negative observations.

8.6 SOME PROPERTIES AND TESTS

8.6.1 Asymmetry right-left tail

8.6.2 Paretianity and moments

Remark 8.5
Given that:

1) the slowly varying class has no higher moments than α, more precisely,

- *if $p > \alpha$, $\mathbb{E}(X^p) = \infty$ if p is even or the distribution has one-tailed support and*

- *$\mathbb{E}(X^p)$ is undefined if p is odd and the distribution has two-tailed support,*

and

2) distributions outside the slowly varying class have all moments $\forall p \in \mathbb{N}^+$, $\mathbb{E}(X^p) < \infty$.
$\exists p \in \mathbb{N}^+$ s.t. $\mathbb{E}(X^p)$ is either undefined or infinite $\Leftrightarrow X \in \mathfrak{P}$.

The rest of the paper examines ways to detect "infinite" moments. Much confusion attends the notion of infinite moments and its identification since by definition sample moments are finite and measurable under the counting measure. We will rely on the nonconvergence of moments. Let $\|\mathbf{X}\|_p$ be the weighted p-norm

$$\|\mathbf{X}\|_p \triangleq \left(\frac{1}{n} \sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

we have the property of power laws:

$$\mathbb{E}(X^p) \not< \infty \Leftrightarrow \|\mathbf{x}\|_p \text{ is not convergent.}$$

We note that, for obvious reasons, belonging to the class of Power Law tails cancels much of the methods in $L - 2$ such as GARCH and similar studies.
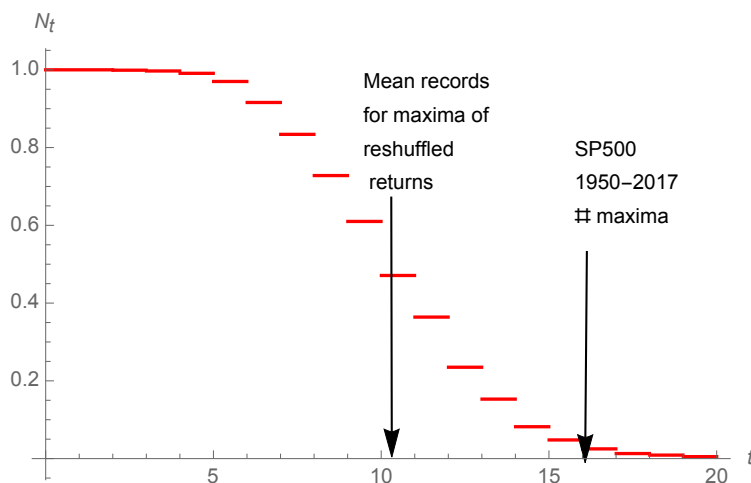
Figure 8.5: *The survival function of the records of positive maxima for the resampled SP500 (10^3 times) by keeping all returns but reshuffling them, thus removing the temporal structure. The mass above 16 (observed number of maxima records for SP500 over the period) is $\frac{1}{40}$.*
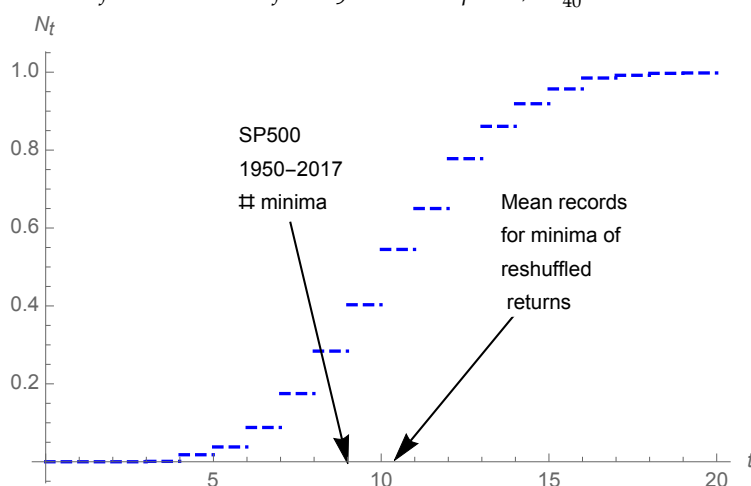
Figure 8.6: *The CDF of the records of negative extrema for the resampled SP500 (10^3 times) reshuffled as above. The mass above 9 (observed number of minima records for SP500 over the period) is $\frac{2}{5}$.*

8.7 CONVERGENCE TESTS

Convergence laws can help us *exclude* some classes of probability distributions.

8.7.1 Test 1: Kurtosis under Aggregation

Result: The verdict as shown in Figure 8.9 is that the one-month kurtosis is not lower than the daily kurtosis and, as we add data, no drop in kurtosis is observed. Further we would expect a drop $\sim N^{-2}$. This allows us to safely eliminate numerous classes, which includes stochastic volatility in its simple formulations such as gamma variance. Next we will get into the technicals of the point and the strength of the evidence.

A typical misunderstanding is as follows. In a note "What can Taleb learn from Markowitz" [180], Jack L. Treynor, one of the founders of portfolio theory, defended the field with the
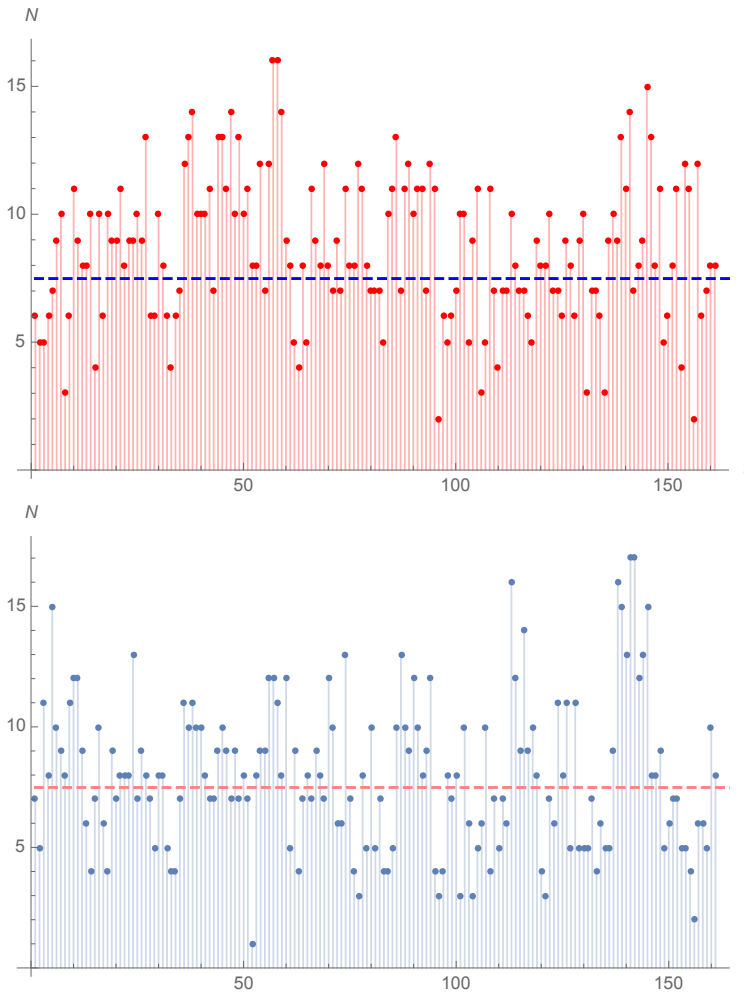
Figure 8.7: *Running shorter period, t = 1000 days of overlapping observations for the records of maxima(top) and minima (bottom), compared to the expected Harmonic number H(1000).*

argument that the data may be fat tailed "short term" but in something called the "long term" things become Gaussian. Sorry, it is not so. (We add the ergodic problem that blurs, if not eliminate, the distinction between long term and short term).

The reason is that, simply we cannot possibly talk about "Gaussian" if kurtosis is infinite. Further, for $\alpha \approx 3$, Central limit operates very slowly, requires n of the order of 10^6 to become acceptable, not what we have in the history of markets. [19]

8.7.2 Test 2: Excess Conditional Expectation

Result: The verdict from this test is that, as we can see in Figure 8.11, that the conditional expectation of X (and $-X$), conditional on X is greater than some arbitrary value K, remains proportional to K.
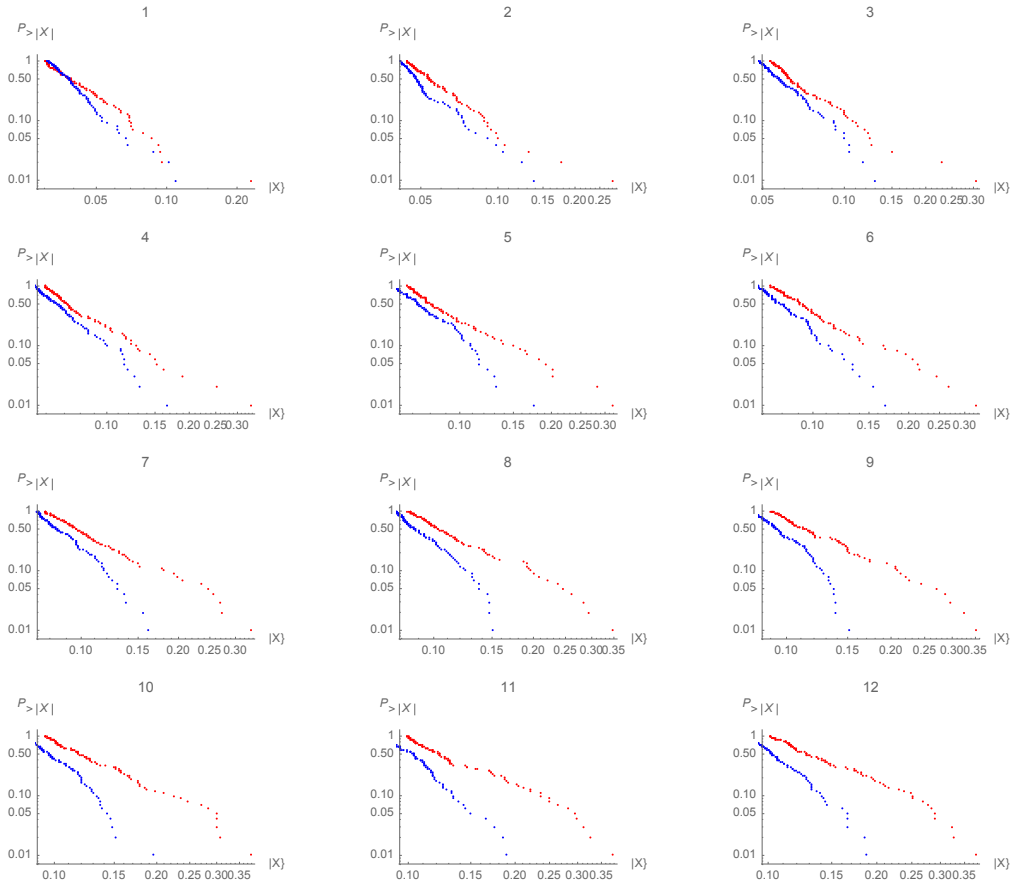
Figure 8.8: *We separate positive and negative logarithmic returns and use overlapping cumulative returns from 1 up to 15. Clearly the negative returns appear to follow a Power Law while the Paretianity of the right one is more questionable.*
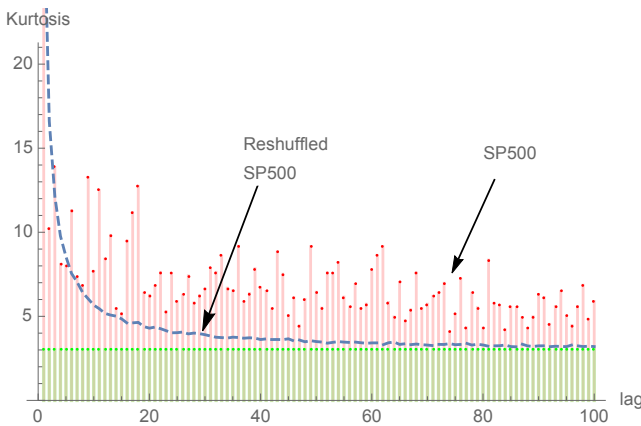
Figure 8.9: *Visual convergence diagnostics for the kurtosis of the SP500 over the past 17000 observations. We compute the kurtosis at different lags for the raw SP500 and reshuffled data. While the 4th norm is not convergent for raw data, it is clearly so for the reshuffled series. We can thus assume that the "fat tailedness" is attributable to the temporal structure of the data, particularly the clustering of its volatility. See Table 6.1 for the expected drop at speed $1/n^2$ for thin-tailed distributions.*

Definition 8.6

Let K be in $\mathbb{R}^+$, the relative excess conditional expectation:

$$\varphi_K^+ \triangleq \frac{\mathbb{E}(X)|_{X>K}}{K}$$
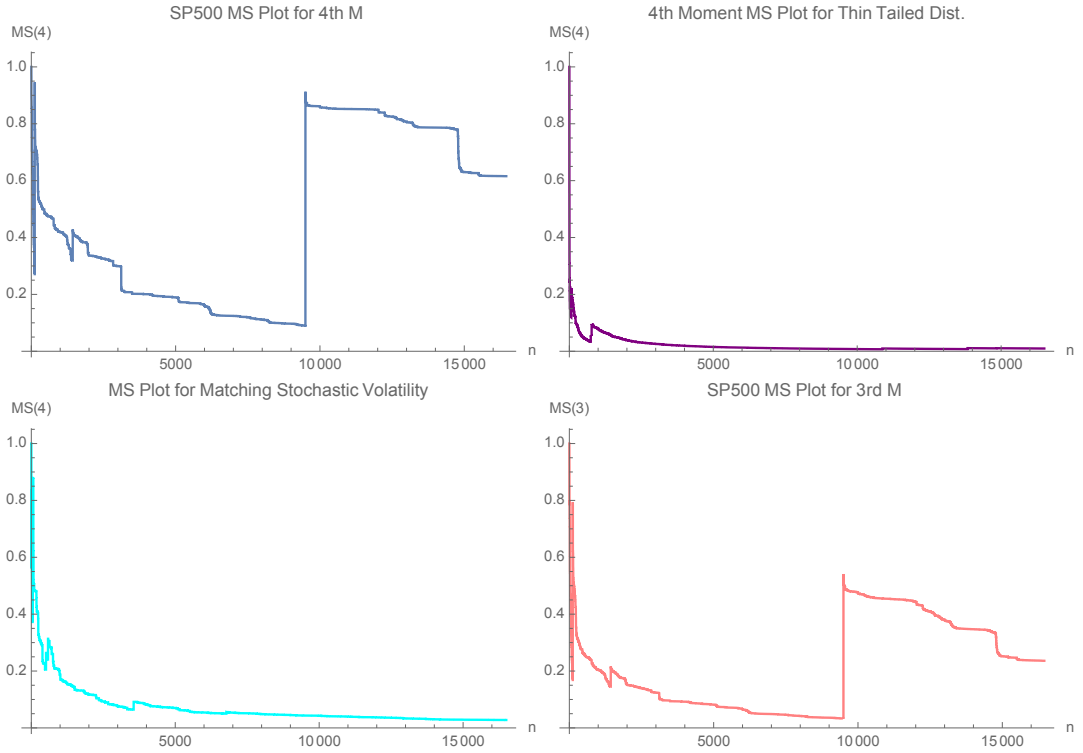
Figure 8.10: *MS Plot (or "law of large numbers for p moments") for p = 4 for the SP500 compared to p = 4 for a Gaussian and stochastic volatility for a matching Kurtosis (30) over the entire period. Convergence, if any, does not take place in any reasonable time. MS Plot for moment p = 3 for the SP500 compared to p = 4 for a Gaussian. We can safely say that the 4th moment is infinite and the 3rd one is indeterminate*
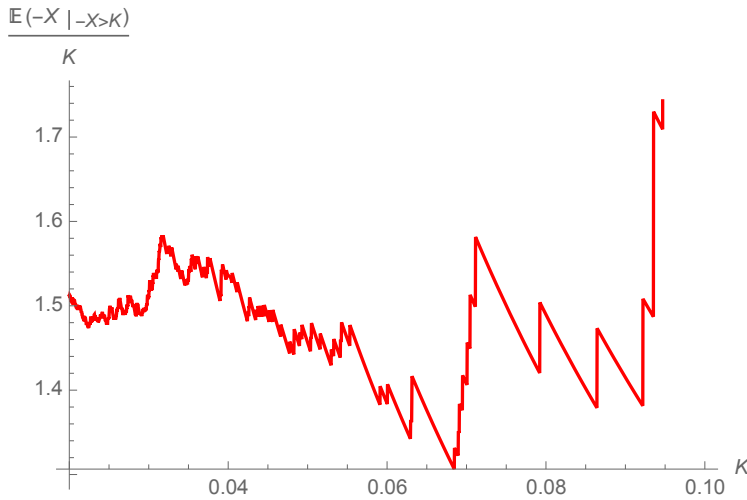
Figure 8.11: *Condexp as test of scalability*

$$\varphi_K^- \triangleq \frac{\mathbb{E}(-X)|_{X>K}}{K}$$

We have

$$\lim_{K\to\infty} \varphi_K = 0$$
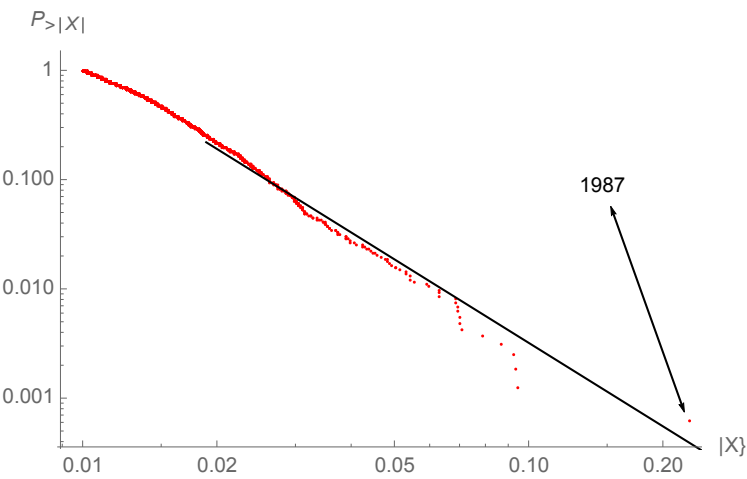
Figure 8.12: *Visual Identification of Paretianity*
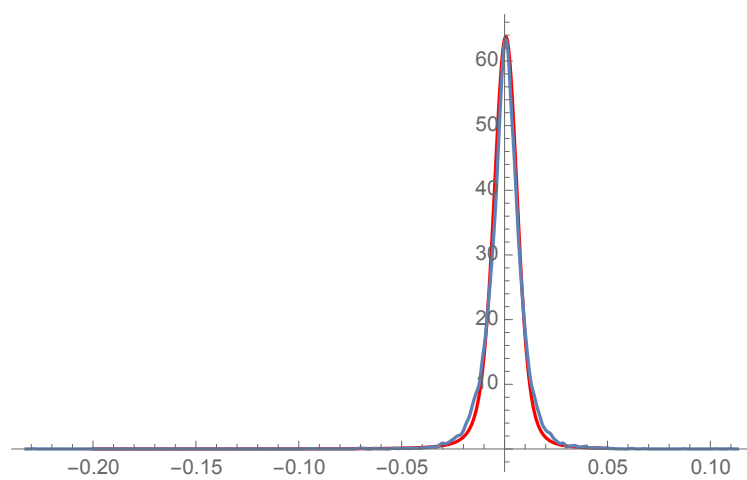
Figure 8.13: *Empirical distribution fits a stable with $\alpha_l = 1.62$*
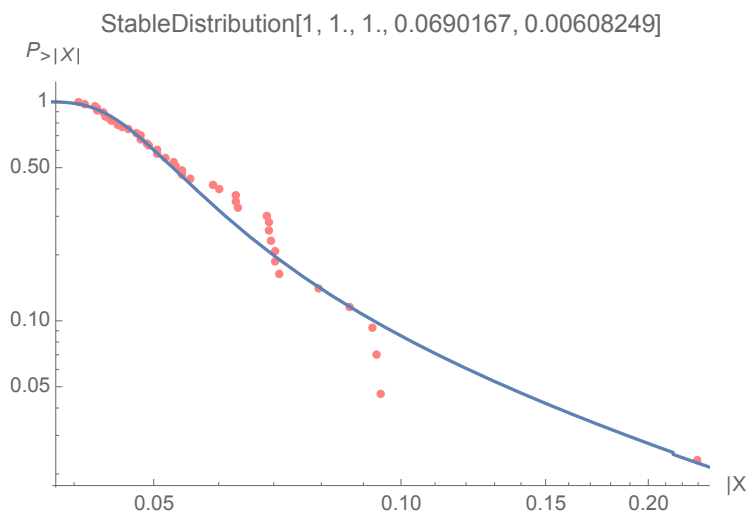
Figure 8.14: *The tails can possibly fit an infinite mean stable $\alpha_l = 1$*

for distributions outside the power-law basin, and
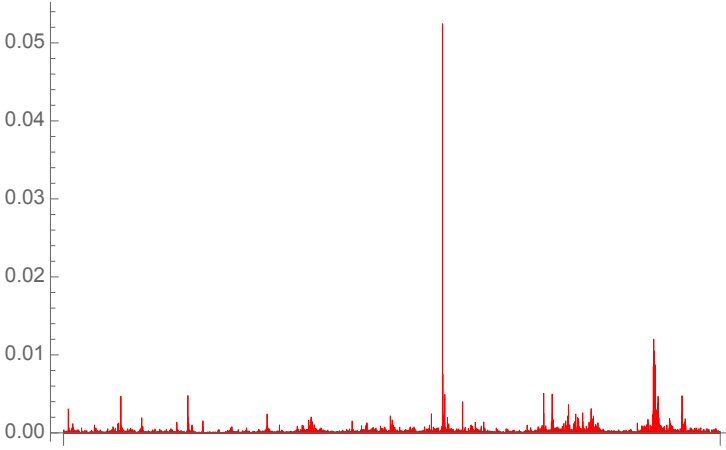
Figure 8.15: *Comparing SP500 squared returns to those from a Standard corresponding GARCH(1,1) for 16500 observations, for illustrative purposes only. A more formal proof comes from the Conditional Expectation test.*
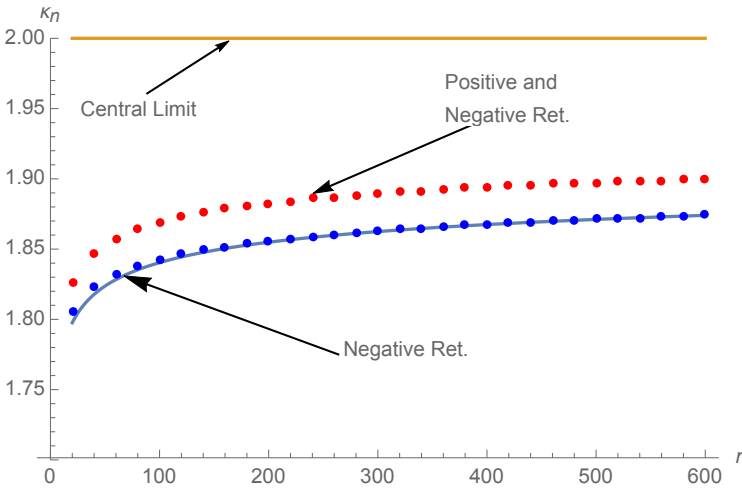
Figure 8.16: *kappa-n estimated empirically.*

$$\lim_{K \to \infty} \varphi_K = k \frac{\alpha}{1 - \alpha}$$

for distribution satisfying Definition 1. Note the van der Wijk's law [32],[159].

Figure 8.11 shows the following: the conditional expectation does not drop for large values, which is incompatible with non-Paretian distributions.

8.7.3   Test 3- Instability of 4th moment

A main argument in [159] is that in 50 years of SP 500 observations, a single one represents >80 % of the Kurtosis. Similar effect are seen with other socioeconomic variables, such as gold, oil, silver other stock markets, soft commodities. Such sample dependence of the kurtosis means that the fourth moment does not have the stability, that is, does not exist.

8.7.4   Test 4: MS Plot

An additional approach consists in examining the behavior of moments in a given sample, is the Maximum-to-Sum plot, or MS plot as it is plotted as in Figure 8.10. The MS Plot
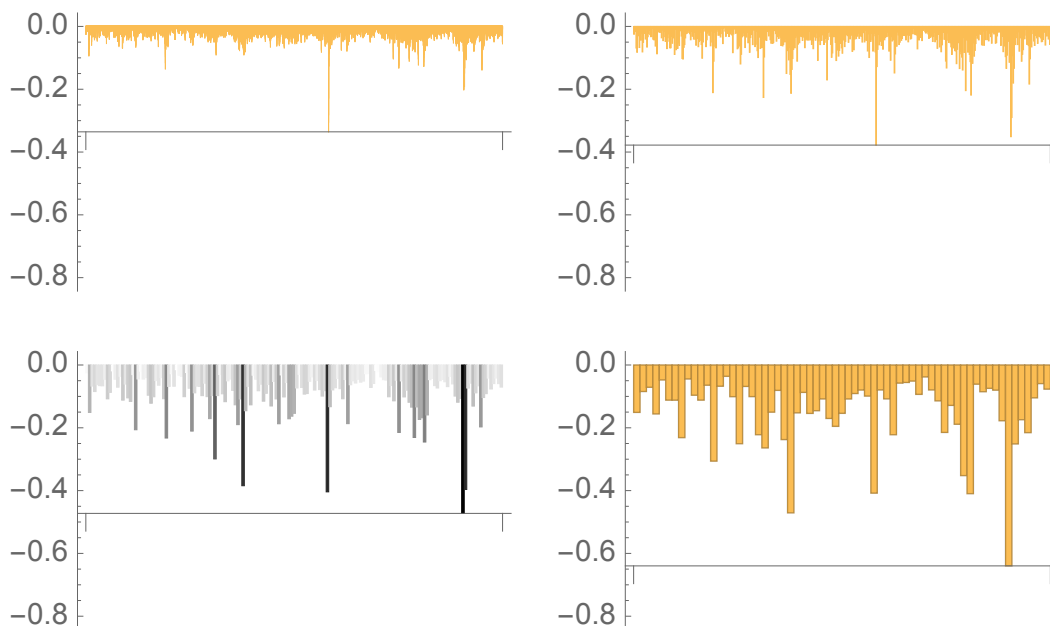
Figure 8.17: *Drawdowns and Scalability*
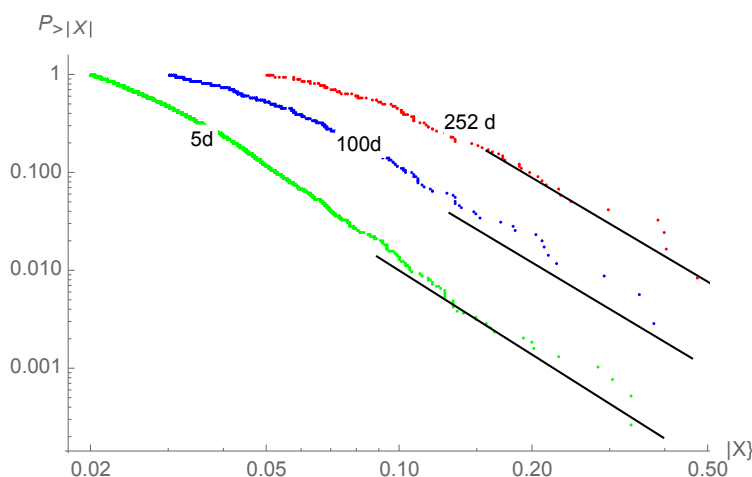
Figure 8.18: *Paretianity of Drawdowns and Scale*

relies on a consequence of the law of large numbers [124]. For a sequence $X_1, X_2, ..., X_n$ of nonnegative i.i.d. random variables, if for $p = 1, 2, 3...,$ $E[X^p] < \infty$, then $R_n^p = M_n^p / S_n^p \to^{a.s.}$ 0 as $n \to \infty$, where $S_n^p = \sum_{i=1}^{n} X_i^p$ is the partial sum, and $M_n^p = \max(X_1^p, ..., X_n^p)$ the partial maximum.

We show by comparison the MS plot for a Gaussian and that for a Student T with a tail exponent of 3. We observe that the SP 500 show the typical characteristics of a steep power law, as in 16,000 observations (50 years) it does not appear to drop to the point of allowing the functioning of the law of large numbers.
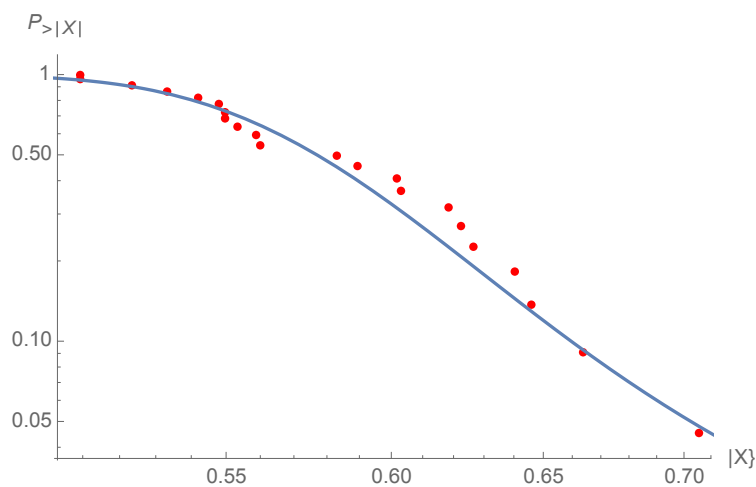
Figure 8.19: *Fitting a Stable Distribution to drawdowns*
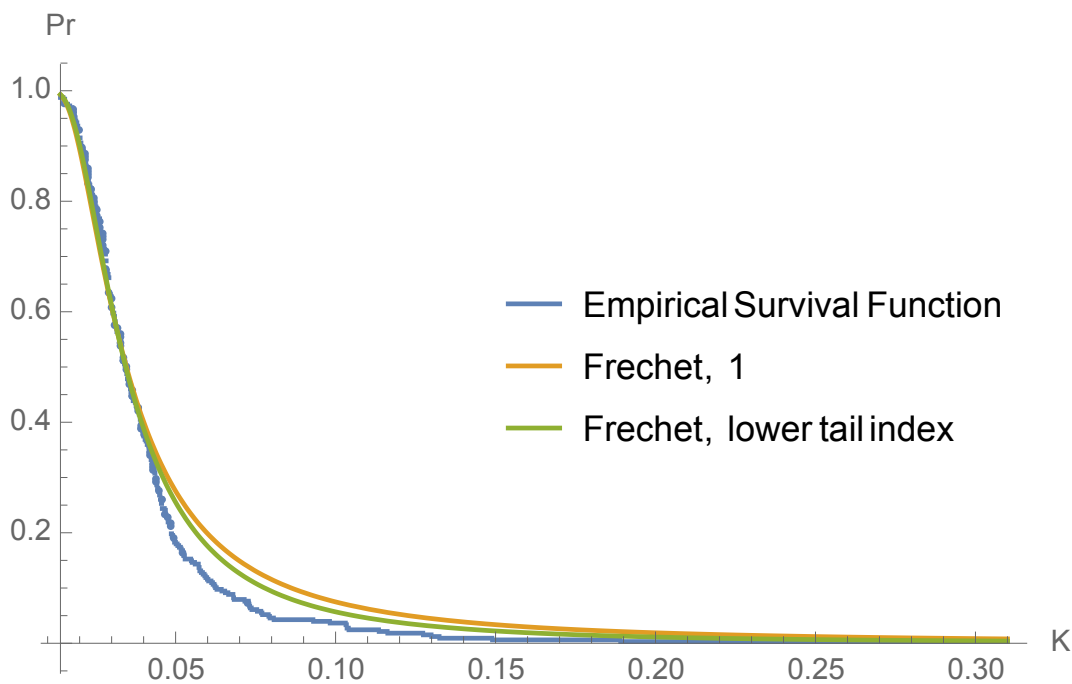
Figure 8.20: *Correcting the empirical distribution function with a Frechet*

8.8 CONCLUSION

This part will be completed.

Part III

ISSUES WITH FAT TAILS

9 | EXPERT CALIBRATION UNDER FAT TAILS

Being a good forecaster doesn't lead to having a good P/L, and vice versa. How someone's forecasting record is not a proxy for abilities to perform in the real world –and can be a reverse indicator under fat tails.

HE DECISION "SCIENCE" literature revolves around the concept of "calibration" of probabilistic assessment –an evaluation of how close someone providing odds of events turns out to be on average (under some operation of the law of large number deemed satisfactory). The methods are highly unempirical except in narrow circumstances of binary payoffs (such as those entailing a "win/lose" outcome) –and generalizing from these payoffs is described as falling for the "Ludic fallacy" [158]. Further, as we saw in Chapter 2, in the real world, people care about expectation, not probability, under condition of survival. Most businesses have severely skewed payoffs, so being calibrated in probability in meaningless there.

Finally, the difference between "binary" and "vanilla" options has been discussed in [156]: the former are bets paying a fixed amount, the latter have full payoff, and one cannot be properly replicated using another, especially under fat tails –meaning performance in one does not translate to performance into the other. And it is highly flawed to separate binary probability into the integral – where F is the probability measure, substituting $\int dF(x)$ for $\int x dF(x)$.

Recall that in the real world, it is P/L that counts, and making "calibration" mistakes where it doesn't matter should be encouraged, not penalized. If a mistake doesn't cost you anything, it is clearly not a mistake. And if it costs you something, and has been present in society for a long time, consider that there may be hidden evolutionary advantages to these types of mistakes –of the following sort: **mistaking a bear for a stone** is worse than **mistaking a stone for a bear**. In risk management, one should never operate in probability space.

In short, probabilistic calibration requires estimation of the zeroth moment while the real world requires all moments (outside of gambling bets where payoffs are truncated), and it is a central property of fat tails that higher moments are explosive (even "infinite") and count more and more.

Worse, the literature operates under the illusion ([50] [98]) that the measure, being probability distribution-free, is robust. Not quite: since probability is a kernel for summation or integration, i.e., something that enters into something else, and not the end product, mistakes on probability will rarely map to mistakes on expectation.

This note will show:

- Metrics of calibration of "right" or "wrong" probabilistic assessment –such as the Brier score– follow a thin-tailed distribution. We will derive the exact distribution of the metric.

- The mean of the n-summed (or averaged) random variable follows what mean by the law of large numbers – which, for power laws, remains a power law regardless of the number of summands.

9.0.1   Background: the robustness of metrics

What is the rationale behind the claim that the probability distribution of probability (sic) is robust? Let us refresh a standard result behind nonparametric discussions and tests, dating from Kolmogorov [100]. The probability integral transform is as follows.Probability integral transform Let X have a continuous distribution for which the cumulative distribution function (CDF) is F_X. Then –in the absence of additional information –the random variable U defined as $U = F_X(X)$ is uniform between 0 and 1.

The proof is as follows: For $t \in [0, 1]$,

$$\mathbb{P}(Y \leq u) = P(F_X(X) \leq u) = P(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u. \tag{9.1}$$

which is the cumulative distribution function of the uniform. This is the case regardless of the probability distribution of X.
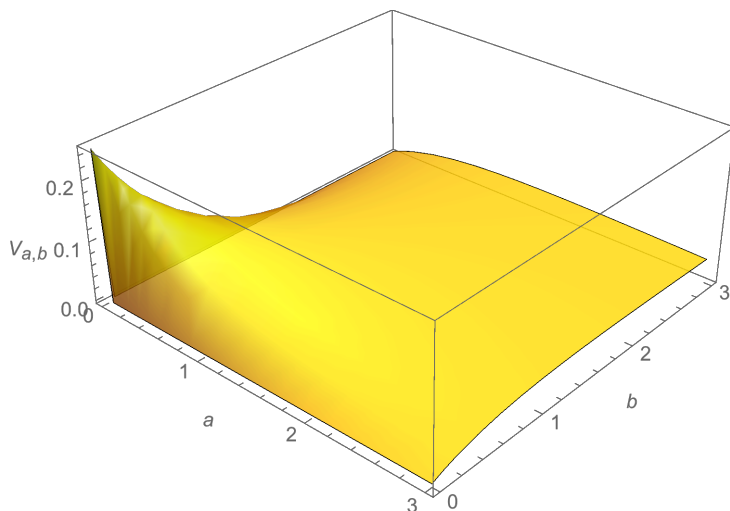
Figure 9.1: *Maximum variance for a beta distribution with support in* $[0, 1]$ *with parameters a and b; as both approach* 0 *we get a "U" shaped function, the Arcsine distribution shown in* 9.3.

9.1   THE BRIER SCORE FOLLOWS A THIN TAILED DISTRIBUTION

This discussion will use for calibration metric the Brier score (DeFinetti, [44]), deemed as the gold standard since it doesn't allow arbitrage and requires perfect probabilistic calibration: someone betting than an event has a probability 1 of occurring will get a perfect score only if the event occurs all the time.

Let λ_n be Brier score for a certain specific type of events:
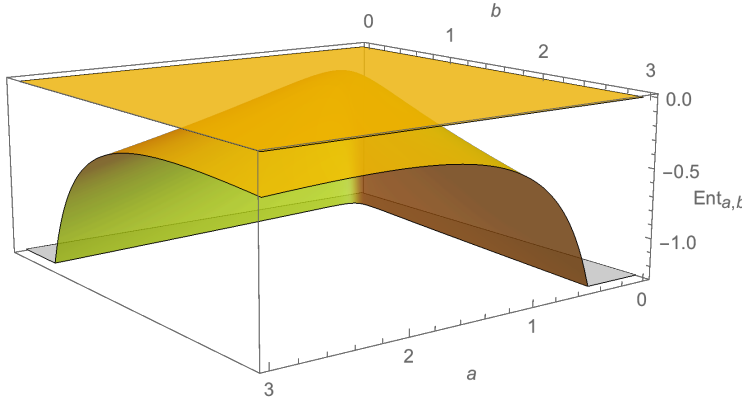
Figure 9.2: *Maximum entropy for a beta distribution with support in $[0,1]$ with parameters a and b . Unlike with the variance, we have the maximum reached with the uniform distribution ($a = b = 1$) and the minimum reached when either is 0. Axes show a, b,and $Ent_{a,b}$*

$$\lambda_n = \frac{1}{n} \sum_{t=1}^{n} (f_t - \mathbb{1}_{At})^2,$$

where $f_t \in [0,1]$ is the probability announced by the forecaster for event t, $\mathbb{1}_{At} \in \{0,1\}$ an indicator function whether the event under consideration took place or not, and n the total number of such forecasting events.

Clearly we are dealing with 1) f_t beta distributed (either as a special case the uniform distribution when purely random, as derived above, or a beta distribution when one has some accuracy, for which the uniform is a special case), and 2) $\mathbb{1}_{At}$ a Bernoulli variable with probability p.

Let us consider the general case.
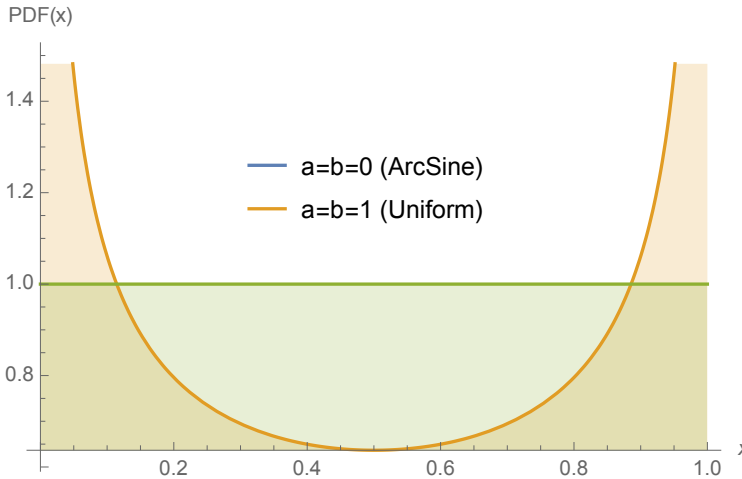
Figure 9.3: *Beta distributions with maximum entropy (Uniform) and maximum variance (Arcsine).*

$$f_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \ 0 < x < 1$$

As we can see in figures 9.1 and 9.2, we have the maximum uncertainty in a set of parameters where $\lim_{a \to 0 \ b \to 0} f_{a,b}$, which produces the Arcsine distribution.

Theorem 9.1 (Convergence in distribution of the Brier score)
Under independence of $(f_1 - \mathbb{1}_{A1}), ldots, (f_n - \mathfrak{n}_{An})$,

$$\lambda_n \xrightarrow{D} \mathcal{N}(\mu, \sigma_n) \text{ where } \mathcal{N} \text{ denotes the Gaussian distribution.}$$

Further, a bit unwieldy but controllable:

$$\mu = \frac{a^2(-p) + a^2 - ap + a + b^2 p + bp}{(a+b)(a+b+1)},$$

$$\sigma_n = (2a(1+a)b(3(1+b) + 2a(3+a+b))$$
$$+ (a-b)(1+a+b)(3+a+b)(2a+a^3 - b(1+b)(2+b) + a^2(3+b)$$
$$- ab(4+b))p - (a-b)^2(1+a+b)^2(2+a+b)(3+a+b)p^2)/(n(a+b)^2(1+a+b)^2(2+a+b)(3+a+b)).$$

It will also be shown that the Brier score has thinner tails than the Gaussian as its kurtosis is lower.

Proof. We start with $y_t = (f_t - \mathbb{1}_{At})$. The characteristic function of f_t, $\Psi_f = {}_1F_1(a; a+b; it)$, where ${}_1F_1(.;.;.)$ is the hypergeometric distribution ${}_1F_1(a; b; z) = \sum_{k=0}^{\infty} \frac{a_k \frac{z^k}{k}!}{b_k}$. The characteristic function of $\mathbb{1}_{At}$, $\Psi_{\mathbb{1}_A} = pe^{it} - p + 1$.

From here we get the characteristic function for $(f_t - \mathbb{1}_{At})^2$

$$\Psi_{y^2}(t) = \sqrt{\pi} 2^{-a-b+1} \Gamma(a+b) \left((1-p) {}_2\tilde{F}_2 \left(\frac{a+1}{2}, \frac{a}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); it \right) \right.$$
$$\left. + p {}_2\tilde{F}_2 \left(\frac{b+1}{2}, \frac{b}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); it \right) \right) \quad (9.2)$$

where ${}_2\tilde{F}_2$ is the generalized hypergeometric function regularized ${}_2\tilde{F}_2(., .; ., .; .) = \frac{{}_2F_2(a;b;z)}{(\Gamma(b_1)...\Gamma(b_q))}$ and ${}_pF_q(a; b; z)$ has series expansion $\sum_{k=0}^{\infty} \frac{(a_1)_k...(a_p)_k}{(b_1)_k...(b_p)_k} z^k/k!$, were $(a)_{(.)}$ is the Pockhammer symbol.

We can proceed to prove directly from there the convergence in distribution for the average:

$$\lim_{n\to\infty} \Psi_{y^2}(t/n)^n = \exp\left(-\frac{it\left(a^2(p-1) + a(p-1) - b(b+1)p \right)}{(a+b)(a+b+1)} \right)$$

which is that of a degenerate Gaussian (Dirac) with location parameter $-\frac{\left(a^2(p-1) + a(p-1) - b(b+1)p \right)}{(a+b)(a+b+1)}$.

We can finally assess the speed of convergence, the rate at which higher moments map to those of a Gaussian distribution: consider the behavior of the 4th cumulant $\kappa_4 = -i \frac{\partial^4 \log \Psi.(.)}{\partial t^4}\Big|_{t\to 0}$:

1) in the maximum entropy case of $a = b = 1$:

$$\kappa_4\big|_{a=1, b=1} = -\frac{32}{4725 n^3}$$

2) In the maximum variance case, using l'Hôpital:

$$\lim_{\substack{a \to 0 \\ b \to 0}} \kappa_4 = -\frac{(p-1)p(6(p-1)p+1)}{n^3}$$

Se we have $\kappa_4 \underset{n \to \infty}{\to} 0$ at rate n^{-3}.   □

Further, we can extract its probability density function of the Brier score:

$$p(z_t) = (\frac{(p-1)z^{a/2}\left(1-\sqrt{z}\right)^b - p\left(1-\sqrt{z}\right)^a z^{b/2}}{2\left(\sqrt{z}-1\right)zB(a,b)}, \ 0 < z < 1.$$

9.2 THE P/L FOLLOWS A DIFFERENT DISTRIBUTION

Now under consideration is the distribution of $\Gamma_n = \frac{1}{n}\sum_n(\omega_i o_i - f_i(x_i))$, where X_i is the random variable, $f_i(x_i)$ a numerical function of the random variable, w_i the weight and o_i the contractual price (or the numerical value) for the exposure.

If the distribution of X_i is in the power law class and f is linear (seen in the absence of any higher derivative than the first), then, by a standard result, the distribution of Γ_n is in the same class. The proof is standard and the convergence of Γ_n as n gets larger is explained in [166], as well as the properties of the deviation from its mean.

Part IV

INEQUALITY ESTIMATORS

10 | GINI ESTIMATION UNDER INFINITE VARIANCE ‡

HIS CHAPTER is about the problems related to the estimation of the Gini index in presence of a fat-tailed data generating process, i.e. one in the stable distribution class with finite mean but infinite variance (i.e. with tail index $\alpha \in (1, 2)$). We show that, in such a case, the Gini coefficient cannot be reliably estimated using conventional nonparametric methods, because of a downward bias that emerges under fat tails. This has important implications for the ongoing discussion about economic inequality.

We start by discussing how the nonparametric estimator of the Gini index undergoes a phase transition in the symmetry structure of its asymptotic distribution, as the data distribution shifts from the domain of attraction of a light-tailed distribution to that of a fat-tailed one, especially in the case of infinite variance. We also show how the nonparametric Gini bias increases with lower values of α. We then prove that maximum likelihood estimation outperforms nonparametric methods, requiring a much smaller sample size to reach efficiency.

Finally, for fat-tailed data, we provide a simple correction mechanism to the small sample bias of the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution.

10.1 INTRODUCTION

Wealth inequality studies represent a field of economics, statistics and econophysics exposed to fat-tailed data generating processes, often with infinite variance [28, 99]. This is not at all surprising if we recall that the prototype of fat-tailed distributions, the Pareto, has been proposed for the first time to model household incomes [125]. However, the fat-tailedness of data can be problematic in the context of wealth studies, as the property of efficiency (and, partially, consistency) does not necessarily hold for many estimators of inequality and concentration [99?].

The scope of this work is to show how fat tails affect the estimation of one of the most celebrated measures of economic inequality, the Gini index [56, 79, 99], often used (and abused) in the econophysics and economics literature as the main tool for describing the distribution and the concentration of wealth around the world [131? ?].

The literature concerning the estimation of the Gini index is wide and comprehensive (e.g. [56?] for a review), however, strangely enough, almost no attention has been paid to

---
o (With A. Fontanari and P. Cirillo), coauthors

its behavior in presence of fat tails, and this is curious if we consider that: 1) fat tails are ubiquitous in the empirical distributions of income and wealth [99, 131], and 2) the Gini index itself can be seen as a measure of variability and fat-tailedness [55, 57, 58, 69].

The standard method for the estimation of the Gini index is nonparametric: one computes the index from the empirical distribution of the available data using Equation (10.5) below. But, as we show in this paper, this estimator suffers from a downward bias when we deal with fat-tailed observations. Therefore our goal is to close this gap by deriving the limiting distribution of the nonparametric Gini estimator in presence of fat tails, and propose possible strategies to reduce the bias. We show how the maximum likelihood approach, despite the risk of model misspecification, needs much fewer observations to reach efficiency when compared to a nonparametric one.[1]

Our results are relevant to the discussion about wealth inequality, recently rekindled by Thomas Piketty in [131], as the estimation of the Gini index under fat tails and infinite variance may cause several economic analyses to be unreliable, if not markedly wrong. Why should one trust a biased estimator?
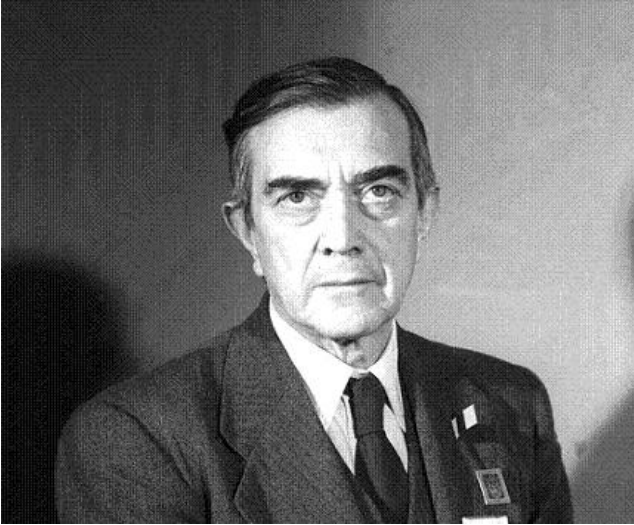
Figure 10.1: *The Italian statistician Corrado Gini, 1884-1965. source: Bocconi*

By fat-tailed data we indicate those data generated by a positive random variable X with cumulative distribution function (c.d.f.) $F(x)$, which is regularly-varying of order α [93], that is, for $\bar{F}(x) := 1 - F(x)$, one has

$$\lim_{x\to\infty} x^{\alpha}\bar{F}(x) = L(x), \tag{10.1}$$

where $L(x)$ is a slowly-varying function such that $\lim_{x\to\infty} \frac{L(cx)}{L(x)} = 1$ with $c > 0$, and where $\alpha > 0$ is called the tail exponent .

Regularly-varying distributions define a large class of random variables whose properties have been extensively studied in the context of extreme value theory [? ?], when dealing with the probabilistic behavior of maxima and minima. As pointed out in [32], regularly-varying and fat-tailed are indeed synonyms. It is known that, if $X_1, ..., X_n$ are i.i.d. observations with a c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (10.1), then their data generating process falls into the maximum domain of attraction of a

---

1 A similar bias also affects the nonparametric measurement of quantile contributions, i.e. those of the type "the top 1% owns x% of the total wealth" [170]. This paper extends the problem to the more widespread Gini coefficient, and goes deeper by making links with the limit theorems.

Fréchet distribution with parameter ρ, in symbols $X \in MDA(\Phi(\rho))$[85]. This means that, for the partial maximum $M_n = \max(X_1, ..., X_n)$, one has

$$P\left(a_n^{-1}(M_n - b_n) \le x\right) \xrightarrow{d} \Phi(\rho) = e^{-x^{-\rho}}, \qquad \rho > 0, \tag{10.2}$$

with $a_n > 0$ and $b_n \in \mathbb{R}$ two normalizing constants. Clearly, the connection between the regularly-varying coefficient α and the Fréchet distribution parameter ρ is given by: $\alpha = \frac{1}{\rho}$ [?].

The Fréchet distribution is one of the limiting distributions for maxima in extreme value theory, together with the Gumbel and the Weibull; it represents the fat-tailed and unbounded limiting case [85]. The relationship between regularly-varying random variables and the Fréchet class thus allows us to deal with a very large family of random variables (and empirical data), and allows us to show how the Gini index is highly influenced by maxima, i.e. extreme wealth, as clearly suggested by intuition [69, 99], especially under infinite variance. Again, this recommends some caution when discussing economic inequality under fat tails.

It is worth remembering that the existence (finiteness) of the moments for a fat-tailed random variable X depends on the tail exponent α, in fact

$$E(X^\delta) < \infty \text{ if } \delta \le \alpha,$$
$$E(X^\delta) = \infty \text{ if } \delta > \alpha. \tag{10.3}$$

In this work, we restrict our focus on data generating processes with finite mean and infinite variance, therefore, according to Equation (10.3), on the class of regularly-varying distributions with tail index $\alpha \in (1, 2)$.

Table 10.1 and Figure 10.2 present numerically and graphically our story, already suggesting its conclusion, on the basis of artificial observations sampled from a Pareto distribution (Equation (10.13) below) with tail parameter α equal to 1.1.

Table 10.1 compares the nonparametric Gini index of Equation (10.5) with the maximum likelihood (ML) tail-based one of Section 10.3. For the different sample sizes in Table 10.1, we have generated 10^8 samples, averaging the estimators via Monte Carlo. As the first column shows, the convergence of the nonparametric estimator to the true Gini value ($g = 0.8333$) is extremely slow and monotonically increasing; this suggests an issue not only in the tail structure of the distribution of the nonparametric estimator but also in its symmetry.

Figure 10.2 provides some numerical evidence that the limiting distribution of the nonparametric Gini index loses its properties of normality and symmetry [68], shifting towards a skewed and fatter-tailed limit, when data are characterized by an infinite variance. As we prove in Section 10.2, when the data generating process is in the domain of attraction of a fat-tailed distribution, the asymptotic distribution of the Gini index becomes a skewed-to-the-right α-stable law. This change of behavior is responsible of the downward bias of the nonparametric Gini under fat tails. However, the knowledge of the new limit allows us to propose a correction for the nonparametric estimator, improving its quality, and thus reducing the risk of badly estimating wealth inequality, with all the possible consequences in terms of economic and social policies [99, 131].

The rest of the paper is organized as follows. In Section 10.2 we derive the asymptotic distribution of the sample Gini index when data possess an infinite variance. In Section 10.3 we deal with the maximum likelihood estimator; in Section 10.4 we provide an illustration with Paretian observations; in Section 10.5 we propose a simple correction based

Table 10.1: *Comparison of the Nonparametric (NonPar) and the Maximum Likelihood (ML) Gini estimators, using Paretian data with tail $\alpha = 1.1$ (finite mean, infinite variance) and different sample sizes. Number of Monte Carlo simulations: 10^8.*

n	Nonpar		ML		Error Ratio[2]
(number of obs.)	Mean	Bias	Mean	Bias	
10^3	0.711	-0.122	0.8333	0	1.4
10^4	0.750	-0.083	0.8333	0	3
10^5	0.775	-0.058	0.8333	0	6.6
10^6	0.790	-0.043	0.8333	0	156
10^7	0.802	-0.031	0.8333	0	10^5+
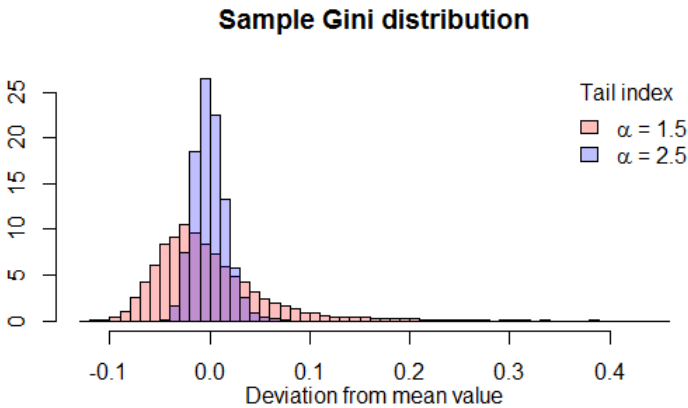
Figure 10.2: *His-tograms for the Gini nonparametric estima-tors for two Paretian (type I) distributions with different tail in-dices, with finite and infinite variance (plots have been centered to ease comparison). Sam-ple size: 10^3. Number of samples: 10^2 for each distribution.*

on the mode-mean distance of the asymptotic distribution of the nonparametric estimator, to take care of its small-sample bias. Section 10.6 closes the paper. A technical Appendix contains the longer proofs of the main results in the work.

10.2 ASYMPTOTICS OF THE NONPARAMETRIC ESTIMATOR UNDER INFINITE VARI-ANCE

We now derive the asymptotic distribution for the nonparametric estimator of the Gini index when the data generating process is fat-tailed with finite mean but infinite variance.

The so-called stochastic representation of the Gini g is

$$g = \frac{1}{2} \frac{\mathbb{E}\left(|X' - X''|\right)}{\mu} \in [0, 1], \tag{10.4}$$

where X' and X'' are i.i.d. copies of a random variable X with c.d.f. $F(x) \in [c, \infty)$, $c > 0$, and with finite mean $\mathbb{E}(X) = \mu$. The quantity $\mathbb{E}\left(|X' - X''|\right)$ is known as the "Gini Mean Difference" (GMD) [?]. For later convenience we also define $g = \frac{\theta}{\mu}$ with $\theta = \frac{\mathbb{E}(|X'-X''|)}{2}$.

The Gini index of a random variable X is thus the mean expected deviation between any two independent realizations of X, scaled by twice the mean [59].

The most common nonparametric estimator of the Gini index for a sample $X_1, ..., X_n$ is defined as

$$G^{NP}(X_n) = \frac{\sum_{1 \leq i < j \leq n} |X_i - X_j|}{(n-1) \sum_{i=1}^{n} X_i}, \tag{10.5}$$

which can also be expressed as

$$G^{NP}(X_n) = \frac{\sum_{i=1}^{n} (2(\frac{i-1}{n-1} - 1) X_{(i)}}{\sum_{i=1}^{n} X_{(i)}} = \frac{\frac{1}{n} \sum_{i=1}^{n} Z_{(i)}}{\frac{1}{n} \sum_{i=1}^{n} X_i}, \tag{10.6}$$

where $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are the ordered statistics of $X_1, ..., X_n$, such that: $X_{(1)} < X_{(2)} < ... < X_{(n)}$ and $Z_{(i)} = 2 \left(\frac{i-1}{n-1} - 1 \right) X_{(i)}$. The asymptotic normality of the estimator in Equation (10.6) under the hypothesis of finite variance for the data generating process is known [99?]. The result directly follows from the properties of the U-statistics and the L-estimators involved in Equation (10.6)

A standard methodology to prove the limiting distribution of the estimator in Equation (10.6), and more in general of a linear combination of order statistics, is to show that, in the limit for $n \to \infty$, the sequence of order statistics can be approximated by a sequence of i.i.d random variables [43, 104]. However, this usually requires some sort of L^2 integrability of the data generating process, something we are not assuming here.

Lemma 10.1 (proved in the Appendix) shows how to deal with the case of sequences of order statistics generated by fat-tailed L^1-only integrable random variables.

Lemma 10.1
Consider the following sequence $R_n = \frac{1}{n} \sum_{i=1}^{n} (\frac{i}{n} - U_{(i)}) F^{-1}(U_{(i)})$ where $U_{(i)}$ are the order statistics of a uniformly distributed i.i.d random sample. Assume that $F^{-1}(U) \in L^1$. Then the following results hold:

$$R_n \xrightarrow{L^1} 0, \tag{10.7}$$

and

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \xrightarrow{L^1} 0, \tag{10.8}$$

with $\alpha \in (1, 2)$ and $L_0(n)$ a slowly-varying function.

10.2.1 A quick recap on α-stable random variables

We here introduce some notation for α-stable distributions, as we need them to study the asymptotic limit of the Gini index.

A random variable X follows an α-stable distribution, in symbols $X \sim S(\alpha, \beta, \gamma, \delta)$, if its characteristic functionis

$$E(e^{itX}) = \begin{cases} e^{-\gamma^\alpha |t|^\alpha (1 - i\beta \operatorname{sign}(t)) \tan(\frac{\pi\alpha}{2}) + i\delta t} & \alpha \neq 1 \\ e^{-\gamma |t| (1 + i\beta \frac{2}{\pi} \operatorname{sign}(t)) \ln |t| + i\delta t} & \alpha = 1 \end{cases},$$

where $\alpha \in (0, 2)$ governs the tail, $\beta \in [-1, 1]$ is the skewness, $\gamma \in \mathbb{R}^+$ is the scale parameter, and $\delta \in \mathbb{R}$ is the location one. This is known as the $S1$ parametrization of α-stable distributions [123, 145].

Interestingly, there is a correspondence between the α parameter of an α-stable random variable, and the α of a regularly-varying random variable as per Equation (10.1): as shown in [68, 123], a regularly-varying random variable of order α is α-stable, with the same tail

coefficient. This is why we do not make any distinction in the use of the α here. Since we aim at dealing with distributions characterized by finite mean but infinite variance, we restrict our focus to $\alpha \in (1, 2)$, as the two α's coincide.

Recall that, for $\alpha \in (1, 2]$, the expected value of an α-stable random variable X is equal to the location parameter δ, i.e. $\mathbb{E}(X) = \delta$. For more details, we refer to [? ?].

The standardized α-stable random variable is expressed as

$$S_{\alpha,\beta} \sim S(\alpha, \beta, 1, 0). \tag{10.9}$$

We note that α-stable distributions are a subclass of infinitely divisible distributions. Thanks to their closure under convolution, they can be used to describe the limiting behavior of (rescaled) partials sums, $S_n = \sum_{i=1}^{n} X_i$, in the General central limit theorem (GCLT) setting [68]. For $\alpha = 2$ we obtain the normal distribution as a special case, which is the limit distribution for the classical CLTs, under the hypothesis of finite variance.

In what follows we indicate that a random variable is in the domain of attraction of an α-stable distribution, by writing $X \in DA(S_\alpha)$. Just observe that this condition for the limit of partial sums is equivalent to the one given in Equation (10.2) for the limit of partial maxima [68?].

10.2.2 The α-stable asymptotic limit of the Gini index

Consider a sample $X_1, ..., X_n$ of i.i.d. observations with a continuous c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (10.1), with tail index $\alpha \in (1, 2)$. The data generating process for the sample is in the domain of attraction of a Fréchet distribution with $\rho \in (\frac{1}{2}, 1)$, given that $\rho = \frac{1}{\alpha}$.

For the asymptotic distribution of the Gini index estimator, as presented in Equation (10.6), when the data generating process is characterized by an infinite variance, we can make use of the following two theorems: Theorem 1 deals with the limiting distribution of the Gini Mean Difference (the numerator in Equation (10.6)), while Theorem 2 extends the result to the complete Gini index. Proofs for both theorems are in the Appendix.

Theorem 1
Consider a sequence $(X_i)_{1 \leq i \leq n}$ of i.i.d random variables from a distribution X on $[c, +\infty)$ with $c > 0$, such that X is in the domain of attraction of an α-stable random variable, $X \in DA(S_\alpha)$, with $\alpha \in (1, 2)$. Then the sample Gini mean deviation (GMD) $\frac{\sum_{i=1}^{n} Z_{(i)}}{n}$ satisfies the following limit in distribution:

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} - \theta \right) \xrightarrow{d} S_{\alpha,1}, \tag{10.10}$$

where $Z_i = (2F(X_i) - 1)X_i$, $\mathbb{E}(Z_i) = \theta$, $L_0(n)$ is a slowly-varying function such that Equation (10.37) holds (see the Appendix), and $S_{\alpha,1}$ is a right-skewed standardized α-stable random variable defined as in Equation (10.9).

Moreover the statistic $\frac{1}{n} \sum_{i=1}^{n} Z_{(i)}$ is an asymptotically consistent estimator for the GMD, i.e. $\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} \xrightarrow{P} \theta$.

Note that Theorem 1 could be restated in terms of the maximum domain of attraction $MDA(\Phi(\rho))$ as defined in Equation (10.2).

Theorem 2
Given the same assumptions of Theorem 1, the estimated Gini index $G^{NP}(X_n) = \frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$ satisfies the following limit in distribution

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(G^{NP}(X_n) - \frac{\theta}{\mu} \right) \xrightarrow{d} Q, \tag{10.11}$$

where $\mathbb{E}(Z_i) = \theta$, $\mathbb{E}(X_i) = \mu$, $L_0(n)$ is the same slowly-varying function defined in Theorem 1 and Q is a right-skewed α-stable random variable $S(\alpha, 1, \frac{1}{\mu}, 0)$.
Furthermore the statistic $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$ is an asymptotically consistent estimator for the Gini index, i.e.
$\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i} \xrightarrow{P} \frac{\theta}{\mu} = g.$

In the case of fat tails with $\alpha \in (1, 2)$, Theorem 2 tells us that the asymptotic distribution of the Gini estimator is always right-skewed notwithstanding the distribution of the underlying data generating process. Therefore heavily fat-tailed data not only induce a fatter-tailed limit for the Gini estimator, but they also change the shape of the limit law, which definitely moves away from the usual symmetric Gaussian. As a consequence, the Gini estimator, whose asymptotic consistency is still guaranteed [104], will approach its true value more slowly, and from below. Some evidence of this was already given in Table 10.1.

10.3 THE MAXIMUM LIKELIHOOD ESTIMATOR

Theorem 2 indicates that the usual nonparametric estimator for the Gini index is not the best option when dealing with infinite-variance distributions, due to the skewness and the fatness of its asymptotic limit. The aim is to find estimators that still preserve their asymptotic normality under fat tails, which is not possible with nonparametric methods, as they all fall into the α-stable Central Limit Theorem case [68?]. Hence the solution is to use parametric techniques.

Theorem 3 shows how, once a parametric family for the data generating process has been identified, it is possible to estimate the Gini index via MLE. The resulting estimator is not just asymptotically normal, but also asymptotically efficient.

In Theorem 3 we deal with random variables X whose distribution belongs to the large and flexible exponential family [147], i.e. whose density can be represented as

$$f_\theta(x) = h(x)e^{(\eta(\theta)T(x) - A(\theta))},$$

with $\theta \in \mathbb{R}$, and where $T(x)$, $\eta(\theta)$, $h(x)$, $A(\theta)$ are known functions.

Theorem 3
Let $X \sim F_\theta$ such that F_θ is a distribution belonging to the exponential family. Then the Gini index obtained by plugging-in the maximum likelihood estimator of θ, $G^{ML}(X_n)_\theta$, is asymptotically normal and efficient. Namely:

$$\sqrt{n} \left(G^{ML}(X_n)_\theta - g_\theta \right) \xrightarrow{D} \mathcal{N} \left(0, g_\theta'^2 I^{-1}(\theta) \right), \tag{10.12}$$

where $g_\theta' = \frac{dg_\theta}{d\theta}$ and $I(\theta)$ is the Fisher Information.

Proof. The result follows easily from the asymptotic efficiency of the maximum likelihood estimators of the exponential family, and the invariance principle of MLE. In particular, the validity of the invariance principle for the Gini index is granted by the continuity and the monotonicity of g_θ with respect to θ. The asymptotic variance is then obtained by application of the delta-method [147]. □

10.4 A PARETIAN ILLUSTRATION

We provide an illustration of the obtained results using some artificial fat-tailed data. We choose a Pareto I [125], with density

$$f(x) = \alpha c^\alpha x^{-\alpha-1}, x \geq c. \tag{10.13}$$

It is easy to verify that the corresponding survival function $\bar{F}(x)$ belongs to the regularly-varying class with tail parameter α and slowly-varying function $L(x) = c^\alpha$. We can therefore apply the results of Section 10.2 to obtain the following corollaries.

Corollary 10.1
Let $X_1, ..., X_n$ be a sequence of i.i.d. observations with Pareto distribution with tail parameter $\alpha \in (1, 2)$. The nonparametric Gini estimator is characterized by the following limit:

$$D_n^{NP} = G^{NP}(X_n) - g \sim S\left(\alpha, 1, \frac{C_\alpha^{-\frac{1}{\alpha}}(\alpha-1)}{n^{\frac{\alpha-1}{\alpha}}}\frac{(\alpha-1)}{\alpha}, 0\right). \tag{10.14}$$

Proof. Without loss of generality we can assume $c = 1$ in Equation (10.13). The results is a mere application of Theorem 2, remembering that a Pareto distribution is in the domain of attraction of α-stable random variables with slowly-varying function $L(x) = 1$. The sequence c_n to satisfy Equation (10.37) becomes $c_n = n^{\frac{1}{\alpha}}C_\alpha^{-\frac{1}{\alpha}}$, therefore we have $L_0(n) = C_\alpha^{-\frac{1}{\alpha}}$, which is independent of n. Additionally the mean of the distribution is also a function of α, that is $\mu = \frac{\alpha}{\alpha-1}$. □

Corollary 10.2
Let the sample $X_1, ..., X_n$ be distributed as in Corollary 10.1, let G_θ^{ML} be the maximum likelihood estimator for the Gini index as defined in Theorem 3. Then the MLE Gini estimator, rescaled by its true mean g, has the following limit:

$$D_n^{ML} = G_\alpha^{ML}(X_n) - g \sim N\left(0, \frac{4\alpha^2}{n(2\alpha-1)^4}\right), \tag{10.15}$$

where N indicates a Gaussian.

Proof. The functional form of the maximum likelihood estimator for the Gini index is known to be $G_\theta^{ML} = \frac{1}{2\alpha^{ML}-1}$ [99]. The result then follows from the fact that the Pareto distribution (with known minimum value x_m) belongs to an exponential family and therefore satisfies the regularity conditions necessary for the asymptotic normality and efficiency of the maximum likelihood estimator. Also notice that the Fisher information for a Pareto distribution is $\frac{1}{\alpha^2}$. □

Now that we have worked out both asymptotic distributions, we can compare the quality of the convergence for both the MLE and the nonparametric case when dealing with

Paretian data, which we use as the prototype for the more general class of fat-tailed observations.

In particular, we can approximate the distribution of the deviations of the estimator from the true value g of the Gini index for finite sample sizes, by using Equations (10.14) and (10.15).

(a) $\alpha = 1.8$

(b) $\alpha = 1.6$

(c) $\alpha = 1.4$

(d) $\alpha = 1.2$
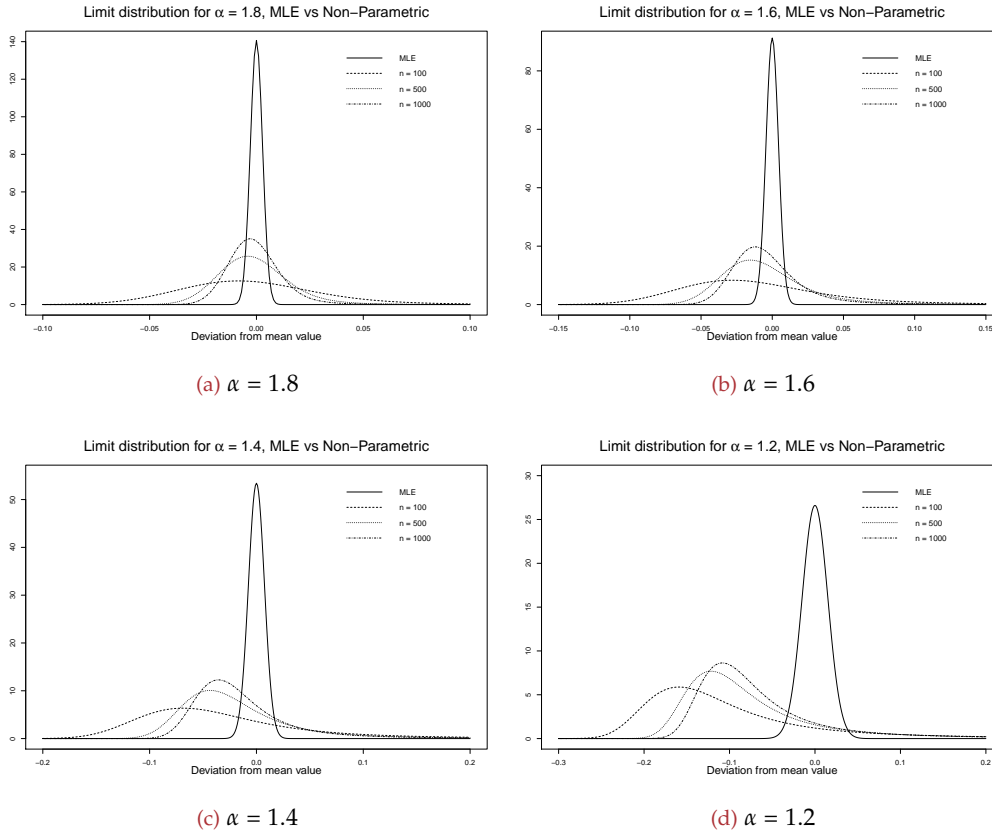
Figure 10.3: *Comparisons between the maximum likelihood and the nonparametric asymptotic distributions for different values of the tail index α. The number of observations for MLE is fixed to $n = 100$. Note that, even if all distributions have mean zero, the mode of the distributions of the nonparametric estimator is different from zero, because of the skewness.*

Figure 10.3 shows how the deviations around the mean of the two different types of estimators are distributed and how these distributions change as the number of observations increases. In particular, to facilitate the comparison between the maximum likelihood and the nonparametric estimators, we fixed the number of observation in the MLE case, while letting them vary in the nonparametric one. We perform this study for different types of tail indices to show how large the impact is on the consistency of the estimator. It is worth noticing that, as the tail index decreases towards 1 (the threshold value for a infinite mean), the mode of the distribution of the nonparametric estimator moves farther away from the mean of the distribution (centered on 0 by definition, given that we are dealing with deviations from the mean). This effect is responsible for the small sample bias observed in applications. Such a phenomenon is not present in the MLE case, thanks to the the normality of the limit for every value of the tail parameter.

We can make our argument more rigorous by assessing the number of observations $\tilde{n}$ needed for the nonparametric estimator to be as good as the MLE one, under different tail scenarios. Let's consider the likelihood-ratio-type function

$$r(c, n) = \frac{P_S(|D_n^{NP}| > c)}{P_N(|D_{100}^{ML}| > c)}, \qquad (10.16)$$

where $P_S(|D_n^{NP}| > c)$ and $P_N(|D_{100}^{ML}| > c)$ are the probabilities (α-stable and Gaussian respectively) of the centered estimators in the nonparametric, and in the MLE cases, of exceeding the thresholds $\pm c$, as per Equations (10.15) and (10.14). In the nonparametric case the number of observations n is allowed to change, while in the MLE case it is fixed to 100. We then look for the value $\tilde{n}$ such that $r(c, \tilde{n}) = 1$ for fixed c.

Table 10.2 displays the results for different thresholds c and tail parameters α. In particular, we can see how the MLE estimator outperforms the nonparametric one, which requires a much larger number of observations to obtain the same tail probability of the MLE with n fixed to 100. For example, we need at least 80×10^6 observations for the nonparametric estimator to obtain the same probability of exceeding the ± 0.02 threshold of the MLE one, when $\alpha = 1.2$.

Table 10.2: *The number of observations $\tilde{n}$ needed for the nonparametric estimator to match the tail probabilities, for different threshold values c and different values of the tail index α, of the maximum likelihood estimator with fixed n = 100.*

α	Threshold c as per Equation (10.16):			
	0.005	0.01	0.015	0.02
1.8	27×10^3	12×10^5	12×10^6	63×10^5
1.5	21×10^4	21×10^4	46×10^5	81×10^7
1.2	33×10^8	67×10^7	20×10^7	80×10^6

Interestingly, the number of observations needed to match the tail probabilities in Equation (10.16) does not vary uniformly with the threshold. This is expected, since as the threshold goes to infinity or to zero, the tail probabilities remain the same for every value of n. Therefore, given the unimodality of the limit distributions, we expect that there will be a threshold maximizing the number of observations needed to match the tail probabilities, while for all the other levels the number of observations will be smaller.

We conclude that, when in presence of fat-tailed data with infinite variance, a plug-in MLE based estimator should be preferred over the nonparametric one.

10.5 SMALL SAMPLE CORRECTION

Theorem 2 can be also used to provide a correction for the bias of the nonparametric estimator for small sample sizes. The key idea is to recognize that, for unimodal distributions, most observations come from around the mode. In symmetric distributions the mode and the mean coincide, thus most observations will be close to the mean value as well, not so for skewed distributions: for right-skewed continuous unimodal distributions the mode is lower than the mean. Therefore, given that the asymptotic distribution of the nonparametric Gini index is right-skewed, we expect that the observed value of the Gini index will be usually lower than the true one (placed at the mean level). We can quantify this difference

(i.e. the bias) by looking at the distance between the mode and the mean, and once this distance is known, we can correct our Gini estimate by adding it back[3].

Formally, we aim to derive a corrected nonparametric estimator $G^C(X_n)$ such that

$$G^C(X_n) = G^{NP}(X_n) + ||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||, \qquad (10.17)$$

where $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ is the distance between the mode m and the mean of the distribution of the nonparametric Gini estimator $G^{NP}(X_n)$.

Performing the type of correction described in Equation (10.17) is equivalent to shifting the distribution of $G^{NP}(X_n)$ in order to place its mode on the true value of the Gini index.

Ideally, we would like to measure this mode-mean distance $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ on the exact distribution of the Gini index to get the most accurate correction. However, the finite distribution is not always easily derivable as it requires assumptions on the parametric structure of the data generating process (which, in most cases, is unknown for fat-tailed data [99]). We therefore propose to use the limiting distribution for the nonparametric Gini obtained in Section 10.2 to approximate the finite sample distribution, and to estimate the mode-mean distance with it. This procedure allows for more freedom in the modeling assumptions and potentially decreases the number of parameters to be estimated, given that the limiting distribution only depends on the tail index and the mean of the data, which can be usually assumed to be a function of the tail index itself, as in the Paretian case where $\mu = \frac{\alpha}{\alpha-1}$.

By exploiting the location-scale property of α-stable distributions and Equation (10.11), we approximate the distribution of $G^{NP}(X_n)$ for finite samples by

$$G^{NP}(X_n) \sim S(\alpha, 1, \gamma(n), g), \qquad (10.18)$$

where $\gamma(n) = \frac{1}{n^{\frac{\alpha-1}{\alpha}}} \frac{L_0(n)}{\mu}$ is the scale parameter of the limiting distribution.

As a consequence, thanks to the linearity of the mode for α-stable distributions, we have

$$||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))|| \approx ||m(\alpha, \gamma(n)) + g - g|| = ||m(\alpha, \gamma(n))||,$$

where $m(\alpha, \gamma(n))$ is the mode function of an α-stable distribution with zero mean.

The implication is that, in order to obtain the correction term, knowledge of the true Gini index is not necessary, given that $m(\alpha, \gamma(n))$ does not depend on g. We then estimate the correction term as

$$\hat{m}(\alpha, \gamma(n)) = \arg \max_x s(x), \qquad (10.19)$$

where $s(x)$ is the numerical density of the associated α-stable distribution in Equation (10.18), but centered on 0. This comes from the fact that, for α-stable distributions, the mode is not available in closed form, but it can be easily computed numerically [123], using the unimodality of the law.

The corrected nonparametric estimator is thus

$$G^C(X_n) = G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)), \qquad (10.20)$$

whose asymptotic distribution is

$$G^C(X_n) \sim S(\alpha, 1, \gamma(n), g + \hat{m}(\alpha, \gamma(n))). \qquad (10.21)$$

---

3 Another idea, which we have tested in writing the paper, is to use the distance between the median and the mean; the performances are comparable.

Note that the correction term $\hat{m}(\alpha, \gamma(n))$ is a function of the tail index α and is connected to the sample size n by the scale parameter $\gamma(n)$ of the associated limiting distribution. It is important to point out that $\hat{m}(\alpha, \gamma(n))$ is decreasing in n, and that $\lim_{n\to\infty} \hat{m}(\alpha, \gamma(n)) \to 0$. This happens because, as n increases, the distribution described in Equation (10.18) becomes more and more centered around its mean value, shrinking to zero the distance between the mode and the mean. This ensures the asymptotic equivalence of the corrected estimator and the nonparametric one. Just observe that

$$
\begin{aligned}
\lim_{n\to\infty} |G(X_n)^C - G^{NP}(X_n)| &= \lim_{n\to\infty} |G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)) - G^{NP}(X_n)| \\
&= \lim_{n\to\infty} |\hat{m}(\alpha, \gamma(n))| \to 0.
\end{aligned}
$$

Naturally, thanks to the correction, $G^C(X_n)$ will always behave better in small samples. Consider also that, from Equation (10.21), the distribution of the corrected estimator has now for mean $g + \hat{m}(\alpha, \gamma(n))$, which converges to the true Gini g as $n \to \infty$.

From a theoretical point of view, the quality of this correction depends on the distance between the exact distribution of $G^{NP}(X_n)$ and its α-stable limit; the closer the two are to each other, the better the approximation. However, given that, in most cases, the exact distribution of $G^{NP}(X_n)$ is unknown, it is not possible to give more details.

From what we have written so far, it is clear that the correction term depends on the tail index of the data, and possibly also on their mean. These parameters, if not assumed to be known a priori, must be estimated. Therefore the additional uncertainty due to the estimation will reflect also on the quality of the correction.

We conclude this Section with the discussion of the effect of the correction procedure with a simple example. In a Monte Carlo experiment, we simulate 1000 Paretian samples of increasing size, from $n = 10$ to $n = 2000$, and for each sample size we compute both the original nonparametric estimator $G^{NP}(X_n)$ and the corrected $G^C(X_n)$. We repeat the experiment for different α's. Figure 10.4 presents the results.

It is clear that the corrected estimators always perform better than the uncorrected ones in terms of absolute deviation from the true Gini value. In particular, our numerical experiment shows that for small sample sizes with $n \leq 1000$ the gain is quite remarkable for all the different values of $\alpha \in (1, 2)$. However, as expected, the difference between the estimators decreases with the sample size, as the correction term decreases both in n and in the tail index α. Notice that, when the tail index equals 2, we obtain the symmetric Gaussian distribution and the two estimators coincide, given that, thanks to the finiteness of the variance, the nonparametric estimator is no longer biased.

10.6 CONCLUSIONS

In this paper we address the issue of the asymptotic behavior of the nonparametric estimator of the Gini index in presence of a distribution with infinite variance, an issue that has been curiously ignored by the literature. The central mistake in the nonparametric methods largely used is to believe that asymptotic consistency translates into equivalent pre-asymptotic properties.

We show that a parametric approach provides better asymptotic results thanks to the properties of maximum likelihood estimation. Hence we strongly suggest that, if the collected data are suspected to be fat-tailed, parametric methods should be preferred.

(a) $\alpha = 1.8$

(b) $\alpha = 1.6$

(c) $\alpha = 1.4$

(d) $\alpha = 1.2$
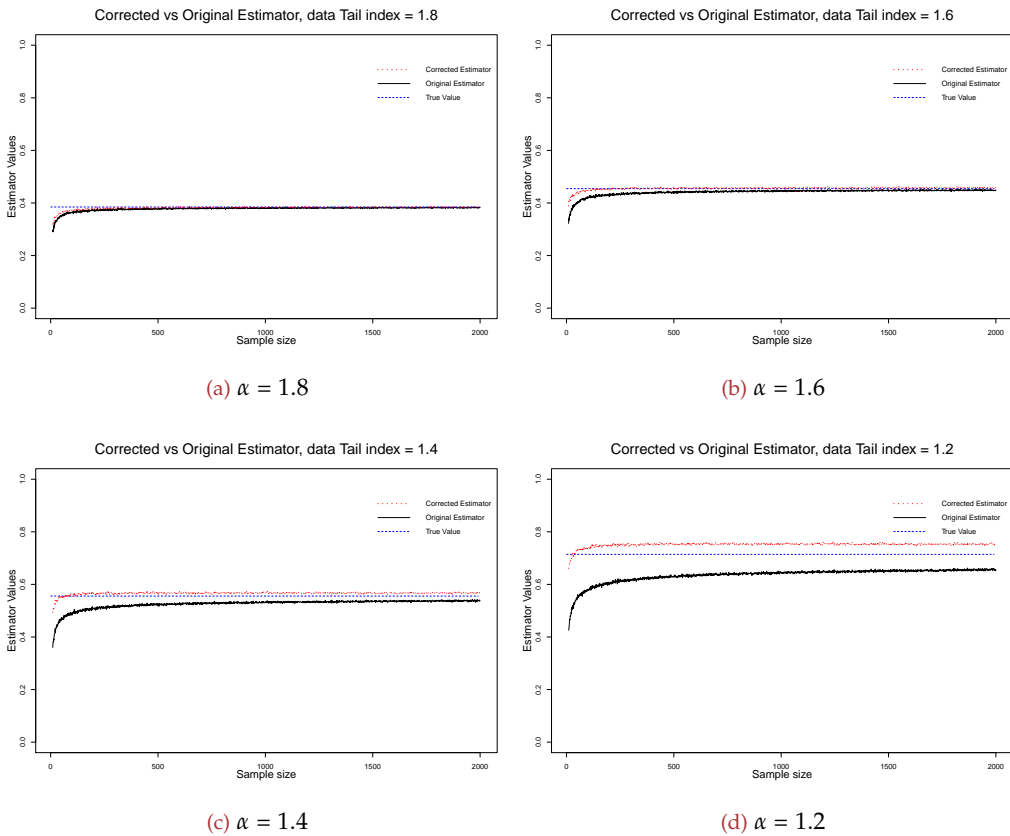
Figure 10.4: *Comparisons between the corrected nonparametric estimator (in red, the one on top) and the usual nonparametric estimator (in black, the one below). For small sample sizes the corrected one clearly improves the quality of the estimation.*

In situations where a fully parametric approach cannot be used, we propose a simple correction mechanism for the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution. Even if the correction works nicely, we suggest caution in its use owing to additional uncertainty from the estimation of the correction term.

TECHNICAL APPENDIX

Proof of Lemma 10.1

Let $U = F(X)$ be the standard uniformly distributed integral probability transform of the random variable X. For the order statistics, we then have [?]: $X_{(i)} \overset{a.s.}{=} F^{-1}(U_{(i)})$. Hence

$$R_n = \frac{1}{n} \sum_{i=1}^{n} (i/n - U_{(i)}) F^{-1}(U_{(i)}). \tag{10.22}$$

Now by definition of empirical c.d.f it follows that

$$R_n = \frac{1}{n} \sum_{i=1}^{n} (F_n(U_{(i)}) - U_{(i)}) F^{-1}(U_{(i)}), \tag{10.23}$$

where $F_n(u) = \frac{1}{n} \sum_{i=1}^{n} 1_{U_i \leq u}$ is the empirical c.d.f of uniformly distributed random variables.

To show that $R_n \overset{L^1}{\longrightarrow} 0$, we are going to impose an upper bound that goes to zero. First we notice that

$$\mathbb{E}|R_n| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|(F_n(U_{(i)}) - U_{(i)}) F^{-1}(U_{(i)})|. \tag{10.24}$$

To build a bound for the right-hand side (r.h.s) of (10.24), we can exploit the fact that, while $F^{-1}(U_{(i)})$ might be just L^1-integrable, $F_n(U_{(i)}) - U_{(i)}$ is L^∞ integrable, therefore we can use Hölder's inequality with $q = \infty$ and $p = 1$. It follows that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|(F_n(U_{(i)}) - U_{(i)}) F^{-1}(U_{(i)})| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})| \mathbb{E}|F^{-1}(U_{(i)})|. \tag{10.25}$$

Then, thanks to the Cauchy-Schwarz inequality, we get

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})| \mathbb{E}|F^{-1}(U_{(i)})|$$

$$\leq \left(\frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})|)^2 \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}(F^{-1}(U_{(i)})))^2 \right)^{\frac{1}{2}}. \tag{10.26}$$

Now, first recall that $\sum_{i=1}^{n} F^{-1}(U_{(i)}) \overset{a.s.}{=} \sum_{i=1}^{n} F^{-1}(U_i)$ with U_i, $i = 1, ..., n$, being an i.i.d sequence, then notice that $\mathbb{E}(F^{-1}(U_i)) = \mu$, so that the second term of Equation (10.26) becomes

$$\mu \left(\frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})|)^2 \right)^{\frac{1}{2}}. \tag{10.27}$$

The final step is to show that Equation (10.27) goes to zero as $n \to \infty$.

We know that F_n is the empirical c.d.f of uniform random variables. Using the triangular inequality the inner term of Equation (10.27) can be bounded as

$$\frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})|)^2 \tag{10.28}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - F(U_{(i)}))|)^2 + \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F(U_{(i)}) - U_{(i)})|)^2.$$

Since we are dealing with uniforms, we known that $F(U) = u$, and the second term in the r.h.s of (10.28) vanishes.

We can then bound $\mathbb{E}(\sup_{U_{(i)}} |(F_n(U_{(i)}) - F(U_{(i)}))|)$ using the so called Vapnik-Chervonenkis (VC) inequality, a uniform bound for empirical processes [21, 42, 183], getting

$$\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - F(U_{(i)}))| \leq \sqrt{\frac{\log(n+1) + \log(2)}{n}}. \tag{10.29}$$

Combining Equation (10.29) with Equation (10.27) we obtain

$$\mu \left(\frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \sup_{U_{(i)}} |(F_n(U_{(i)}) - U_{(i)})|)^2 \right)^{\frac{1}{2}} \leq \mu \sqrt{\frac{\log(n+1) + \log(2)}{n}}, \tag{10.30}$$

which goes to zero as $n \to \infty$, thus proving the first claim.

For the second claim, it is sufficient to observe that the r.h.s of (10.30) still goes to zero when multiplied by $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}$ if $\alpha \in (1, 2)$.

Proof of Theorem 1

The first part of the proof consists in showing that we can rewrite Equation (10.10) as a function of i.i.d random variables in place of order statistics, to be able to apply a Central Limit Theorem (CLT) argument.

Let's start by considering the sequence

$$\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} = \frac{1}{n} \sum_{i=1}^{n} \left(2 \frac{i-1}{n-1} - 1 \right) F^{-1}(U_{(i)}). \tag{10.31}$$

Using the integral probability transform $X \overset{d}{=} F^{-1}(U)$ with U standard uniform, and adding and removing $\frac{1}{n} \sum_{i=1}^{n} (2U_{(i)} - 1) F^{-1}(U_{(i)})$, the r.h.s. in Equation (10.31) can be rewritten as

$$\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} = \frac{1}{n} \sum_{i=1}^{n} (2U_{(i)} - 1) F^{-1}(U_{(i)}) + \frac{1}{n} \sum_{i=1}^{n} 2 \left(\frac{i-1}{n-1} - U_{(i)} \right) F^{-1}(U_{(i)}). \tag{10.32}$$

Then, by using the properties of order statistics [43] we obtain the following almost sure equivalence

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n}(2U_i - 1)F^{-1}(U_i) + \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1} - U_{(i)}\right)F^{-1}(U_{(i)}). \qquad (10.33)$$

Note that the first term in the r.h.s of (10.33) is a function of i.i.d random variables as desired, while the second term is just a reminder, therefore

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n,$$

with $Z_i = (2U_i - 1)F^{-1}(U_i)$ and $R_n = \frac{1}{n}\sum_{i=1}^{n}(2(\frac{i-1}{n-1} - U_{(i)}))F^{-1}(U_{(i)})$.

Given Equation (10.10) and exploiting the decomposition given in (10.33) we can rewrite our claim as

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} - \theta\right) = \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta\right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}R_n. \qquad (10.34)$$

From the second claim of the Lemma 10.1 and Slutsky Theorem, the convergence in Equation (10.10) can be proven by looking at the behavior of the sequence

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta\right), \qquad (10.35)$$

where $Z_i = (2U_i - 1)F^{-1}(U_i) = (2F(X_i) - 1)X_i$. This reduces to proving that Z_i is in the fat tails domain of attraction.

Recall that by assumption $X \in DA(S_\alpha)$ with $\alpha \in (1,2)$. This assumption enables us to use a particular type of CLT argument for the convergence of the sum of fat-tailed random variables. However, we first need to prove that $Z \in DA(S_\alpha)$ as well, that is $P(|Z| > z) \sim L(z)z^{-\alpha}$, with $\alpha \in (1,2)$ and $L(z)$ slowly-varying.

Notice that

$$P(|\tilde{Z}| > z) \leq P(|Z| > z) \leq P(2X > z),$$

where $\tilde{Z} = (2U - 1)X$ and $U \perp X$. The first bound holds because of the positive dependence between X and $F(X)$ and it can be proven rigorously by noting that $2UX \leq 2F(X)X$ by the so-called re-arrangement inequality [87]. The upper bound conversely is trivial.

Using the properties of slowly-varying functions, we have $P(2X > z) \sim 2^\alpha L(z)z^{-\alpha}$. To show that $\tilde{Z} \in DA(S_\alpha)$, we use the Breiman's Theorem, which ensure the stability of the α-stable class under product, as long as the second random variable is not too fat-tailed [190].

To apply the Theorem we re-write $P(|\tilde{Z}| > z)$ as

$$P(|\tilde{Z}| > z) = P(\tilde{Z} > z) + P(-\tilde{Z} > z) = P(\tilde{U}X > z) + P(-\tilde{U}X > z),$$

where $\tilde{U}$ is a standard uniform with $\tilde{U} \perp X$.

We focus on $P(\tilde{U}X > z)$ since the procedure is the same for $P(-\tilde{U}X > z)$. We have

$$P(\tilde{U}X > z) = P(\tilde{U}X > z|\tilde{U} > 0)P(\tilde{U} > 0) + P(\tilde{U}X > z|\tilde{U} \leq 0)P(\tilde{U} \leq 0),$$

for $z \to +\infty$.

Now, we have that $P(\tilde{U}X > z | \tilde{U} \leq 0) \to 0$, while, by applying Breiman's Theorem, $P(\tilde{U}X > z | \tilde{U} > 0)$ becomes

$$P(\tilde{U}X > z | \tilde{U} > 0) \to E(\tilde{U}^\alpha | U > 0)P(X > z)P(U > 0).$$

Therefore

$$P(|\tilde{Z}| > z) \to \frac{1}{2}E(\tilde{U}^\alpha | U > 0)P(X > z) + \frac{1}{2}E((-\tilde{U})^\alpha | U \leq 0)P(X > z).$$

From this

$$P(|\tilde{Z}| > z) \to \frac{1}{2}P(X > z)[E(\tilde{U})^\alpha | U > 0) + E((-\tilde{U}^\alpha | U \leq 0)]$$

$$= \frac{2^\alpha}{1-\alpha}P(X > z) \sim \frac{2^\alpha}{1-\alpha}L(z)z^{-\alpha}.$$

We can then conclude that, by the squeezing Theorem [?],

$$P(|Z| > z) \sim L(z)z^{-\alpha},$$

as $z \to \infty$. Therefore $Z \in DA(S_\alpha)$.

We are now ready to invoke the Generalized Central Limit Theorem (GCLT) [?] for the sequence Z_i, i.e.

$$nc_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}Z_i - \mathbb{E}(Z_i)\right) \xrightarrow{d} S_{\alpha,\beta}. \tag{10.36}$$

with $\mathbb{E}(Z_i) = \theta$, $S_{\alpha,\beta}$ a standardized α-stable random variable, and where c_n is a sequence which must satisfy

$$\lim_{n\to\infty}\frac{nL(c_n)}{c_n^\alpha} = \frac{\Gamma(2-\alpha)|\cos(\frac{\pi\alpha}{2})|}{\alpha - 1} = C_\alpha. \tag{10.37}$$

Notice that c_n can be represented as $c_n = n^{\frac{1}{\alpha}}L_0(n)$, where $L_0(n)$ is another slowly-varying function possibly different from $L(n)$.

The skewness parameter β is such that

$$\frac{P(Z > z)}{P(|Z| > z)} \to \frac{1+\beta}{2}.$$

Recalling that, by construction, $Z \in [-c, +\infty)$, the above expression reduces to

$$\frac{P(Z > z)}{P(Z > z) + P(-Z > z)} \to \frac{P(Z > z)}{P(Z > z)} = 1 \to \frac{1+\beta}{2}, \tag{10.38}$$

therefore $\beta = 1$. This, combined with Equation (10.34), the result for the reminder R_n of Lemma 10.1 and Slutsky Theorem, allows us to conclude that the same weak limits holds for the ordered sequence of $Z_{(i)}$ in Equation (10.10) as well.

Proof of Theorem 2

The first step of the proof is to show that the ordered sequence $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$, characterizing the Gini index, is equivalent in distribution to the i.i.d sequence $\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i}$. In order to prove this, it is sufficient to apply the factorization in Equation (10.33) to Equation (10.11), getting

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i} - \frac{\theta}{\mu} \right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \frac{n}{\sum_{i=1}^{n} X_i}. \tag{10.39}$$

By Lemma 10.1 and the application of the continuous mapping and Slutsky Theorems, the second term in Equation (10.39) goes to zero at least in probability. Therefore to prove the claim it is sufficient to derive a weak limit for the following sequence

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)} \left(\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i} - \frac{\theta}{\mu} \right). \tag{10.40}$$

Expanding Equation (10.40) and recalling that $Z_i = (2F(X_i) - 1)X_i$, we get

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{n}{\sum_{i=1}^{n} X_i} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right) \right). \tag{10.41}$$

The term $\frac{n}{\sum_{i=1}^{n} X_i}$ in Equation (10.41) converges in probability to $\frac{1}{\mu}$ by an application of the continuous mapping Theorem, and the fact that we are dealing with positive random variables X. Hence it will contribute to the final limit via Slutsky Theorem.

We first start by focusing on the study of the limit law of the term

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^{n} X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right). \tag{10.42}$$

Set $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ and note that $\mathbb{E}(\hat{Z}_i) = 0$, since $\mathbb{E}(Z_i) = \theta$ and $\mathbb{E}(X_i) = \mu$.

In order to apply a GCLT argument to characterize the limit distribution of the sequence $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i$ we need to prove that $\hat{Z} \in DA(S_\alpha)$. If so then we can apply GCLT to

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{\sum_{i=1}^{n} \hat{Z}_i}{n} - \mathbb{E}(\hat{Z}_i) \right). \tag{10.43}$$

Note that, since $\mathbb{E}(\hat{Z}_i) = 0$, Equation (10.43) equals Equation (10.42).

To prove that $\hat{Z} \in DA(S_\alpha)$, remember that $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ is just $Z_i = X_i(2F(X_i) - 1)$ shifted by $\frac{\theta}{\mu}$. Therefore the same argument used in Theorem 1 for Z applies here to show that $\hat{Z} \in DA(S_\alpha)$. In particular we can point out that $\hat{Z}$ and Z (therefore also X) share the same α and slowly-varying function $L(n)$.

Notice that by assumption $X \in [c, \infty)$ with $c > 0$ and we are dealing with continuous distributions, therefore $\hat{Z} \in [-c(1 + \frac{\theta}{\mu}), \infty)$. As a consequence the left tail of $\hat{Z}$ does not contribute to changing the limit skewness parameter β, which remains equal to 1 (as for Z) by an application of Equation (10.38).

Therefore, by applying the GCLT we finally get

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)} \left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu} \right) \xrightarrow{d} \frac{1}{\mu} S(\alpha, 1, 1, 0). \tag{10.44}$$

We conclude the proof by noting that, as proven in Equation (10.39), the weak limit of the Gini index is characterized by the i.i.d sequence of $\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i}$ rather than the ordered one, and that an α-stable random variable is closed under scaling by a constant [145].

11 | ON THE SUPER–ADDITIVITY AND ESTIMATION BIASES OF QUANTILE CONTRIBUTIONS ‡

AMPLE MEASURES[a] 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.

---

a With R. Douady

11.1 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 [109] and [110] 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 $X \in [x_{\min}, \infty)$,

Figure 11.1: *The young Vil-fredo Pareto, before he discov-ered power laws.*

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

For a given sample $(X_k)_{1 \le k \le n}$, 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 :

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

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.

11.2 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:

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{11.1}$$

where $L : [x_{\min}, +\infty) \rightarrow (0, +\infty)$ is a slowly varying function, defined as $\lim_{x\rightarrow+\infty} \frac{L(kx)}{L(x)} = 1$ 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,[40],[41] Singh-Maddala distribution [148], or straight Pareto.

For exponents $1 \leq \alpha \leq 2$, as observed in [?], 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 $\alpha = 1$.

11.2.1 Bias and Convergence

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:

$$\kappa_q = \frac{\int_{h(q)}^\infty x\, \phi(x)dx}{\int_{x_{min}}^\infty x\, \phi(x)dx} = \left(\frac{h(q)}{x_{\min}}\right)^{1-\alpha} = q^{\frac{\alpha-1}{\alpha}} \tag{11.2}$$

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

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

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 11.1 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

---

1 This value, which is lower than the estimated exponents one can find in the literature – around 2 – is, following [64], a lower estimate which cannot be excluded from the observations.

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 11.1: *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.

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

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

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 11.1

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

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

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

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.
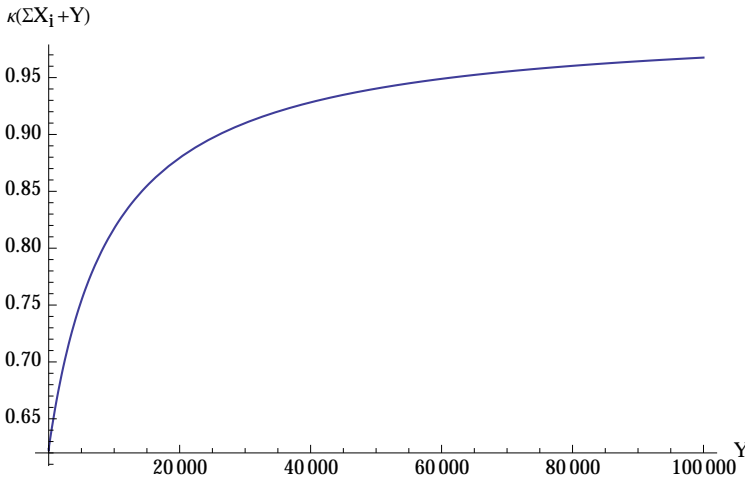
Figure 11.2: *Effect of additional observations on κ*
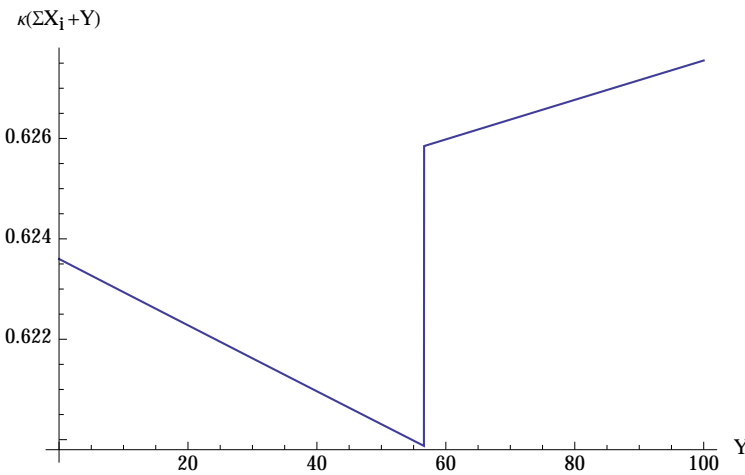
Figure 11.3: *Effect of additional observations on κ, we can see convexity on both sides of h except for values of no effect to the left of h, an area of order $1/n$*

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 $(\hat{\alpha}, \hat{\lambda})$ and only then, estimating the theoretical tail contribution $\kappa_q(\hat{\alpha}, \hat{\lambda})$. Falk

[64] 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 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.

11.3 AN INEQUALITY ABOUT AGGREGATING INEQUALITY

For the estimation of the mean of a fat-tailed r.v. $(X)^j_i$, 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 $\widehat{\kappa}_q(n_i)$.

Theorem 11.1
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 subsamples, 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 $(X_1, \ldots, X_m, X'_1, \ldots, X'_n)$.

---

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

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:

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

and, similarly $A' = \max_{J' \subset \{1,...,n\}, |J'| = qn} \sum_{i \in J'} X'_i$ and $A'' = \max_{J'' \subset \{1,...,m+n\}, |J''| = q(m+n)} \sum_{i \in J''} X_i$, 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 $T = X_{[mq]}$ be the cut-off point (where $[mq]$ 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:

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

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

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 $X = (X_1, \ldots, X_n)$ of i.i.d. variables with the same distribution as X.

Theorem 11.2

For any $n \in \mathbb{N}$, we have:

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

and

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

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. □

11.4 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:

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

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 15.1:

$$\mathbb{P}(X > x) \sim \sum_{j=1}^{N} \omega_i L_i(x) x^{-\alpha_j} \tag{11.3}$$

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:

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

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.

11.5   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 11.1, $\widehat{\kappa}_q$ and S are positively correlated.

For the case in which the random variables under concern are wealth, we observe as in Figure 11.4 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.)
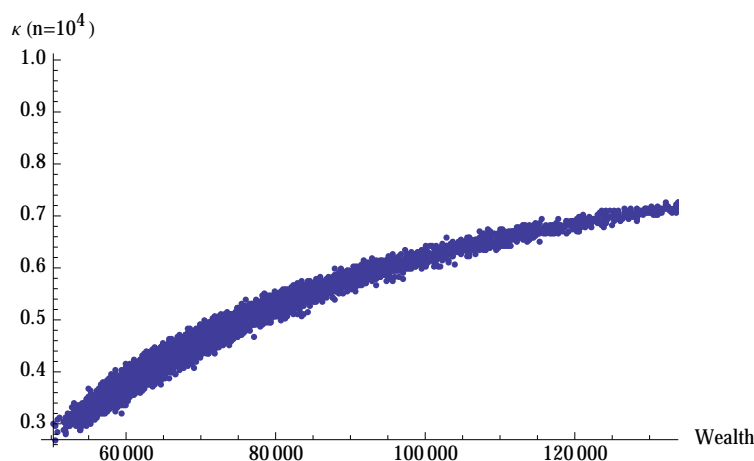
Figure 11.4: *Effect of additional wealth on $\widehat{\kappa}$*

11.6   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]

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 [?].

---

3 Accumulated wealth is typically thicker tailed than income, see [72].

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 [131], 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 [134] 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.

11.6.1 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, tend to fail with fat-tailed data, see [?]. 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.

11.6.2 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

---

4 *Financial Times*, May 23, 2014 "Piketty findings undercut by errors" by Chris Giles.
5 Using Richardson's data, [134]: "(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.

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

---

6 Also note that, in addition to the centile estimation problem, some authors such as [132] 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 [?].

Part V

SHADOW MOMENTS PAPERS

12 | ON THE SHADOW MOMENTS OF APPARENTLY INFINITE–MEAN PHENOMENA (WITH P. CIRILLO)‡

THIS CHAPTER proposes an approach to compute the conditional moments of fat-tailed phenomena that, only looking at data, could be mistakenly considered as having infinite mean. This type of problems manifests itself when a random variable Y has a heavy-tailed distribution with an extremely wide yet bounded support.

We introduce the concept of dual distribution, by means of a log-transformation that smoothly removes the upper bound. The tail of the dual distribution can then be studied using extreme value theory, without making excessive parametric assumptions, and the estimates one obtains can be used to study the original distribution and compute its moments by reverting the transformation.

The central difference between our approach and a simple truncation is in the smoothness of the transformation between the original and the dual distribution, allowing use of extreme value theory.

War casualties, operational risk, environment blight, complex networks and many other econophysics phenomena are possible fields of application.

12.1 INTRODUCTION

Consider a heavy-tailed random variable Y with finite support $[L, H]$. W.l.o.g. set $L >> 0$ for the lower bound, while for upper one H, assume that its value is remarkably large, yet finite. It is so large that the probability of observing values in its vicinity is extremely small, so that in data we tend to find observations only below a certain $M << H < \infty$.

Figure 12.1 gives a graphical representation of the problem. For our random variable Y with remote upper bound H the real tail is represented by the continuous line. However, if we only observe values up to $M << H$, and - willing or not - we ignore the existence of H, which is unlikely to be seen, we could be inclined to believe the the tail is the dotted one, the apparent one. The two tails are indeed essentially indistinguishable for most cases, as the divergence is only evident when we approach H.

Now assume we want to study the tail of Y and, since it is fat-tailed and despite $H < \infty$, we take it to belong to the so-called Fréchet class[1]. In extreme value theory [124], a distribution F of a random variable Y is said to be in the Fréchet class if $\bar{F}(y) = 1 - F(y) =$

---

[1] Note that treating Y as belonging to the Fréchet class is a mistake. If a random variable has a finite upper bound, it cannot belong to the Fréchet class, but rather to the Weibull class [85].

$y^{-\alpha}L(y)$, where $L(y)$ is a slowly varying function. In other terms, the Fréchet class is the class of all distributions whose right tail behaves as a power law.

Looking at the data, we could be led to believe that the right tail is the dotted line in Figure 12.1, and our estimation of α shows it be smaller than 1. Given the properties of power laws, this means that $E[Y]$ is not finite (as all the other higher moments). This also implies that the sample mean is essentially useless for making inference, in addition to any considerations about robustness [114]. But if H is finite, this cannot be true: all the moments of a random variable with bounded support are finite.

A solution to this situation could be to fit a parametric model, which allows for fat tails and bounded support, such as for example a truncated Pareto [1]. But what happens if Y only shows a Paretian behavior in the upper tail, and not for the whole distribution? Should we fit a mixture model?

In the next section we propose a simple general solution, which does not rely on strong parametric assumptions.
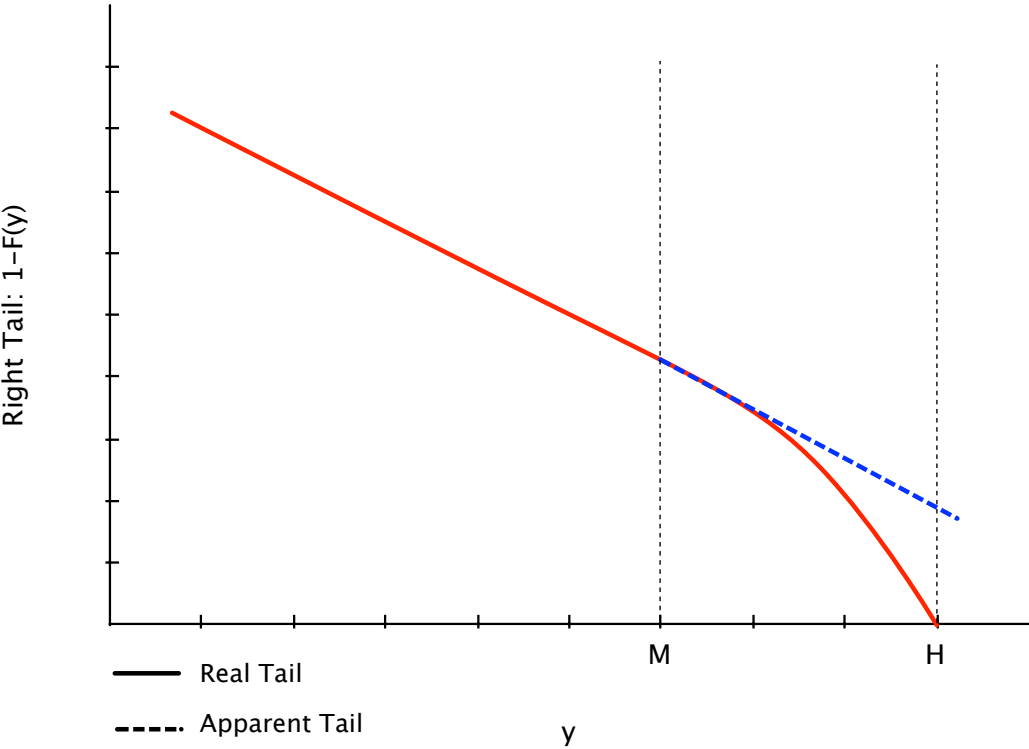
Figure 12.1: *Graphical representation of what may happen if one ignores the existence of the finite upper bound H, since only M is observed.*

12.2 THE DUAL DISTRIBUTION

Instead of altering the tails of the distribution we find it more convenient to transform the data and rely on distributions with well known properties. In Figure 12.1, the real and the apparent tails are indistinguishable to a great extent. We can use this fact to our advantage, by transforming Y to remove its upper bound H, so that the new random variable Z - the dual random variable - has the same tail as the apparent tail. We can then estimate the

shape parameter α of the tail of Z and come back to Y to compute its moments or, to be more exact, to compute its excess moments, the conditional moments above a given threshold, view that we will just extract the information from the tail of Z.

Take Y with support $[L, H]$, and define the function

$$\varphi(Y) = L - H \log \left(\frac{H - Y}{H - L} \right).$$ (12.1)

We can verify that φ is "smooth": $\varphi \in C^\infty$, $\varphi^{-1}(\infty) = H$, and $\varphi^{-1}(L) = \varphi(L) = L$. Then $Z = \varphi(Y)$ defines a new random variable with lower bound L and an infinite upper bound. Notice that the transformation induced by $\varphi(\cdot)$ does not depend on any of the parameters of the distribution of Y.

By construction, $z = \varphi(y) \approx y$ for very large values of H. This means that for a very large upper bound, unlikely to be touched, the results we get for the tail of Y and $Z = \varphi(Y)$ are essentially the same, until we do not reach H. But while Y is bounded, Z is not. Therefore we can safely model the unbounded dual distribution of Z as belonging to the Fréchet class, study its tail, and then come back to Y and its moments, which under the dual distribution of Z could not exist.[2]

The tail of Z can be studied in different ways, see for instance [124] and [65]. Our suggestions is to rely on the so-called de Pickands, Balkema and de Haan's Theorem [85]. This theorem allows us to focus on the right tail of a distribution, without caring too much about what happens below a given threshold threshold u. In our case $u \geq L$.

Consider a random variable Z with distribution function G, and call G_u the conditional df of Z above a given threshold u. We can then define the r.v. W, representing the rescaled excesses of Z over the threshold u, so that

$$G_u(w) = P(Z - u \leq w | Z > u) = \frac{G(u + w) - G(u)}{1 - G(u)},$$

for $0 \leq w \leq z_G - u$, where z_G is the right endpoint of G.

Pickands, Balkema and de Haan have showed that for a large class of distribution functions G, and a large u, G_u can be approximated by a Generalized Pareto distribution, i.e. $G_u(w) \to GPD(w; \xi, \sigma)$, as $u \to \infty$ where

$$GPD(w; \xi, \sigma) = \begin{cases} 1 - (1 + \xi \frac{w}{\sigma})^{-1/\xi} & if \ \xi \neq 0 \\ 1 - e^{-\frac{w}{\sigma}} & if \ \xi = 0 \end{cases}, \ w \geq 0.$$ (12.2)

The parameter ξ, known as the shape parameter, and corresponding to $1/\alpha$, governs the fatness of the tails, and thus the existence of moments. The moment of order p of a Generalized Pareto distributed random variable only exists if and only if $\xi < 1/p$, or $\alpha > p$ [124]. Both ξ and σ can be estimated using MLE or the method of moments [85].[3]

12.3 BACK TO y: THE SHADOW MEAN (OR POPULATION MEAN)

With f and g, we indicate the densities of Y and Z.

---

2 Note that the use of logarithmic transformation is quite natural in the context of utility.

3 There are alternative methods to face finite (or concave) upper bounds, i.e., the use of tempered power laws (with exponential dampening)[137] or stretched exponentials [102]; while being of the same nature as our exercise, these methods do not allow for immediate applications of extreme value theory or similar methods for parametrization.

We know that $Z = \varphi(Y)$, so that $Y = \varphi^{-1}(Z) = (L - H)e^{\frac{L-Z}{H}} + H$.

Now, let's assume we found $u = L^* \geq L$, such that $G_u(w) \approx \text{GPD}(w; \xi, \sigma)$. This implies that the tail of Y, above the same value L^* that we find for Z, can be obtained from the tail of Z, i.e. G_u.

First we have

$$\int_{L^*}^{\infty} g(z)\,dz = \int_{L^*}^{\varphi^{-1}(\infty)} f(y)\,dy. \tag{12.3}$$

And we know that

$$g(z; \xi, \sigma) = \frac{1}{\sigma}\left(1 + \frac{\xi z}{\sigma}\right)^{-\frac{1}{\xi}-1}, \qquad z \in [L^*, \infty). \tag{12.4}$$

Setting $\alpha = \xi^{-1}$, we get

$$f(y; \alpha, \sigma) = \frac{H\left(1 + \frac{H(\log(H-L)-\log(H-y))}{\alpha\sigma}\right)^{-\alpha-1}}{\sigma(H-y)}, \quad y \in [L^*, H], \tag{12.5}$$

or, in terms of distribution function,

$$F(y; \alpha, \sigma) = 1 - \left(1 + \frac{H(\log(H-L)-\log(H-y))}{\alpha\sigma}\right)^{-\alpha}. \tag{12.6}$$

Clearly, given that φ is a one-to-one transformation, the parameters of f and g obtained by maximum likelihood methods will be the same —the likelihood functions of f and g differ by a scaling constant.

We can derive the shadow mean[4] of Y, conditionally on $Y > L^*$, as

$$E[Y|Y > L^*] = \int_{L^*}^{H} y\,f(y; \alpha, \sigma)\,dy, \tag{12.7}$$

obtaining

$$E[Y|Z > L^*] = (H - L^*)e^{\frac{\alpha\sigma}{H}}\left(\frac{\alpha\sigma}{H}\right)^{\alpha}\Gamma\left(1 - \alpha, \frac{\alpha\sigma}{H}\right) + L^*. \tag{12.8}$$

The conditional mean of Y above $L^* \geq L$ can then be estimated by simply plugging in the estimates $\hat{\alpha}$ and $\hat{\sigma}$, as resulting from the GPD approximation of the tail of Z. It is worth noticing that if $L^* = L$, then $E[Y|Y > L^*] = E[Y]$, i.e. the conditional mean of Y above Y is exactly the mean of Y.

Naturally, in a similar way, we can obtain the other moments, even if we may need numerical methods to compute them.

Our method can be used in general, but it is particularly useful when, from data, the tail of Y appears so fat that no single moment is finite, as it is often the case when dealing with operational risk losses, the degree distribution of large complex networks, or other econophysical phenomena.

For example, assume that for Z we have $\xi > 1$. Then both $E[Z|Z > L^*]$ and $E[Z]$ are not finite[5]. Figure 12.1 tells us that we might be inclined to assume that also $E[Y]$ is infinite -

---

4 We call the population average –as opposed to the sample one – "shadow", as it is not immediately visible from the data.

5 Remember that for a GPD random variable Z, $E[Z^p] < \infty$ iff $\xi < 1/p$.

and this is what the data are likely to tell us if we estimate $\hat{\xi}$ from the tail[6] of Y. But this cannot be true because $H < \infty$, and even for $\xi > 1$ we can compute the expected value $E[Y|Z > L^*]$ using equation (12.8).

Value-at-Risk and Expected Shortfall

Thanks to equation (12.6), we can compute by inversion the quantile function of Y when $Y \geq L^*$, that is

$$Q(p; \alpha, \sigma, H, L) = e^{-\gamma(p)}\left(L^* e^{\frac{\alpha\sigma}{H}} + He^{\gamma(p)} - He^{\frac{\alpha\sigma}{H}}\right), \tag{12.9}$$

where $\gamma(p) = \frac{\alpha\sigma(1-p)^{-1/\alpha}}{H}$ and $p \in [0,1]$. Again, this quantile function is conditional on Y being larger than L^*.

From equation (12.9), we can easily compute the Value-at-Risk (*VaR*) of $Y|Y \geq L^*$ for whatever confidence level. For example, the 95% *VaR* of Y, if Y represents operational losses over a 1-year time horizon, is simply $VaR^Y_{0.95} = Q(0.95; \alpha, \sigma, H, L)$.

Another quantity we might be interested in when dealing with the tail risk of Y is the so-called expected shortfall (*ES*), that is $E[Y|Y > u \geq L^*]$. This is nothing more than a generalization of equation (12.8).

We can obtain the expected shortfall by first computing the mean excess function of $Y|Y \geq L^*$, defined as

$$e_u(Y) = E[Y - u|Y > u] = \frac{\int_u^\infty (u - y)f(y; \alpha, \sigma)dy}{1 - F(u)},$$

for $y \geq u \geq L^*$. Using equation (12.5), we get

$$
\begin{aligned}
e_u(Y) \quad = \quad & (H - L)e^{\frac{\alpha\sigma}{H}}\left(\frac{\alpha\sigma}{H}\right)^\alpha \left(\frac{H\log\left(\frac{H-L}{H-u}\right)}{\alpha\sigma} + 1\right)^\alpha \times \\
& \Gamma\left(1 - \alpha, \frac{\alpha\sigma}{H} + \log\left(\frac{H-L}{H-u}\right)\right).
\end{aligned} \tag{12.10}
$$

The Expected Shortfall is then simply computed as

$$E[Y|Y > u \geq L^*] = e_u(Y) + u.$$

As in finance and risk management, ES and VaR can be combined. For example we could be interested in computing the 95% ES of Y when $Y \geq L^*$. This is simply given by $VaR^Y_{0.95} + e_{VaR^Y_{0.95}}(Y)$.

12.4 COMPARISON TO OTHER METHODS

There are three ways to go about explicitly cutting a Paretian distribution in the tails (not counting methods to stretch or "temper" the distribution).

---

6 Because of the similarities between $1 - F(y)$ and $1 - G(z)$, at least up until M, the GPD approximation will give two statistically undistinguishable estimates of ξ for both tails [124].

1) The first one consists in hard truncation, i.e. in setting a single endpoint for the distribution and normalizing. For instance the distribution would be normalized between L and H, distributing the excess mass across all points.

2) The second one would assume that H is an absorbing barrier, that all the realizations of the random variable in excess of H would be compressed into a Dirac delta function at H –as practiced in derivative models. In that case the distribution would have the same density as a regular Pareto except at point H.

3) The third is the one presented here.

The same problem has cropped up in quantitative finance over the use of truncated normal (to correct for Bachelier's use of a straight Gaussian) vs. logarithmic transformation (Sprenkle, 1961 [150]), with the standard model opting for logarithmic transformation and the associated one-tailed lognormal distribution. Aside from the additivity of log-returns and other such benefits, the models do not produce a "cliff", that is an abrupt change in density below or above, with the instability associated with risk measurements on non-smooth function.

As to the use of extreme value theory, Breilant et al. (2014)[?] go on to truncate the distribution by having an excess in the tails with the transformation $Y^{-\alpha} \to (Y^{-\alpha} - H^{-\alpha})$ and apply EVT to the result. Given that the transformation includes the estimated parameter, a new MLE for the parameter α is required. We find issues with such a non-smooth transformation. The same problem occurs as with financial asset models, particularly the presence an abrupt "cliff" below which there is a density, and above which there is none. The effect is that the expectation obtained in such a way will be higher than ours, particularly at values of $\alpha < 1$, as seen in Figure ??.

We can demonstrate the last point as follows. Assume we observe distribution is Pareto that is in fact truncated but treat it as a Pareto. The density is $f(x) = \frac{1}{\sigma} \left(\frac{x-L}{\alpha\sigma} + 1 \right)^{-\alpha-1}, x \in$ $[L, \infty)$. The truncation gives $g(x) = \frac{\left(\frac{x-L}{\alpha\sigma} + 1 \right)^{-\alpha-1}}{\sigma(1 - \alpha^\alpha\sigma^\alpha(\alpha\sigma+H-L)^{-\alpha})}, x \in [L, H]$.

Moments of order p of the truncated Pareto (i.e. what is seen from realizations of the process), $M(p)$ are:

$$M(p) = \alpha e^{-i\pi p}(\alpha\sigma)^\alpha(\alpha\sigma - L)^{p-\alpha}$$

$$\frac{\left(B_{\frac{H}{L-\alpha\sigma}}(p+1, -\alpha) - B_{\frac{L}{L-\alpha\sigma}}(p+1, -\alpha) \right)}{\left(\frac{\alpha\sigma}{\alpha\sigma+H-L} \right)^\alpha - 1} \tag{12.11}$$

where $B(.,.)$ is the Euler Beta function, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1} dt$.

We end up with $r(H, \alpha)$, the ratio of the mean of the soft truncated to that of the truncated Pareto.

$$r(H, \alpha) = e^{-\frac{\alpha}{H}} \left(\frac{\alpha}{H} \right)^\alpha \left(\frac{\alpha}{\alpha+H} \right)^{-\alpha} \left(\frac{\alpha+H}{\alpha} \right)^{-\alpha}$$

$$\frac{\left(-\left(\frac{\alpha+H}{\alpha} \right)^\alpha + H + 1 \right)}{(\alpha-1) \left(\left(\frac{\alpha}{H} \right)^\alpha - \left(\frac{\alpha+H}{H} \right)^\alpha \right) E_\alpha \left(\frac{\alpha}{H} \right)} \tag{12.12}$$

where $E_\alpha \left(\frac{\alpha}{H} \right)$ is the exponential integral $e_\alpha z = \int_1^\infty \frac{e^{t(-\alpha)}}{t^n} dt$.

$$\frac{E[X_{\text{smooth}}]}{E[X_{\text{truncated}}]}$$
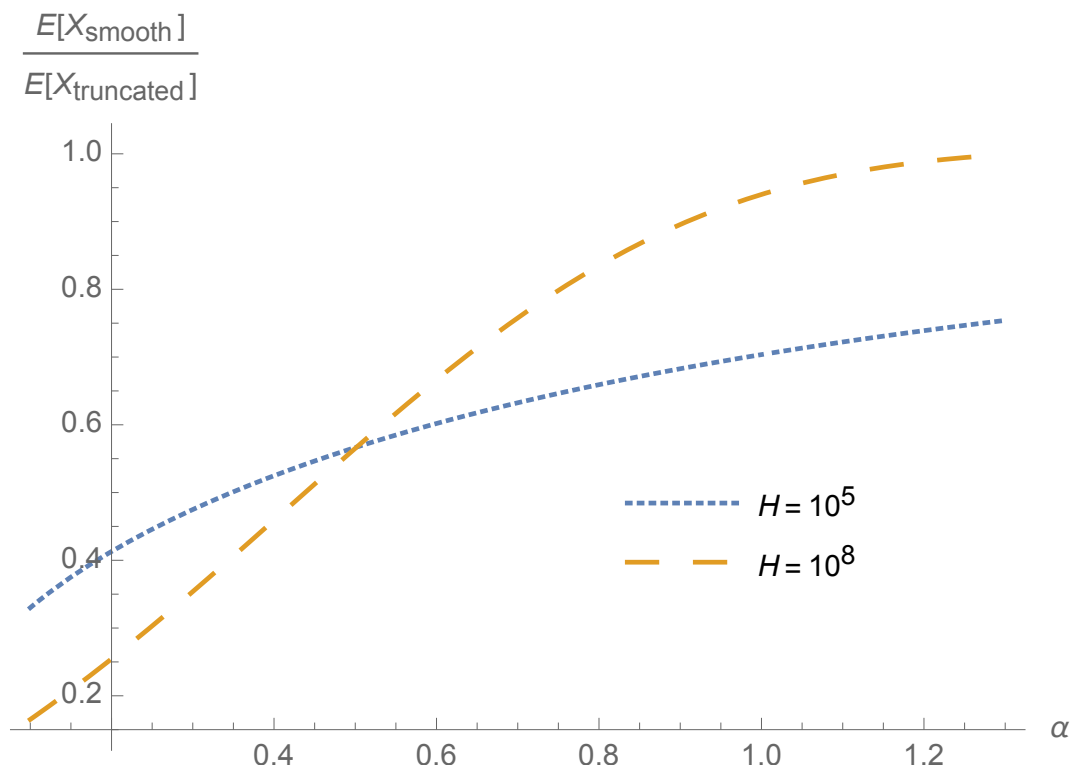
Figure 12.2: *Ratio of the expectation of smooth transformation to truncated.*

12.5 APPLICATIONS

Operational risk The losses for a firm are bounded by the capitalization, with well-known maximum losses.

Capped Reinsurance Contracts Reinsurance contracts almost always have caps (i.e., a maximum claim); but a reinsurer can have many such contracts on the same source of risk and the addition of the contract pushes the upper bound in such a way as to cause larger potential cumulative harm.

Violence While wars are extremely fat-tailed, the maximum effect from any such event cannot exceed the world's population.

Credit risk A loan has a finite maximum loss, in a way similar to reinsurance contracts.

City size While cities have been shown to be Zipf distributed, the size of a given city cannot exceed that of the world's population.

Environmental harm While these variables are exceedingly fat-tailed, the risk is confined by the size of the planet (or the continent on which they take place) as a firm upper bound.

Complex networks   The number of connections is finite.

Company size   The sales of a company is bound by the GDP.

Earthquakes   The maximum harm from an earthquake is bound by the energy.

Hydrology   The maximum level of a flood can be determined.

13 | ON THE TAIL RISK OF VIOLENT CONFLICT AND ITS UNDERESTIMATION (WITH P. CIRILLO)‡

We examine all possible statistical pictures of violent conflicts over common era history with a focus on dealing with incompleteness and unreliability of data. We apply methods from extreme value theory on log-transformed data to remove compact support, then, owing to the boundedness of maximum casualties, retransform the data and derive expected means. We find the estimated mean likely to be at least three times larger than the sample mean, meaning severe underestimation of the severity of conflicts from naive observation. We check for robustness by sampling between high and low estimates and jackknifing the data. We study inter-arrival times between tail events and find (first-order) memorylessless of events. The statistical pictures obtained are at variance with the claims about "long peace".

13.1 INTRODUCTION/SUMMARY
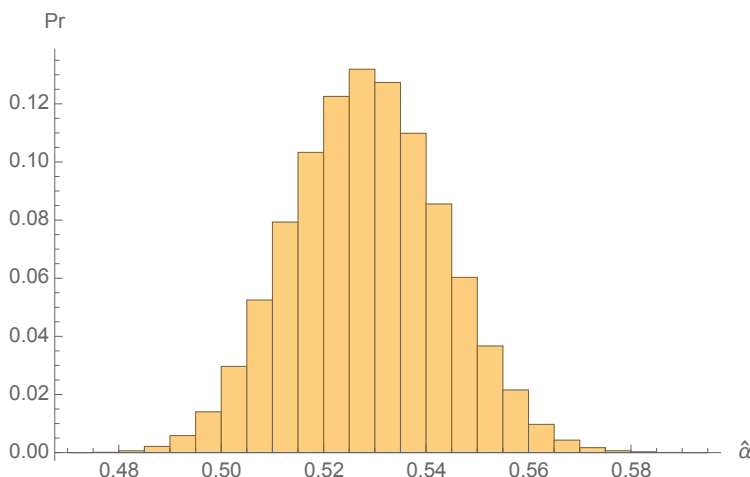
Figure 13.1: *Values of the tail exponent α from Hill estimator obtained across 100,000 different rescaled casualty numbers uniformly selected between low and high estimates of conflict. The exponent is slightly (but not meaningfully) different from the Maximum Likelihood for all data as we focus on top 100 deviations.*

This study is as much about new statistical methodologies with fat tailed (and unreliable data), as well as bounded random variables with local Power Law behavior, as it is about the properties of violence.[1]
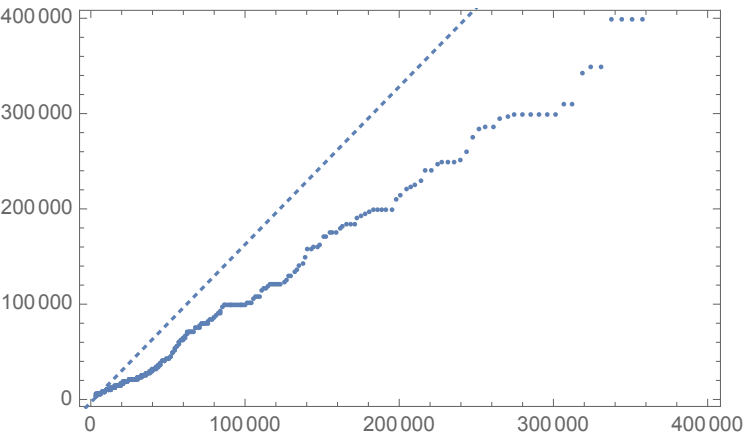
Figure 13.2: *Q-Q plot of the rescaled data in the near-tail plotted against a Pareto II-Lomax Style distribution.*
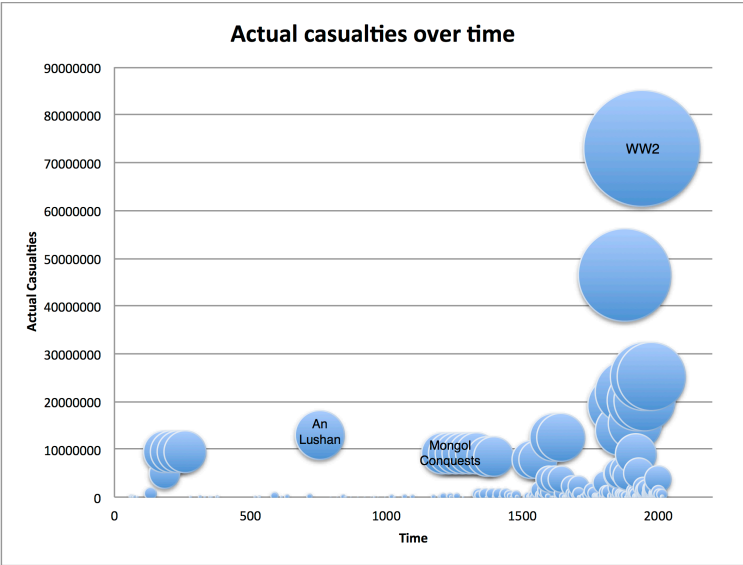
Figure 13.3: *Death toll from "named conflicts" over time. Conflicts lasting more than 25 years are disaggregated into two or more conflicts, each one lasting 25 years.*
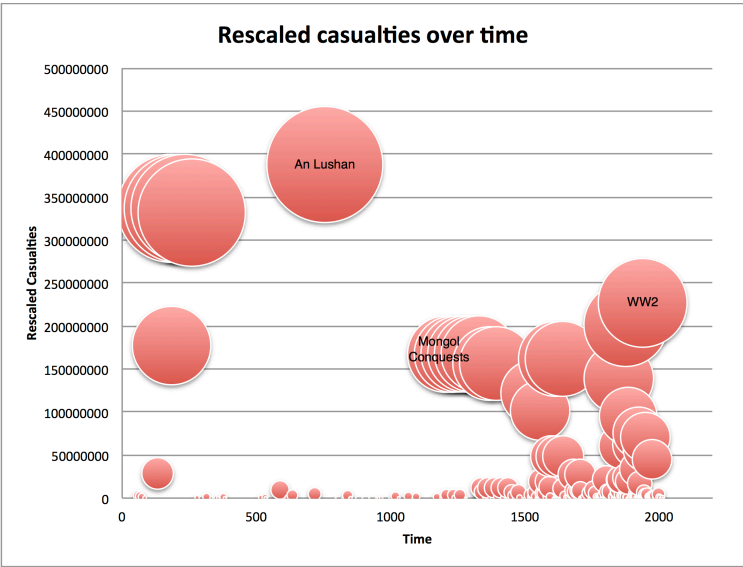
Figure 13.4: *Rescaled death toll of armed conflict and regimes over time. Data are rescaled w.r.t. today's world population. Conflicts lasting more than 25 years are disaggregated into two or more conflicts, each one lasting 25 years.*
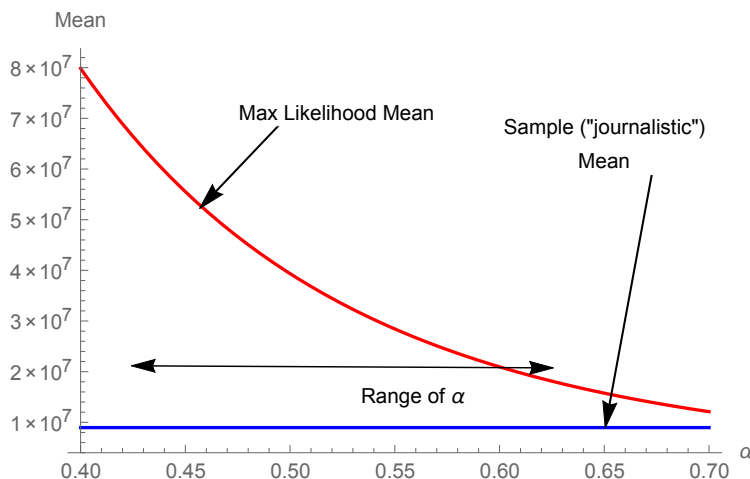
Figure 13.5: *Observed "journalistic" mean compared to MLE mean (derived from rescaling back the data to compact support) for different values of α (hence for permutations of the pair (σ_α, α)). The "range of α is the one we get from possible variations of the data from bootstrap and reliability simulations.*

Violence is much more severe than it seems from conventional analyses and the prevailing "long peace" theory which claims that violence has declined. Adapting methods from extreme value theory, and adjusting for errors in reporting of conflicts and historical estimates of casualties, we look at the various statistical pictures of violent conflicts, with focus for the parametrization on those with more than 50k victims (in equivalent ratio of today's population, which would correspond to $\approx$ 5k in the 18th C.). Contrary to current discussions, *all* statistical pictures thus obtained show that 1) the risk of violent conflict has not been decreasing, but is rather underestimated by techniques relying on naive year-on-year changes in the mean, or using sample mean as an estimator of the true mean of an extremely fat-tailed phenomenon; 2) armed conflicts have memoryless inter-arrival times, thus incompatible with the idea of a time trend. Our analysis uses 1) raw data, as recorded and estimated by historians; 2) a naive transformation, used by certain historians and sociologists, which rescales past conflicts and casualties with respect to the actual population; 3) more importantly, a log transformation to account for the fact that the number of casualties in a conflict cannot be larger than the world population. (This is similar to the transformation of data into log-returns in mathematical finance in order to use distributions with support on the real line.)

All in all, among the different classes of data (raw and rescaled), we observe that 1) casualties are Power Law distributed.[2] In the case of log-rescaled data we observe $.4 \leq \alpha \leq .7$, thus indicating an extremely fat-tailed phenomenon with an undefined mean (a result that is robustly obtained); 2) the inter-arrival times of conflicts above the 50k threshold follow a homogeneous Poisson process, indicating no particular trend, and therefore contradicting a popular narrative about the decline of violence; 3) the true mean to be expected in the future, and the most compatible with the data, though highly stochastic, is $\approx 3\times$ higher than past mean.

Further, we explain: 1) how the mean (in terms of expected casualties) is severely underestimated by conventional data analyses as the observed mean is not an estimator of true mean (unlike the tail exponent that provides a picture with smaller noise); 2) how misconceptions arise from the deceiving lengthy (and volatile) inter-arrival times between large conflicts.

---

[2] Many earlier studies have found Paretianity in data, [?],[27]. Our study, aside from the use of extreme value techniques, reliability bootstraps, and compact support transformations, varies in both calibrations and interpretation.

To remedy the inaccuracies of historical numerical assessments, we provide a standard bootstrap analysis of our estimates, in addition to Monte Carlo checks for unreliability of wars and absence of events from currently recorded history.

13.2   SUMMARY STATISTICAL DISCUSSION

13.2.1   Results

Paretian tails   Peak-Over-Threshold methods show (both raw and rescaled variables) exhibit strong Paretian tail behavior, with survival probability $P(X > x) \sim \lambda(x)x^{-\alpha}$, where $\lambda : [L, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x \to +\infty} \frac{\lambda(kx)}{\lambda(x)} = 1$ for any $k > 0$.

We parametrize G(.), a Generalized Pareto Distribution (GPD) , see Table 13.4, $G(x) = 1 - (1 + \xi y/\beta)^{-1/\xi}$, with $\xi \approx 1.88, \pm.14$ for rescaled data which corresponds to a tail $\alpha = \frac{1}{\xi} = .53, \pm.04$.

Memorylessness of onset of conflicts   Tables 13.2 and 13.3 show inter-arrival times, meaning one can wait more than a hundred years for an event such as WWII without changing one's expectation. There is no visible autocorrelation, no statistically detectable temporal structure (i.e. we cannot see the imprint of a self-exciting process), see Figure 13.8.

Full distribution(s)   Rescaled data fits a Lomax-Style distribution with same tail as obtained by POT, with strong goodness of fit. For events with casualties $> L = 10K, 25K, 50K, etc.$ we fit different Pareto II (Lomax) distributions with corresponding tail α (fit from GPD), with scale $\sigma = 84,360$, i.e., with density $\frac{\alpha \left(\frac{-L+\sigma+x}{\sigma} \right)^{-\alpha-1}}{\sigma}, x \geq L$.

We also consider a wider array of statistical "pictures" from pairs α, σ_α across the data from potential alternative values of α, with recalibration of maximum likelihood σ, see Figure 13.5.

Difference between sample mean and maximum likelihood mean   : Table 13.1 shows the true mean using the parametrization of the Pareto distribution above and inverting the transformation back to compact support. "True" or maximum likelihood, or "statistical" mean is between 3 and 4 times observed mean.

This means the "journalistic" observation of the mean, aside from the conceptual mistake of relying on sample mean, underestimates the true mean by at least 3 times and higher future observations would not allow the conlusion that violence has "risen".

13.2.2   Conclusion

History as seen from tail analysis is far more risky, and conflicts far more violent than acknowledged by naive observation of behavior of averages in historical time series.

Table 13.1: *Sample means and estimated maximum likelihood mean across minimum values L –Rescaled data.*

L	Sample Mean	ML Mean	Ratio
10K	9.079×10^6	3.11×10^7	3.43
25K	9.82×10^6	3.62×10^7	3.69
50K	1.12×10^7	4.11×10^7	3.67
100K	1.34×10^7	4.74×10^7	3.53
200K	1.66×10^7	6.31×10^7	3.79
500K	2.48×10^7	8.26×10^7	3.31

Table 13.2: *Average inter-arrival times and their mean absolute deviation for events with more than 1, 2, 5 and 10 million casualties, using actual estimates.*

Threshold	Average	MAD
1	26.71	31.66
2	42.19	47.31
5	57.74	68.60
10	101.58	144.47

Table 13.3: *Average inter-arrival times and their mean absolute deviation for events with more than 1, 2, 5, 10, 20, and 50 million casualties, using rescaled amounts.*

Threshold	Average	MAD
1	11.27	12.59
2	16.84	18.13
5	26.31	27.29
10	37.39	41.30
20	48.47	52.14
50	67.88	78.57

Table 13.4: *Estimates (and standard errors) of the Generalized Pareto Distribution parameters for casualties over a 50k threshold. For both actual and rescaled casualties, we also provide the number of events lying above the threshold (the total number of events in our data is 99).*

Data	Nr. Excesses	ζ	β
Raw Data	307	1.5886	3.6254
		(0.1467)	(0.8191)
Naive Rescaling	524	1.8718	14.3254
		(0.1259)	(2.1111)
Log-rescaling	524	1.8717	14.3261
		(0.1277)	(2.1422)

13.3 METHODOLOGICAL DISCUSSION

13.3.1 Rescaling method

We remove the compact support to be able to use power laws as follows (see Taleb(2015) [?]). Using X_t as the r.v. for number of incidences from conflict at times t, consider first a naive rescaling of $X'_t = \frac{X_t}{H_t}$, where H_t is the total human population at period t. See appendix for methods of estimation of H_t.

Next, with today's maximum population H and L the naively rescaled minimum for our definition of conflict, we introduce a smooth rescaling function $\varphi : [L, H] \rightarrow [L, \infty)$ satisfying:

  i φ is "smooth": $\varphi \in C^\infty$,

  ii $\varphi^{-1}(\infty) = H$,

  iii $\varphi^{-1}(L) = \varphi(L) = L$.

In particular, we choose:

$$\varphi(x) = L - H \log \left(\frac{H - x}{H - L} \right). \tag{13.1}$$

We can perform appropriate analytics on $x_r = \varphi(x)$ given that it is unbounded, and properly fit Power Law exponents. Then we can rescale back for the properties of X. Also notice that the $\varphi(x) \approx x$ for very large values of H. This means that for a very large upper bound, the results we will get for x and $\varphi(x)$ will be essentially the same. The big difference is only from a philosophical/methodological point of view, in the sense that we remove the upper bound (unlikely to be reached).

In what follows we will use the naively rescaled casualties as input for the $\varphi(\cdot)$ function.

We pick $H = P_{t_0}$ for the exercise.

The distribution of x can be rederived as follows from the distribution of x_r:

$$\int_L^\infty f(x_r) \, dx_r = \int_L^{\varphi^{-1}(\infty)} g(x) \, dx, \tag{13.2}$$

where $\varphi^{-1}(u) = (L - H)e^{\frac{L-u}{H}} + H$

In this case, from the Pareto-Lomax selected:

$$f(x_r) = \frac{\alpha \left(\frac{-L + \sigma + x_r}{\sigma} \right)^{-\alpha - 1}}{\sigma}, \quad x_r \in [L, \infty) \tag{13.3}$$

$$g(x) = \frac{\alpha H \left(\frac{\sigma - H \log\left(\frac{H-x}{H-L} \right)}{\sigma} \right)^{-\alpha - 1}}{\sigma(H - x)}, \quad x \in [L, H],$$

which verifies $\int_L^H x \, g(x) \, dx = 1$. Hence the expectation

$$\mathbb{E}_g(x; L, H, \sigma, \alpha) = \int_L^H x \, g(x) \, dx, \tag{13.4}$$

$$\mathbb{E}_g(X; L, H, \sigma, \alpha) = \alpha H \left(\frac{1}{\alpha} - \frac{(H - L)e^{\sigma/H} E_{\alpha+1} \left(\frac{\sigma}{H} \right)}{H} \right) \tag{13.5}$$

where $E_.(.)$ is the exponential integral $E_n z = \int_1^\infty \frac{e^{t(-z)}}{t^n} \, \mathrm{d}t$.

Note that we rely on the invariance property:

Remark 13.1

If $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ, then for an absolutely continuous function ϕ, $\phi(\hat{\theta})$ is the MLE estimator of $\phi(\theta)$.

For further details see [147].

13.3.2 Expectation by Conditioning (less rigorous)

We would be replacing a smooth function in C^∞ by a Heaviside step function, that is the indicator function $\mathbb{1} : \mathbb{R} \to \{0, 1\}$, written as $\mathbb{1}_{X \in [L, H]}$:

$$\mathbb{E}(\mathbb{1}_{X \in [L,H]}) = \frac{\int_L^H x \, f(x) \, \mathrm{d}x}{\int_L^H f(x) \, \mathrm{d}x}$$

which for the Pareto Lomax becomes:

$$\mathbb{E}(\mathbb{1}_{X \in [L,H]}) = \frac{\frac{\alpha \sigma^\alpha (H-L)}{\sigma^\alpha - (H-L+\sigma)^\alpha} + (\alpha - 1)L + \sigma}{\alpha - 1} \tag{13.6}$$

13.3.3 Reliability of data and effect on tail estimates

Data from violence is largely anecdotal, spreading via citations, often based on some vague estimate, without anyone's ability to verify the assessments using period sources. An event that took place in the seventh century, such as the an Lushan rebellion, is "estimated" to have killed 26 million people, with no precise or reliable methodology to allow us to trust the number. The independence war of Algeria has various estimates, some from France, others from the rebels, and nothing scientifically or professionally obtained.

As said earlier, in this paper, we use different data: raw data, naively rescaled data w.r.t. the current world population, and log-rescaled data to avoid the theoretical problem of the upper bound.

For some observations, together with the estimated number of casualties, as resulting from historical sources, we also have a lower and upper bound available. Let X_t be the number of casualties in a given conflict at time t. In principle, we can define triplets like

- $\{X_t, X_t^l, X_t^u\}$ for the actual estimates (raw data), where X_t^l and X_t^u represent the lower and upper bound, if available.
- $\{Y_t = X_t \frac{P_{2015}}{P_t}, Y_t^l = X_t^l \frac{P_{2015}}{P_t}, Y_t^u = X_t^u \frac{P_{2015}}{P_t}\}$ for the naively rescaled data, where P_{2015} is the world population in 2015 and P_t is the population at time $t = 1, ..., 2014$.
- $\{Z_t = \varphi(Y_t), Z_t^l = \varphi(Y_t^l), Z_t^u = \varphi(Y_t^u)\}$ for the log-rescaled data.

To prevent possible criticism about the use of middle estimates, when bounds are present, we have decided to use the following Monte Carlo procedure (for more details [139]), obtaining no significant different in the estimates of all the quantities of interest (like the tail exponent $\alpha = 1/\xi$):

1. For each event X for which bounds are present, we have assumed casualties to be uniformly distributed between the lower and the upper bound, i.e. $X \sim U(X^l, X^u)$. The choice of the uniform distribution is to keep things simple. All other bounded

distributions would in fact generate the same results in the limit, thanks to the central limit theorem.

2. We have then generated a large number of Monte Carlo replications, and in each replication we have assigned a random value to each event X according to $U(X^l, X^u)$.

3. For each replication we have computed the statistics of interest, typically the tail exponent, obtaining values that we have later averaged.

This procedure has shown that the precision of estimates does not affect the tail of the distribution of casualties, as the tail exponent is rather stable.

For those events for which no bound is given, the options were to use them as they are, or to perturb them by creating fictitious bounds around them (and then treat them as the other bounded ones in the Monte Carlo replications). We have chosen the second approach.

The above also applies to Y_t and Z_t.

Note that the tail α derived from an average is different from an average alpha across different estimates, which is the reason we perform the various analyses across estimates.

Technical comment   These simulations are largely looking for a "stochastic alpha" bias from errors and unreliability of data (Chapter x). With a sample size of n, a parameter $\hat{\theta}_m$ will be the average parameter obtained across a large number of Monte Carlo runs. Let X_i be a given Monte Carlo simulated vector indexed by i and X_μ is the middle estimate between high and low bounds. Since, with $\frac{1}{m}\sum_{\le m}\|X_j\|_1 = \|X_\mu\|_1$ across Monte Carlo runs but $\forall j, \|X_j\|_1 \ne \|X_\mu\|_1$, $\hat{\theta}_m = \frac{1}{m}\sum_{\le m}\hat{\theta}(X_j) \ne \hat{\theta}(X_\mu)$. For instance, consider the maximum likelihood estimation of a Paretian tail, $\hat{\alpha}(X_i) \triangleq n\left(\sum_{1\le i\le n}\log\left(\frac{x_i}{L}\right)\right)^{-1}$. With $\Delta \ge x_m$, define

$$\hat{\alpha}(X_i \sqcup \Delta) \triangleq \frac{1}{2}\left(\frac{n}{\sum_{i=1}^n \log\left(\frac{x_i}{L}\right) - \log\left(\frac{\Delta}{L}\right)} + \frac{n}{\sum_{i=1}^n \log\left(\frac{x_i}{L}\right) + \log\left(\frac{\Delta}{L}\right)}\right)$$

which, owing to the concavity of the logarithmic function, gives the inequality

$$\forall \Delta \ge x_m, \hat{\alpha}(X_i \sqcup \Delta) \ge \hat{\alpha}(X_i).$$

13.3.4   Definition of an "event"

"Named" conflicts are an arbitrary designation that, often, does not make sense statistically: a conflict can have two or more names; two or more conflicts can have the same name, and we found no satisfactory hierarchy between war and conflict. For uniformity, we treat events as the shorter of event or its disaggregation into units with a maximum duration of 25 years each. Accordingly, we treat Mongolian wars, which lasted more than a century and a quarter, as more than a single event. It makes little sense otherwise as it would be the equivalent of treating the period from the Franco-Prussian war to WW II as "German(ic) wars", rather than multiple events because these wars had individual names in contemporary sources. Effectively the main sources such as the *Encyclopedia of War* [129] list numerous conflicts in place of "Mongol Invasions" –the more sophisticated the historians in a given area, the more likely they are to break conflicts into different "named" events and, depending on historians, Mongolian wars range between 12 and 55 conflicts.

What controversy about the definition of a "name" can be, once again, solved by boot-strapping. Our conclusion, incidentally, is invariant to the bundling or unbundling of the Mongolian wars.

Further, in the absence of a clearly defined protocol in historical studies, it has been hard to disentangle direct death from wars and those from less direct effects on populations (say blocades, famine). For instance the First Jewish War has confused historians as an estimated 30K death came from the war, and a considerably higher (between 350K and the number 1M according to Josephus) from the famine or civilian casualties.

13.3.5   Missing events

We can assume that there are numerous wars that are not part of our sample, even if we doubt that such events are in the "tails" of the distribution, given that large conflicts are more likely to be reported by historians. Further, we also assume that their occurrence is random across the data (in the sense that they do not have an effect on clustering).

But we are aware of a bias from differential in both accuracy and reporting across time: events are more likely to be recorded in modern times than in the past. Raising the minimum value L the number of such "missed" events and their impact are likely to drop rapidly. Indeed, as a robustness check, raising the bar to a minimum $L = 500K$ does not change our analysis.

A simple jackknife procedure, performed by randomly removing a proportion of events from the sample and repeating analyses, shows us the dependence of our analysis on missing events, dependence that we have found to be insignificant, when focusing on the tail of the distribution of casualties. In other words, given that we are dealing with extremes, if removing 30% of events and checking the effects on parameters produce no divergence from initial results, then we do not need to worry of having missed 30% of events, as missing events are not likely to cause thinning of the tails.[3]

13.3.6   Survivorship Bias

We did not take into account of the survivorship biases in the analysis, assuming it to be negligible before 1960, as the probability of a conflict affecting all of mankind was negligible. Such probability (and risk) became considerably higher since, especially because of nuclear and other mass destruction weapons.

13.4   DATA ANALYSIS

Figures 13.3 and 13.4 graphically represent our data: the number of casualties over time. Figure 13.3 refers to the estimated actual number of victims, while Figure 13.4 shows the rescaled amounts, obtained by rescaling the past observation with respect to the world population in 2015 (around 7.2 billion people)[4]. Figures 13.3 might suggest an increase in the death toll of armed conflicts over time, thus supporting the idea that war violence has

---

3 The opposite is not true, which is at the core of the Black Swan asymmetry: such procedure does not remedy the missing of tail, "Black Swan" events from the record. A single "Black Swan" event can considerably fatten the tail. In this case the tail is fat enough and no missing information seems able to make it thinner.

4 Notice that, in equation (13.1), for H = 7.2 billion, $\varphi(x) \approx x$. Therefore Figure 13.4 is also representative for log-rescaled data.

increased. Figure 13.4, conversely, seems to suggest a decrease in the (rescaled) number of victims, especially in the last hundred years, and possibly in violence as well. In what follows we show that both interpretations are surely naive, because they do not take into consideration the fact that we are dealing with extreme events.

13.4.1 Peaks over Threshold

Given the fat-tailed nature of the data, which can be easily observed with some basic graphical tools like histograms on the logs and QQplots (Figure 13.6 shows the QQplot of actual casualties against an exponential distribution: the clear concavity is a signal of fat-tailed distribution), it seems appropriate to use a well-known method of extreme value theory to model war casualties over time: the Peaks-over-Threshold or POT [124].

According to the POT method, excesses of an i.i.d. sequence over a high threshold u (that we have to identify) occur at the times of a homogeneous Poisson process, while the excesses themselves can be modeled with a Generalized Pareto Distribution (GPD). Arrival times and excesses are assumed to be independent of each other.

In our case, assuming the independence of the war events does not seem a strong assumption, given the time and space separation among them. Regarding the other assumptions, on the contrary, we have to check them.

We start by identifying the threshold u above which the GPD approximation may hold. Different heuristic tools can be used for this purpose, from Zipf plot to mean excess function plots, where one looks for the linearity which is typical of fat-tailed phenomena [32, 60]. Figure 13.7 shows the mean excess function plot for actual casualties[5]: an upward trend is clearly present, already starting with a threshold equal to 5k victims. For the goodness of fit, it might be appropriate to choose a slightly larger threshold, like $u = 50k$[6].
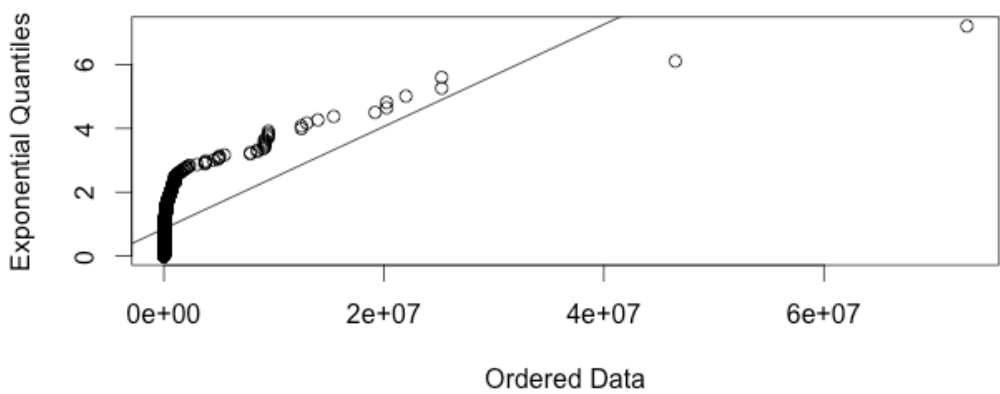
Figure 13.6: *QQplot of actual casualties against standard exponential quantile. The concave curvature of data points is a clear signal of heavy tails.*

---

5 Similar results hold for the rescaled amounts (naive and log). For the sake of brevity we always show plots for one of the two variables, unless a major difference is observed.
6 This idea has also been supported by subsequent goodness-of-fit tests.
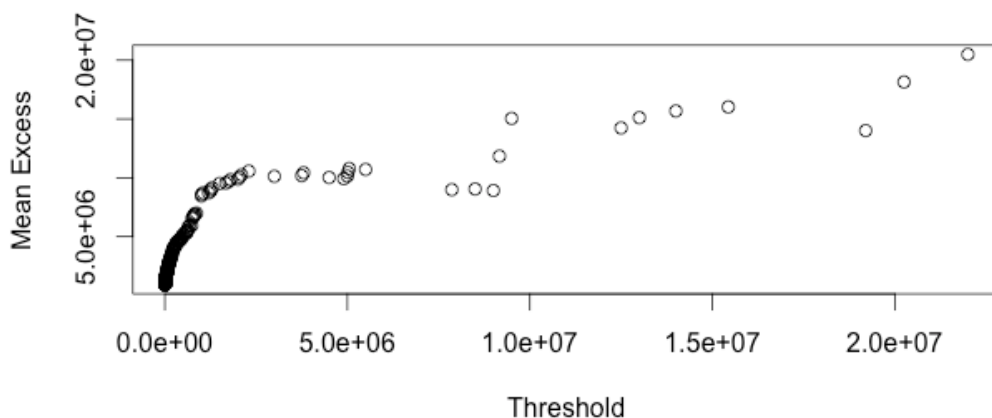
Figure 13.7: *Mean excess function plot (MEPLOT) for actual casualties. An upward trend - almost linear in the first part of the graph - is present, suggesting the presence of a fat right tail. The variability of the mean excess function for higher thresholds is due to the small number of observation exceeding those thresholds and should not be taken into consideration.*

13.4.2 Gaps in Series and Autocorrelation

To check whether events over time occur according to a homogeneous Poisson process, a basic assumption of the POT method, we can look at the distribution of the inter-arrival times or gaps, which should be exponential. Gaps should also show no autocorrelation.
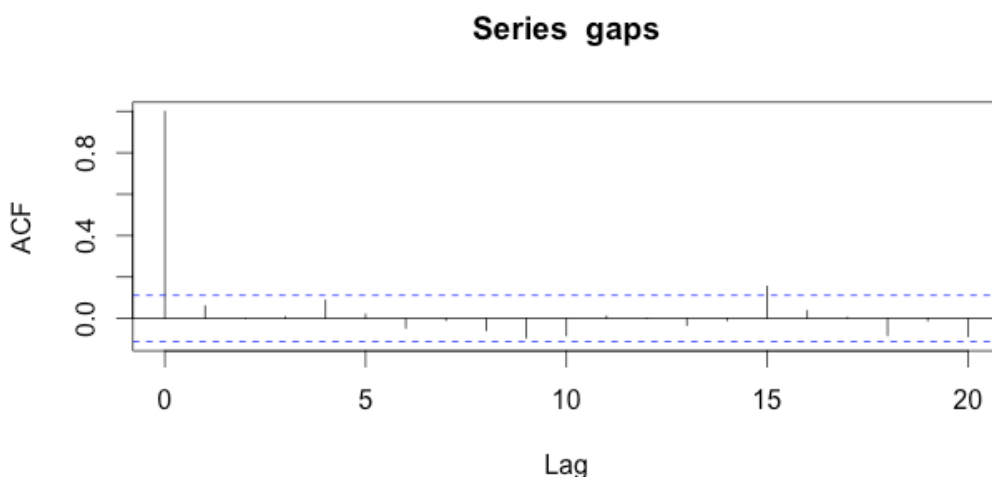
Figure 13.8: *ACF plot of gaps for actual casualties, no significant autocorrelation is visible.*

Figure 13.8 clearly shows the absence of autocorrelation. The plausibility of an exponential distribution for the inter-arrival times can be positively checked using both heuristic and analytic tools. Here we omit the positive results for brevity.

However, in order to provide some extra useful information, in Tables 13.2 and 13.3 we provide some basic statistics about the inter-arrival times for very catastrophic events in terms of casualties[7]. The simple evidence there contained should already be sufficient to underline how unreliable can be the statement that war violence has been decreasing over time. For an events with more than 10 million victims, if we refer to actual estimates, the average time delay is 101.58 years, with a mean absolute deviation of 144.47 years[8]. This means that it is totally plausible that in the last few years we have not observed such a large event. It could simply happen tomorrow or some time in the future. This also means that every trend extrapolation makes no great sense for this type of extreme events. Finally, we have to consider that an event as large as WW2 happened only once in 2014 years, if we deal with actual casualties (for rescaled casualties we can consider the An Lushan rebellion); in this case the possible waiting time is even longer.

13.4.3 Tail analysis

Given that the POT assumptions about the Poisson process seem to be confirmed by data, it is finally the time to fit a Generalized Pareto Distribution to the exceedances.

Consider a random variable X with df F, and call F_u the conditional df of X above a given threshold u. We can then define a r.v. Y, representing the rescaled excesses of X over the threshold u, getting [124]

$$F_u(y) = P(X - u \le y | X > u) = \frac{F(u + y) - F(u)}{1 - F(u)}$$

for $0 \le y \le x_F - u$, where x_F is the right endpoint of the underlying distribution F. Pickands [130], Balkema and de Haan [6], [7] and [8] showed that for a large class of underlying distribution functions F (following in the so-called domain of attraction of the GEV distribution [124]), and a large u, F_u can be approximated by a Generalized Pareto distribution: $F_u(y) \to G(y)$, as $u \to \infty$ where

$$G(y) = \begin{cases} 1 - (1 + \xi y / \beta)^{-1/\xi} & if\ \xi \ne 0 \\ 1 - e^{-y/\beta} & if\ \xi = 0. \end{cases} \tag{13.7}$$

It can be shown that the GPD distribution is a distribution interpolating between the exponential distribution (for $\xi = 0$) and a class of Pareto distributions. We refer to [124] for more details.

The parameters in (13.7) can be estimated using methods like maximum likelihood or probability weighted moments [124]. The goodness of fit can then be tested using bootstrap-based tests [186].

Table 13.4 contains our mle estimates for actual and rescaled casualties above a 50k victims threshold. This threshold is in fact the one providing the best compromise between goodness of fit and a sufficient number of observation, so that standard errors are reliable. The actual and both the rescaled data show two different sets of estimates, but their interpretation is strongly consistent. For this reason we just focus on actual casualties for the discussion.

---

7 Table 13.2 does not show the average delay for events with 20M(50M) or more casualties. This is due to the limited amount of these observations in actual, non-rescaled data. In particular, all the events with more than 20 million victims have occurred during the last 150 years, and the average inter-arrival time is below 20 years. Are we really living in more peaceful world?

8 In case of rescaled amounts, inter-arrival times are shorter, but the interpretation is the same

The parameter ξ is the most important for us: it is the parameter governing the fatness of the right tail. A ξ greater than 1 (we have 1.5886) signifies that no moment is defined for our Generalized Pareto: a very fat-tailed situation. Naturally, in the sample, we can compute all the moments we are interested in, but from a theoretical point of view they are completely unreliable and their interpretation is extremely flawed (a very common error though). According to our fitting, very catastrophic events are not at all improbable. It is worth noticing that the estimates is significant, given that its standard error is 0.1467.

Figures 13.9 and 13.10 compare our fittings to actual data. In both figures it is possible to see the goodness of the GPD fitting for most of the observations above the 50k victims threshold. Some problems arise for the very large events, like WW2 and the An Lushan rebellion [9]. In this case it appears that our fitting expects larger events to have happened. This is a well-known problem for extreme data [124]. The very large event could just be behind the corner.

Similarly, events with 5 to 10 million victims (not at all minor ones!) seem to be slightly more frequent than what is expected by our GPD fitting. This is another signal of the extreme character of war casualties, which does not allow for the extrapolation of simplistic trends.
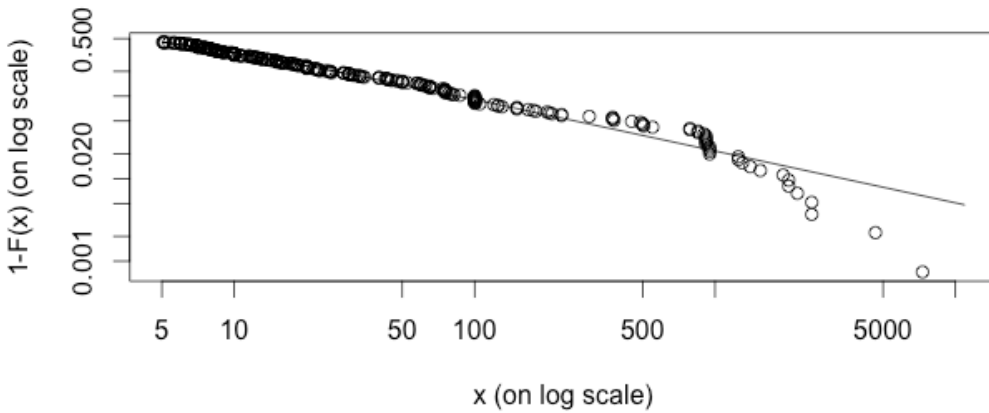
Figure 13.9: *GPD tail fitting to actual casualties' data (in 10k). Parameters as per Table 13.4, first line.*

13.4.4   An alternative view on maxima

Another method is the block-maxima approach of extreme value theory. In this approach data are divided into blocks, and within each block only the maximum value is taken into consideration. The Fisher-Tippet theorem [124] then guarantees that the normalized maxima converge in distribution to a Generalized Extreme Value Distribution, or GEV.

$$GEV(x;\xi) = \begin{cases} \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right) & \xi \neq 0 \\ \exp\left(-\exp(-x)\right) & \xi = 0 \end{cases}, \quad 1+\xi x > 0$$

[9] If we remove the two largest events from the data, the GPD hypothesis cannot be rejected at the 5% significance level.
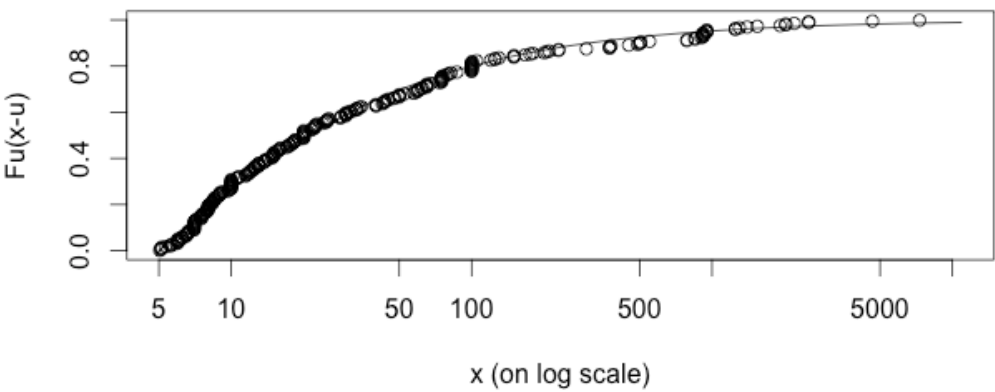
Figure 13.10: *GPD cumulative distribution fitting to actual casualties' data (in 10k). Parameters as per Table 13.4, first line.*

This distribution is naturally related to the GPD, and we refer to [124] for more details.

If we divide our data into 100-year blocks, we obtain 21 observation (the last block is the residual one from 2001 to 2014). Maximum likelihood estimations give a ζ larger than 2, indicating that we are in the so-called Fréchet maximum domain of attraction, compatible with very heavy-tailed phenomena. A value of ζ greater than 2 under the GEV distribution further confirms the idea of the absence of moments, a clear signal of a very heavy right tail.

13.4.5 Full Data Analysis

Naturally, being aware of limitations, we can try to fit all our data, while for casualties in excess of 10000, we fit the Pareto Distribution from Equation 13.3 with $\alpha \approx 0.53$ throughout. The goodness of fit for the "near tail" (L=10K) can be see in Figure 13.2. Similar results to Figure 13.2 are seen for different values in table below, all with the same goodness of fit.

L	σ
$10K$	$84,260$
$25K$	$899,953$
$50K$	$116,794$
$100K$	$172,733$
$200K$	$232,358$
$500K$	$598,292$

The different possible values of the mean in Equation 13.4 can be calculated across different set values of α, with one single degree of freedom: the corresponding σ is a MLE estimate using such α as fixed: for a sample size n, and x_i the observations higher than L, $\sigma_\alpha = \left\{ \sigma : \frac{\alpha n}{\sigma} - (\alpha + 1) \sum_{i=1}^n \frac{1}{x_i - L + \sigma} = 0, \sigma > 0 \right\}$.

The sample average for $L = 10K$ is 9.12×10^6, across 100K simulations, with the spread in values showed in Figure 13.15.

The "true" mean from Equation 13.4 yields $3.1 * 10^7$, and we repeated for $L =$10K, 20K, 50K, 100K, 200K, and 500K, finding ratios of true estimated mean to observed safely between 3 and 4., see Table 13.1. Notice that this value for the mean of ≈ 3.5 times the observed sample mean is only a general guideline, since, being stochastic, does not reveal any precise information other than prevent us from taking the naive mean estimation seriously.

For under fat tails, the mean derived from estimates of α is more rigorous and has a smaller error, since the estimate of α is asymptotically Gaussian while the average of a power law, even when it exists, is considerably more stochastic. See the discussion on "slowness of the law of large numbers" in [?] in connection with the point.

We get the mean by truncation for L=10K a bit lower, under equation 13.6; around 1.8835×10^7.

We finally note that, for values of L considered, 96 % of conflicts with more than 10,000 victims are below the mean: where m is the mean,

$$P(X < m) = 1 - \left(1 - \frac{H \log\left(\alpha e^{\sigma/H} E_{\alpha+1}\left(\frac{\sigma}{H}\right)\right)}{\sigma}\right)^{-\alpha}.$$

13.5 ADDITIONAL ROBUSTNESS AND RELIABILITY TESTS

13.5.1 Bootstrap for the GPD

In order to check our sensitivity to the quality/precision of our data, we have decided to perform some bootstrap analysis. For both raw data and the rescaled ones we have generated 100K new samples by randomly selecting 90% of the observations, with replacement. Figures 13.11, 13.12 and 13.13 show the stability of our ξ estimates. In particular $\xi > 0$ in all samples, indicating the extreme fat-tailedness of the number of victims in armed conflicts. The ξ estimates in Table 13.4 appear to be good approximations for our GPD real shape parameters, notwithstanding imprecisions and missing observations in the data.

Raw data: 100k bootstrap samples
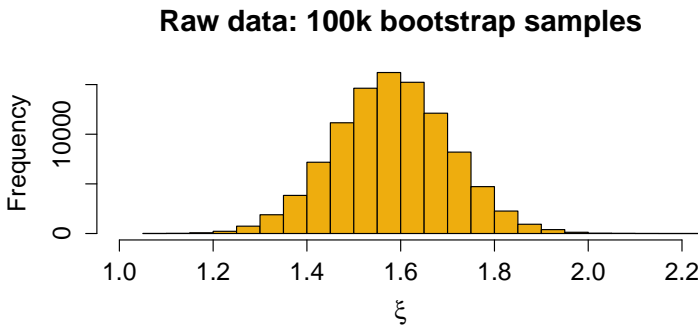
Figure 13.11: ξ parameter's distribution over 100K bootstrap samples for actual data. Each sample is randomly selected with replacement using 90% of the original observations.

13.5.2 Perturbation across bounds of estimates

We performed analyses for the "near tail" using the Monte Carlo techniques discussed in section 13.3.3. We look at second order "p-values", that is the sensitivity of the p-values

Naively rescaled data: 100k bootstrap samples
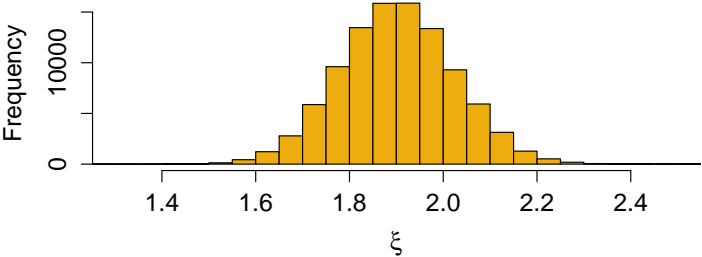
Figure 13.12: ξ parameter's distribution over 100K bootstrap samples for naively rescaled data. Each sample is randomly selected with replacement using 90% of the original observations.

Log–rescaled data: 100k bootstrap samples
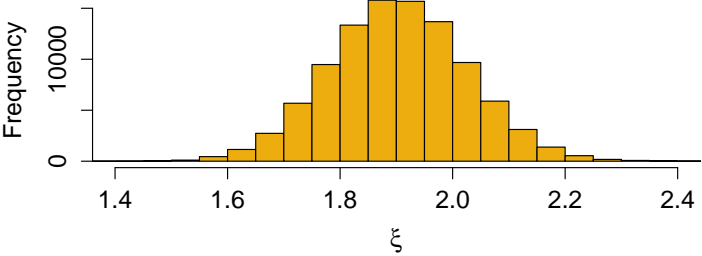
Figure 13.13: ξ parameter's distribution over 100K bootstrap samples for log-rescaled data. Each sample is randomly selected with replacement using 90% of the original observations.

across different estimates in Figure 13.14 –practically all results meet the same statistical significance and goodness of fit.

In addition, we look at values of both the sample means and the alpha-derived MLE mean across permutations, see Figures 13.15 and 13.16.
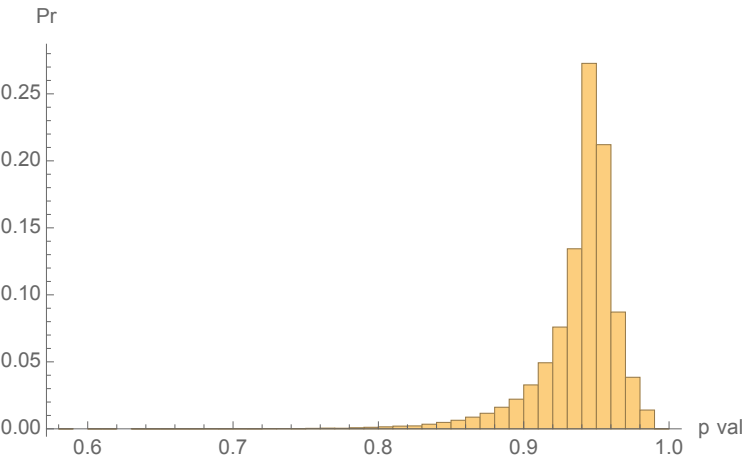
Figure 13.14: P-Values of Pareto-Lomax across 100K combinations. This is not to ascertain the p-value, rather to check the robustness by looking at the variations across permutations of estimates.

13.6 CONCLUSION: IS THE WORLD MORE UNSAFE THAN IT SEEMS?
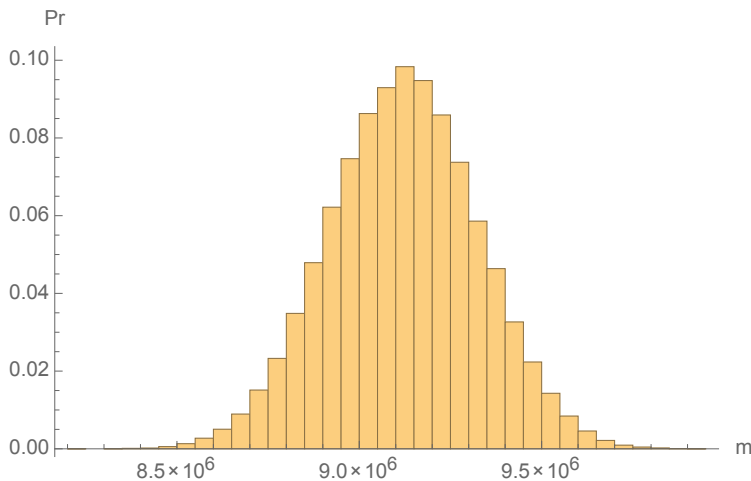
Figure 13.15: *Rescaled sample mean across 100K estimates between high-low.*
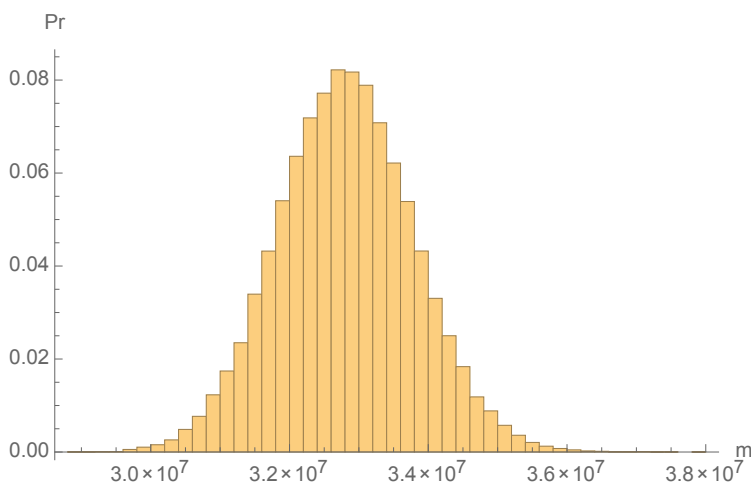
Figure 13.16: *Rescaled MLE mean across 100K estimates between high-low.*
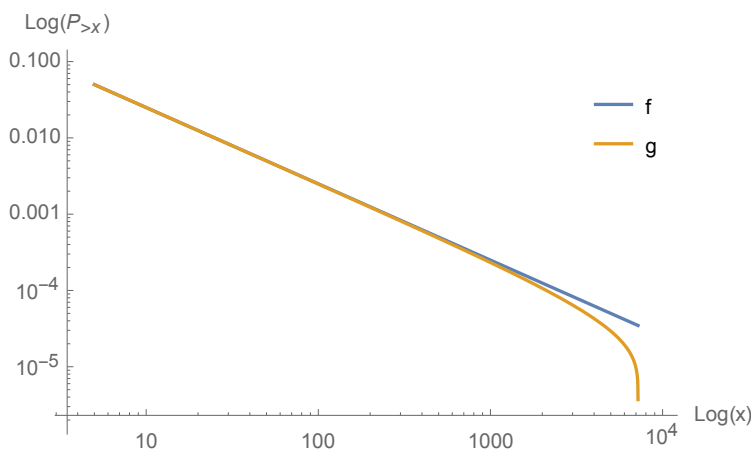
Figure 13.17: *Loglogplot comparison of f and g, showing a pasting-boundary style capping around H.*

To put our conclusion in the simplest of terms: the occurrence of events that would raise the average violence by a multiple of 3 would not cause us to rewrite this paper, nor to change the parameters calibrated within.

- Indeed, from statistical analysis alone, the world is more unsafe than casually examined numbers. Violence is underestimated by journalistic nonstatistical looks at the mean and lack of understanding of the stochasticity of under inter-arrival times.

- The transformation into compact support allowed us to perform the analyses and gauge such underestimation which , if noisy, gives us an idea of the underestimation and its bounds.

- In other words, a large event and even a rise in observed mean violence would not be inconsistent with statistical properties, meaning it would justify a "nothing has changed" reaction.

- We avoided discussions of homicide since we limited L to values $> 10,000$, but its rate doesn't appear to have a particular bearing on the tails. It could be a drop in the bucket. It obeys different dynamics. We may have observed lower rate of homicide in societies but most risks of death come from violent conflict. (Casualties from homicide by rescaling from the rate 70 per 100k, gets us 5.04×10^6 casualties per annum at today's population. A drop to minimum levels stays below the difference between errors on the mean of violence from conflicts with higher than 10,000 casualties.)

- We ignored survivorship bias in the data analysis (that is, the fact that *had the world been more violent, we wouldn't be here to talk about it*). Adding it would increase the risk. The presence of tail effects today makes further analysis require taking it into account. Since 1960, a single conflict –which almost happened– has the ability to reach the max casualties, something we did not have before.(We can rewrite the model with one of fragmentation of the world, constituted of "separate" isolated n independent random variables X_i, each with a maximum value H_i, with the total $\sum_n \omega_i H_i = H$, with all $w_i > 0$, $\sum_n \omega_i = 1$. In that case the maximum (that is worst conflict) could require the joint probabilities that all $X_1, X_2, \cdots X_n$ are near their maximum value, which, under subexponentiality, is an event of much lower probability than having a single variable reach its maximum.)

G | WHAT ARE THE CHANCES OF A THIRD WORLD WAR?*,†

HIS IS FROM AN ARTICLE that is part of the debate with public intellectuals who claim that violence have dropped "from data", without realizing that science is hard; significance requires further data under fat tails and more careful examination. We responded; our response provides a way to summarize the main problem with naive empiricism under fat tails.

In a recent issue of *Significance* Mr. Peter McIntyre asked what the chances are that World War III will occur this century. Prof. Michael Spagat wrote that nobody knows, nobody can really answer–and we totally agree with him on this. Then he adds that "a really huge war is possible but, in my view, extremely unlikely." To support his statement, Prof. Spagat relies partly on the popular science work of Prof. Steven Pinker, expressed in The Better Angels of our Nature and journalistic venues. Prof. Pinker claims that the world has experienced a long-term decline in violence, suggesting a structural change in the level of belligerence of humanity.

It is unfortunate that Prof. Spagat, in his answer, refers to our paper (this volume, Chapter 13), which is part of a more ambitious project we are working on related to fat-tailed variables.

What characterizes fat tailed variables? They have their properties (such as the mean) dominated by extreme events, those "in the tails". The most popularly known version is the "Pareto 80/20".

We show that, simply, data do not support the idea of a structural change in human belligerence. So Prof. Spagat's first error is to misread our claim: we are making neither pessimistic nor optimistic declarations: we just believe that statisticians should abide by the foundations of statistical theory and avoid telling data what to say.

Let us go back to first principles.

Foundational Principles

Fundamentally, statistics is about ensuring people do not build scientific theories from hot air, that is without significant departure from random. Otherwise, it is patently "fooled by randomness".

Further, for fat tailed variables, the conventional mechanism of the law of large numbers is considerably slower and significance requires more data and longer periods. Ironically, there are claims that can be done on little data: inference is asymmetric under fat-tailed domains. *We require more data to assert that there are no Black Swans than to assert that*

Figure G.1: *After Napoleon, there was a lull in Europe. Until nationalism came to change the story.*

there are Black Swans hence we would need much more data to claim a drop in violence than to claim a rise in it.

Finally, statements that are not deemed statistically significant –and shown to be so – should never be used to construct scientific theories.

These foundational principles are often missed because, typically, social scientists' statistical training is limited to mechanistic tools from thin tailed domains [2]. In physics, one can often claim evidence from small data sets, bypassing standard statistical methodologies, simply because the variance for these variables is low. The higher the variance, the more data one needs to make statistical claims. For fat-tails, the variance is typically high and underestimated in past data.

The second –more serious –error Spagat and Pinker made is to believe that tail events and the mean are somehow different animals, not realizing that the mean includes these tail events. ***For fat-tailed variables, the mean is almost entirely determined by extremes. If you are uncertain about the tails, then you are uncertain about the mean***. It is thus incoherent to say that violence has dropped but maybe not the risk of tail events; it would be like saying that someone is "extremely virtuous except during the school shooting episode when he killed 30 students".

Robustness

Our study tried to draw the most robust statistical picture of violence, relying on methods from extreme value theory and statistical methods adapted to fat tails. We also put robustness checks to deal with the imperfection of data collected some thousand years ago: our results need to hold even if a third (or more) of the data were wrong.

Inter-arrival times

We show that the inter-arrival times among major conflicts are extremely long, and consistent with a homogenous Poisson process: therefore no specific trend can be established:

we as humans can not be deemed as less belligerent than usual. For a conflict generating at least 10 million casualties, an event less bloody than WW1 or WW2, the waiting time is on average 136 years, with a mean absolute deviation of 267 (or 52 years and 61 deviations for data rescaled to today's population). The seventy years of what is called the "Long Peace" are clearly not enough to state much about the possibility of WW3 in the near future.

Underestimation of the mean

We also found that the average violence observed in the past underestimates the true statistical average by at least half. Why? Consider that about 90-97% of the observations fall below the mean, which requires some corrections with the help of extreme value theory. (Under extreme fat tails, the statistical mean can be closer to the past maximum observation than sample average.)

A common mistake

Similar mistakes have been made in the past. In 1860, one H.T. Buckle[1] used the same unstatistical reasoning as Pinker and Spagat.

> That this barbarous pursuit is, in the progress of society, steadily declining, must be evident, even to the most hasty reader of European history. If we compare one country with another, we shall find that for a very long period wars have been becoming less frequent; and now so clearly is the movement marked, that, until the late commencement of hostilities, we had remained at peace for nearly forty years: a circumstance unparalleled (...) The question arises, as to what share our moral feelings have had in bringing about this great improvement.

Moral feelings or not, the century following Mr. Buckle's prose turned out to be the most murderous in human history.

We conclude by saying that we find it fitting –and are honored –to expose fundamental statistical mistakes in a journal called *Significance*, as the problem is precisely about significance and conveying notions of statistical rigor to the general public.

Gini

---

1 Buckle, H.T. (1858) *History of Civilization in England, Vol. 1*, London: John W. Parker and Son.

Part VI

METAPROBABILITY PAPERS

14 | HOW FAT TAILS EMERGE FROM RECURSIVE EPISTEMIC UNCERTAINTY[†]

> **T**HE OPPOSITE OF CENTRAL LIMIT:[a] With the Central Limit Theorem, we start with a distribution and end with a Gaussian. The opposite is more likely to be true. Recall how we fattened the tail of the Gaussian by stochasticizing the variance? Now let us use the same metaprobability method, putting additional layers of uncertainty.
>
> ---
> a A version of this chapter was presented at Benoit Mandelbrot's Scientific Memorial on April 29, 2011, in New Haven, CT.

The Regress Argument (Error about Error)   The main problem behind *The Black Swan* is the limited understanding of model (or representation) 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, an inability to continuously reapplying the thinking all the way to its limit (*particularly when they provide no reason to stop*). Again, there is no problem with stopping the recursion, provided it is accepted as a declared *a priori* that escapes quantitative and statistical methods.

Epistemic not statistical re-derivation of power laws: Note that 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 work is entirely epistemic, based on standard philosophical doubts and regress arguments.

14.1 METHODS AND DERIVATIONS

14.1.1 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, Derman, 1994, Dupire, 1994, Hull and White, 1997) –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
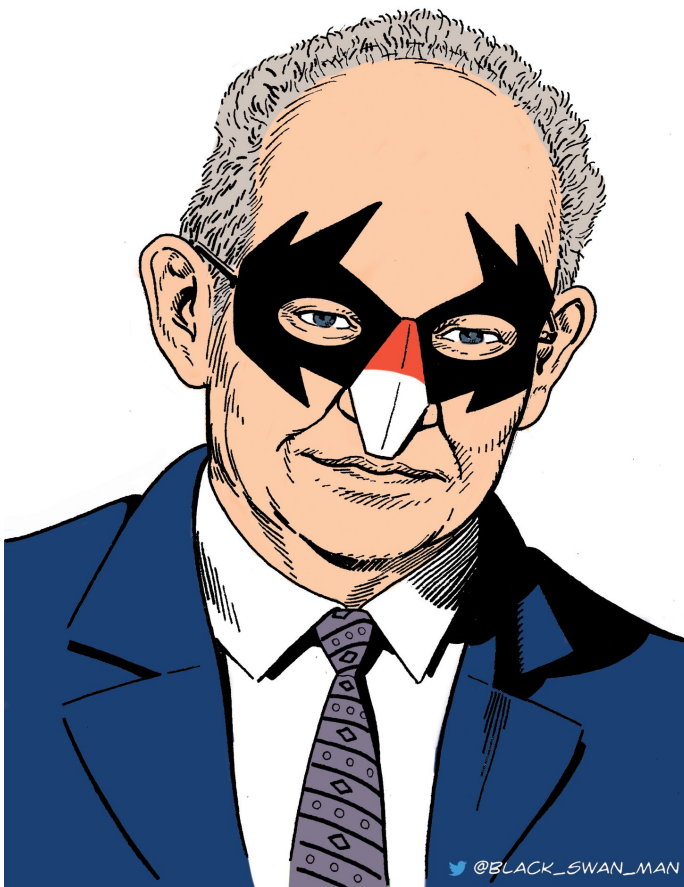
Figure 14.1: *A version of this chapter was presented at Benoit Mandelbrot's memorial.*

the uncertainty rate, and so forth. There is no reason to have certainty anywhere in the process.

14.1.2 Higher order integrals in the Standard Gaussian Case

We start with the case of a Gaussian and focus the uncertainty on the assumed standard deviation. Define $\phi(\mu,\sigma,x)$ as the Gaussian PDF for value x with mean μ and standard deviation σ.

A 2nd order stochastic standard deviation is the integral of ϕ across values of $\sigma \in \mathbb{R}^+$, under the measure $f(\bar{\sigma},\sigma_1,\sigma)$, with σ_1 its scale parameter (our approach to trach the error of the error), not necessarily its standard deviation; the expected value of σ_1 is $\overline{\sigma_1}$.

$$f(x)_1 = \int_0^\infty \phi(\mu,\sigma,x) f(\bar{\sigma},\sigma_1,\sigma) \, \mathrm{d}\sigma$$

Generalizing to the Nth order, the density function $f(x)$ becomes

$$f(x)_N = \int_0^\infty \ldots \int_0^\infty \phi(\mu, \sigma, x) f\left(\bar{\sigma}, \sigma_1, \sigma\right) f\left(\overline{\sigma_1}, \sigma_2, \sigma_1\right) \ldots f\left(\overline{\sigma_{N-1}}, \sigma_N, \sigma_{N-1}\right) d\sigma \, d\sigma_1 \, d\sigma_2 \ldots d\sigma_N$$

(14.1)

The problem is that this approach is parameter-heavy and requires the specifications of the subordinated distributions (in finance, the lognormal has been traditionally used for σ^2 (or Gaussian for the ratio $\text{Log}[\frac{\sigma_t^2}{\sigma^2}]$ since the direct use of a Gaussian allows for negative values). We would need to specify a measure f for each layer of error rate. Instead this can be approximated by using the mean deviation for σ, as we will see next.

Discretization using nested series of two-states for σ- a simple multiplicative process

We saw in the last chapter a quite effective simplification to capture the convexity, the ratio of (or difference between) $\phi(\mu,\sigma,x)$ and $\int_0^\infty \phi(\mu, \sigma, x) f\left(\bar{\sigma}, \sigma_1, \sigma\right) d\sigma$ (the first order standard deviation) by using a weighted average of values of σ, say, for a simple case of one-order stochastic volatility:

$$\sigma(1 \pm a(1))$$

with $0 \le a(1) < 1$, where $a(1)$ is the proportional mean absolute deviation for σ, in other word the measure of the absolute error rate for σ. We use $\frac{1}{2}$ as the probability of each state. Unlike the earlier situation we are not preserving the variance, rather the STD.

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

$$f(x)_1 = \frac{1}{2}\left(\phi(\mu, \sigma\left(1 + a(1)\right), x) + \phi(\mu, \sigma(1 - a(1)), x)\right)$$

(14.2)

Now assume uncertainty about the error rate a(1), expressed by a(2), in the same manner as before. Thus in place of a(1) we have $\frac{1}{2}$ a(1)(1$\pm$ a(2)).

The second order stochastic standard deviation:

$$f(x)_2 = \frac{1}{4}\left(\phi\left(\mu, \sigma(1 + a(1)(1 + a(2))), x\right) + \right.$$

$$\left.\phi\left(\mu, \sigma(1 - a(1)(1 + a(2))), x\right) + \phi(\mu, \sigma(1 + a(1)(1 - a(2))), x) + \phi\left(\mu, \sigma(1 - a(1)(1 - a(2))), x\right)\right)$$

(14.3)

and the Nth order:

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

where M_i^N is the ith scalar (line) of the matrix M^N $(2^N \times 1)$

$$M^N = \left(\prod_{j=1}^{N}(a(j)\mathbf{T}_{i,j} + 1)\right)_{i=1}^{2^N}$$
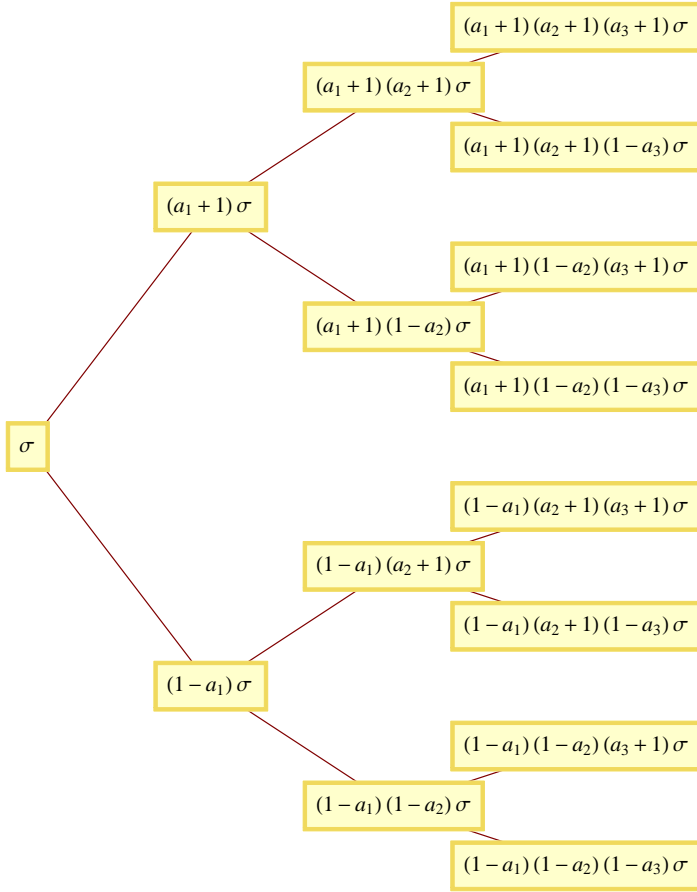
Figure 14.2: *Three levels of error rates for σ following a multiplicative process*

and $\mathbf{T}_{i,j}$ the element of i^{th}line and j^{th}column of the matrix of the exhaustive combination of n-Tuples of the set $\{-1, 1\}$,that is the sequences of n length $(1, 1, 1, ...)$ representing all combinations of 1 and -1.

for N=3,

$$
T = \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & -1 \\
1 & -1 & 1 \\
1 & -1 & -1 \\
-1 & 1 & 1 \\
-1 & 1 & -1 \\
-1 & -1 & 1 \\
-1 & -1 & -1
\end{pmatrix}
$$

and

$$M^3 = \begin{pmatrix} (1 - a(1))(1 - a(2))(1 - a(3)) \\ (1 - a(1))(1 - a(2))(a(3) + 1) \\ (1 - a(1))(a(2) + 1)(1 - a(3)) \\ (1 - a(1))(a(2) + 1)(a(3) + 1) \\ (a(1) + 1)(1 - a(2))(1 - a(3)) \\ (a(1) + 1)(1 - a(2))(a(3) + 1) \\ (a(1) + 1)(a(2) + 1)(1 - a(3)) \\ (a(1) + 1)(a(2) + 1)(a(3) + 1) \end{pmatrix}$$

So $M_1^3 = \{(1 - a(1))(1 - a(2))(1 - a(3))\}$, etc.
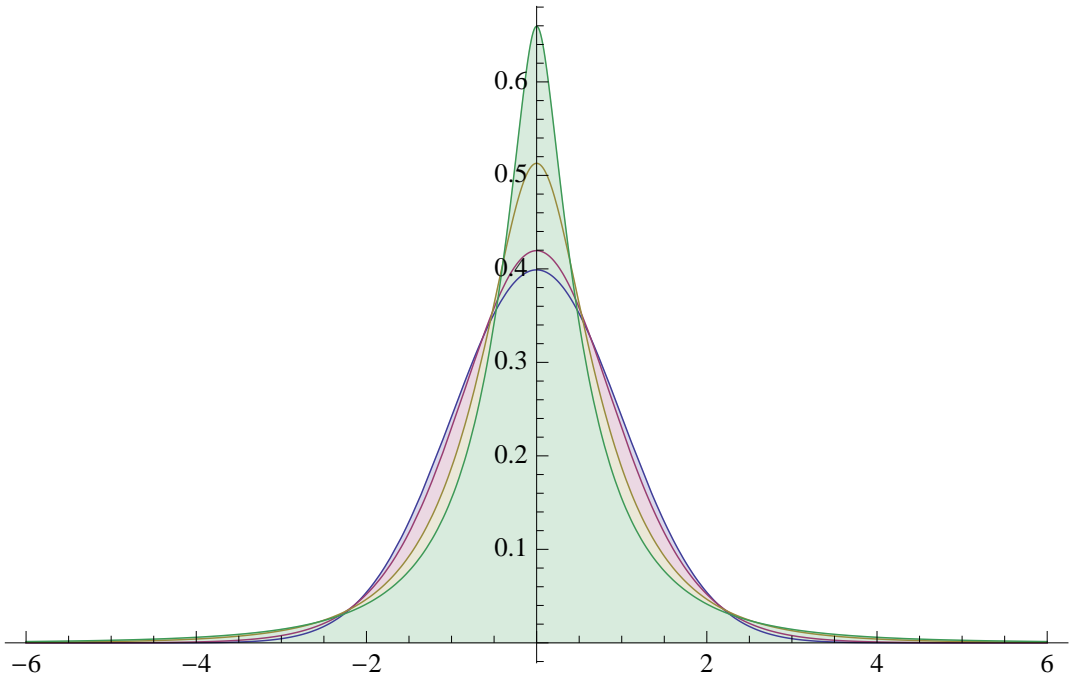
Figure 14.3: *Thicker tails (higher peaks) for higher values of N; here N = 0, 5, 10, 25, 50, all values of $a = \frac{1}{10}$*

Note that the various error rates $a(i)$ are not similar to sampling errors, but rather projection of error rates into the future. They are, to repeat, *epistemic*.

The Final Mixture Distribution The mixture weighted average distribution (recall that ϕ is the ordinary Gaussian PDF with mean μ, std σ for the random variable x).

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

It could be approximated by a lognormal distribution for σ and the corresponding V as its own variance. But it is precisely the V that interest us, and V depends on how higher order errors behave.

Next let us consider the different regimes for higher order errors.

REGIME 1 (EXPLOSIVE): CASE OF A CONSTANT PARAMETER a

Special case of constant a: Assume that $a(1)=a(2)=...a(N)=a$, i.e. the case of flat proportional error rate a. The Matrix M collapses into a conventional binomial tree for the dispersion at the level N.

$$f(x|\mu,\sigma,M,N) = 2^{-N} \sum_{j=0}^{N} \binom{N}{j} \phi\left(\mu, \sigma(a+1)^{j}(1-a)^{N-j}, x\right) \tag{14.4}$$

Because of the linearity of the sums, when a is constant, we can use the binomial distribution as weights for the moments (note again the artificial effect of constraining the first moment μ in the analysis to a set, certain, and known *a priori*).

$$\begin{pmatrix} & \text{Moment} \\ 1 & \mu \\ 2 & \sigma^2 \left(a^2+1\right)^N + \mu^2 \\ 3 & 3\mu\sigma^2 \left(a^2+1\right)^N + \mu^3 \\ 4 & 6\mu^2\sigma^2 \left(a^2+1\right)^N + \mu^4 + 3\left(a^4+6a^2+1\right)^N \sigma^4 \end{pmatrix}$$

For clarity, we simplify the table of moments, with $\mu=0$

$$\begin{pmatrix} & \text{Moment} \\ 1 & 0 \\ 2 & \left(a^2+1\right)^N \sigma^2 \\ 3 & 0 \\ 4 & 3\left(a^4+6a^2+1\right)^N \sigma^4 \\ 5 & 0 \\ 6 & 15\left(a^6+15a^4+15a^2+1\right)^N \sigma^6 \\ 7 & 0 \\ 8 & 105\left(a^8+28a^6+70a^4+28a^2+1\right)^N \sigma^8 \end{pmatrix}$$

Note again the oddity that in spite of the explosive nature of higher moments, the expectation of the absolute value of x is both independent of a and N, since the perturbations of σ do not affect the first absolute moment $= \sqrt{\frac{2}{\pi}}\sigma$ (that is, the initial assumed σ). The situation would be different under addition of x.

Every recursion multiplies the variance of the process by $(1+a^2)$. The process is similar to a stochastic volatility model, with the standard deviation (not the variance) following a lognormal distribution, the volatility of which grows with M, hence will reach infinite variance at the limit.

Consequences

For a constant $a > 0$, and in the more general case with variable a where $a(n) \geq a(n-1)$, the moments explode.

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 $N \to \infty$. 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- In these conditions, we need to use power laws for epistemic reasons, or, at least, distributions outside the $\mathcal{L}^2$ norm, regardless of observations of past data.

$$a=\frac{1}{10}, \ N=0,5,10,25,50$$
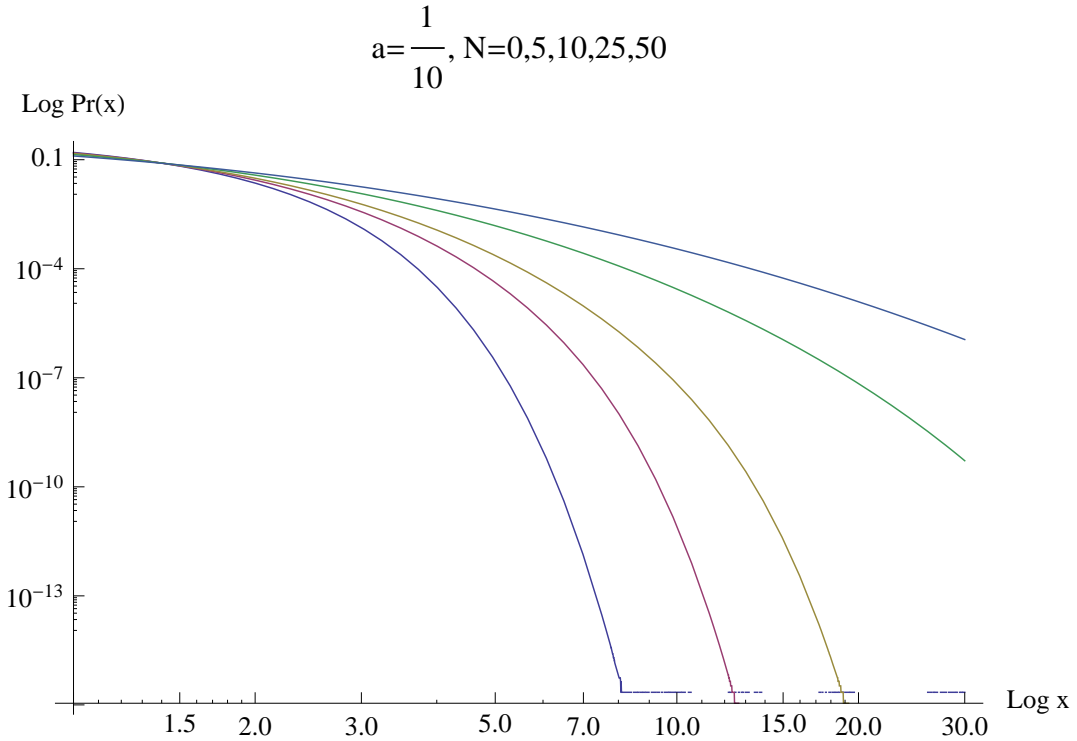
Figure 14.4: *LogLog Plot of the probability of exceeding x showing power law-style flattening as N rises. Here all values of a= 1/10*

Note that we need an *a priori* reason (in the philosophical sense) to cutoff the N somewhere, hence bound the expansion of the second moment.

Convergence to Properties Similar 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 $a(1)=a(2)=...=a(N)=\frac{1}{10}$) how the density approaches that of a Power Law (just like the Lognormal distribution at higher variance), as shown in flatter density on the LogLog plot. The probabilities keep rising in the tails as we add layers of uncertainty until they seem to reach the boundary of the power law, while ironically the first moment remains invariant.

The same effect takes place as a increases towards 1, as at the limit the tail exponent P>x approaches 1 but remains >1.

14.1.3   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 exceedant probability,that is, the probability of exceeding K, given N, for parameter a constant:

$$P > K|N = \sum_{j=0}^{N} 2^{-N-1} \begin{pmatrix} N \\ j \end{pmatrix}$$

$$\mathrm{erfc} \left(\frac{K}{\sqrt{2}\sigma(a+1)^j(1-a)^{N-j}} \right) \quad (14.5)$$

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

Convexity effect   The next Table shows the ratio of exceedant probability under different values of N divided by the probability in the case of a standard Gaussian.

Table 14.1: *Case of a = $\frac{1}{10}$*

N	$\frac{P>3,N}{P>3,N=0}$	$\frac{P>5,N}{P>5,N=0}$	$\frac{P>10,N}{P>10,N=0}$
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

Table 14.2: *Case of a = $\frac{1}{100}$*

N	$\frac{P>3,N}{P>3,N=0}$	$\frac{P>5,N}{P>5,N=0}$	$\frac{P>10,N}{P>10,N=0}$
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}

14.2   REGIME 2: CASES OF DECAYING PARAMETERS $a(n)$

As we said, we may have (actually we need to have) *a priori* reasons to decrease the parameter a or stop N somewhere. When the higher order of $a(i)$ decline, then the moments tend to be capped (the inherited tails will come from the lognormality of σ).

14.2.1   Regime 2-a;"bleed" of higher order error

Take a "bleed" of higher order errors at the rate λ, $0 \le \lambda < 1$, such as $a(N) = \lambda\, a(N\text{-}1)$, hence $a(N) = \lambda^N a(1)$, with $a(1)$ the conventional intensity of stochastic standard deviation. Assume μ=0.

With N=2 , the second moment becomes:

$$M_2(2) = \left(a(1)^2 + 1\right) \sigma^2 \left(a(1)^2\lambda^2 + 1\right)$$

With N=3,

$$M_2(3) = \sigma^2 \left(1 + a(1)^2\right) \left(1 + \lambda^2 a(1)^2\right) \left(1 + \lambda^4 a(1)^2\right)$$

finally, for the general N:

$$M_3(N) = \left(a(1)^2 + 1\right) \sigma^2 \prod_{i=1}^{N-1} \left(a(1)^2 \lambda^{2i} + 1\right) \qquad (14.6)$$

We can reexpress 14.6 using the Q-Pochhammer symbol $(a;q)_N = \prod_{i=1}^{N-1} \left(1 - aq^i\right)$

$$M_2(N) = \sigma^2 \left(-a(1)^2; \lambda^2\right)_N$$

Which allows us to get to the limit

$$\lim_{N \to \infty} M_2(N) = \sigma^2 \frac{\left(\lambda^2; \lambda^2\right)_2 \left(a(1)^2; \lambda^2\right)_\infty}{\left(\lambda^2 - 1\right)^2 \left(\lambda^2 + 1\right)}$$

As to the fourth moment:

By recursion:

$$M_4(N) = 3\sigma^4 \prod_{i=0}^{N-1} \left(6a(1)^2 \lambda^{2i} + a(1)^4 \lambda^{4i} + 1\right)$$

$$M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2} - 3\right) a(1)^2; \lambda^2\right)_N$$

$$\left(-\left(3 + 2\sqrt{2}\right) a(1)^2; \lambda^2\right)_N (14.7)$$

$$\lim_{N \to \infty} M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2} - 3\right) a(1)^2; \lambda^2\right)_\infty$$

$$\left(-\left(3 + 2\sqrt{2}\right) a(1)^2; \lambda^2\right)_\infty (14.8)$$

So the limiting second moment for $\lambda=.9$ and $a(1)=.2$ is just $1.28 \, \sigma^2$, a significant but relatively benign convexity bias. The limiting fourth moment 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 that of a Gaussian.

14.2.2 Regime 2-b; Second Method, a Non Multiplicative Error Rate

For N recursions,

$$\sigma(1 \pm (a(1)(1 \pm (a(2)(1 \pm a(3)(\dots))))$$

$$P(x, \mu, \sigma, N) = \frac{1}{L} \sum_{i=1}^{L} f\left(x, \mu, \sigma \left(1 + \left(\mathbf{T}^N . \mathbf{A}^N\right)_i\right)\right)$$

$(\mathbf{M}^N . \mathbf{T} + 1)_i)$ is the ith component of the $(N \times 1)$ dot product of $\mathbf{T}^N$ the matrix of Tuples in (xx) , L the length of the matrix, and A contains the parameters

$$A^N = \left(a^j\right)_{j=1,\dots N}$$

So for instance, for $N = 3$, $\mathbf{T} = \left(1, a, a^2, a^3\right)$

$$\mathbf{A}^3\,\mathbf{T}^3 = \begin{pmatrix} a^3 + a^2 + a \\ -a^3 + a^2 + a \\ a^3 - a^2 + a \\ -a^3 - a^2 + a \\ a^3 + a^2 - a \\ -a^3 + a^2 - a \\ a^3 - a^2 - a \\ -a^3 - a^2 - a \end{pmatrix}$$

The moments are as follows:

$$M_1(N) = \mu$$

$$M_2(N) = \mu^2 + 2\sigma$$

$$M_4(N) = \mu^4 + 12\mu^2\sigma + 12\sigma^2 \sum_{i=0}^{N} a^{2i}$$

At the limit:

$$\lim_{N \to \infty} M_4(N) = \frac{12\sigma^2}{1 - a^2} + \mu^4 + 12\mu^2\sigma$$

which is very mild.

15 | STOCHASTIC TAIL EXPONENT FOR ASYMMETRIC POWER LAWS[†]

WE EXAMINE random variables in the power law/slowly varying class with stochastic tail exponent , the exponent α having its own distribution. We show the effect of stochasticity of α on the expectation and higher moments of the random variable. For instance, the moments of a right-tailed or right-asymmetric variable, when finite, increase with the variance of α; those of a left-asymmetric one decreases. The same applies to conditional shortfall (CVar), or mean-excess functions.

We prove the general case and examine the specific situation of lognormally distributed $\alpha \in [b, \infty), b > 1$.

The stochasticity of the exponent induces a significant bias in the estimation of the mean and higher moments in the presence of data uncertainty. This has consequences on sampling error as uncertainty about α translates into a higher expected mean.

The bias is conserved under summation, even upon large enough a number of summands to warrant convergence to the stable distribution. We establish inequalities related to the asymmetry.

We also consider the situation of capped power laws (i.e. with compact support), and apply it to the study of violence by Cirillo and Taleb (2016). We show that uncertainty concerning the historical data increases the true mean.

15.1 BACKGROUND

stochastic volatility has been introduced heuristically in mathematical finance by traders looking for biases on option valuation, where a Gaussian distribution is considered to have several possible variances, either locally or at some specific future date. Options far from the money (i.e. concerning tail events) increase in value with uncertainty on the variance of the distribution, as they are convex to the standard deviation.

This led to a family of models of Brownian motion with stochastic variance (see review in Gatheral [73]) and proved useful in tracking the distributions of the underlying and the effect of the nonGaussian character of random processes on functions of the process (such as option prices).

---

o Conference: Extremes and Risks in Higher Dimensions, Lorentz Center, Leiden, The Netherlands, September 2016.

Just as options are convex to the scale of the distribution, we find many situations where expectations are convex to the Power Law tail exponent . This note examines two cases:

- The standard power laws, one-tailed or asymmetric.
- The pseudo-power law, where a random variable appears to be a Power Law but has compact support, as in the study of violence [34] where wars have the number of casualties capped at a maximum value.

15.2 ONE TAILED DISTRIBUTIONS WITH STOCHASTIC ALPHA

15.2.1 General Cases

Definition 15.1
Let X be a random variable belonging to the class of distributions with a "power law" right tail, that is support in $[x_0, +\infty)$, $\in \mathbb{R}$:

Subclass $\mathfrak{P}_1$:

$$\{X : \mathbb{P}(X > x) = L(x)x^{-\alpha}, \frac{\partial^q L(x)}{\partial x^q} = 0 \text{ for } q \geq 1\} \tag{15.1}$$

We note that x_o can be negative by shifting, so long as $x_0 > -\infty$.

Class $\mathfrak{P}$:

$$\{X : \mathbb{P}(X > x) \sim L(x) x^{-\alpha}\} \tag{15.2}$$

where $\sim$ means that the limit of the ratio or rhs to lhs goes to 1 as $x \to \infty$. $L : [x_{\min}, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x \to +\infty} \frac{L(kx)}{L(x)} = 1$ for any $k > 0$. $L'(x)$ is monotone. The constant $\alpha > 0$.

We further assume that:

$$\lim_{x \to \infty} L'(x) x = 0 \tag{15.3}$$

$$\lim_{x \to \infty} L''(x) x = 0 \tag{15.4}$$

We have

$$\mathfrak{P}_1 \subset \mathfrak{P}$$

We note that the first class corresponds to the Pareto distributions (with proper shifting and scaling), where L is a constant and $\mathfrak{P}$ to the more general one-sided power laws.

15.2.2 Stochastic Alpha Inequality

Throughout the rest of the paper we use for notation X' for the stochastic alpha version of X, the constant α case.

Proposition 15.1
Let $p = 1, 2, ..., X'$ be the same random variable as X above in $\mathfrak{P}_1$ (the one-tailed regular variation class), with $x_0 \geq 0$, except with stochastic α with all realizations $> p$ that preserve the mean $\bar{\alpha}$,

$$\mathbb{E}(X'^p) \geq \mathbb{E}(X^p).$$

Proposition 15.2

Let K be a threshold. With X in the $\mathfrak{P}$ class, we have the expected conditional shortfall (CVar):

$$\lim_{K \to \infty} \mathbb{E}(X'|_{X'>K}) \geq \lim_{K \to \infty} \mathbb{E}(X|_{X>K}).$$

The sketch of the proof is as follows.

We remark that $\mathbb{E}(X^p)$ is convex to α, in the following sense. Let X_{α_i} be the random variable distributed with constant tail exponent α_i, with $\alpha_i > p, \forall i$, and ω_i be the normalized positive weights: $\sum_i \omega_i = 1, 0 \leq |\omega_i| \leq 1, \sum_i \omega_i \alpha_i = \bar{\alpha}$. By Jensen's inequality:

$$\omega_i \sum_i \mathbb{E}(X^p_{\alpha_i}) \geq \mathbb{E}(\sum_i (\omega_i X^p_{\alpha_i})).$$

As the classes are defined by their survival functions, we first need to solve for the corresponding density: $\varphi(x) = \alpha x^{-\alpha-1} L(x, \alpha) - x^{-\alpha} L^{(1,0)}(x, \alpha)$ and get the normalizing constant.

$$L(x_0, \alpha) = x_0^\alpha - \frac{2x_0 L^{(1,0)}(x_0, \alpha)}{\alpha - 1} - \frac{2x_0^2 L^{(2,0)}(x_0, \alpha)}{(\alpha - 1)(\alpha - 2)}, \tag{15.5}$$

$\alpha \neq 1, 2$ when the first and second derivative exist, respectively. The slot notation $L^{(p,0)}(x_0, \alpha)$ is short for $\frac{\partial^p L(x,\alpha)}{\partial x^p}|_{x=x_0}$.

By the Karamata representation theorem, [14],[174], a function L on $[x_0, +\infty)$ is slowly moving (Definition) if and only if it can be written in the form

$$L(x) = \exp\left(\int_{x_0}^x \frac{\epsilon(t)}{t} \, dt\right) + \eta(x)$$

where $\eta(.)$ is a bounded measurable function converging to a finite number as $x \to +\infty$, and $\epsilon(x)$ is a bounded measurable function converging to zero as $x \to +\infty$.

Accordingly, $L'(x)$ goes to 0 as $x \to \infty$. (We further assumed in 15.3 and 15.4 that $L'(x)$ goes to 0 faster than x and $L''(x)$ goes to 0 faster than x^2). Integrating by parts,

$$\mathbb{E}(X^p) = x_0^p + p \int_{x_0}^\infty x^{p-1} \, d\bar{F}(x)$$

where $\bar{F}$ is the survival function in Eqs. 15.1 and 15.2. Integrating by parts three additional times and eliminating derivatives of $L(.)$ of higher order than 2:

$$\mathbb{E}(X^p) = \frac{x_0^{p-\alpha} L(x_0, \alpha)}{p - \alpha} - \frac{x_0^{p-\alpha+1} L^{(1,0)}(x_0, \alpha)}{(p - \alpha)(p - \alpha + 1)} + \frac{x_0^{p-\alpha+2} L^{(2,0)}(x_0, \alpha)}{(p - \alpha)(p - \alpha + 1)(p - \alpha + 2)} \tag{15.6}$$

which, for the special case of X in $\mathfrak{P}_1$ reduces to:

$$\mathbb{E}(X^p) = x_0^p \frac{\alpha}{\alpha - p} \tag{15.7}$$

As to Proposition 2, we can approach the proof from the property that $\lim_{x \to \infty} L'(x) = 0$. This allows a proof of van der Mijk's law that Paretian inequality is invariant to the threshold in the tail, that is $\frac{\mathbb{E}(X|_{X>K})}{K}$ converges to a constant as $K \to +\infty$. Equation 15.6 presents the exact conditions on the functional form of $L(x)$ for the convexity to extend to sub-classes between $\mathfrak{P}_1$ and $\mathfrak{P}$.

Our results hold to distributions that are transformed by shifting and scaling, of the sort: $x \mapsto x - \mu + x_0$ (Pareto II), or with further transformations to Pareto types II and IV.

We note that the representation $\mathfrak{P}_1$ uses the same parameter, x_0, for both scale and minimum value, as a simplification.

We can verify that the expectation from Eq. 15.7 is convex to α: $\frac{\partial \mathbb{E}(X^p)}{\partial \alpha^2} = x_0^p \frac{2}{(\alpha-1)^3}$.

15.2.3   Approximations for the Class $\mathfrak{P}$

For $\mathfrak{P} \setminus \mathfrak{P}_1$, our results hold when we can write an approximation the expectation of X as a constant multiplying the integral of $x^{-\alpha}$, namely

$$\mathbb{E}(X) \approx k \frac{\nu(\alpha)}{\alpha - 1} \tag{15.8}$$

where k is a positive constant that does not depend on α and $\nu(.)$ is approximated by a linear function of α (plus a threshold). The expectation will be convex to α.

Example: Student T Distribution   For the Student T distribution with tail α, the "sophisticated" slowly varying function in common use for symmetric power laws in quantitative finance, the half-mean or the mean of the one-sided distribution (i.e. with support on $\mathbb{R}^+$ becomes

$$2\nu(\alpha) = 2 \frac{\sqrt{\alpha}\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha}{2}\right)} \approx \alpha \frac{(1 + \log(4))}{\pi},$$

where $\Gamma(.)$ is the gamma function.

15.3   SUMS OF POWER LAWS

As we are dealing from here on with convergence to the stable distribution, we consider situations of $1 < \alpha < 2$, hence $p = 1$ and will be concerned solely with the mean.

We observe that the convexity of the mean is invariant to summations of Power Law distributed variables as X above. The Stable distribution has a mean that in conventional parameterizations does not appear to depend on α –but in fact depends on it.

Let Y be distributed according to a Pareto distribution with density $f(y) \triangleq \alpha \lambda^\alpha y^{-\alpha-1}, y \geq \lambda > 0$ and with its tail exponent $1 < \alpha < 2$. Now, let $Y_1, Y_2, \ldots Y_n$ be identical and independent copies of Y. Let $\chi(t)$ be the characteristic function for $f(y)$. We have $\chi(t) = \alpha(-it)^\alpha \Gamma(-\alpha, -it)$, where $\gamma(.,.)$ is the incomplete gamma function. We can get the mean from the characteristic function of the average of n summands $\frac{1}{n}(Y_1 + Y_2 + \ldots Y_n)$, namely $\chi(\frac{t}{n})^n$. Taking the first derivative:

$$-i\frac{\partial \chi(\frac{t}{n})^n}{\partial t} = (-i)^{\alpha(n-1)} n^{1-\alpha n} \alpha^n \lambda^{\alpha(n-1)} t^{\alpha(n-1)-1} \Gamma\left(-\alpha, \right. \tag{15.9}$$

$$\left. -\frac{it\lambda}{n}\right)^{n-1} \left((-i)^\alpha \alpha \lambda^\alpha t^\alpha \Gamma\left(-\alpha, -\frac{it\lambda}{n}\right) - n^\alpha e^{\frac{i\lambda t}{n}}\right)$$

and

$$\lim_{n\to\infty} -i \frac{\partial \chi(\frac{t}{n})^n}{\partial t}\bigg|_{t=0} = \lambda \frac{\alpha}{\alpha - 1} \tag{15.10}$$

Thus we can see how the converging asymptotic distribution for the average will have for mean the scale times $\frac{\alpha}{\alpha-1}$, which does not depends on n.

Let $\chi^S(t)$ be the characteristic function of the corresponding stable distribution $S_{\alpha,\beta,\mu,\sigma}$, from the distribution of an infinitely summed copies of Y. By the Lévy continuity theorem, we have

- $\frac{1}{n}\Sigma_{i\leq n}Y_i \xrightarrow{\mathcal{D}} S$, with distribution $S_{\alpha,\beta,\mu,\sigma}$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution

  and

- $\chi^S(t) = \lim_{n\to\infty} \chi(t/n)^n$

are equivalent.

So we are dealing with the standard result [193],[145], for exact Pareto sums [191], replacing the conventional μ with the mean from above:

$$\chi^S(t) = \exp\left(i\left(\lambda \frac{\alpha t}{\alpha - 1} + |t|^\alpha \left(\beta \tan\left(\frac{\pi\alpha}{2}\right) \text{sgn}(t) + i \right) \right) \right).$$

15.4 ASYMMETRIC STABLE DISTRIBUTIONS

We can verify by symmetry that, effectively, flipping the distribution in subclasses $\mathfrak{P}_1$ and $\mathfrak{P}_2$ around y_0 to make it negative yields a negative value of the mean d higher moments, hence degradation from stochastic α.

The central question becomes:

Remark 15.1 (Preservation of Asymmetry)
A normalized sum in $\mathfrak{P}_1$ one-tailed distribution with expectation that depends on α of the form in Eq. 15.8 will necessarily converge in distribution to an asymmetric stable distribution $S_{\alpha,\beta,\mu,1}$, with $\beta \neq 0$.

Remark 15.2
Let Y' be Y under mean-preserving stochastic α. The convexity effect, or $\text{sgn}\left(\mathbb{E}(Y') - \mathbb{E}(Y)\right) = \text{sgn}(\beta)$.

The sketch of the proof is as follows. Consider two slowly varying functions as in 15.1, each on one side of the tails. We have $L(y) = \mathbb{1}_{y<y_\theta} L^-(y) + \mathbb{1}_{y\geq y_\theta} L^+(y)$:

$$\begin{cases} L^+(y), L : [y_\theta, +\infty], & \lim_{y\to\infty} L^+(y) = c \\ \\ L^-(y), L : [-\infty, y_\theta], & \lim_{y\to-\infty} L^-(y) = d. \end{cases}$$

From [145],

$$\text{if } \begin{cases} \mathbb{P}(X > x) \sim cx^{-\alpha}, x \to +\infty \\ \\ \mathbb{P}(X < x) \sim d|x|^{-\alpha}, x \to +\infty, \end{cases} \quad \text{then } Y \text{ converges in distribution to } S_{\alpha,\beta,\mu,1} \text{ with the}$$

coefficient $\beta = \frac{c-d}{c+d}$.

We can show that the mean can be written as $(\lambda_+ - \lambda_-)\frac{\alpha}{\alpha-1}$ where:

$$\lambda_+ \geq \lambda_- \text{ if } \int_{y_\theta}^{\infty} L^+(y)dy, \geq \int_{-\infty}^{y_\theta} L^-(y)dy$$

15.5   PARETO DISTRIBUTION WITH LOGNORMALLY DISTRIBUTED α

Now assume α is following a shifted Lognormal distribution with mean α_0 and minimum value b, that is, $\alpha - b$ follows a Lognormal $\mathcal{LN}\left(\log(\alpha_0) - \frac{\sigma^2}{2}, \sigma\right)$. The parameter b allows us to work with a lower bound on the tail exponent in order to satisfy finite expectation. We know that the tail exponent will eventually converge to b but the process may be quite slow.

Proposition 15.3
Assuming finite expectation for X' and for exponent the lognormally distributed shifted variable $\alpha - b$ with law $\mathcal{LN}\left(\log(\alpha_0) - \frac{\sigma^2}{2}, \sigma\right)$, $b \geq 1$ mininum value for α, and scale λ:

$$\mathbb{E}(Y') = \mathbb{E}(Y) + \lambda \frac{(e^{\sigma^2} - b)}{\alpha_0 - b} \tag{15.11}$$

We need $b \geq 1$ to avoid problems of infinite expectation.

Let $\phi(y, \alpha)$ be the density with stochastic tail exponent. With $\alpha > 0, \alpha_0 > b, b \geq 1, \sigma > 0, Y \geq \lambda > 0$,

$$
\begin{aligned}
\mathbb{E}(Y) &= \int_b^{\infty} \int_L^{\infty} y\phi(y; \alpha) \, dy \, d\alpha \\
&= \int_b^{\infty} \lambda \frac{\alpha}{\alpha - 1} \frac{1}{\sqrt{2\pi}\sigma(\alpha - b)} \\
&\quad \exp\left(-\frac{\left(\log(\alpha - b) - \log(\alpha_0 - b) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}\right) d\alpha \\
&= \frac{\lambda\left(\alpha_0 + e^{\sigma^2} - b\right)}{\alpha_0 - b} .
\end{aligned}
\tag{15.12}
$$

Approximation of the density

With $b = 1$ (which is the lower bound for b),we get the density with stochastic α:

$$\phi(y; \alpha_0, \sigma) = \lim_{k\to\infty} \frac{1}{Y^2} \sum_{i=0}^{k} \frac{1}{i!} L(\alpha_0 - 1)^i e^{\frac{1}{2}i(i-1)\sigma^2} (\log(\lambda) - \log(y))^{i-1}(i + \log(\lambda) - \log(y)) \tag{15.13}$$

This result is obtained by expanding α around its lower bound b (which we simplified to $b = 1$) and integrating each summand.

15.6 PARETO DISTRIBUTION WITH GAMMA DISTRIBUTED ALPHA

Proposition 15.4
Assuming finite expectation for X' scale λ, and for exponent a gamma distributed shifted variable $\alpha - 1$ with law $\varphi(.)$, mean α_0 and variance s^2, all values for α greater than 1:

$$\mathbb{E}(X') = \mathbb{E}(X') + \frac{s^2}{(\alpha_0 - 1)(\alpha_0 - s - 1)(\alpha_0 + s - 1)} \tag{15.14}$$

Proof.

$$\varphi(\alpha) = \frac{e^{-\frac{(\alpha-1)(\alpha_0-1)}{s^2}} \left(\frac{s^2}{(\alpha-1)(\alpha_0-1)} \right)^{-\frac{(\alpha_0-1)^2}{s^2}}}{(\alpha-1)\Gamma\left(\frac{(\alpha_0-1)^2}{s^2} \right)}, \quad \alpha > 1 \tag{15.15}$$

$$\int_1^\infty \alpha \lambda^\alpha x^{-\alpha-1} \varphi(\alpha)\, d\alpha \tag{15.16}$$

$$= \int_1^\infty \frac{\alpha \left(e^{-\frac{(\alpha-1)(\alpha_0-1)}{s^2}} \left(\frac{s^2}{(\alpha-1)(\alpha_0-1)} \right)^{-\frac{(\alpha_0-1)^2}{s^2}} \right)}{(\alpha - 1)\left((\alpha - 1)\Gamma\left(\frac{(\alpha_0-1)^2}{s^2} \right) \right)}\, d\alpha$$

$$= \frac{1}{2} \left(\frac{1}{\alpha_0 + s - 1} + \frac{1}{\alpha_0 - s - 1} + 2 \right)$$

$\square$

15.7 THE BOUNDED POWER LAW IN CIRILLO AND TALEB (2016)

In [34] and [33], the studies make use of bounded power laws, applied to violence and operational risk, respectively. Although with $\alpha < 1$ the variable Z has finite expectations owing to the upper bound.

The methods offered were a smooth transformation of the variable as follows: we start with $z \in [L, H], L > 0$ and transform it into $x \in [L, \infty)$, the latter legitimately being Power Law distributed.

So the smooth logarithmic transformation):

$$x = \varphi(z) = L - H \log\left(\frac{H - z}{H - L} \right),$$

and

$$f(x) = \frac{\left(\frac{x-L}{\alpha\sigma} + 1 \right)^{-\alpha-1}}{\sigma}.$$

We thus get the distribution of Z which will have a finite expectation for all positive values of α.

$$\frac{\partial^2 E(Z)}{\partial \alpha^2} = \frac{1}{H^3}(H - L)\left(e^{\frac{\alpha\sigma}{H}} \left(2H^3 G_{3,4}^{4,0}\left(\frac{\alpha\sigma}{H}\Big|\ \begin{matrix} \alpha + 1, \alpha + 1, \alpha + 1 \\ 1, \alpha, \alpha, \alpha \end{matrix} \right) \right. \right.$$
$$\left. - 2H^2(H + \sigma)G_{2,3}^{3,0}\left(\frac{\alpha\sigma}{H}\Big|\ \begin{matrix} \alpha + 1, \alpha + 1 \\ 1, \alpha, \alpha \end{matrix} \right) \right. \quad (15.17)$$
$$\left. + \sigma\left(\alpha\sigma^2 + (\alpha + 1)H^2 + 2\alpha H\sigma \right) E_\alpha\left(\frac{\alpha\sigma}{H} \right) \right) - H\sigma(H + \sigma) \right)$$

which appears to be positive in the range of numerical perturbations in [34].[1] At such a low level of α, around $\frac{1}{2}$, the expectation is extremely convex and the bias will be accordingly extremely pronounced.

This convexity has the following practical implication. Historical data on violence over the past two millennia, is fundamentally unreliable [34]. Hence an imprecision about the tail exponent , from errors embedded in the data, need to be present in the computations. The above shows that uncertainty about α, is more likely to make the "true" statistical mean (that is the mean of the process as opposed to sample mean) higher than lower, hence supports the statement that more uncertainty increases the estimation of violence.

15.8 ADDITIONAL COMMENTS

The bias in the estimation of the mean and shortfalls from uncertainty in the tail exponent can be added to analyses where data is insufficient, unreliable, or simply prone to forgeries.

In additional to statistical inference, these result can extend to processes, whether a compound Poisson process with power laws subordination [151] (i.e. a Poisson arrival time and a jump that is Power Law distributed) or a Lévy process. The latter can be analyzed by considering successive "slice distributions" or discretization of the process [37]. Since the expectation of a sum of jumps is the sum of expectation, the same convexity will appear as the one we got from Eq. 15.8.

15.9 ACKNOWLEDGMENTS

---

1 $G_{3,4}^{4,0}\left(\frac{\alpha\sigma}{H}\Big|\ \begin{matrix} \alpha + 1, \alpha + 1, \alpha + 1 \\ 1, \alpha, \alpha, \alpha \end{matrix} \right)$ is the Meijer G function.

Part VII

TAILS FOR BOUNDED RANDOM VARIABLES

16 | THE META-DISTRIBUTION OF STANDARD P-VALUES‡

ᴇ ᴘʀᴇsᴇɴᴛ an exact probability distribution (meta-distribution) for p-values across ensembles of statistically identical phenomena, as well as the distribution of the minimum p-value among m independents tests. We derive the distribution for small samples $2 < n \leq n^* \approx 30$ as well as the limiting one as the sample size n becomes large. We also look at the properties of the "power" of a test through the distribution of its inverse for a given p-value and parametrization.

P-values are shown to be extremely skewed and volatile, regardless of the sample size n, and vary greatly across repetitions of exactly same protocols under identical stochastic copies of the phenomenon; such volatility makes the minimum p value diverge significantly from the "true" one. Setting the power is shown to offer little remedy unless sample size is increased markedly or the p-value is lowered by at least one order of magnitude.

The formulas allow the investigation of the stability of the reproduction of results and "p-hacking" and other aspects of meta-analysis.

From a probabilistic standpoint, neither a p-value of .05 nor a "power" at .9 appear to make the slightest sense.

Assume that we know the "true" p-value, p_s, what would its realizations look like across various attempts on statistically identical copies of the phenomena? By true value p_s, we mean its expected value by the law of large numbers across an m ensemble of possible samples for the phenomenon under scrutiny, that is $\frac{1}{m} \sum_{\leq m} p_i \xrightarrow{P} p_s$ (where $\xrightarrow{P}$ denotes convergence in probability). A similar convergence argument can be also made for the corresponding "true median" p_M. The main result of the paper is that the the distribution of n small samples can be made explicit (albeit with special inverse functions), as well as its parsimonious limiting one for n large, with no other parameter than the median value p_M. We were unable to get an explicit form for p_s but we go around it with the use of the median. Finally, the distribution of the minimum p-value under can be made explicit, in a parsimonious formula allowing for the understanding of biases in scientific studies.

It turned out, as we can see in Fig. 16.2 the distribution is extremely asymmetric (right-skewed), to the point where 75% of the realizations of a "true" p-value of .05 will be <.05 (a borderline situation is 3× as likely to pass than fail a given protocol), and, what is worse, 60% of the true p-value of .12 will be below .05.
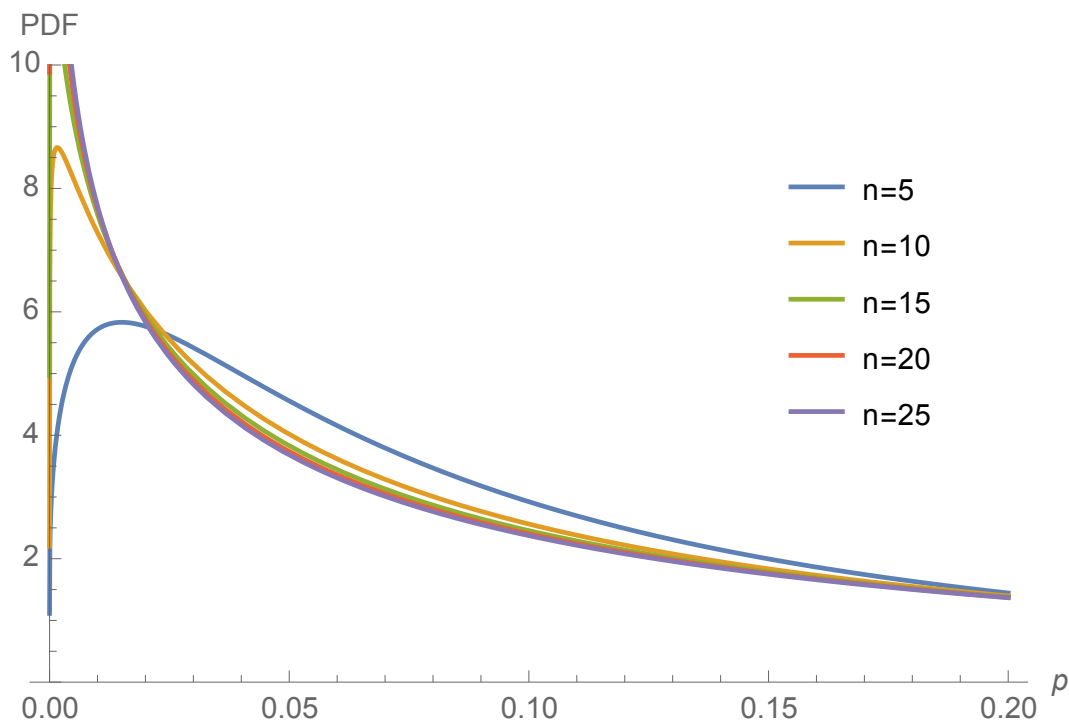
Figure 16.1: *The different values for Equ. 16.1 showing convergence to the limiting distribution.*

Although with compact support, the distribution exhibits the attributes of extreme fat-tailedness. For an observed p-value of, say, .02, the "true" p-value is likely to be >.1 (and very possibly close to .2), with a standard deviation >.2 (sic) and a mean deviation of around .35 (sic, sic). Because of the excessive skewness, measures of dispersion in $\mathcal{L}^1$ and $\mathcal{L}^2$ (and higher norms) vary hardly with p_s, so the standard deviation is not proportional, meaning an in-sample .01 p-value has a significant probability of having a true value > .3.

So clearly we don't know what we are talking about when we talk about p-values.

Earlier attempts for an explicit meta-distribution in the literature were found in [91] and [144], though for situations of Gaussian subordination and less parsimonious parametrization. The severity of the problem of *significance of the so-called "statistically significant"* has been discussed in [75] and offered a remedy via Bayesian methods in [95], which in fact recommends the same tightening of standards to p-values $\approx$ .01. But the gravity of the extreme skewness of the distribution of p-values is only apparent when one looks at the meta-distribution.

For notation, we use n for the sample size of a given study and m the number of trials leading to a p-value.

16.1 PROOFS AND DERIVATIONS

Proposition 16.1
Let P be a random variable $\in [0,1]$) corresponding to the sample-derived one-tailed p-value from

the paired T-test statistic (unknown variance) with median value $\mathbb{M}(P) = p_M \in [0, 1]$ *derived from a sample of n size. The distribution across the ensemble of statistically identical copies of the sample has for PDF*

$$\varphi(p; p_M) = \begin{cases} \varphi(p; p_M)_L & \text{for } p < \frac{1}{2} \\ \varphi(p; p_M)_H & \text{for } p > \frac{1}{2} \end{cases}$$

$$\varphi(p; p_M)_L = \lambda_p^{\frac{1}{2}(-n-1)}$$

$$\sqrt{-\frac{\lambda_p (\lambda_{p_M} - 1)}{(\lambda_p - 1) \lambda_{p_M} - 2\sqrt{(1 - \lambda_p) \lambda_p}\sqrt{(1 - \lambda_{p_M}) \lambda_{p_M}} + 1}}$$

$$\left(\frac{1}{\frac{1}{\lambda_p} - \frac{2\sqrt{1-\lambda_p}\sqrt{\lambda_{p_M}}}{\sqrt{\lambda_p}\sqrt{1-\lambda_{p_M}}} + \frac{1}{1-\lambda_{p_M}} - 1} \right)^{n/2}$$

$$\varphi(p; p_M)_H = \left(1 - \lambda_p'\right)^{\frac{1}{2}(-n-1)}$$

$$\left(\frac{\left(\lambda_p' - 1\right) \left(\lambda_{p_M} - 1\right)}{\lambda_p' \left(-\lambda_{p_M}\right) + 2\sqrt{\left(1 - \lambda_p'\right) \lambda_p'}\sqrt{\left(1 - \lambda_{p_M}\right) \lambda_{p_M}} + 1} \right)^{\frac{n+1}{2}} \quad (16.1)$$

where $\lambda_p = I_{2p}^{-1}\left(\frac{n}{2}, \frac{1}{2}\right)$, $\lambda_{p_M} = I_{1-2p_M}^{-1}\left(\frac{1}{2}, \frac{n}{2}\right)$, $\lambda_p' = I_{2p-1}^{-1}\left(\frac{1}{2}, \frac{n}{2}\right)$, and $I_{(.)}^{-1}(.,.)$ is the inverse beta regularized function.

Remark 16.1
For $p=\frac{1}{2}$ *the distribution doesn't exist in theory, but does in practice and we can work around it with the sequence* $p_{m_k} = \frac{1}{2} \pm \frac{1}{k}$, *as in the graph showing a convergence to the Uniform distribution on* $[0, 1]$ *in Figure 16.3. Also note that what is called the "null" hypothesis is effectively a set of measure 0.*

Proof. Let Z be a random normalized variable with realizations ζ, from a vector $\vec{v}$ of n realizations, with sample mean m_v, and sample standard deviation s_v, $\zeta = \frac{m_v - m_h}{\frac{s_v}{\sqrt{n}}}$ (where m_h is the level it is tested against), hence assumed to $\sim$ Student T with n degrees of freedom, and, crucially, supposed to deliver a mean of $\bar{\zeta}$,

$$f(\zeta; \bar{\zeta}) = \frac{\left(\frac{n}{(\bar{\zeta}-\zeta)^2+n}\right)^{\frac{n+1}{2}}}{\sqrt{n}B\left(\frac{n}{2}, \frac{1}{2}\right)}$$

where $B(.,.)$ is the standard beta function. Let $g(.)$ be the one-tailed survival function of the Student T distribution with zero mean and n degrees of freedom:

$$g(\zeta) = \mathbb{P}(Z > \zeta) = \begin{cases} \frac{1}{2} I_{\frac{n}{\zeta^2+n}}\left(\frac{n}{2}, \frac{1}{2}\right) & \zeta \geq 0 \\ \frac{1}{2}\left(I_{\frac{\zeta^2}{\zeta^2+n}}\left(\frac{1}{2}, \frac{n}{2}\right) + 1 \right) & \zeta < 0 \end{cases}$$

where $I_{(.,.)}$ is the incomplete Beta function.

We now look for the distribution of $g \circ f(\zeta)$. Given that $g(.)$ is a legit Borel function, and naming p the probability as a random variable, we have by a standard result for the transformation:

$$\varphi(p, \bar{\zeta}) = \frac{f\left(g^{(-1)}(p)\right)}{|g'\left(g^{(-1)}(p)\right)|}$$

We can convert $\bar{\zeta}$ into the corresponding median survival probability because of symmetry of Z. Since one half the observations fall on either side of $\bar{\zeta}$, we can ascertain that the transformation is median preserving: $g(\bar{\zeta}) = \frac{1}{2}$, hence $\varphi(p_M, .) = \frac{1}{2}$. Hence we end up having $\{\bar{\zeta} : \frac{1}{2} I_{\frac{n}{\bar{\zeta}^2+n}}\left(\frac{n}{2}, \frac{1}{2}\right) = p_M\}$ (positive case) and $\{\bar{\zeta} : \frac{1}{2}\left(I_{\frac{\bar{\zeta}^2}{\bar{\zeta}^2+n}}\left(\frac{1}{2}, \frac{n}{2}\right) + 1\right) = p_M\}$ (negative case). Replacing we get Eq.16.1 and Proposition 16.1 is done.

□

We note that n does not increase significance, since p-values are computed from normalized variables (hence the universality of the meta-distribution); a high n corresponds to an increased convergence to the Gaussian. For large n, we can prove the following proposition:

Proposition 16.2
Under the same assumptions as above, the limiting distribution for $\varphi(.)$:

$$\lim_{n \to \infty} \varphi(p; p_M) = e^{-erfc^{-1}(2p_M)\left(erfc^{-1}(2p_M) - 2erfc^{-1}(2p)\right)} \tag{16.2}$$

where erfc(.) is the complementary error function and $erfc(.)^{-1}$ its inverse.

The limiting CDF $\Phi(.)$

$$\Phi(k; p_M) = \frac{1}{2} erfc\left(erf^{-1}(1 - 2k) - erf^{-1}(1 - 2p_M)\right) \tag{16.3}$$

Proof. For large n, the distribution of $Z = \frac{m_v}{\frac{s_v}{\sqrt{n}}}$ becomes that of a Gaussian, and the one-tailed survival function $g(.) = \frac{1}{2} erfc\left(\frac{\zeta}{\sqrt{2}}\right)$, $\zeta(p) \to \sqrt{2} erfc^{-1}(p)$. □

This limiting distribution applies for paired tests with known or assumed sample variance since the test becomes a Gaussian variable, equivalent to the convergence of the T-test (Student T) to the Gaussian when n is large.

Remark 16.2
For values of p close to 0, φ in Equ. 16.2 can be usefully calculated as:

$$\varphi(p; p_M) = \sqrt{2\pi} p_M \sqrt{\log\left(\frac{1}{2\pi p_M^2}\right)}$$

$$e^{\sqrt{-\log\left(2\pi \log\left(\frac{1}{2\pi p^2}\right)\right) - 2\log(p)} \sqrt{-\log\left(2\pi \log\left(\frac{1}{2\pi p_M^2}\right)\right) - 2\log(p_M)}}$$

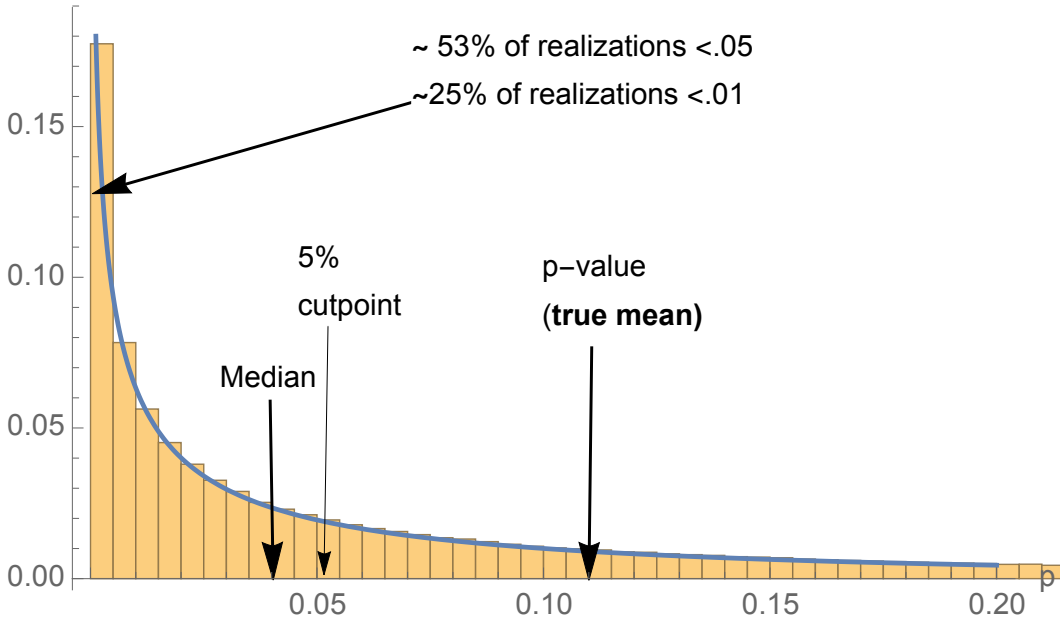$$+ O(p^2). \tag{16.4}$$

PDF/Frequ.

Figure 16.2: *The probability distribution of a one-tailed p-value with expected value .11 generated by Monte Carlo (histogram) as well as analytically with φ(.) (the solid line). We draw all possible subsamples from an ensemble with given properties. The excessive skewness of the distribution makes the average value considerably higher than most observations, hence causing illusions of "statistical significance".*

The approximation works more precisely for the band of relevant values $0 < p < \frac{1}{2\pi}$.

From this we can get numerical results for convolutions of φ using the Fourier Transform or similar methods.

We can and get the distribution of the minimum p-value per m trials across statistically identical situations thus get an idea of "p-hacking", defined as attempts by researchers to get the lowest p-values of many experiments, or try until one of the tests produces statistical significance.

Proposition 16.3
The distribution of the minimum of m observations of statistically identical p-values becomes (under the limiting distribution of proposition 16.2):

$$\varphi_m(p; p_M) = m\, e^{erfc^{-1}(2p_M)\left(2erfc^{-1}(2p) - erfc^{-1}(2p_M)\right)}$$

$$\left(1 - \frac{1}{2}erfc\left(erfc^{-1}(2p) - erfc^{-1}(2p_M)\right)\right)^{m-1} \quad (16.5)$$

Proof. $P\left(p_1 > p, p_2 > p, \ldots, p_m > p\right) = \bigcap_{i=1}^{n} \Phi(p_i) = \bar{\Phi}(p)^m$. Taking the first derivative we get the result. □

Outside the limiting distribution: we integrate numerically for different values of m as shown in figure 16.4. So, more precisely, for m trials, the expectation is calculated as:

$$\mathbb{E}(p_{min}) = \int_0^1 -m\, \varphi(p; p_M)\left(\int_0^p \varphi(u, .)\, \mathrm{d}u\right)^{m-1} \mathrm{d}p$$
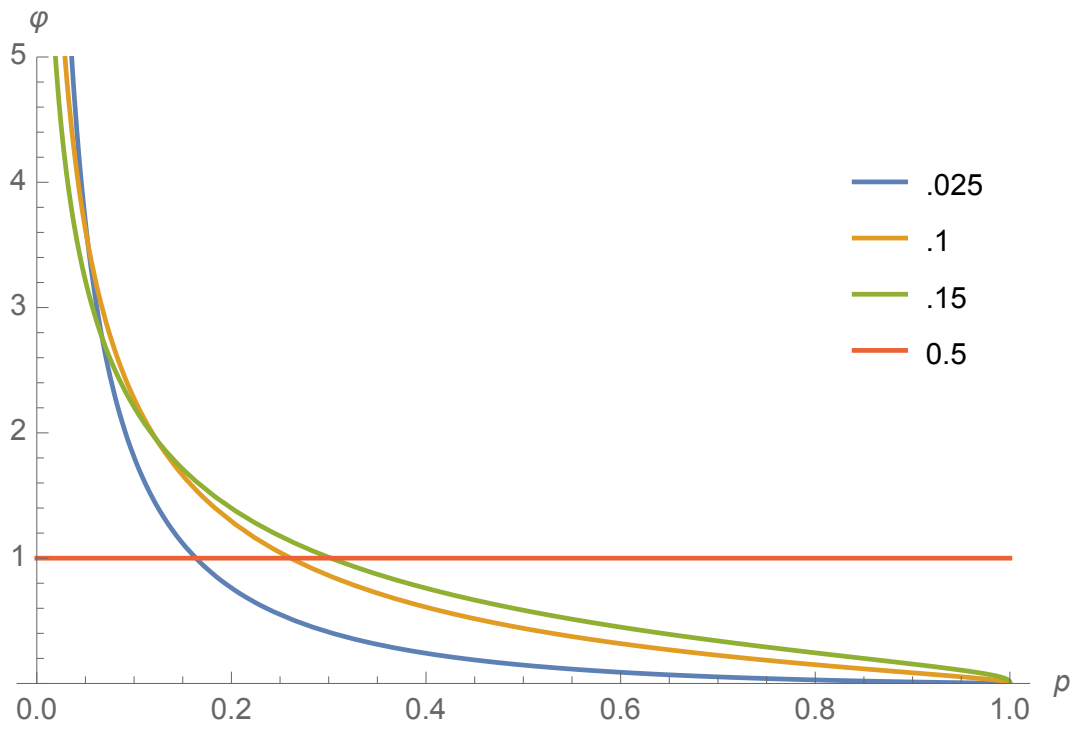
Figure 16.3: *The probability distribution of p at different values of* p_M. *We observe how* $p_M = \frac{1}{2}$ *leads to a uniform distribution.*
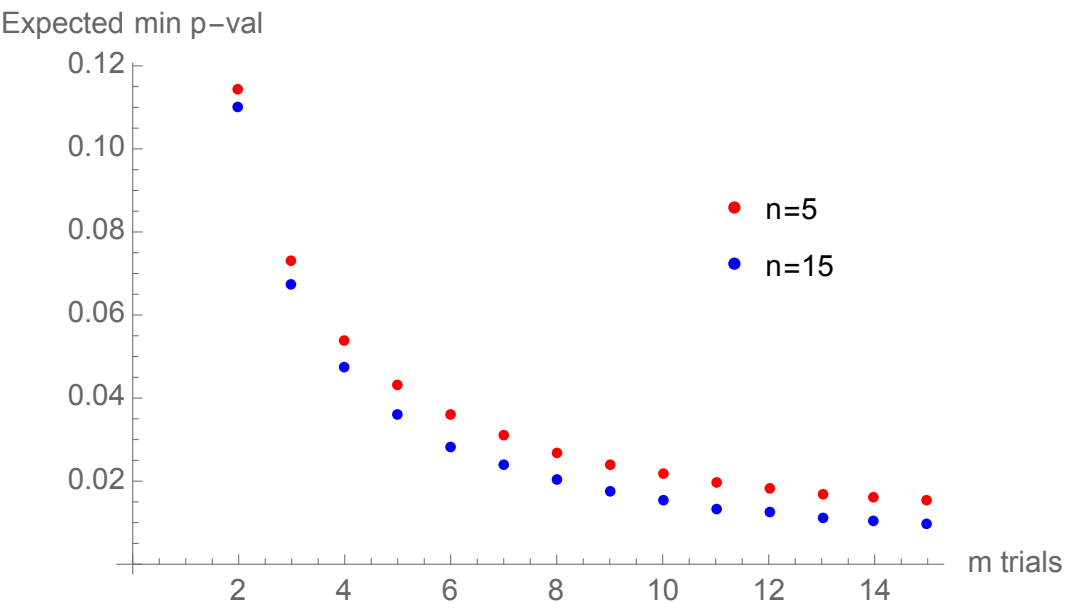

Figure 16.4: *The "p-hacking" value across m trials for* $p_M = .15$ *and* $p_s = .22$.

16.2 INVERSE POWER OF TEST

Let β be the power of a test for a given p-value p, for random draws X from unobserved parameter θ and a sample size of n. To gauge the reliability of β as a true measure of power, we perform an inverse problem:

$$\beta \longrightarrow X_{\theta,p,n}$$
$$\Delta \uparrow \quad \diagup$$
$$\beta^{-1}(X)$$

Proposition 16.4
Let β_c be the projection of the power of the test from the realizations assumed to be student T distributed and evaluated under the parameter θ. We have

$$\Phi(\beta_c) = \begin{cases} \Phi(\beta_c)_L & \text{for } \beta_c < \frac{1}{2} \\ \Phi(\beta_c)_H & \text{for } \beta_c > \frac{1}{2} \end{cases}$$

where

$$\Phi(\beta_c)_L = \sqrt{1-\gamma_1}\gamma_1^{-\frac{n}{2}}$$
$$\frac{\left(-\dfrac{\gamma_1}{2\sqrt{\frac{1}{\gamma_3}-1}\sqrt{-(\gamma_1-1)\gamma_1}-2\sqrt{-(\gamma_1-1)\gamma_1}+\gamma_1\left(2\sqrt{\frac{1}{\gamma_3}-1}-\frac{1}{\gamma_3}\right)-1}\right)^{\frac{n+1}{2}}}{\sqrt{-(\gamma_1-1)\gamma_1}} \quad (16.6)$$

$$\Phi(\beta_c)_H = \sqrt{\gamma_2}\,(1-\gamma_2)^{-\frac{n}{2}}\,B\left(\frac{1}{2},\frac{n}{2}\right)$$
$$\frac{\left(\dfrac{1}{\frac{-2\left(\sqrt{-(\gamma_2-1)\gamma_2}+\gamma_2\right)\sqrt{\frac{1}{\gamma_3}-1}+2\sqrt{\frac{1}{\gamma_3}-1}+2\sqrt{-(\gamma_2-1)\gamma_2}-1}{\gamma_2-1}}+\frac{1}{\gamma_3}\right)^{\frac{n+1}{2}}}{\sqrt{-(\gamma_2-1)\gamma_2}B\left(\frac{n}{2},\frac{1}{2}\right)} \quad (16.7)$$

where $\gamma_1 = I_{2\beta_c}^{-1}\left(\frac{n}{2},\frac{1}{2}\right)$, $\gamma_2 = I_{2\beta_c-1}^{-1}\left(\frac{1}{2},\frac{n}{2}\right)$, *and* $\gamma_3 = I_{(1,2p_s-1)}^{-1}\left(\frac{n}{2},\frac{1}{2}\right)$.

16.3 APPLICATION AND CONCLUSION

- One can safely see that under such stochasticity for the realizations of p-values and the distribution of its minimum, to get what people mean by 5% confidence (and the inferences they get from it), they need a p-value of at least one order of magnitude smaller.

- Attempts at replicating papers, such as the open science project [36], should consider a margin of error in *its own* procedure and a pronounced bias towards favorable results (Type-I error). There should be no surprise that a previously deemed significant test fails during replication –in fact it is the replication of results deemed significant at a close margin that should be surprising.

- The "power" of a test has the same problem unless one either lowers p-values or sets the test at higher levels, such at .99.

ACKNOWLEDGMENT

17 | ELECTION PREDICTIONS AS MARTINGALES: AN ARBITRAGE APPROACH‡

WE EXAMINE the effect of uncertainty on binary outcomes, with application to elections. A standard result in quantitative finance is that when the volatility of the underlying security increases, arbitrage pressures push the corresponding binary option to trade closer to 50%, and become less variable over the remaining time to expiration. Counterintuitively, the higher the uncertainty of the underlying security, the lower the volatility of the binary option. This effect should hold in all domains where a binary price is produced – yet we observe severe violations of these principles in many areas where binary forecasts are made, in particular those concerning the U.S. presidential election of 2016. We observe stark errors among political scientists and forecasters, for instance with 1) assessors giving the candidate D. Trump between 0.1% and 3% chances of success , 2) jumps in the revisions of forecasts from 48% to 15%, both made while invoking uncertainty.

Conventionally, the quality of election forecasting has been assessed statically by De Finetti's method, which consists in minimizing the Brier score , a metric of divergence from the final outcome (the standard for tracking the accuracy of probability assessors across domains, from elections to weather). No intertemporal evaluations of changes in estimates appear to have been imposed outside the quantitative finance practice and literature. Yet De Finetti's own principle is that a probability should be treated like a two-way "choice" price, which is thus violated by conventional practice.

In this paper we take a dynamic, continuous-time approach based on the principles of quantitative finance and argue that a probabilistic estimate of an election outcome by a given "assessor" needs be treated like a tradable price, that is, as a binary option value subjected to arbitrage boundaries (particularly since binary options are actually used in betting markets). Future revised estimates need to be compatible with martingale pricing, otherwise intertemporal arbitrage is created, by "buying" and "selling" from the assessor.

A mathematical complication arises as we move to continuous time and apply the standard martingale approach: namely that as a probability forecast, the underlying security lives in $[0, 1]$. Our approach is to create a dual (or "shadow") martingale process Y, in an interval $[L, H]$ from an arithmetic Brownian motion, X in $(-\infty, \infty)$ and price elections accordingly. The dual process Y can for example represent the numerical votes needed for success. A complication is that, because of the transformation from X to Y, if Y is a martingale, X cannot be a martingale (and vice-versa).

The process for Y allows us to build an arbitrage relationship between the volatility of a probability estimate and that of the underlying variable, e.g. the vote number. Thus we are
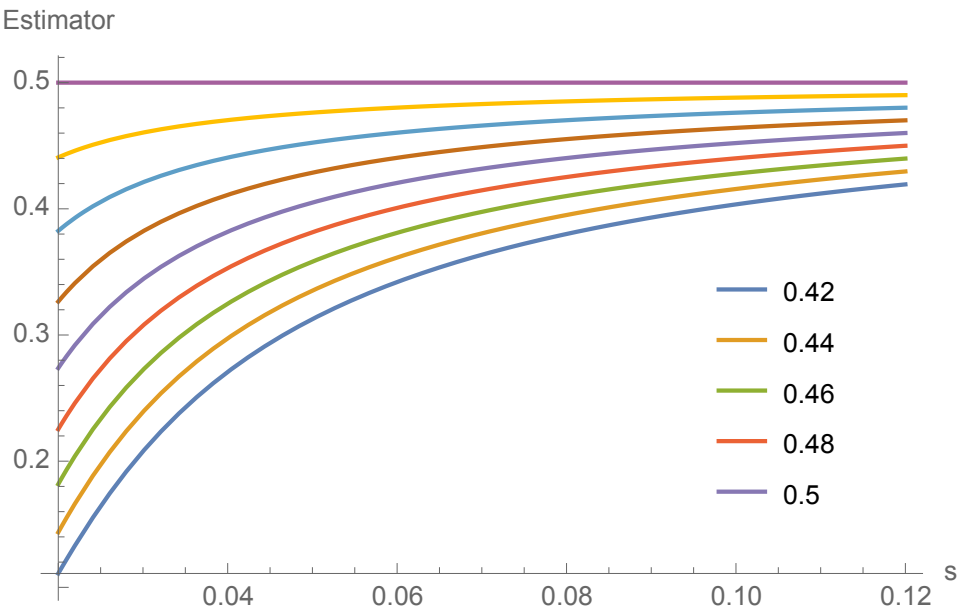
Figure 17.1: *Election arbitrage "estimation" (i.e., valuation) at different expected proportional votes $Y \in [0,1]$, with s the expected volatility of Y between present and election results. We can see that under higher uncertainty, the estimation of the result gets closer to 0.5, and becomes insensitive to estimated electoral margin.*
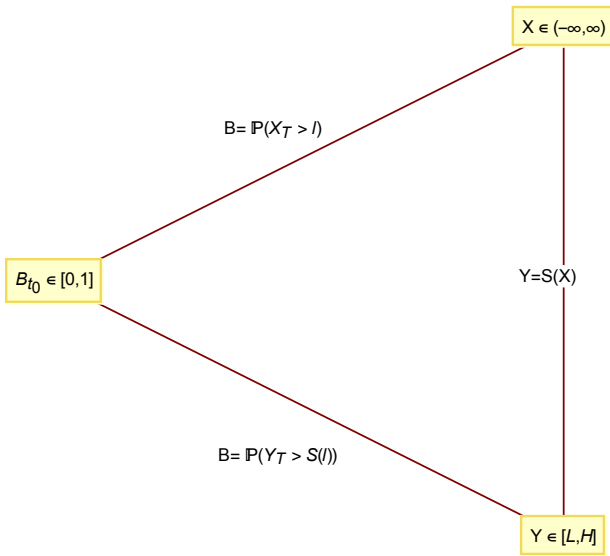

Figure 17.2: *X is an open non observable random variable (a shadow variable of sorts) on $\mathbb{R}$, Y, its mapping into "votes" or "electoral votes" via a sigmoidal function S(.), which maps one-to-one, and the binary as the expected value of either using the proper corresponding distribution.*

able to show that when there is a high uncertainty about the final outcome, 1) indeed, the

arbitrage value of the forecast (as a binary option) gets closer to 50% and 2) the estimate should not undergo large changes even if polls or other bases show significant variations.[1]

The pricing links are between 1) the binary option value (that is, the forecast probability), 2) the estimation of Y and 3) the volatility of the estimation of Y over the remaining time to expiration (see Figures 17.1 and 17.2).

17.0.1 Main results

For convenience, we start with our notation.

Notation

Y_0 the observed estimated proportion of votes expressed in $[0, 1]$ at time t_0. These can be either popular or electoral votes, so long as one treats them with consistency.

T period when the irrevocable final election outcome Y_T is revealed, or expiration.

t_0 present evaluation period, hence $T - t_0$ is the time until the final election, expressed in years.

s annualized volatility of Y, or uncertainty attending outcomes for Y in the remaining time until expiration. We assume s is constant without any loss of generality –but it could be time dependent.

$B(.)$ "forecast probability", or estimated continuous-time arbitrage evaluation of the election results, establishing arbitrage bounds between $B(.)$, Y_0 and the volatility s.

Main results

$$B(Y_0, \sigma, t_0, T) = \frac{1}{2} \text{erfc} \left(\frac{l - \text{erf}^{-1}(2Y_0 - 1) e^{\sigma^2(T - t_0)}}{\sqrt{e^{2\sigma^2(T - t_0)} - 1}} \right), \tag{17.1}$$

where

$$\sigma \approx \frac{\sqrt{\log \left(2\pi s^2 e^{2 \text{erf}^{-1}(2Y_0 - 1)^2} + 1 \right)}}{\sqrt{2}\sqrt{T - t_0}}, \tag{17.2}$$

l is the threshold needed (defaults to .5), and erfc(.) is the standard complementary error function, 1-erf(.), with $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$. □

We find it appropriate here to answer the usual comment by statisticians and people operating outside of mathematical finance: "why not simply use a Beta-style distribution for Y?". The answer is that 1) the main purpose of the paper is establishing (arbitrage-free) time consistency in binary forecasts, and 2) we are not aware of a continuous time stochastic process that accommodates a beta distribution or a similarly bounded conventional one.

---

[1] A central property of our model is that it prevents $B(.)$ from varying more than the estimated Y: in a two candidate contest, it will be capped (floored) at Y if lower (higher) than .5. In practice, we can observe probabilities of winning of 98% vs. 02% from a narrower spread of estimated votes of 47% vs. 53%; our approach prevents, under high uncertainty, the probabilities from diverging away from the estimated votes. But it remains conservative enough to not give a higher proportion.

17.0.2 Organization

The remaining parts of the paper are organized as follows. First, we show the process for Y and the needed transformations from a specific Brownian motion. Second, we derive the arbitrage relationship used to obtain equation (17.1). Finally, we discuss De Finetti's approach and show how a martingale valuation relates to minimizing the conventional standard in the forecasting industry, namely the Brier Score.

A comment on absence of closed form solutions for σ We note that for Y we lack a closed form solution for the integral reflecting the total variation: $\int_{t_0}^{T} \frac{\sigma}{\sqrt{\pi}} e^{-\text{erf}^{-1}(2y_s-1)^2} ds$, though the corresponding one for X is computable. Accordingly, we have relied on propagation of uncertainty methods to obtain a closed form solution for the probability density of Y, though not explicitly its moments as the logistic normal integral does not lend itself to simple expansions [135].

Time slice distributions for X and Y The time slice distribution is the probability density function of Y from time t, that is the one-period representation, starting at t with $y_0 = \frac{1}{2} + \frac{1}{2}\text{erf}(x_0)$. Inversely, for X given y_0, the corresponding x_0, X may be found to be normally distributed for the period $T - t_0$ with

$$\mathbb{E}(X, T) = X_0 e^{\sigma^2(T-t_0)},$$

$$\mathbb{V}(X, T) = \frac{e^{2\sigma^2(T-t_0)} - 1}{2}$$

and a kurtosis of 3. By probability transformation we obtain φ, the corresponding distribution of Y with initial value y_0 is given by

$$\varphi(y; y_0, T) = \frac{1}{\sqrt{e^{2\sigma^2(t-t_0)} - 1}} \exp\left\{ \text{erf}^{-1}(2y - 1)^2 \right.$$

$$\left. - \frac{1}{2}\left(\coth\left(\sigma^2 t\right) - 1\right)\left(\text{erf}^{-1}(2y - 1) - \text{erf}^{-1}(2y_0 - 1)e^{\sigma^2(t-t_0)}\right)^2\right\}$$

(17.3)

and we have $\mathbb{E}(Y_t) = Y_0$.

As to the variance, $\mathbb{E}(Y^2)$, as mentioned above, does not lend itself to a closed-form solution derived from $\varphi(.)$, nor from the stochastic integral; but it can be easily estimated from the closed form distribution of X using methods of propagation of uncertainty for the first two moments (the delta method).

Since the variance of a function f of a finite moment random variable X can be approximated as $V(f(X)) = f'(\mathbb{E}(X))^2 V(X)$:

$$\left.\frac{\partial S^{-1}(y)}{\partial y}\right|_{y=Y_0} s^2 \approx \frac{e^{2\sigma^2(T-t_0)} - 1}{2}$$

$$s \approx \sqrt{\frac{e^{-2\text{erf}^{-1}(2Y_0-1)^2}\left(e^{2\sigma^2(T-t_0)} - 1\right)}{2\pi}}.$$

(17.4)

Likewise for calculations in the opposite direction, we find

$$\sigma \approx \frac{\sqrt{\log\left(2\pi s^2 e^{2\mathrm{erf}^{-1}(2Y_0-1)^2} + 1\right)}}{\sqrt{2}\sqrt{T-t_0}},$$

which is (17.2) in the presentation of the main result.

Note that expansions including higher moments do not bring a material increase in precision – although s is highly nonlinear around the center, the range of values for the volatility of the total or, say, the electoral college is too low to affect higher order terms in a significant way, in addition to the boundedness of the sigmoid-style transformations.

17.0.3 A Discussion on Risk Neutrality

We apply risk neutral valuation, for lack of conviction regarding another way, as a default option. Although Y may not necessarily be tradable, adding a risk premium for the process involved in determining the arbitrage valuation would necessarily imply a negative one for the other candidate(s), which is hard to justify. Further, option values or binary bets, need to satisfy a no Dutch Book argument (the De Finetti form of no-arbitrage) (see [?]), i.e. properly priced binary options interpreted as probability forecasts give no betting "edge" in all outcomes without loss. Finally, any departure from risk neutrality would degrade the Brier score (about which, below) as it would represent a diversion from the final forecast.

Also note the absence of the assumptions of financing rate usually present in financial discussions.

17.1 THE BACHELIER–STYLE VALUATION

Let $F(.)$ be a function of a variable X satisfying

$$dX_t = \sigma^2 X_t dt + \sigma\, dW_t. \tag{17.5}$$

We wish to show that X has a simple Bachelier option price $B(.)$. The idea of no arbitrage is that a continuously made forecast must itself be a martingale.
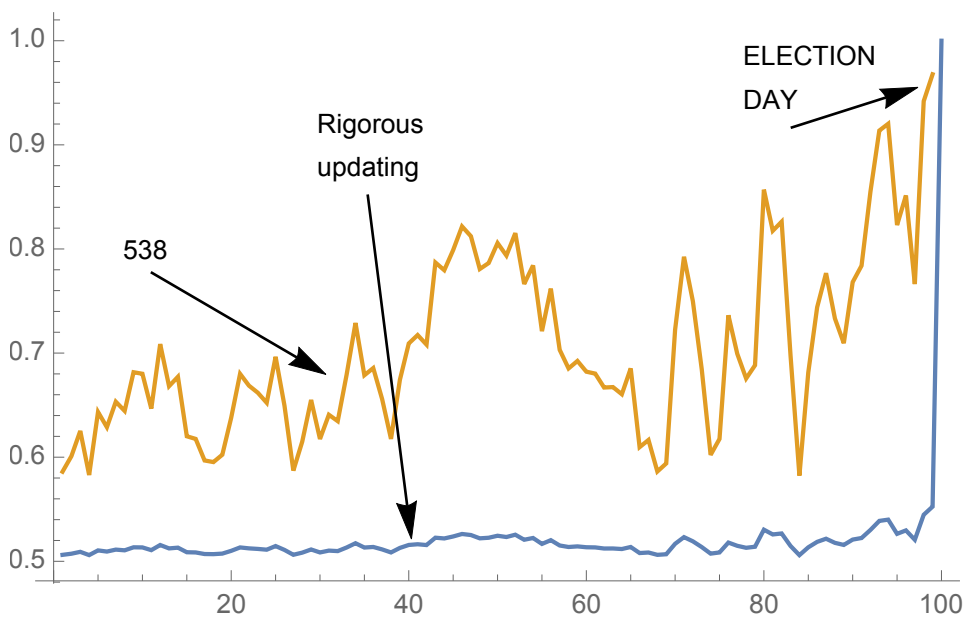
Applying Itô's Lemma to $F \triangleq B$ for X satisfying (17.5) yields

$$dF = \left[\sigma^2 X \frac{F}{X} + \frac{1}{2}\sigma^2 \frac{{}^2 F}{X^2} + \frac{F}{t}\right] dt + \sigma \frac{F}{X} dW$$

so that, since $\frac{F}{t} \triangleq 0$, F must satisfy the partial differential equation

$$\frac{1}{2}\sigma^2 \frac{{}^2 F}{X^2} + \sigma^2 X \frac{F}{X} + \frac{F}{t} = 0, \tag{17.6}$$

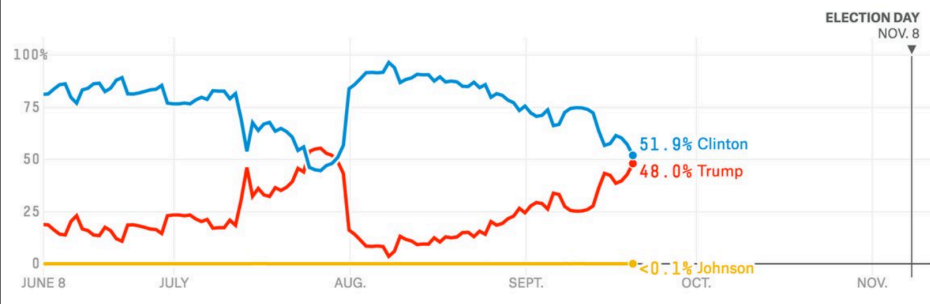which is the driftless condition that makes B a martingale.

Figure 17.3: *Theoretical approach (top) vs practice (bottom). Shows how the estimation process cannot be in sync with the volatility of the estimation of (electoral or other) votes as it violates arbitrage boundaries*

For a binary (call) option, we have for terminal conditions $B(X, t) \triangleq F, F_T = \theta(x - l)$, where $\theta(.)$ is the Heaviside theta function and l is the threshold:

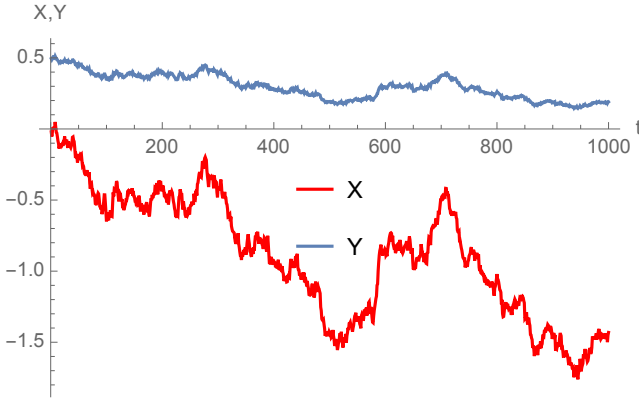$$\theta(x) := \begin{cases} 1, & x \geq l \\ 0, & x < l \end{cases}$$

with initial condition x_0 at time t_0 and terminal condition at T given by:

$$\frac{1}{2}\mathrm{erfc}\left(\frac{x_0 e^{\sigma^2 t} - l}{\sqrt{e^{2\sigma^2 t} - 1}}\right)$$

which is, simply, the survival function of the Normal distribution parametrized under the process for X.

Likewise we note from the earlier argument of one-to one (one can use Borel set arguments) that

$$\theta(y) := \begin{cases} 1, & y \geq S(l) \\ 0, & y < S(l), \end{cases}$$

so we can price the alternative process $B(Y, t) = \mathbb{P}(Y > \frac{1}{2})$ (or any other similarly obtained threshold l, by pricing

$$B(Y_0, t_0) = \mathbb{P}(x > S^{-1}(l)).$$

The pricing from the proportion of votes is given by:

$$B(Y_0, \sigma, t_0, T) = \frac{1}{2}\mathrm{erfc}\left(\frac{l - \mathrm{erf}^{-1}(2Y_0 - 1)e^{\sigma^2(T - t_0)}}{\sqrt{e^{2\sigma^2(T - t_0)} - 1}}\right),$$

the main equation (17.1), which can also be expressed less conveniently as

$$B(y_0, \sigma, t_0, T) = \frac{1}{\sqrt{e^{2\sigma^2 t} - 1}} \int_l^1 \exp\left(\mathrm{erf}^{-1}(2y - 1)^2 \right.$$
$$\left. - \frac{1}{2}\left(\coth\left(\sigma^2 t\right) - 1\right)\left(\mathrm{erf}^{-1}(2y - 1) - \mathrm{erf}^{-1}(2y_0 - 1)e^{\sigma^2 t}\right)^2\right) dy$$

17.2 BOUNDED DUAL MARTINGALE PROCESS

Y_T is the terminal value of a process on election day. It lives in $[0, 1]$ but can be generalized to the broader $[L, H]$, $L, H \in [0, \infty)$. The threshold for a given candidate to win is fixed at l. Y can correspond to raw votes, electoral votes, or any other metric. We assume that Y_t is an intermediate realization of the process at t, either produced synthetically from polls (corrected estimates) or other such systems.

Next, we create, for an unbounded arithmetic stochastic process, a bounded "dual" stochastic process using a sigmoidal transformation. It can be helpful to map processes such as a bounded electoral process to a Brownian motion, or to map a bounded payoff to an unbounded one, see Figure 17.2.

Proposition 17.1
Under sigmoidal style transformations $S : x \mapsto y, \mathbb{R} \to [0,1]$ of the form a) $\frac{1}{2} + \frac{1}{2}erf(x)$, or b) $\frac{1}{1+\exp(-x)}$, if X is a martingale, Y is only a martingale for $Y_0 = \frac{1}{2}$, and if Y is a martingale, X is only a martingale for $X_0 = 0$.

Proof. The proof is sketched as follows. From Itô's lemma, the drift term for dX_t becomes 1) $\sigma^2 X(t)$, or 2) $\frac{1}{2}\sigma^2 \text{Tanh}\left(\frac{X(t)}{2}\right)$, where σ denotes the volatility, respectively with transformations of the forms a) of X_t and b) of X_t under a martingale for Y. The drift for dY_t becomes: 1) $\frac{\sigma^2 e^{-erf^{-1}(2Y-1)^2} erf^{-1}(2Y-1)}{\sqrt{\pi}}$ or 2) $\frac{1}{2}\sigma^2 Y(Y-1)(2Y-1)$ under a martingale for X. □

We therefore select the case of Y being a martingale and present the details of the transformation a). The properties of the process have been developed by Carr [24]. Let X be the arithmetic Brownian motion (17.5), with X-dependent drift and constant scale σ:

$$dX_t = \sigma^2 X_t dt + \sigma dW_t, \quad 0 < t < T < +\infty.$$

We note that this has similarities with the Ornstein-Uhlenbeck process normally written $dX_t = \theta(\mu - X_t)dt + \sigma dW$, except that we have $\mu = 0$ and violate the rules by using a negative mean reversion coefficient, rather more adequately described as "mean repelling", $\theta = -\sigma^2$.

We map from $X \in (-\infty, \infty)$ to its dual process Y as follows. With $S : \mathbb{R} \to [0,1]$, $Y = S(x)$,

$$S(x) = \frac{1}{2} + \frac{1}{2}erf(x)$$

the dual process (by unique transformation since S is one to one, becomes, for $y \triangleq S(x)$, using Ito's lemma (since S(.) is twice differentiable and $S/t = 0$):

$$dS = \left(\frac{1}{2}\sigma^2\frac{\partial^2 S}{\partial x^2} + X\sigma^2\frac{\partial S}{\partial x}\right) dt + \sigma\frac{\partial S}{\partial x}dW$$

which with zero drift can be written as a process

$$dY_t = s(Y)dW_t,$$

for all $t > \tau, \mathbb{E}(Y_t|Y_\tau) = Y_\tau$. and scale

$$s(Y) = \frac{\sigma}{\sqrt{\pi}}e^{-erf^{-1}(2y-1)^2}$$

which as we can see in Figure 17.5, $s(y)$ can be approximated by the quadratic function $y(1-y)$ times a constant.

We can recover equation (17.5) by inverting, namely $S^{-1}(y) = erf^{-1}(2y-1)$, and again applying Itô's Lemma. As a consequence of gauge invariance option prices are identical whether priced on X or Y, even if one process has a drift while the other is a martingale . In other words, one may apply one's estimation to the electoral threshold, or to the more
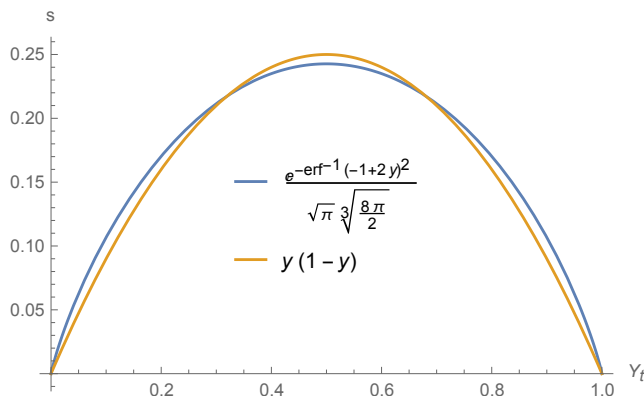
Figure 17.5: *The instantaneous volatility of Y as a function of the level of Y for two different methods of transformations of X, which appear to not be substantially different. We compare to the quadratic form* $y - y^2$ *scaled by a constant* $\frac{1}{\sqrt[3]{\frac{8\pi}{2}}}$. *The volatility declines as we move away from* $\frac{1}{2}$ *and collapses at the edges, thus maintaining Y in* $(0,1)$. *For simplicity we assumed* $\sigma = t = 1$.

complicated X with the same results. And, to summarize our method, pricing an option on X is familiar, as it is exactly a Bachelier-style option price.

17.3 RELATION TO DE FINETTI'S PROBABILITY ASSESSOR

Figure 17.6: *Bruno de Finetti (1906-1985). A probabilist, philosopher, and insurance mathematician, he formulated the Brier score for probabilistic assessment which we show is compatible dynamically with a martingale. Source: DeFinetti.org*

This section provides a brief background for the conventional approach to probability assessment. The great De Finetti [44] has shown that the "assessment" of the "probability" of the realization of a random variable in $\{0, 1\}$ requires a nonlinear loss function – which makes his definition of *probabilistic assessment* differ from that of the P/L of a trader engaging in binary bets.

Assume that a betting agent in an n-repeated two period model, t_0 and t_1, produces a strategy $\mathfrak{S}$ of bets $b_{0,i} \in [0,1]$ indexed by $i = 1, 2, \ldots, n$, with the realization of the binary r.v. $\mathbb{1}_{t_1,i}$. If we take the absolute variation of his P/L over n bets, it will be

$$L_1(\mathfrak{S}) = \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{1}_{t_1,i} - b_{t_0,i} \right|.$$

For example, assume that $\mathbb{E}(\mathbb{1}_{t_1}) = \frac{1}{2}$. Betting on the probability, here $\frac{1}{2}$, produces a loss of $\frac{1}{2}$ in expectation, which is the same as betting either 0 or 1 – hence not favoring the agent to bet on the exact probability.

If we work with the same random variable and non-time-varying probabilities, the L^1 metric would be appropriate:

$$L_1(\mathfrak{S}) = \frac{1}{n} \left| \mathbb{1}_{t_1,i} - \sum_{i=1}^{n} b_{t_0,i} \right|.$$

De Finetti proposed a "Brier score" type function, a quadratic loss function in $\mathcal{L}^2$:

$$L_2(\mathfrak{S}) = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}_{t_1,i} - b_{t_0,i})^2,$$

the minimum of which is reached for $b_{t_0,i} = \mathbb{E}(\mathbb{1}_{t_1})$.

In our world of continuous time derivative valuation, where, in place of a two period lattice model, we are interested, for the same final outcome at t_1, in the stochastic process b_t, $t_0 \geq t \geq t_1$, the arbitrage "value" of a bet on a binary outcome needs to match the expectation, hence, again, we map to the Brier score – *by an arbitrage argument*. Although there is no quadratic loss function involved, the fact that the bet is a function of a martingale, which is required to be itself a martingale, i.e. that the conditional expectation remains invariant to time, does not allow an arbitrage to take place. A "high" price can be "shorted" by the arbitrageur, a "low" price can be "bought", and so on repeatedly. The consistency between bets at period t and other periods $t + \Delta t$ enforces the probabilistic discipline. In other words, someone can "buy" from the forecaster then "sell" back to him, generating a positive expected "return" if the forecaster is out of line with martingale valuation.

As to the current practice by forecasters, although some election forecasters appear to be aware of the need to minimize their Brier score , the idea that the revisions of estimates should also be subjected to martingale valuation is not well established.

17.4 CONCLUSION AND COMMENTS

As can be seen in Figure 17.1, a binary option reveals more about uncertainty than about the true estimation, a result well known to traders, see [156].

In the presence of more than 2 candidates, the process can be generalized with the following heuristic approximation. Establish the stochastic process for $Y_{1,t}$, and just as $Y_{1,t}$ is a process in $[0,1]$, $Y_{2,t}$ is a process $\in (Y_{1,t}, 1]$, with $Y_{3,t}$ the residual $1 - Y_{2,t} - Y_{1,t}$, and more generally $Y_{n-1,t} \in (Y_{n_2,t}, 1]$ and $Y_{n,t}$ is the residual $Y_n = 1 - \sum_{i=1}^{n-1} Y_{i,t}$. For n candidates, the nth is the residual.

17.5   ACKNOWLEDGEMENTS

Part VIII

OPTION TRADING AND PRICING UNDER FAT TAILS

18 | FINANCIAL THEORY'S FAILURES IN OPTION PRICING

†

ᴇᴛ ᴜs ᴅɪsᴄᴜss why option theory as according to the so-called "neoclassical economics" fails in the real world. How does financial theory price financial products? The principal difference in paradigm between the one presented by Bachelier in 1900, [4] and the modern finance one known as Black-Scholes-Merton [16] and [116] lies in the following.

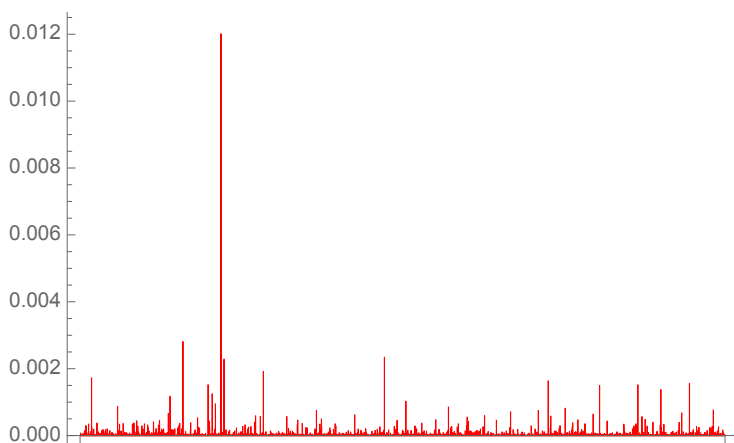Figure 18.1: *The hedging errors for an option portfolio (under a daily revision regime) over 3000 days, under a constant volatility Student T with tail exponent α = 3. Technically the errors should not converge in finite time as their distribution has infinite variance.*
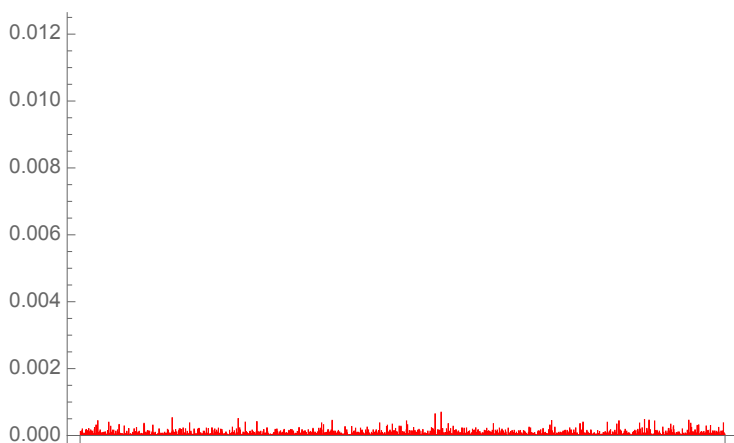

Figure 18.2: *Hedging errors for an option portfolio (daily revision) under an equivalent (rather fictional) "Black-Scholes" world.*
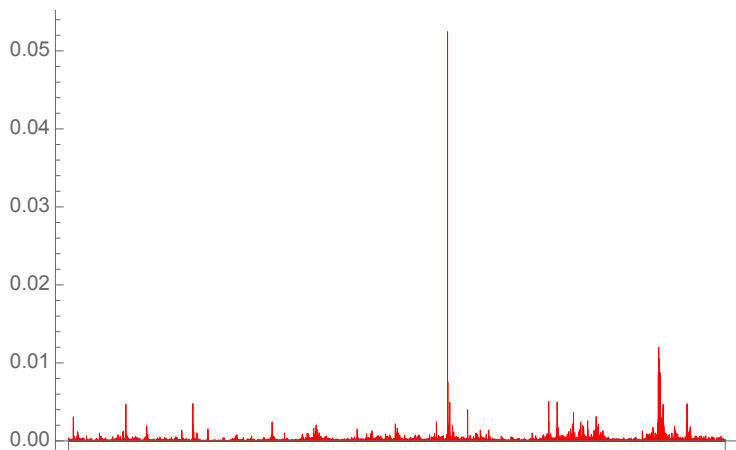
Figure 18.3: *Portfolio Hedging errors including the stock market crash of 1987.*

Bachelier's model is based on an actuarial expectation of final payoffs –not dynamic hedging. It means you can use any distribution! A more formal proof using measure theory is provided in Chapter 19 so for now let us just get the intuition without too much mathematics.

The same method was later used by a series of researchers, such as Sprenkle [150] in 1964, Boness, [18] in 1964, Kassouf and Thorp, [178] in 1967, Thorp, [175] (only published in 1973).

They all encountered the following problem: how to produce a risk parameter –a risky asset discount rate – to make it compatible with portfolio theory? The Capital Asset Pricing Model requires that securities command an expected rate of return in proportion to their riskiness. In the Black-Scholes-Merton approach, an option price is derived from continuous-time dynamic hedging, and only in properties obtained from continuous time dynamic hedging –we will describe dynamic hedging in some details further down. Thanks to such a method, an option collapses into a deterministic payoff and provides returns independent of the market; hence it does not require any risk premium.

The problem we have with the Black-Scholes-Merton approach is that the requirements for dynamic hedging are extremely idealized, requiring the following strict conditions. The operator is assumed to be able to buy and sell in a frictionless market, incurring no transaction costs. The procedure does not allow for the price impact of the order flow –if an operator sells a quantity of shares, it should not have consequences on the subsequent price. The operator knows the probability distribution, which is the Gaussian, with fixed and constant parameters through time (all parameters do not change). Finally, the most significant restriction: no scalable jumps. In a subsequent revision [Merton, 1976] allows for jumps but these are deemed to be Poisson arrival time, and fixed or, at the worst, Gaussian. The framework does not allow the use of power laws both in practice and mathematically. Let us examine the mathematics behind the stream of dynamic hedges in the Black-Scholes-Merton equation.

Assume the risk-free interest rate $r=0$ with no loss of generality. The canonical Black-Scholes-Merton model consists in selling a call and purchasing shares of stock that provide a hedge against instantaneous moves in the security. Thus the portfolio π locally "hedged" against exposure to the first moment of the distribution is the following:

$$\pi = -C + \frac{\partial C}{\partial S} S \qquad (18.1)$$

where C is the call price, and S the underlying security.

Take the change in the values of the portfolio

$$\Delta \pi = -\Delta C + \frac{\partial C}{\partial S} \Delta S \tag{18.2}$$

By expanding around the initial values of S, we have the changes in the portfolio in discrete time. Conventional option theory applies to the Gaussian in which all orders higher than $(\Delta S)^2$ and Δt disappears rapidly.

$$\Delta \pi = -\frac{\partial C}{\partial t} \Delta t - \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \Delta S^2 + O \left(\Delta S^3 \right) \tag{18.3}$$

Taking expectations on both sides, we can see from (3) very strict requirements on moment finiteness: *all* moments need to converge. If we include another term, $-\frac{1}{6} \frac{\partial^3 C}{\partial S^3} \Delta S^3$, it may be of significance in a probability distribution with significant cubic or quartic terms. Indeed, although the nth derivative with respect to S can decline very sharply, for options that have a strike K away from the center ot the distribution, it remains that the moments are rising disproportionately fast for that to carry a mitigating effect.

So here we mean all moments need to be finite and losing in impact –no approximation. Note here that the jump diffusion model (Merton,1976) does not cause much trouble since it has all the moments. And the annoyance is that a power law will have every moment higher than α infinite, causing the equation of the Black-Scholes-Merton portfolio to fail.

As we said, the logic of the Black-Scholes-Merton so-called solution thanks to Ito's lemma was that the portfolio collapses into a deterministic payoff. But let us see how quickly or effectively this works in practice.

18.0.1 The actual replication process:

The payoff of a call should be replicated with the following stream of dynamic hedges, the limit of which can be seen here, between t and T

$$\underset{\Delta t \to 0}{\text{Lim}} \left(\sum_{i=1}^{n=T/\Delta t} \frac{\partial C}{\partial S} |_{S=S_{t+(i-1)\Delta t}, t=t+(i-1)\Delta t,} \left(S_{t+i\Delta t} - S_{t+(i-1)\Delta t} \right) \right) \tag{18.4}$$

We break up the period into n increments Δt. Here the hedge ratio $\frac{\partial C}{\partial S}$ is computed as of time $t + (i-1) \Delta t$, but we get the nonanticipating difference between the price at the time the hedge was initiatied and the resulting price at $t+ i \Delta t$.

This is supposed to make the payoff deterministic *at the limit of* $\Delta t \to 0$. In the Gaussian world, this would be an Ito-McKean integral.

18.0.2 Failure:How hedging errors can be prohibitive.

As a consequence of the mathematical property seen above, hedging errors in an cubic α appear to be indistinguishable from those from an infinite variance process. Furthermore such error has a disproportionaly large effect on strikes away from the money.

In short: dynamic hedging in a power law world removes no risk.

The next chapter will use measure theory to show why options can still be risk-neutral.

19 | UNIQUE OPTION PRICING MEASURE WITH NEITHER DYNAMIC HEDGING NOR COMPLETE MARKETS‡

E PRESENT THE proof that under simple assumptions, such as constraints of Put-Call Parity, the probability measure for the valuation of a European option has the mean derived from the forward price which can, but does not have to be the risk-neutral one, under any general probability distribution, bypassing the Black-Scholes-Merton dynamic hedging argument, and without the requirement of complete markets and other strong assumptions. We confirm that the heuristics used by traders for centuries are both more robust, more consistent, and more rigorous than held in the economics literature. We also show that options can be priced using infinite variance (finite mean) distributions.

19.1 BACKGROUND

Option valuations methodologies have been used by traders for centuries, in an effective way (Haug and Taleb,

There have been a couple of predecessors to the present thesis that Put-Call parity is sufficient constraint to enforce some structure at the level of the mean of the underlying distribution, such as Derman and Taleb (2005), Haug and Taleb (2010). These approaches were heuristic, robust though deemed hand-waving (Ruffino and Treussard, [143]). In addition they showed that operators need to use the risk-neutral mean. What this paper does is

- It goes beyond the "handwaving" with formal proofs.

- It uses a completely distribution-free, expectation-based approach and proves the risk-neutral argument without dynamic hedging, and without any distributional assumption.

- Beyond risk-neutrality, it establishes the case of a unique pricing distribution for option prices in the absence of such argument. The forward (or future) price can embed expectations and deviate from the arbitrage price (owing to, say, regulatory or other limitations) yet the options can still be priced at a distibution corresponding to the mean of such a forward.

- It shows how one can *practically* have an option market without "completeness" and without having the theorems of financial economics hold.

These are done with solely two constraints: "horizontal", i.e. put-call parity, and "vertical", i.e. the different valuations across strike prices deliver a probability measure which is shown to be unique. The only economic assumption made here is that the forward exits, is tradable — in the absence of such unique forward price it is futile to discuss standard option pricing. We also require the probability measures to correspond to distributions with finite first moment.

Preceding works in that direction are as follows. Breeden and Litzenberger [22] and Dupire [54], show how option spreads deliver a unique probability measure; there are papers establishing broader set of arbitrage relations between options such as Carr and Madan [26][1].

However 1) none of these papers made the bridge between calls and puts via the forward, thus translating the relationships from arbitrage relations between options delivering a probability distribution into the necessity of lining up to the mean of the distribution of the forward, hence the risk-neutral one (in case the forward is arbitraged.) 2) Nor did any paper show that in the absence of second moment (say, infinite variance), we can price options very easily. Our methodology and proofs make no use of the variance. 3) Our method is vastly simpler, more direct, and robust to changes in assumptions.

We make no assumption of general market completeness. Options are not redundant securities and remain so. Table 1 summarizes the gist of the paper.[2] [3]

19.2 PROOF

Define $C(S_{t_0}, K, t)$ and $P(S_{t_0}, K, t)$ as European-style call and put with strike price K, respectively, with expiration t, and S_0 as an underlying security at times t_0, $t \geq t_0$, and S_t the possible value of the underlying security at time t.

19.2.1 Case 1: Forward as risk-neutral measure

Define $r = \frac{1}{t-t_0} \int_{t_0}^{t} r_s ds$, the return of a risk-free money market fund and $\delta = \frac{1}{t-t_0} \int_{t_0}^{t} \delta_s ds$ the payout of the asset (continuous dividend for a stock, foreign interest for a currency).

We have the arbitrage forward price F_t^Q:

$$F_t^Q = S_0 \frac{(1+r)^{(t-t_0)}}{(1+\delta)^{(t-t_0)}} \approx S_0 \, e^{(r-\delta)(t-t_0)} \tag{19.1}$$

---

1 See also Green and Jarrow [83] and Nachman [119]. We have known about the possibility of risk neutral pricing without dynamic hedging since Harrison and Kreps [88] but the theory necessitates extremely strong –and severely unrealistic –assumptions, such as strictly complete markets and a multiperiod pricing kernel

2 The famed Hakkanson paradox is as follows: if markets are complete and options are redudant, why would someone need them? If markets are incomplete, we may need options but how can we price them? This discussion may have provided a solution to the paradox: markets are incomplete *and* we can price options.

3 Option prices are not unique in the absolute sense: the premium over intrinsic can take an entire spectrum of values; it is just that the put-call parity constraints forces the measures used for puts and the calls to be the same and to have the same expectation as the forward. As far as securities go, options are securities on their own; they just have a strong link to the forward.

Table 19.1: *Main practical differences between the dynamic hedging argument and the static Put-Call parity with spreading across strikes.*

	Black-Scholes Merton	Put-Call Parity with Spreading
Type	Continuous rebalancing.	Interpolative static hedge.
Limit	Law of large numbers in time (horizontal).	Law of large numbers across strikes (vertical).
Market Assumptions	1) Continuous Markets, no gaps, no jumps.	1) Gaps and jumps acceptable. Possibility of continuous Strikes, or acceptable number of strikes.
	2) Ability to borrow and lend underlying asset for all dates.	2) Ability to borrow and lend underlying asset for single forward date.
	3) No transaction costs in trading asset.	3) Low transaction costs in trading options.
Probability Distribution	Requires all moments to be finite. Excludes the class of slowly varying distributions	Requires finite 1st moment (infinite variance is acceptable).
Market Completeness	Achieved through dynamic completeness	Not required (in the traditional sense)
Realism of Assumptions	Low	High
Convergence	Uncertain; one large jump changes expectation	Robust
Fitness to Reality	Only used after "fudging" standard deviations per strike.	Portmanteau, using specific distribution adapted to reality

by arbitrage, see Keynes 1924. We thus call F_t^Q the future (or forward) price obtained by arbitrage, at the risk-neutral rate. Let F_t^P be the future requiring a risk-associated "expected return" m, with expected forward price:

$$F_t^P = S_0(1+m)^{(t-t_0)} \approx S_0 e^{m(t-t_0)}. \tag{19.2}$$

Remark: *By arbitrage, all tradable values of the forward price given S_{t_0} need to be equal to F_t^Q.*

"Tradable" here does not mean "traded", only subject to arbitrage replication by "cash and carry", that is, borrowing cash and owning the secutity yielding d if the embedded forward return diverges from r.

19.2.2 Derivations

In the following we take F as having dynamics on its own –irrelevant to whether we are in case 1 or 2 –hence a unique probability measure Q.

Define $\Omega = [0, \infty) = A_K \cup A_K^c$ where $A_K = [0, K]$ and $A_K^c = (K, \infty)$.

Consider a class of standard (simplified) probability spaces (Ω, μ_i) indexed by i, where μ_i is a probability measure, i.e., satisfying $\int_\Omega d\mu_i = 1$.

Theorem 19.1
For a given maturity T, there is a unique measure μ_Q that prices European puts and calls by expectation of terminal payoff.

This measure can be risk-neutral in the sense that it prices the forward F_t^Q, but does not have to be and imparts rate of return to the stock embedded in the forward.

Lemma 19.1
For a given maturity T, there exist two measures μ_1 and μ_2 for European calls and puts of the same maturity and same underlying security associated with the valuation by expectation of terminal payoff, which are unique such that, for any call and put of strike K, we have:

$$C = \int_\Omega f_C \, d\mu_1, \tag{19.3}$$

and

$$P = \int_\Omega f_P \, d\mu_2, \tag{19.4}$$

respectively, and where f_C and f_P are $(S_t - K)^+$ and $(K - S_t)^+$ respectively.

Proof. For clarity, set r and δ to 0 without a loss of generality. By Put-Call Parity Arbitrage, a positive holding of a call ("long") and negative one of a put ("short") replicates a tradable forward; because of P/L variations, using positive sign for long and negative sign for short:

$$C(S_{t_0}, K, t) - P(S_{t_0}, K, t) + K = F_t^P \tag{19.5}$$

necessarily since F_t^P is tradable.

Put-Call Parity holds for all strikes, so:

$$C(S_{t_0}, K + \Delta K, t) - P(S_{t_0}, K + \Delta K, t) + K + \Delta K = F_t^P \tag{19.6}$$

for all $K \in \Omega$

Now a Call spread in quantities $\frac{1}{\Delta K}$, expressed as

$$C(S_{t_0}, K, t) - C(S_{t_0}, K + \Delta K, t),$$

delivers \$1 if $S_t > K + \Delta K$ (that is, corresponds to the indicator function $\mathbf{1}_{S > K + \Delta K}$), 0 if $S_t \le K$ (or $\mathbf{1}_{S > K}$), and the quantity times $S_t - K$ if $K < S_t \le K + \Delta K$, that is, between 0 and \$1 (see Breeden and Litzenberger, 1978[22]). Likewise, consider the converse argument for a put, with $\Delta K < S_t$.

At the limit, for $\Delta K \to 0$

$$\frac{\partial C(S_{t_0}, K, t)}{\partial K} = -P(S_t > K) = -\int_{A_K^c} d\mu_1. \tag{19.7}$$

By the same argument:

$$\frac{\partial P(S_{t_0}, K, t)}{\partial K} = \int_{A_K} d\mu_2 = 1 - \int_{A_K^c} d\mu_2. \tag{19.8}$$

As semi-closed intervals generate the whole Borel σ-algebra on Ω, this shows that μ_1 and μ_2 are unique.

$\square$

Lemma 19.2
The probability measures of puts and calls are the same, namely for each Borel set A in Ω, $\mu_1(A) = \mu_2(A)$.

Proof. Combining Equations 19.5 and 19.6, dividing by $\frac{1}{\Delta K}$ and taking $\Delta K \to 0$:

$$-\frac{\partial C(S_{t_0}, K, t)}{\partial K} + \frac{\partial P(S_{t_0}, K, t)}{\partial K} = 1 \tag{19.9}$$

for all values of K, so

$$\int_{A_K^c} d\mu_1 = \int_{A_K^c} d\mu_2, \tag{19.10}$$

hence $\mu_1(A_K) = \mu_2(A_K)$ for all $K \in [0, \infty)$. This equality being true for any semi-closed interval, it extends to any Borel set.

$\square$

Lemma 19.3
Puts and calls are required, by static arbitrage, to be evaluated at same as risk-neutral measure μ_Q as the tradable forward.

Proof.
$$F_t^P = \int_\Omega F_t \, d\mu_Q; \tag{19.11}$$

from Equation 19.5

$$\int_\Omega f_C(K) \, d\mu_1 - \int_\Omega f_P(K) \, d\mu_1 = \int_\Omega F_t \, d\mu_Q - K \tag{19.12}$$

Taking derivatives on both sides, and since $f_C - f_P = S_0 + K$, we get the Radon-Nikodym derivative:

$$\frac{d\mu_Q}{d\mu_1} = 1 \tag{19.13}$$

for all values of K.

$\square$

19.3 CASE WHERE THE FORWARD IS NOT RISK NEUTRAL

Consider the case where F_t is observable, tradable, and use it solely as an underlying security with dynamics on its own. In such a case we can completely ignore the dynamics of the nominal underlying S, or use a non-risk neutral "implied" rate linking cash to forward, $m^* = \frac{\log\left(\frac{F}{S_0}\right)}{t-t_0}$. the rate m can embed risk premium, difficulties in financing, structural or regulatory impediments to borrowing, with no effect on the final result.

In that situation, it can be shown that the exact same results as before apply, by remplacing the measure μ_Q by another measure μ_{Q^*}. Option prices remain unique [4].

19.4 COMMENT

We have replaced the complexity and intractability of dynamic hedging with a simple, more benign interpolation problem, and explained the performance of pre-Black-Scholes option operators using simple heuristics and rules, bypassing the structure of the theorems of financial economics.

Options can remain non-redundant and markets incomplete: we are just arguing here for a form of arbitrage pricing (which includes risk-neutral pricing at the level of the expectation of the probability measure), nothing more. But this is sufficient for us to use any probability distribution with finite first moment, which includes the Lognormal, which recovers Black Scholes.

A final comparison. In dynamic hedging, missing a single hedge, or encountering a single gap (a tail event) can be disastrous —as we mentioned, it requires a series of assumptions beyond the mathematical, in addition to severe and highly unrealistic constraints on the mathematical. Under the class of fat tailed distributions, increasing the frequency of the hedges does not guarantee reduction of risk. Further, the standard dynamic hedging argument requires the exact specification of the *risk-neutral* stochastic process between t_0 and t, something econometrically unwieldy, and which is generally reverse engineered from the price of options, as an arbitrage-oriented interpolation tool rather than as a representation of the process.

Here, in our Put-Call Parity based methodology, our ability to track the risk neutral distribution is guaranteed by adding strike prices, and since probabilities add up to 1, the degrees of freedom that the recovered measure μ_Q has in the gap area between a strike price K and the next strike up, $K + \Delta K$, are severely reduced, since the measure in the interval is constrained by the difference $\int_{A_K}^c d\mu - \int_{A_{K+\Delta K}}^c d\mu$. In other words, no single gap between strikes can significantly affect the probability measure, even less the first moment, unlike with dynamic hedging. In fact it is no different from standard kernel smoothing methods for statistical samples, but applied to the distribution across strikes.[5]

The assumption about the presence of strike prices constitutes a natural condition: conditional on having a *practical* discussion about options, options strikes need to exist. Further, as it is the experience of the author, market-makers can add over-the-counter strikes at will, should they need to do so.

---

4 We assumed 0 discount rate for the proofs; in case of nonzero rate, premia are discounted at the rate of the arbitrage operator

5 For methods of interpolation of implied probability distribution between strikes, see Avellaneda et al.[3].

ACKNOWLEDGMENT

20 | OPTION TRADERS NEVER USE THE BLACK–SCHOLES–MERTON FORMULA[*,‡]

PTION TRADERS use a heuristically derived pricing formula which they adapt by fudging and changing the tails and skewness by varying one parameter, the standard deviation of a Gaussian. Such formula is popularly called BlackScholesMerton owing to an attributed eponymous discovery (though changing the standard deviation parameter is in contradiction with it). However, we have historical evidence that: (1) the said Black, Scholes and Merton did not invent any formula, just found an argument to make a well known (and used) formula compatible with the economics establishment, by removing the risk parameter through dynamic hedging, (2) option traders use (and evidently have used since 1902) sophisticated heuristics and tricks more compatible with the previous versions of the formula of Louis Bachelier and Edward O. Thorp (that allow a broad choice of probability distributions) and removed the risk parameter using put-call parity, (3) option traders did not use the BlackScholesMerton formula or similar formulas after 1973 but continued their bottom-up heuristics more robust to the high impact rare event. The chapter draws on historical trading methods and 19th and early 20th century references ignored by the finance literature. It is time to stop using the wrong designation for option pricing.

20.1 BREAKING THE CHAIN OF TRANSMISSION

For us, practitioners, theories should arise from practice [1]. This explains our concern with the scientific notion that practice should fit theory. Option hedging, pricing, and trading is neither philosophy nor mathematics. It is a rich craft with traders learning from traders (or traders copying other traders) and tricks developing under evolution pressures, in a bottom-up manner. It is technë, not ëpistemë. Had it been a science it would not have survived for the empirical and scientific fitness of the pricing and hedging theories offered are, we will see, at best, defective and unscientific (and, at the worst, the hedging methods create more risks than they reduce). Our approach in this paper is to ferret out historical evidence of technë showing how option traders went about their business in the past.

  Options, we will show, have been extremely active in the pre-modern finance world. Tricks and heuristically derived methodologies in option trading and risk management of derivatives books have been developed over the past century, and used quite effectively by

---

1 For us, in this discussion, a practitioner is deemed to be someone involved in repeated decisions about option hedging, not a support quant who writes pricing software or an academic who provides consulting advice.

operators. In parallel, many derivations were produced by mathematical researchers. The economics literature, however, did not recognize these contributions, substituting the re-discoveries or subsequent reformulations done by (some) economists. There is evidence of an attribution problem with Black-Scholes-Merton option formula, which was developed, used, and adapted in a robust way by a long tradition of researchers and used heuristi-cally by option book runners. Furthermore, in a case of scientific puzzle, the exact formula called Black-Sholes-Merton was written down (and used) by Edward Thorp which, para-doxically, while being robust and realistic, has been considered unrigorous. This raises the following: 1) The Black-Scholes-Merton innovation was just a neoclassical finance ar-gument, no more than a thought experiment [2], 2) We are not aware of traders using their argument or their version of the formula.

It is high time to give credit where it belongs.

20.2 BLACK–SCHOLES WAS AN ARGUMENT

Option traders call the formula they use the Black-Scholes-Merton formula without being aware that by some irony, of all the possible options formulas that have been produced in the past century, what is called the Black-Scholes-Merton formula (after Black and Scholes, 1973, and Merton, 1973) is the one the furthest away from what they are using. In fact of the formulas written down in a long history it is the only formula that is fragile to jumps and tail events.

First, something seems to have been lost in translation: Black and Scholes [17] and Merton [117] actually never came up with a new option formula, but only an theoretical economic argument built on a new way of deriving, rather re-deriving, an already existing and well known formula. The argument, we will see, is extremely fragile to assumptions. The foundations of option hedging and pricing were already far more firmly laid down before them. The Black-Scholes-Merton argument, simply, is that an option can be hedged using a certain methodology called dynamic hedging and then turned into a risk-free instrument, as the portfolio would no longer be stochastic. Indeed what Black, Scholes and Merton did was marketing, finding a way to make a well-known formula palatable to the economics establishment of the time, little else, and in fact distorting its essence.

Such argument requires strange far-fetched assumptions: some liquidity at the level of transactions, knowledge of the probabilities of future events (in a neoclassical Arrow-Debreu style) , and, more critically, a certain mathematical structure that requires thin-tails, or mild randomness, on which, later[3]. The entire argument is indeed, quite strange and rather inapplicable for someone clinically and observation-driven standing outside con-ventional neoclassical economics. Simply, the dynamic hedging argument is dangerous in practice as it subjects you to blowups; it makes no sense unless you are concerned with neoclassical economic theory. The Black-Scholes-Merton argument and equation flow a top-down general equilibrium theory, built upon the assumptions of operators working in full knowledge of the probability distribution of future outcomes in addition to a collection of assumptions that, we will see, are highly invalid mathematically, the main one being the

---

2 Here we question the notion of confusing thought experiments in a hypothetical world, of no predictive power, with either science or practice. The fact that the Black-Scholes-Merton argument works in a Platonic world and appears to be elegant does not mean anything since one can always produce a Platonic world in which a certain equation works, or in which a rigorous proof can be provided, a process called reverse-engineering.

3 Of all the misplaced assumptions of Black Scholes that cause it to be a mere thought experiment, though an extremely elegant one, a flaw shared with modern portfolio theory, is the certain knowledge of future delivered variance for the random variable (or, equivalently, all the future probabilities). This is what makes it clash with practice the rectification by the market fattening the tails is a negation of the Black-Scholes thought experiment.

ability to cut the risks using continuous trading which only works in the very narrowly special case of thin-tailed distributions. But it is not just these flaws that make it inapplicable: option traders do not buy theories, particularly speculative general equilibrium ones, which they find too risky for them and extremely lacking in standards of reliability. A normative theory is, simply, not good for decision-making under uncertainty (particularly if it is in chronic disagreement with empirical evidence). People may take decisions based on speculative theories, but avoid the fragility of theories in running their risks.

Yet professional traders, including the authors (and, alas, the Swedish Academy of Science) have operated under the illusion that it was the Black-Scholes-Merton formula they actually used we were told so. This myth has been progressively reinforced in the literature and in business schools, as the original sources have been lost or frowned upon as anecdotal (Merton [118]).
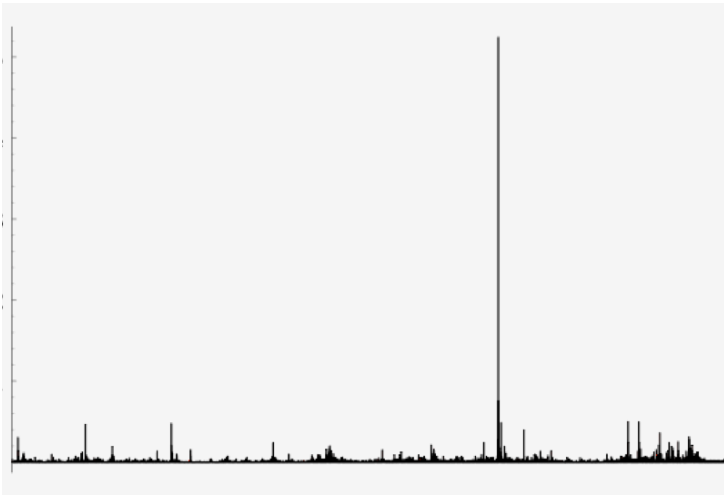

Figure 20.1: *The typical "risk reduction" performed by the Black-Scholes-Merton argument. These are the variations of a dynamically hedged portfolio (and a quite standard one). BSM indeed "smoothes" out variations but exposes the operator to massive tail events reminiscent of such blowups as LTCM. Other option formulas are robust to the rare event and make no such claims.*

This discussion will present our real-world, ecological understanding of option pricing and hedging based on what option traders actually do and did for more than a hundred years.

This is a very general problem. As we said, option traders develop a chain of transmission of technë, like many professions. But the problem is that the chain is often broken as universities do not store the acquired skills by operators. Effectively plenty of robust heuristically derived implementations have been developed over the years, but the economics establishment has refused to quote them or acknowledge them. This makes traders need to relearn matters periodically. Failure of dynamic hedging in 1987, by such firm as Leland OBrien Rubinstein, for instance, does not seem to appear in the academic literature published after the event (Merton, [118], Rubinstein,[142], Ross [141]); to the contrary dynamic hedging is held to be a standard operation [4].

There are central elements of the real world that can escape them academic research without feedback from practice (in a practical and applied field) can cause the diversions we witness between laboratory and ecological frameworks. This explains why some many finance academics have had the tendency to make smooth returns, then blow up using their own theories[5]. We started the other way around, first by years of option trading doing

---

4 For instance how mistakes never resurface into the consciousness, Mark Rubinstein was awarded in 1995 the Financial Engineer of the Year award by the International Association of Financial Engineers. There was no mention of portfolio insurance and the failure of dynamic hedging.

5 For a standard reaction to a rare event, see the following: "Wednesday is the type of day people will remember in quant-land for a very long time," said Mr. Rothman, a University of Chicago Ph.D. who ran a quantitative fund

million of hedges and thousands of option trades. This in combination with investigating the forgotten and ignored ancient knowledge in option pricing and trading we will explain some common myths about option pricing and hedging. There are indeed two myths:

- That we had to wait for the Black-Scholes-Merton options formula to trade the product, price options, and manage option books. In fact the introduction of the Black, Scholes and Merton argument increased our risks and set us back in risk management. More generally, it is a myth that traders rely on theories, even less a general equilibrium theory, to price options.

- That we use the Black-Scholes-Merton options pricing formula. We, simply dont.

In our discussion of these myth we will focus on the bottom-up literature on option theory that has been hidden in the dark recesses of libraries. And that addresses only recorded matters not the actual practice of option trading that has been lost.

20.3   MYTH 1: TRADERS DID NOT "PRICE" OPTIONS BEFORE BLACK–SCHOLES

It is assumed that the Black-Scholes-Merton theory is what made it possible for option traders to calculate their delta hedge (against the underlying) and to price options. This argument is highly debatable, both historically and analytically.

Options were actively trading at least already in the 1600 as described by Joseph De La Vega implying some form of technë, a heuristic method to price them and deal with their exposure. De La Vega describes option trading in the Netherlands, indicating that operators had some expertise in option pricing and hedging. He diffusely points to the put-call parity, and his book was not even meant to teach people about the technicalities in option trading. Our insistence on the use of Put-Call parity is critical for the following reason: The Black-Scholes-Mertons claim to fame is removing the necessity of a risk-based drift from the underlying security to make the trade risk-neutral. But one does not need dynamic hedging for that: simple put call parity can suffice (Derman and Taleb, 2005), as we will discuss later. And it is this central removal of the risk-premium that apparently was behind the decision by the Nobel committee to grant Merton and Scholes the (then called) Bank of Sweden Prize in Honor of Alfred Nobel: Black, Merton and Scholes made a vital contribution by showing that it is in fact not necessary to use any risk premium when valuing an option. This does not mean that the risk premium disappears; instead it is already included in the stock price. It is for having removed the effect of the drift on the value of the option, using a thought experiment, that their work was originally cited, something that was mechanically present by any form of trading and converting using far simpler techniques.

Options have a much richer history than shown in the conventional literature. Forward contracts seems to date all the way back to Mesopotamian clay tablets dating all the way back to 1750 B.C. Gelderblom and Jonker [74] show that Amsterdam grain dealers had used options and forwards already in 1550.

In the late 1800 and the early 1900 there were active option markets in London and New York as well as in Paris and several other European exchanges. Markets it seems, were active and extremely sophisticated option markets in 1870. Kairys and Valerio (1997)

---

discuss the market for equity options in USA in the 1870s, indirectly showing that traders were sophisticated enough to price for tail events[6].

There was even active option arbitrage trading taking place between some of these markets. There is a long list of missing treatises on option trading: we traced at least ten German treatises on options written between the late 1800s and the hyperinflation episode[7]

---

6 The historical description of the market is informative until Kairys and Valerio [96] try to gauge whether options in the 1870s were underpriced or overpriced (using Black-Scholes-Merton style methods). There was one tail-event in this period, the great panic of September 1873. Kairys and Valerio find that holding puts was profitable, but deem that the market panic was just a one-time event : However, the put contracts benefit from the financial panic that hit the market in September, 1873. Viewing this as a one-time event, we repeat the analysis for puts excluding any unexpired contracts written before the stock market panic. Using references to the economic literature that also conclude that options in general were overpriced in the 1950s 1960s and 1970s they conclude: "Our analysis shows that option contracts were generally overpriced and were unattractive for retail investors to purchase. They add: Empirically we find that both put and call options were regularly overpriced relative to a theoretical valuation model." These results are contradicted by the practitioner Nelson (1904): the majority of the great option dealers who have found by experience that it is the givers, and not the takers, of option money who have gained the advantage in the long run.

21 | FOUR LESSONS FROM JEFF HOLMAN'S MISTAKES IN THE DISCUSSION OF ANTIFRAGILE[*,‡]

We discuss Jeff Holman's comments in *Quantitative Finance* to illustrate four critical errors students should learn to avoid: 1) Mistaking tails (4th moment) for volatility (2nd moment), 2) Missing Jensen's Inequality, 3) Analyzing the hedging wihout the underlying, 4) The necessity of a numéraire in finance.

The review of *Antifragile* by Mr Holman (dec 4, 2013) is replete with factual, logical, and analytical errors. We will only list here the critical ones, and ones with generality to the risk management and quantitative finance communities; these should be taught to students in quantitative finance as central mistakes to avoid, so beginner quants and risk managers can learn from these fallacies.

21.1 CONFLATION OF SECOND AND FOURTH MOMENTS

It is critical for beginners not to fall for the following elementary mistake. Mr Holman gets the relation of the VIX (volatility contract) to betting on "tail events" backwards. Let us restate the notion of "tail events" in the *Incerto* (that is the four books on uncertainty of which *Antifragile* is the last installment): it means a disproportionate role of the tails in defining the properties of distribution, which, mathematically, means a smaller one for the "body".[1]

Mr Holman seems to get the latter part of the attributes of fattailedness in reverse. It is an error to mistake the VIX for tail events. The VIX is mostly affected by at-the-money options which corresponds to the center of the distribution, closer to the second moment not the fourth (at-the-money options are actually linear in their payoff and correspond to the conditional first moment). As explained about seventeen years ago in *Dynamic Hedging* (Taleb, 1997) (see appendix), in the discussion on such tail bets, or "fourth moment bets", betting on the disproportionate role of tail events of fattailedness is done by *selling* the around-the-money-options (the VIX) and purchasing options in the tails, in order to extract the second moment and achieve neutrality to it (sort of becoming "market neutral"). Such a neutrality requires some type of "short volatility" in the body because higher kurtosis means lower action in the center of the distribution.

A more mathematical formulation is in the technical version of the *Incerto* : fat tails means "higher peaks" for the distribution as, the fatter the tails, the more markets spend time be-

tween $\mu - \sqrt{\frac{1}{2}\left(5 - \sqrt{17}\right)}\sigma$ and $\mu + \sqrt{\frac{1}{2}\left(5 - \sqrt{17}\right)}\sigma$ where σ is the standard deviation and

---

1 The point is staring at every user of spreadsheets: kurtosis, or scaled fourth moment, the standard measure of fattailedness, entails normalizing the fourth moment by the square of the variance.

μ the mean of the distribution (we used the Gaussian here as a base for ease of presentation but the argument applies to all unimodal distributions with "bell-shape" curves, known as semiconcave). And "higher peaks" means less variations that are not tail events, more quiet times, not less. For the consequence on option pricing, the reader might be interested in a quiz I routinely give students after the first lecture on derivatives: "What happens to at-the-money options when one fattens the tails?", the answer being that they should drop in value. [2]

Effectively, but in a deeper argument, in the QF paper (Taleb and Douady 2013), our measure of fragility has an opposite sensitivity to events around the center of the distribution, since, by an argument of survival probability, what is fragile is sensitive to tail shocks and, critically, should not vary in the body (otherwise it would be broken).

21.2   MISSING JENSEN'S INEQUALITY IN ANALYZING OPTION RETURNS

Here is an error to avoid at all costs in discussions of volatility strategies or, for that matter, anything in finance. Mr Holman seems to miss the existence of Jensen's inequality, which is the entire point of owning an option, a point that has been belabored in *Antifragile*. One manifestation of missing the convexity effect is a critical miscalculation in the way one can naively assume options respond to the VIX.

> "A $1 investment on January 1, 2007 in a strategy of buying and rolling short-term VIX futures would have peaked at $4.84 on November 20, 2008 -and then subsequently lost 99% of its value over the next four and a half years, finishing under $0.05 as of May 31, 2013." [3]

This mistake in the example given underestimates option returns by up to... several orders of magnitude. Mr Holman analyzes the performance a tail strategy using investments in financial options by using the VIX (or VIX futures) as proxy, which is mathematically erroneous owing to second- order effects, as the link is tenuous (it would be like evaluating investments in ski resorts by analyzing temperature futures). Assume a periodic rolling of an option strategy: an option 5 STD away from the money [4] gains 16 times in value if its implied volatility goes up by 4, but only loses its value if volatility goes to 0. For a 10 STD it is 144 times. And, to show the acceleration, assuming these are traded, a 20 STD options by around 210,000 times[5]. There is a second critical mistake in the discussion: Mr Holman's calculations here exclude the payoff from actual in-the-moneyness.

One should remember that the VIX is not a price, but an inverse function, an index derived from a price: one does not buy "volatility" like one can buy a tomato; operators buy options coreponding to such inverse function and there are severe, very severe nonlinearities in the effect. Although more linear than tail options, the VIX is still convex to actual market volatility, somewhere between variance and standard deviation, since a strip of

---

2 **Technical Point: Where Does the Tail Start?** As we saw in 3.3, for a general class of symmetric distributions with power laws, the tail starts at: $\pm \frac{\sqrt{\frac{5\alpha+\sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}}{\sqrt{2}}s$, with α infinite in the stochastic volatility Gaussian case and s the standard deviation. The "tail" is located between around 2 and 3 standard deviations. This flows from the heuristic definition of fragility as second order effect: the part of the distribution is convex to errors in the estimation of the scale. But in practice, because historical measurements of STD will be biased lower because of small sample effects (as we repeat fat tails accentuate small sample effects), the deviations will be > 2-3 STDs.

3 In the above discussion Mr Holman also shows evidence of dismal returns on index puts which, as we said before, respond to volatility not tail events. These are called, in the lingo, "sucker puts".

4 We are using implied volatility as a benchmark for its STD.

5 An event this author witnessed, in the liquidation of Victor Niederhoffer, options sold for $.05 were purchased back at up to $38, which bankrupted Refco, and, which is remarkable, without the options getting close to the money: it was just a panic rise in implied volatility.

options spanning all strikes should deliver the variance (Gatheral,2006). The reader can go through a simple exercise. Let's say that the VIX is "bought" at 10% -that is, the component options are purchased at a combination of volatilities that corresponds to a VIX at that level. Assume returns are in squares. Because of nonlinearity, the package could benefit from an episode of 4% volatility followed by an episode of 15%, for an average of 9.5%; Mr Holman believes or wants the reader to believe that this 0.5 percentage point should be treated as a loss when in fact second order un-evenness in volatility changes are more relevant than the first order effect.

21.3 THE INSEPARABILITY OF INSURANCE AND INSURED

One should never calculate the cost of insurance without offsetting it with returns generated from packages than one would not have purchased otherwise.

Even had he gotten the sign right on the volatility, Mr Holman in the example above analyzes the performance of a strategy buying options to protect a tail event without adding the performance of the portfolio itself, like counting the cost side of the insurance without the performance of what one is insuring that would not have been bought otherwise. Over the same period he discusses the market rose more than 100%: a healthy approach would be to compare dollar-for-dollar what an investor would have done (and, of course, getting rid of this "VIX" business and focusing on very small dollars invested in tail options that would allow such an aggressive stance). Many investors (such as this author) would have stayed out of the market, or would not have added funds to the market, without such an insurance.

21.4 THE NECESSITY OF A NUMÉRAIRE IN FINANCE

There is a deeper analytical error.

A barbell is defined as a bimodal investment strategy, presented as investing a portion of your portfolio in what is explicitly defined as a "numéraire repository of value" (*Antifragile*), and the rest in risky securities (*Antifragile* indicates that such numéraire would be, among other things, inflation protected). Mr Holman goes on and on in a nihilistic discourse on the absence of such riskless numéraire (of the sort that can lead to such sophistry as "he is saying one is safer on *terra firma* than at sea, but what if there is an earthquake?").

The familiar Black and Scholes derivation uses a riskless asset as a baseline; but the literature since around 1977 has substituted the notion of "cash" with that of a numéraire , along with the notion that one can have different currencies, which technically allows for changes of probability measure. A numéraire is defined as *the unit to which all other units relate*. (Practically, the numéraire is a basket the variations of which do not affect the welfare of the investor.) Alas, without numéraire, there is no probability measure, and no quantitative in quantitative finance, as one needs a unit to which everything else is brought back to. In this (emotional) discourse, Mr Holton is not just rejecting the barbell per se, but any use of the expectation operator with any economic variable, meaning he should go attack the tens of thousand research papers and the existence of the journal *Quantitative Finance* itself.

Clearly, there is a high density of other mistakes or incoherent statements in the outpour of rage in Mr Holman's review; but I have no doubt these have been detected by the

Quantitative Finance reader and, as we said, the object of this discussion is the prevention of analytical mistakes in quantitative finance.

To conclude, this author welcomes criticism from the finance community that are not straw man arguments, or, as in the case of Mr Holmam, violate the foundations of the field itself.
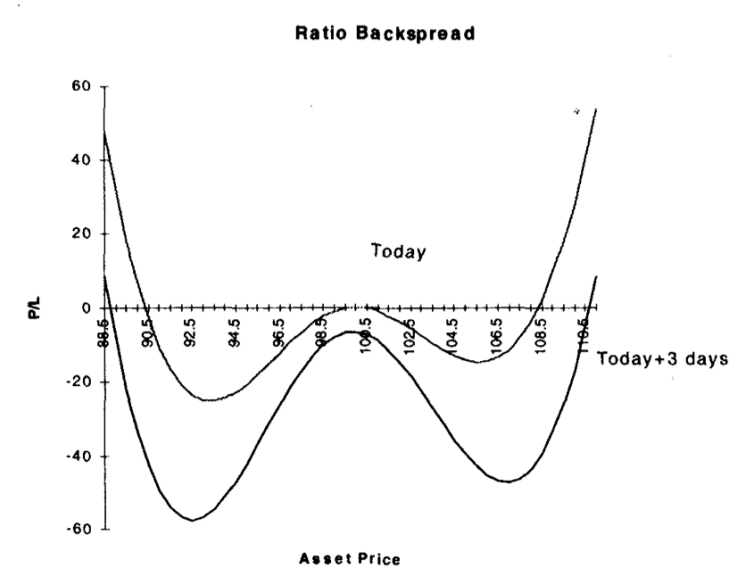

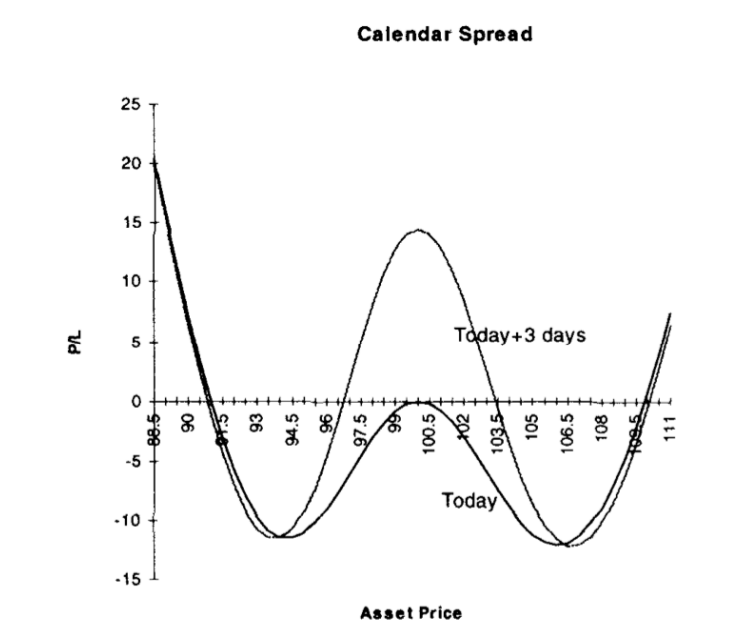Figure 21.1: *First Method to Extract the Fourth Moment, from Dynamic Hedging, 1997.*

Figure 21.2: *Second Method to Extract the Fourth Moment , from Dynamic Hedging, 1997.*

21.5 APPENDIX (BETTING ON TAILS OF DISTRIBUTION)

From *Dynamic Hedging*, pages 264-265:

> *A fourth moment bet is long or short the volatility of volatility. It could be achieved either with out-of-the-money options or with calendars. Example: A ratio "backspread" or reverse spread is a method that includes the buying of out-of-the-money options in large amounts and the selling of smaller amounts of at-the-money but making sure the trade satisfies the "credit" rule (i.e., the trade initially generates a positive cash flow). The credit rule is more difficult to interpret when one uses in-the-money options. In that case, one should deduct the present value of the intrinsic part of every option using the put-call parity rule to equate them with out-of-the-money.*
>
> *The trade shown in Figure 1 was accomplished with the purchase of both out-of-the-money puts and out-of-the-money calls and the selling of smaller amounts of at-the-money straddles of the same maturity.*
>
> *Figure 2 shows the second method, which entails the buying of 60- day options in some amount and selling 20-day options on 80% of the amount. Both trades show the position benefiting from the fat tails and the high peaks. Both trades, however, will have different vega sensitivities, but close to flat modified vega.*

See The Body, The Shoulders, and The Tails from section 3.3 where we assume tails start at the level of convexity of the segment of the probability distribution to the scale of the distribution.

22 | TAIL RISK CONSTRAINTS AND MAXIMUM ENTROPY (WITH D. GEMAN AND H. GEMAN)‡

ORTFOLIO SELECTION in the financial literature has essentially been analyzed under two central assumptions: full knowledge of the joint probability distribution of the returns of the securities that will comprise the target portfolio; and investors' preferences are expressed through a utility function. In the real world, operators build portfolios under risk constraints which are expressed both by their clients and regulators and which bear on the maximal loss that may be generated over a given time period at a given confidence level (the so-called *Value at Risk* of the position). Interestingly, in the finance literature, a serious discussion of how much or little is known from a probabilistic standpoint about the multi-dimensional density of the assets' returns seems to be of limited relevance.

Our approach in contrast is to highlight these issues and then adopt throughout a framework of entropy maximization to represent the real world ignorance of the "true" probability distributions, both univariate and multivariate, of traded securities' returns. In this setting, we identify the optimal portfolio under a number of downside risk constraints. Two interesting results are exhibited: (i) the left- tail constraints are sufficiently powerful to override all other considerations in the conventional theory; (ii) the "barbell portfolio" (maximal certainty/ low risk in one set of holdings, maximal uncertainty in another), which is quite familiar to traders, naturally emerges in our construction.

22.1 LEFT TAIL RISK AS THE CENTRAL PORTFOLIO CONSTRAINT

Customarily, when working in an institutional framework, operators and risk takers principally use regulatorily mandated tail-loss limits to set risk levels in their portfolios (obligatorily for banks since Basel II). They rely on stress tests, stop-losses, value at risk (VaR), expected shortfall (–i.e., the expected loss conditional on the loss exceeding VaR, also known as CVaR), and similar loss curtailment methods, rather than utility. In particular, the margining of financial transactions is calibrated by clearing firms and exchanges on tail losses, seen both probabilistically and through stress testing. (In the risk-taking terminology, a stop loss is a mandatory order that attempts terminates all or a portion of the exposure upon a trigger, a certain pre-defined nominal loss. Basel II is a generally used name for recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision. The Value-at-risk, VaR, is defined as a threshold loss value K such that the probability that the loss on the portfolio over the given time horizon exceeds

this value is ϵ. A stress test is an examination of the performance upon an arbitrarily set deviation in the underlying variables.) The information embedded in the choice of the constraint is, to say the least, a meaningful statistic about the appetite for risk and the shape of the desired distribution.

Operators are less concerned with portfolio variations than with the drawdown they may face over a time window. Further, they are in ignorance of the joint probability distribution of the components in their portfolio (except for a vague notion of association and hedges), but can control losses organically with allocation methods based on maximum risk. (The idea of substituting variance for risk can appear very strange to practitioners of risk-taking. The aim by Modern Portfolio Theory at lowering variance is inconsistent with the preferences of a rational investor, regardless of his risk aversion, since it also minimizes the variability in the profit domain –except in the very narrow situation of certainty about the future mean return, and in the far-fetched case where the investor can only invest in variables having a symmetric probability distribution, and/or only have a symmetric pay-off. Stop losses and tail risk controls violate such symmetry.) The conventional notions of utility and variance may be used, but not directly as information about them is embedded in the tail loss constaint.

Since the stop loss, the VaR (and expected shortfall) approaches and other risk-control methods concern only one segment of the distribution, the negative side of the loss domain, we can get a dual approach akin to a portfolio separation, or "barbell-style" construction, as the investor can have opposite stances on different parts of the return distribution. Our definition of barbell here is the mixing of two extreme properties in a portfolio such as a linear combination of maximal conservatism for a fraction w of the portfolio, with $w \in (0, 1)$, on one hand and maximal (or high) risk on the $(1 - w)$ remaining fraction.

Historically, finance theory has had a preference for parametric, less robust, methods. The idea that a decision-maker has clear and error-free knowledge about the distribution of future payoffs has survived in spite of its lack of practical and theoretical validity –for instance, correlations are too unstable to yield precise measurements. It is an approach that is based on distributional and parametric certainties, one that may be useful for research but does not accommodate responsible risk taking. (Correlations are unstable in an unstable way, as joint returns for assets are not elliptical, see Bouchaud and Chicheportiche (2012) [30].)

There are roughly two traditions: one based on highly parametric decision-making by the economics establishment (largely represented by Markowitz [112]) and the other based on somewhat sparse assumptions and known as the Kelly criterion (Kelly, 1956 [97], see Bell and Cover, 1980 [11].) (In contrast to the minimum-variance approach, Kelly's method, developed around the same period as Markowitz, requires no joint distribution or utility function. In practice one needs the ratio of expected profit to worst-case return dynamically adjusted to avoid ruin. Obviously, model error is of smaller consequence under the Kelly criterion: Thorp (1969) [176], Haigh (2000) [86], Mac Lean, Ziemba and Blazenko [107]. For a discussion of the differences between the two approaches, see Samuelson's objection to the Kelly criterion and logarithmic sizing in Thorp 2010 [177].) Kelly's method is also related to left- tail control due to proportional investment, which automatically reduces the portfolio in the event of losses; but the original method requires a hard, nonparametric worst-case scenario, that is, securities that have a lower bound in their variations, akin to a gamble in a casino, which is something that, in finance, can only be accomplished through binary options. The Kelly criterion, in addition, requires some precise knowledge of future returns such as the mean. Our approach goes beyond the latter method in accommodating more uncertainty about the returns, whereby an operator can only control his left-tail via derivatives and other forms of insurance or dynamic portfolio construction based on stop-

losses. (Xu, Wu, Jiang, and Song (2014) [189] contrast mean variance to maximum entropy and uses entropy to construct robust portfolios.)

In a nutshell, we hardwire the curtailments on loss but otherwise assume maximal uncertainty about the returns. More precisely, we equate the return distribution with the *m*aximum entropy extension of constraints expressed as statistical expectations on the left-tail behavior as well as on the expectation of the return or log-return in the non-danger zone. (Note that we use *S*hannon entropy throughout. There are other information measures, such as Tsallis entropy [181] , a generalization of Shannon entropy, and Renyi entropy, [94] , some of which may be more convenient computationally in special cases. However, Shannon entropy is the best known and has a well-developed maximization framework.)

Here, the "left-tail behavior" refers to the hard, explicit, institutional constraints discussed above. We describe the shape and investigate other properties of the resulting so-called *m*axent distribution. In addition to a mathematical result revealing the link between acceptable tail loss (VaR) and the expected return in the Gaussian mean-variance framework, our contribution is then twofold: 1) an investigation of the shape of the distribution of returns from portfolio construction under more natural constraints than those imposed in the mean-variance method, and 2) the use of stochastic entropy to represent residual uncertainty.

VaR and CVaR methods are not error free –parametric VaR is known to be ineffective as a risk control method on its own. However, these methods can be made robust using constructions that, upon paying an insurance price, no longer depend on parametric assumptions. This can be done using derivative contracts or by organic construction (clearly if someone has 80% of his portfolio in numéraire securities, the risk of losing more than 20% is zero independent from all possible models of returns, as the fluctuations in the numéraire are not considered risky). We use "pure robustness" or both VaR and zero shortfall via the "hard stop" or insurance, which is the special case in our paper of what we called earlier a "barbell" construction.

It is worth mentioning that it is an old idea in economics that an investor can build a portfolio based on two distinct risk categories, see Hicks (1939) [**?**]. Modern Portfolio Theory proposes the mutual fund theorem or "separation" theorem, namely that all investors can obtain their desired portfolio by mixing two mutual funds, one being the riskfree asset and one representing the optimal mean-variance portfolio that is tangent to their constraints; see Tobin (1958) [179], Markowitz (1959) [113], and the variations in Merton (1972) [115], Ross (1978) [140]. In our case a riskless asset is the part of the tail where risk is set to exactly zero. Note that the risky part of the portfolio needs to be minimum variance in traditional financial economics; for our method the exact opposite representation is taken for the risky one.

22.1.1 The Barbell as seen by E.T. Jaynes

Our approach to constrain only what can be constrained (in a robust manner) and to maximize entropy elsewhere echoes a remarkable insight by E.T. Jaynes in "How should we use entropy in economics?" [92]:

> "It may be that a macroeconomic system does not move in response to (or at least not solely in response to) the forces that are supposed to exist in current theories; it may simply move in the direction of increasing entropy as constrained by the conservation laws imposed by Nature and Government."

22.2 REVISITING THE MEAN VARIANCE SETTING

Let $\vec{X} = (X_1, ..., X_m)$ denote m asset returns over a given single period with joint density $g(\vec{x})$, mean returns $\vec{\mu} = (\mu_1, ..., \mu_m)$ and $m \times m$ covariance matrix Σ: $\Sigma_{ij} = \mathbb{E}(X_i X_j) - \mu_i \mu_j, 1 \leq i, j \leq m$. Assume that $\vec{\mu}$ and Σ can be reliably estimated from data.

The return on the portolio with weights $\vec{w} = (w_1, ..., w_m)$ is then

$$X = \sum_{i=1}^{m} w_i X_i,$$

which has mean and variance

$$\mathbb{E}(X) = \vec{w}\vec{\mu}^T, \quad V(X) = \vec{w}\Sigma\vec{w}^T.$$

In standard portfolio theory one minimizes $V(X)$ over all $\vec{w}$ subject to $\mathbb{E}(X) = \mu$ for a fixed desired average return μ. Equivalently, one maximizes the expected return $\mathbb{E}(X)$ subject to a fixed variance $V(X)$. In this framework variance is taken as a substitute for risk.

To draw connections with our entropy-centered approach, we consider two standard cases:

(1) **Normal World**: The joint distribution $g(\vec{x})$ of asset returns is multivariate Gaussian $N(\vec{\mu}, \Sigma)$. Assuming normality is equivalent to assuming $g(\vec{x})$ has maximum (Shannon) entropy among all multivariate distributions with the given first- and second-order statistics $\vec{\mu}$ and Σ. Moreover, for a fixed mean $\mathbb{E}(X)$, minimizing the variance $V(X)$ is equivalent to minimizing the entropy (uncertainty) of X. (This is true since joint normality implies that X is univariate normal for any choice of weights and the entropy of a $\mathcal{N}(\mu, \sigma^2)$ variable is $H = \frac{1}{2}(1 + \log(2\pi\sigma^2))$.) This is natural in a world with complete information. (The idea of entropy as mean uncertainty is in Philippatos and Wilson (1972) [128]; see Zhou –et al. (2013) [192] for a review of entropy in financial economics and Georgescu-Roegen (1971) [77] for economics in general.)

(2) **Unknown Multivariate Distribution**: Since we assume we can estimate the second-order structure, we can still carry out the Markowitz program, –i.e., choose the portfolio weights to find an optimal mean-variance performance, which determines $\mathbb{E}(X) = \mu$ and $V(X) = \sigma^2$. However, we do not know the distribution of the return X. Observe that *a*ssuming X is normally distributed $\mathcal{N}(\mu, \sigma^2)$ is equivalent to assuming the entropy of X is *m*aximized since, again, the normal maximizes entropy at a given mean and variance, see [128].

Our strategy is to generalize the second scenario by replacing the variance σ^2 by two left-tail value-at-risk constraints and to model the portfolio return as the maximum entropy extension of these constraints together with a constraint on the overall performance or on the growth of the portfolio in the non-danger zone.

22.2.1 Analyzing the Constraints

Let X have probability density $f(x)$. In everything that follows, let $K < 0$ be a normalizing constant chosen to be consistent with the risk-taker's wealth. For any $\epsilon > 0$ and $v_- < K$, the *v*alue-at-risk constraints are:

(1) **T**ail probability:

$$\mathbb{P}(X \leq K) = \int_{-\infty}^{K} f(x)\,dx = \epsilon.$$

(2) **E**xpected shortfall (CVaR):

$$\mathbb{E}(X|X \leq K) = \nu_-.$$

Assuming (1) holds, constraint (2) is equivalent to

$$\mathbb{E}(XI_{(X \leq K)}) = \int_{-\infty}^{K} x f(x)\,dx = \epsilon \nu_-.$$

Given the value-at-risk parameters $\theta = (K, \epsilon, \nu_-)$, let $\Omega_{var}(\theta)$ denote the set of probability densities f satisfying the two constraints. Notice that $\Omega_{var}(\theta)$ is convex: $f_1, f_2 \in \Omega_{var}(\theta)$ implies $\alpha f_1 + (1 - \alpha) f_2 \in \Omega_{var}(\theta)$. Later we will add another constraint involving the overall mean.

22.3 REVISITING THE GAUSSIAN CASE

Suppose we assume X is Gaussian with mean μ and variance σ^2. In principle it should be possible to satisfy the VaR constraints since we have two free parameters. Indeed, as shown below, the left-tail constraints determine the mean and variance; see Figure 22.1. However, satisfying the VaR constraints imposes interesting restrictions on μ and σ and leads to a natural inequality of a "no free lunch" style.
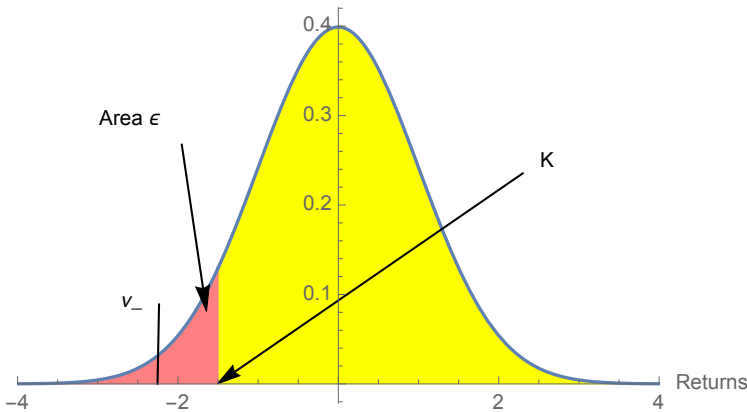
Figure 22.1: *By setting K (the value at risk), the probability ϵ of exceeding it, and the shortfall when doing so, there is no wiggle room left under a Gaussian distribution: σ and μ are determined, which makes construction according to portfolio theory less relevant.*

Let $\eta(\epsilon)$ be the ϵ-quantile of the standard normal distribution, –i.e., $\eta(\epsilon) = \Phi^{-1}(\epsilon)$, where Φ is the c.d.f. of the standard normal density $\phi(x)$. In addition, set

$$B(\epsilon) = \frac{1}{\epsilon \eta(\epsilon)} \phi(\eta(\epsilon)) = \frac{1}{\sqrt{2\pi} \epsilon \eta(\epsilon)} \exp\left\{-\frac{\eta(\epsilon)^2}{2}\right\}.$$

Proposition 22.1
If $X \sim N(\mu, \sigma^2)$ and satisfies the two VaR constraints, then the mean and variance are given by:

$$\mu = \frac{\nu_- + KB(\epsilon)}{1 + B(\epsilon)}, \quad \sigma = \frac{K - \nu_-}{\eta(\epsilon)(1 + B(\epsilon))}.$$

Moreover, $B(\epsilon) < -1$ and $\lim_{\epsilon \downarrow 0} B(\epsilon) = -1$.

The proof is in the Appendix. The VaR constraints lead directly to two linear equations in μ and σ:

$$\mu + \eta(\epsilon)\sigma = K, \quad \mu - \eta(\epsilon)B(\epsilon)\sigma = \nu_-.$$

Consider the conditions under which the VaR constraints allow a *positive* mean return $\mu = \mathbb{E}(X) > 0$. First, from the above linear equation in μ and σ in terms of $\eta(\epsilon)$ and K, we see that σ increases as ϵ increases for any fixed mean μ, and that $\mu > 0$ if and only if $\sigma > \frac{K}{\eta(\epsilon)}$, –i.e., we must accept a lower bound on the variance which increases with ϵ, which is a reasonable property. Second, from the expression for μ in Proposition 1, we have

$$\mu > 0 \iff |\nu_-| > KB(\epsilon).$$

Consequently, the only way to have a positive expected return is to accommodate a sufficiently large risk expressed by the various tradeoffs among the risk parameters θ satisfying the inequality above. (This type of restriction also applies more generally to symmetric distributions since the left tail constraints impose a structure on the location and scale. For instance, in the case of a Student T distribution with scale s, location m, and tail exponent α, the same linear relation between s and m applies: $s = (K - m)\kappa(\alpha)$, where $\kappa(\alpha) = -\dfrac{i\sqrt{I_{2\epsilon}^{-1}\left(\frac{\alpha}{2},\frac{1}{2}\right)}}{\sqrt{\alpha}\sqrt{I_{2\epsilon}^{-1}\left(\frac{1}{2},\frac{1}{2}\right)-1}}$, where I^{-1} is the inverse of the regularized incomplete beta function I, and s the solution of $\epsilon = \frac{1}{2}I_{\frac{\alpha s^2}{(k-m)^2+\alpha s^2}}\left(\frac{\alpha}{2},\frac{1}{2}\right)$.)

22.3.1 A Mixture of Two Normals

In many applied sciences, a mixture of two normals provides a useful and natural extension of the Gaussian itself; in finance, the Mixture Distribution Hypothesis (denoted as MDH in the literature) refers to a mixture of two normals and has been very widely investigated (see for instance Richardson and Smith (1995) [138]). H. Geman and T. Ané (1996) [2] exhibit how an infinite mixture of normal distributions for stock returns arises from the introduction of a "stochastic clock" accounting for the uneven arrival rate of information flow in the financial markets. In addition, option traders have long used mixtures to account for fat tails, and to examine the sensitivity of a portfolio to an increase in kurtosis ("DvegaDvol"); see Taleb (1997) [156]. Finally, Brigo and Mercurio (2002) [23] use a mixture of two normals to calibrate the skew in equity options.

Consider the mixture

$$f(x) = \lambda N(\mu_1, \sigma_1^2) + (1 - \lambda)N(\mu_2, \sigma_2^2).$$

An intuitively simple and appealing case is to fix the overall mean μ, and take $\lambda = \epsilon$ and $\mu_1 = \nu_-$, in which case μ_2 is constrained to be $\frac{\mu - \epsilon\nu_-}{1-\epsilon}$. It then follows that the left-tail constraints are approximately satisfied for σ_1, σ_2 sufficiently small. Indeed, when $\sigma_1 = \sigma_2 \approx 0$, the density is effectively composed of two spikes (small variance normals) with the left one centered at ν_- and the right one centered at at $\frac{\mu - \epsilon\nu_-}{1-\epsilon}$. The extreme case is a Dirac function on the left, as we see next.

Dynamic Stop Loss, A Brief Comment One can set a level K below which there is no mass, with results that depend on accuracy of the execution of such a stop. The distribution to the right of the stop-loss no longer looks like the standard Gaussian, as it builds positive

skewness in accordance to the distance of the stop from the mean. We limit any further discussion to the illustrations in Figure **??**.

22.4 MAXIMUM ENTROPY

From the comments and analysis above, it is clear that, in practice, the density f of the return X is unknown; in particular, no theory provides it. Assume we can adjust the portfolio parameters to satisfy the VaR constraints, and perhaps another constraint on the expected value of some function of X (e.g., the overall mean). We then wish to compute probabilities and expectations of interest, for example $\mathbb{P}(X > 0)$ or the probability of losing more than $2K$, or the expected return given $X > 0$. One strategy is to make such estimates and predictions under the most unpredictable circumstances consistent with the constraints. That is, use the *maximum entropy extension* (MEE) of the constraints as a model for $f(x)$.

The "differential entropy" of f is $h(f) = -\int f(x)\ln f(x)\,dx$. (In general, the integral may not exist.) Entropy is concave on the space of densities for which it is defined. In general, the MEE is defined as

$$f_{MEE} = \arg\max_{f\in\Omega} h(f)$$

where Ω is the space of densities which satisfy a set of constraints of the form $E\phi_j(X) = c_j, j = 1, ..., J$. Assuming Ω is non-empty, it is well-known that f_{MEE} is unique and (away from the boundary of feasibility) is an exponential distribution in the constraint functions, –i.e., is of the form

$$f_{MEE}(x) = C^{-1}\exp\left(\sum_j \lambda_j\phi_j(x)\right)$$

where $C = C(\lambda_1, ..., \lambda_M)$ is the normalizing constant. (This form comes from differentiating an appropriate functional $J(f)$ based on entropy, and forcing the integral to be unity and imposing the constraints with Lagrange mult1ipliers.) In the special cases below we use this characterization to find the *MEE* for our constraints.

In our case we want to maximize entropy subject to the VaR constraints together with any others we might impose. Indeed, the VaR constraints alone do not admit an MEE since they do not restrict the density $f(x)$ for $x > K$. The entropy can be made arbitrarily large by allowing f to be identically $C = \frac{1-\epsilon}{N-K}$ over $K < x < N$ and letting $N \to \infty$. Suppose, however, that we have adjoined one or more constraints on the behavior of f which are compatible with the VaR constraints in the sense that the set of densities Ω satisfying all the constraints is non-empty. Here Ω would depend on the VaR parameters $\theta = (K, \epsilon, v_-)$ together with those parameters associated with the additional constraints.

22.4.1 Case A: Constraining the Global Mean

The simplest case is to add a constraint on the mean return, –i.e., fix $\mathbb{E}(X) = \mu$. Since $\mathbb{E}(X) = \mathbb{P}(X \leq K)\mathbb{E}(X|X \leq K) + \mathbb{P}(X > K)\mathbb{E}(X|X > K)$, adding the mean constraint is equivalent to adding the constraint

$$\mathbb{E}(X|X > K) = v_+$$

where v_+ satisfies $\epsilon v_- + (1 - \epsilon)v_+ = \mu$.

Define

$$f_-(x) = \begin{cases} \frac{1}{(K-\nu_-)} \exp\left[-\frac{K-x}{K-\nu_-}\right] & \text{if } x < K, \\ 0 & \text{if } x \geq K. \end{cases}$$

and

$$f_+(x) = \begin{cases} \frac{1}{(\nu_+-K)} \exp\left[-\frac{x-K}{\nu_+-K}\right] & \text{if } x > K, \\ 0 & \text{if } x \leq K. \end{cases}$$

It is easy to check that both f_- and f_+ integrate to one. Then

$$f_{MEE}(x) = \epsilon f_-(x) + (1-\epsilon)f_+(x)$$

is the MEE of the three constraints. First, evidently

1. $\int_{-\infty}^{K} f_{MEE}(x)\,dx = \epsilon$;
2. $\int_{-\infty}^{K} x f_{MEE}(x)\,dx = \epsilon \nu_-$;
3. $\int_{K}^{\infty} x f_{MEE}(x)\,dx = (1-\epsilon)\nu_+$.

Hence the constraints are satisfied. Second, f_{MEE} has an exponential form in our constraint functions:

$$f_{MEE}(x) = C^{-1} \exp\left[-(\lambda_1 x + \lambda_2 I_{(x \leq K)} + \lambda_3 x I_{(x \leq K)})\right].$$

The shape of f_- depends on the relationship between K and the expected shortfall ν_-. The closer ν_- is to K, the more rapidly the tail falls off. As $\nu_- \to K$, f_- converges to a unit spike at $x = K$ (Figures 22.3 and 22.4).
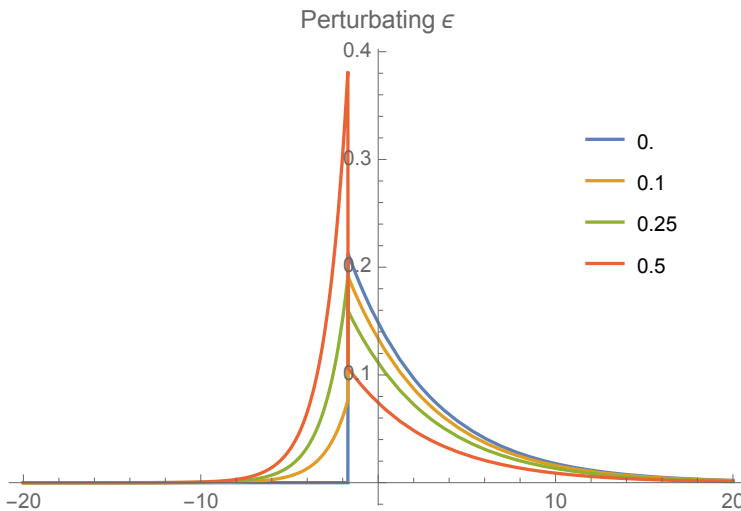
Figure 22.3: *Case A: Effect of different values of ϵ on the shape of the distribution.*
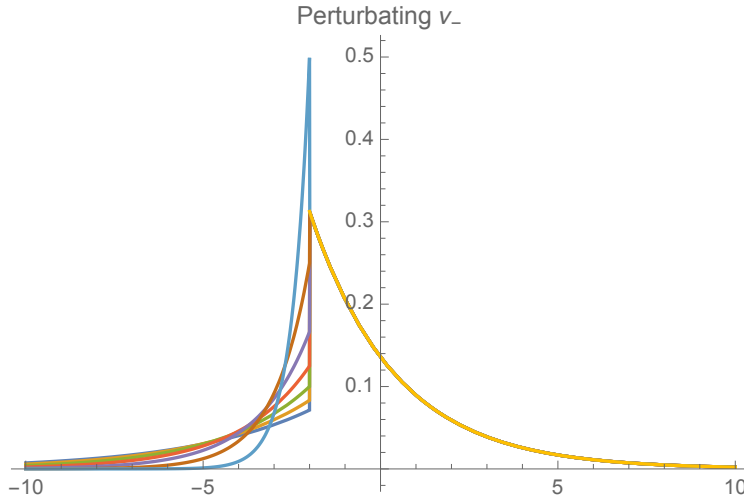
Figure 22.4: *Case A: Effect of different values of v_- on the shape of the distribution.*

22.4.2 Case B: Constraining the Absolute Mean

If instead we constrain the absolute mean, namely

$$E|X| = \int |x| f(x) \, \mathrm{d}x = \mu,$$

then the MEE is somewhat less apparent but can still be found. Define $f_-(x)$ as above, and let

$$f_+(x) = \begin{cases} \frac{\lambda_1}{2 - \exp(\lambda_1 K)} \exp(-\lambda_1 |x|) & \text{if } x \geq K, \\ 0 & \text{if } x < K. \end{cases}$$

Then λ_1 can be chosen such that

$$\epsilon v_- + (1 - \epsilon) \int_K^\infty |x| f_+(x) \, \mathrm{d}x = \mu.$$

22.4.3 Case C: Power Laws for the Right Tail

If we believe that actual returns have "fat tails," in particular that the right tail decays as a Power Law rather than exponentially (as with a normal or exponential density), than we can add this constraint to the VaR constraints instead of working with the mean or absolute mean. In view of the exponential form of the MEE, the density $f_+(x)$ will have a power law, namely

$$f_+(x) = \frac{1}{C(\alpha)} (1 + |x|)^{-(1+\alpha)}, x \geq K,$$

for $\alpha > 0$ if the constraint is of the form

$$E\left(\log(1 + |X|) | X > K\right) = A.$$

Moreover, again from the MEE theory, we know that the parameter is obtained by minimizing the logarithm of the normalizing function. In this case, it is easy to show that

$$C(\alpha) = \int_K^\infty (1 + |x|)^{-(1+\alpha)} \, dx = \frac{1}{\alpha}(2 - (1 - K)^{-\alpha}).$$

It follows that A and α satisfy the equation

$$A = \frac{1}{\alpha} - \frac{\log(1 - K)}{2(1 - K)^\alpha - 1}.$$

We can think of this equation as determining the decay rate α for a given A or, alternatively, as determining the constraint value A necessary to obtain a particular Power Law α.

The final MEE extension of the VaR constraints together with the constraint on the log of the return is then:

$$f_{MEE}(x) \quad = \quad \epsilon I_{(x \leq K)} \frac{1}{(K - v_-)} \exp\left[-\frac{K - x}{K - v_-}\right] + (1 - \epsilon) I_{(x>K)} \frac{(1 + |x|)^{-(1+\alpha)}}{C(\alpha)},$$

(see Figures 22.5 and 22.6).
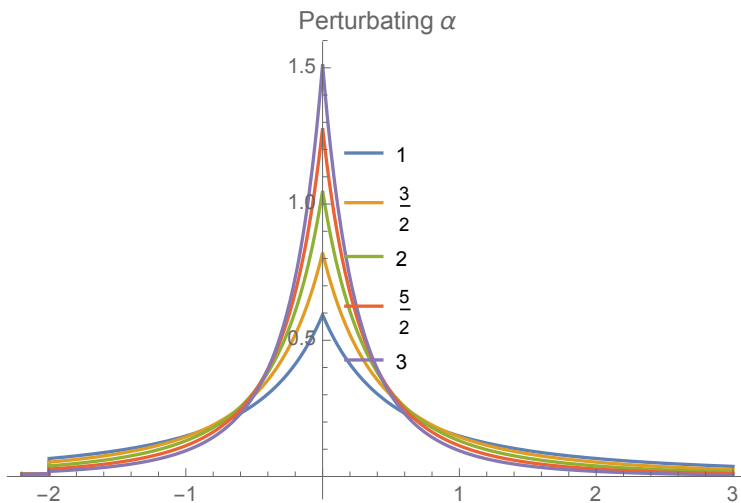
Figure 22.5: *Case C: Effect of different values of on the shape of the fat-tailed maximum entropy distribution.*
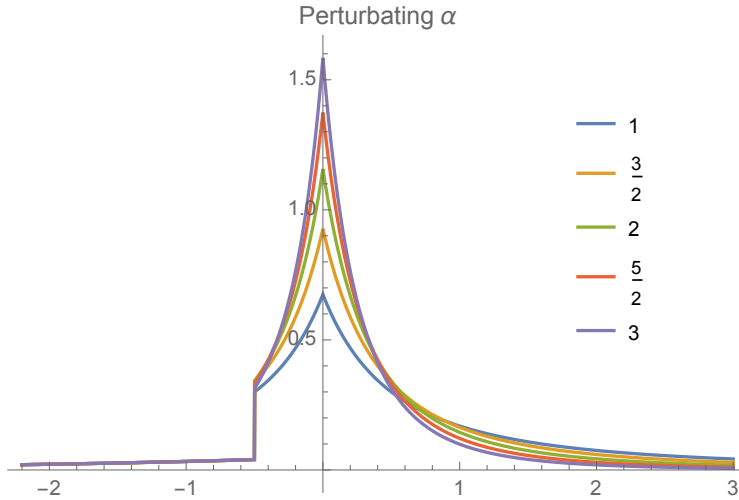
Figure 22.6: *Case C: Effect of different values of on the shape of the fat-tailed maximum entropy distribution (closer K).*

22.4.4   Extension to a Multi-Period Setting: A Comment

Consider the behavior in multi-periods. Using a naive approach, we sum up the performance as if there was no response to previous returns. We can see how Case A approaches the regular Gaussian, but not Case C (Figure 22.7).

For case A the characteristic functioncan be written:

$$\Psi^A(t) = \frac{e^{iKt}(t(K - \nu_-\epsilon + \nu_+(\epsilon - 1)) - i)}{(Kt - \nu_-t - i)(-1 - it(K - \nu_+))}$$

So we can derive from convolutions that the function $\Psi^A(t)^n$ converges to that of an n-summed Gaussian. Further, the characteristic function of the limit of the average of strategies, namely

$$\lim_{n \to \infty} \Psi^A(t/n)^n = e^{it(\nu_+ + \epsilon(\nu_- - \nu_+))}, \tag{22.1}$$

is the characteristic function of the Dirac delta, visibly the effect of the law of large numbers delivering the same result as the Gaussian with mean $\nu_+ + \epsilon(\nu_- - \nu_+)$.

As to the Power Law in Case C, convergence to Gaussian only takes place for $\alpha \geq 2$, and rather slowly.
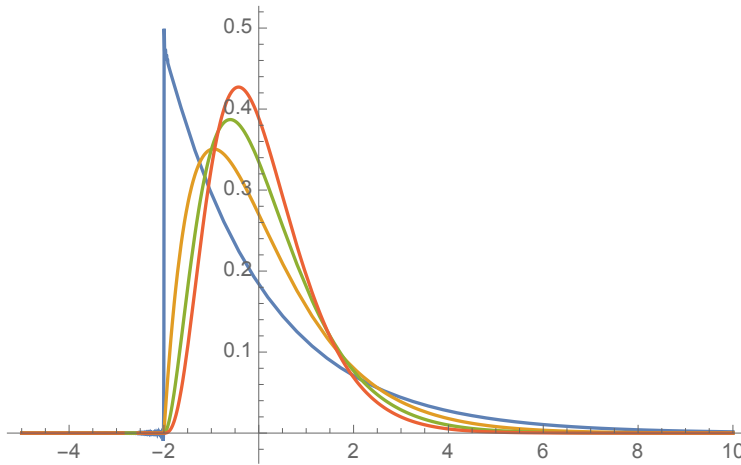
Figure 22.7: *Average return for multiperiod naive strategy for Case A, that is, assuming independence of "sizing", as position size does not depend on past performance. They aggregate nicely to a standard Gaussian, and (as shown in Equation (22.1)), shrink to a Dirac at the mean value.*

22.5 COMMENTS AND CONCLUSION

We note that the stop loss plays a larger role in determining the stochastic properties than the portfolio composition. Simply, the stop is not triggered by individual components, but by variations in the total portfolio. This frees the analysis from focusing on individual portfolio components when the tail –via derivatives or organic construction– is all we know and can control.

To conclude, most papers dealing with entropy in the mathematical finance literature have used minimization of entropy as an optimization criterion. For instance, Fritelli (2000) [71] exhibits the unicity of a "minimal entropy martingale measure" under some conditions and shows that minimization of entropy is equivalent to maximizing the expected exponential utility of terminal wealth. We have, instead, and outside any utility criterion, proposed entropy maximization as the recognition of the uncertainty of asset distributions. Under VaR and Expected Shortfall constraints, we obtain in full generality a "barbell portfolio" as the optimal solution, extending to a very general setting the approach of the two-fund separation theorem.

Appendix A

Proof of Proposition 1: Since $X \sim N(\mu, \sigma^2)$, the tail probability constraint is

$$\epsilon = \mathbb{P}(X < K) = \mathbb{P}(Z < \frac{K - \mu}{\sigma}) = \Phi(\frac{K - \mu}{\sigma}).$$

By definition, $\Phi(\eta(\epsilon)) = \epsilon$. Hence,

$$K = \mu + \eta(\epsilon)\sigma \tag{22.2}$$

For the shortfall constraint,

$$
\begin{aligned}
\mathbb{E}(X; X < k) &= \int_{-\infty}^{K} \frac{x}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} \, dx \\
&= \mu\epsilon + \sigma \int_{-\infty}^{(K-\mu)/\sigma)} x\phi(x) \, dx \\
&= \mu\epsilon - \frac{\sigma}{\sqrt{2\pi}} \exp -\frac{(K-\mu)^2}{2\sigma^2}
\end{aligned}
$$

Since, $\mathbb{E}(X; X < K) = \epsilon v_-$, and from the definition of $B(\epsilon)$, we obtain

$$
v_- = \mu - \eta(\epsilon)B(\epsilon)\sigma \tag{22.3}
$$

Solving (22.2) and (22.3) for μ and σ^2 gives the expressions in Proposition 1.

Finally, by symmetry to the "upper tail inequality" of the standard normal, we have, for $x < 0$, $\Phi(x) \leq \frac{\phi(x)}{-x}$. Choosing $x = \eta(\epsilon) = \Phi^{-1}(\epsilon)$ yields $\epsilon = P(X < \eta(\epsilon)) \leq -\epsilon B(\epsilon)$ or $1 + B(\epsilon) \leq 0$. Since the upper tail inequality is asymptotically exact as $x \to \infty$ we have $B(0) = -1$, which concludes the proof.

Part IX

BIBLIOGRAPHY AND INDEX

BIBLIOGRAPHY

[1] Inmaculada B Aban, Mark M Meerschaert, and Anna K Panorska. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277, 2006.

[2] Thierry Ané and Hélyette Geman. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance*, 55(5):2259–2284, 2000.

[3] Marco Avellaneda, Craig Friedman, Richard Holmes, and Dominick Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Mathematical Finance*, 4(1):37–64, 1997.

[4] Louis Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 1900.

[5] Kevin P Balanda and HL MacGillivray. Kurtosis: a critical review. *The American Statistician*, 42(2):111–119, 1988.

[6] August A Balkema and Laurens De Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974.

[7] August A Balkema and Laurens De Haan. Limit distributions for order statistics. i. *Theory of Probability & Its Applications*, 23(1):77–92, 1978.

[8] August A Balkema and Laurens de Haan. Limit distributions for order statistics. ii. *Theory of Probability & Its Applications*, 23(2):341–358, 1979.

[9] Shaul K Bar-Lev, Idit Lavi, and Benjamin Reiser. Bayesian inference for the power law process. *Annals of the Institute of Statistical Mathematics*, 44(4):623–639, 1992.

[10] Norman C Beaulieu, Adnan A Abu-Dayya, and Peter J McLane. Estimating the distribution of a sum of independent lognormal random variables. *Communications, IEEE Transactions on*, 43(12):2869, 1995.

[11] Robert M Bell and Thomas M Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, 5(2):161–166, 1980.

[12] Shlomo Benartzi and Richard Thaler. Heuristics and biases in retirement savings behavior. *Journal of Economic perspectives*, 21(3):81–104, 2007.

[13] Shlomo Benartzi and Richard H Thaler. Naive diversification strategies in defined contribution saving plans. *American economic review*, 91(1):79–98, 2001.

[14] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27. Cambridge university press, 1989.

[15] Giulio Biroli, J-P Bouchaud, and Marc Potters. On the top eigenvalue of heavy-tailed random matrices. *EPL (Europhysics Letters)*, 78(1):10001, 2007.

[16] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. 81:637–654, May–June 1973.

[17] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[18] A.J. Boness. Elements of a theory of stock-option value. 72:163–175, 1964.

[19] Jean-Philippe Bouchaud, Marc Mézard, Marc Potters, et al. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256, 2002.

[20] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[21] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[22] D. T. Breeden and R. H. Litzenberger. Price of state-contigent claimes implicit in option prices. 51:621–651, 1978.

[23] Damiano Brigo and Fabio Mercurio. Lognormal-mixture dynamics and calibration to market volatility smiles. *International Journal of Theoretical and Applied Finance*, 5(04):427–446, 2002.

[24] Peter Carr. Bounded brownian motion. *NYU Tandon School of Engineering*, 2017.

[25] Peter Carr, Hélyette Geman, Dilip B Madan, and Marc Yor. Stochastic volatility for lévy processes. *Mathematical finance*, 13(3):345–382, 2003.

[26] Peter Carr and Dilip Madan. Optimal positioning in derivative securities. 2001.

[27] Lars-Erik Cederman. Modeling the size of wars: from billiard balls to sandpiles. *American Political Science Review*, 97(01):135–150, 2003.

[28] Bikas K Chakrabarti, Anirban Chakraborti, Satya R Chakravarty, and Arnab Chatterjee. *Econophysics of income and wealth distributions*. Cambridge University Press, 2013.

[29] Shaohua Chen, Hong Nie, and Benjamin Ayers-Glassey. Lognormal sum approximation with a variant of type iv pearson distribution. *IEEE Communications Letters*, 12(9), 2008.

[30] Rémy Chicheportiche and Jean-Philippe Bouchaud. The joint distribution of stock returns is not elliptical. *International Journal of Theoretical and Applied Finance*, 15(03), 2012.

[31] VP Chistyakov. A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probability & Its Applications*, 9(4):640–648, 1964.

[32] Pasquale Cirillo. Are your data really pareto distributed? *Physica A: Statistical Mechanics and its Applications*, 392(23):5947–5962, 2013.

[33] Pasquale Cirillo and Nassim Nicholas Taleb. Expected shortfall estimation for apparently infinite-mean models of operational risk. *Quantitative Finance*, pages 1–10, 2016.

[34] Pasquale Cirillo and Nassim Nicholas Taleb. On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, 452:29–45, 2016.

[35] Pasquale Cirillo and Nassim Nicholas Taleb. What are the chances of war? *Significance*, 13(2):44–45, 2016.

[36] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[37] Rama Cont and Peter Tankov. *Financial modelling with jump processes*, volume 2. CRC press, 2003.

[38] Harald Cramér. *On the mathematical theory of risk*. Centraltryckeriet, 1930.

[39] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[40] Camilo Dagum. Inequality measures between income distributions with applications. *Econometrica*, 48(7):1791–1803, 1980.

[41] Camilo Dagum. *Income distribution models*. Wiley Online Library, 1983.

[42] Anirban DasGupta. *Probability for statistics and machine learning: fundamentals and advanced topics*. Springer Science & Business Media, 2011.

[43] Herbert A David and Haikady N Nagaraja. Order statistics. 2003.

[44] Bruno De Finetti. *Philosophical Lectures on Probability: collected, edited, and annotated by Alberto Mura*, volume 340. Springer Science & Business Media, 2008.

[45] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

[46] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2007.

[47] Marco Di Renzo, Fabio Graziosi, and Fortunato Santucci. Further results on the approximation of log-normal power sum via pearson type iv distribution: a general formula for log-moments computation. *IEEE Transactions on Communications*, 57(4), 2009.

[48] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[49] Persi Diaconis and Sandy Zabell. Closed form summation for classical distributions: variations on a theme of de moivre. *Statistical Science*, pages 284–302, 1991.

[50] Cornelius Frank Dietrich. *Uncertainty, calibration and probability: the statistics of scientific and industrial measurement*. Routledge, 2017.

[51] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.19 of 2018-06-22. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[52] Daniel Dufresne. Sums of lognormals. In *Proceedings of the 43rd actuarial research conference*. University of Regina, 2008.

[53] Daniel Dufresne et al. The log-normal approximation in financial and other computations. *Advances in Applied Probability*, 36(3):747–773, 2004.

[54] Bruno Dupire. Pricing with a smile. 7(1), 1994.

[55] Iddo Eliazar. Inequality spectra. *Physica A: Statistical Mechanics and its Applications*, 469:824–847, 2017.

[56] Iddo Eliazar and Morrel H Cohen. On social inequality: Analyzing the rich–poor disparity. *Physica A: Statistical Mechanics and its Applications*, 401:148–158, 2014.

[57] Iddo Eliazar and Igor M Sokolov. Maximization of statistical heterogeneity: From shannon's entropy to gini's index. *Physica A: Statistical Mechanics and its Applications*, 389(16):3023–3038, 2010.

[58] Iddo I Eliazar and Igor M Sokolov. Gini characterization of extreme-value statistics. *Physica A: Statistical Mechanics and its Applications*, 389(21):4462–4472, 2010.

[59] Iddo I Eliazar and Igor M Sokolov. Measuring statistical evenness: A panoramic overview. *Physica A: Statistical Mechanics and its Applications*, 391(4):1323–1353, 2012.

[60] Paul Embrechts. *Modelling extremal events: for insurance and finance*, volume 33. Springer, 1997.

[61] Paul Embrechts and Charles M Goldie. On convolution tails. *Stochastic Processes and their Applications*, 13(3):263–278, 1982.

[62] Paul Embrechts, Charles M Goldie, and Noël Veraverbeke. Subexponentiality and infinite divisibility. *Probability Theory and Related Fields*, 49(3):335–347, 1979.

[63] M Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):247–271, 1909.

[64] Michael Falk et al. On testing the extreme value index via the pot-method. *The Annals of Statistics*, 23(6):2013–2035, 1995.

[65] Michael Falk, Jürg Hüsler, and Rolf-Dieter Reiss. *Laws of small numbers: extremes and rare events*. Springer Science & Business Media, 2010.

[66] Kai-Tai Fang. Elliptically contoured distributions. *Encyclopedia of Statistical Sciences*, 2006.

[67] Doyne James Farmer and John Geanakoplos. Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates. 2009.

[68] William Feller. 1971an introduction to probability theory and its applications, vol. 2.

[69] Andrea Fontanari, Pasquale Cirillo, and Cornelis W Oosterlee. From concentration profiles to concentration maps. new tools for the study of loss distributions. *Insurance: Mathematics and Economics*, 78:13–29, 2018.

[70] Shane Frederick, George Loewenstein, and Ted O'donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

[71] Marco Frittelli. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Mathematical finance*, 10(1):39–52, 2000.

[72] Xavier Gabaix. Power laws in economics and finance. Technical report, National Bureau of Economic Research, 2008.

[73] Jim Gatheral. *The Volatility Surface: a Practitioner's Guide*. John Wiley & Sons, 2006.

[74] Oscar Gelderblom and Joost Jonker. Amsterdam as the cradle of modern futures and options trading, 1550-1650. *William Goetzmann and K. Geert Rouwenhorst*, 2005.

[75] Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

[76] Donald Geman, Hélyette Geman, and Nassim Nicholas Taleb. Tail risk constraints and maximum entropy. *Entropy*, 17(6):3724, 2015.

[77] Nicholas Georgescu-Roegen. The entropy law and the economic process, 1971. *Cambridge, Mass*, 1971.

[78] Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, New York, 1999.

[79] Corrado Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[80] BV Gnedenko and AN Kolmogorov. *Limit Distributions for Sums of Independent Random Variables (1954)*.

[81] Charles M Goldie. Subexponential distributions and dominated-variation tails. *Journal of Applied Probability*, pages 440–442, 1978.

[82] Daniel Goldstein and Nassim Taleb. We don't quite know what we are talking about when we talk about volatility. *Journal of Portfolio Management*, 33(4), 2007.

[83] Richard C Green, Robert A Jarrow, et al. Spanning and completeness in markets with contingent claims. *Journal of Economic Theory*, 41(1):202–210, 1987.

[84] Emil Julius Gümbel. Statistics of extremes. 1958.

[85] Laurens Haan and Ana Ferreira. Extreme value theory: An introduction. *Springer Series in Operations Research and Financial Engineering (*, 2006.

[86] John Haigh. The kelly criterion and bet comparisons in spread betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(4):531–539, 2000.

[87] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

[88] J Michael Harrison and David M Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic theory*, 20(3):381–408, 1979.

[89] Espen Gaarder Haug and Nassim Nicholas Taleb. Option traders use (very) sophisticated heuristics, never the black–scholes–merton formula. *Journal of Economic Behavior & Organization*, 77(2):97–106, 2011.

[90] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[91] HM James Hung, Robert T O'Neill, Peter Bauer, and Karl Kohne. The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, pages 11–22, 1997.

[92] E.T. Jaynes. How should we use entropy in economics? 1991.

[93] Hedegaard Anders Jessen and Thomas Mikosch. Regularly varying functions. *Publications de l'Institut Mathematique*, 80(94):171–192, 2006.

[94] Petr Jizba, Hagen Kleinert, and Mohammad Shefaat. Rényi's information transfer between financial time series. *Physica A: Statistical Mechanics and its Applications*, 391(10):2971–2989, 2012.

[95] Valen E Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

[96] Joseph P Kairys Jr and NICHOLAS VALERIO III. The market for equity options in the 1870s. *The Journal of Finance*, 52(4):1707–1723, 1997.

[97] John L Kelly. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.

[98] Gideon Keren. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273, 1991.

[99] Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*, volume 470. John Wiley & Sons, 2003.

[100] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[101] Samuel Kotz and Norman Johnson. *Encyclopedia of Statistical Sciences*. Wiley, 2004.

[102] Jean Laherrere and Didier Sornette. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[103] David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

[104] Deli Li, M Bhaskara Rao, and RJ Tomkins. The law of the iterated logarithm and central limit theorem for l-statistics. Technical report, PENNSYLVANIA STATE UNIV UNIVERSITY PARK CENTER FOR MULTIVARIATE ANALYSIS, 1997.

[105] Filip Lundberg. *I. Approximerad framställning af sannolikhetsfunktionen. II. Återförsäkring af kollektivrisker. Akademisk afhandling... af Filip Lundberg,...* Almqvist och Wiksells boktryckeri, 1903.

[106] HL MacGillivray and Kevin P Balanda. Mixtures, myths and kurtosis. *Communications in Statistics-Simulation and Computation*, 17(3):789–802, 1988.

[107] LC MacLean, William T Ziemba, and George Blazenko. Growth versus security in dynamic investment analysis. *Management Science*, 38(11):1562–1585, 1992.

[108] Spyros Makridakis and Nassim Taleb. Decision making and planning under low levels of predictability, 2009.

[109] Benoit Mandelbrot. The pareto-levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.

[110] Benoit Mandelbrot. The stable paretian income distribution when the apparent exponent is near two. *International Economic Review*, 4(1):111–115, 1963.

[111] Benoît B Mandelbrot and Nassim Nicholas Taleb. Random jump, not random walk, 2010.

[112] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

[113] Harry M Markowitz. *Portfolio selection: efficient diversification of investments*, volume 16. Wiley, 1959.

[114] RARD Maronna, Douglas Martin, and Victor Yohai. *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.

[115] Robert C Merton. An analytic derivation of the efficient portfolio frontier. *Journal of financial and quantitative analysis*, 7(4):1851–1872, 1972.

[116] Robert C. Merton. The relationship between put and call prices: Comment. 28(1):183–184, 1973.

[117] Robert C. Merton. Theory of rational option pricing. 4:141–183, Spring 1973.

[118] Robert C Merton and Paul Anthony Samuelson. Continuous-time finance. 1992.

[119] David C Nachman. Spanning and completeness with options. *The review of financial studies*, 1(3):311–328, 1988.

[120] Hansjörg Neth and Gerd Gigerenzer. Heuristics: Tools for an uncertain world. *Emerging trends in the social and behavioral sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 2015.

[121] Donald J Newman. *A problem seminar*. Springer Science & Business Media, 2012.

[122] Hong Nie and Shaohua Chen. Lognormal sum approximation with type iv pearson distribution. *IEEE Communications Letters*, 11(10), 2007.

[123] John P Nolan. Parameterizations and modes of stable distributions. *Statistics & probability letters*, 38(2):187–195, 1998.

[124] T. Mikosch P. Embrechts, C. Kluppelberg. *Modelling Extremal Events*. Springer, 2003.

[125] Vilfredo Pareto. La courbe des revenus. *Travaux de Sciences Sociales*, pages 299–345, 1896 (1964).

[126] O. Peters and M. Gell-Mann. Evaluating gambles using dynamics. *Chaos*, 26(2), 2016.

[127] T Pham-Gia and TL Hung. The mean and median absolute deviations. *Mathematical and Computer Modelling*, 34(7-8):921–936, 2001.

[128] George C Philippatos and Charles J Wilson. Entropy, market risk, and the selection of efficient portfolios. *Applied Economics*, 4(3):209–220, 1972.

[129] Charles Phillips and Alan Axelrod. *Encyclopedia of Wars:(3-Volume Set).* Infobase Pub., 2004.

[130] James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.

[131] Thomas Piketty. Capital in the 21st century, 2014.

[132] Thomas Piketty and Emmanuel Saez. The evolution of top incomes: a historical and international perspective. Technical report, National Bureau of Economic Research, 2006.

[133] Iosif Pinelis. Characteristic function of the positive part of a random variable and related results, with applications. *Statistics & Probability Letters*, 106:281–286, 2015.

[134] Steven Pinker. *The better angels of our nature: Why violence has declined*. Penguin, 2011.

[135] Dan Pirjol. The logistic-normal integral and its generalizations. *Journal of Computational and Applied Mathematics*, 237(1):460–469, 2013.

[136] EJG Pitman. Subexponential distribution functions. *J. Austral. Math. Soc. Ser. A*, 29(3):337–347, 1980.

[137] Svetlozar T Rachev, Young Shin Kim, Michele L Bianchi, and Frank J Fabozzi. *Financial models with Lévy processes and volatility clustering*, volume 187. John Wiley & Sons, 2011.

[138] Matthew Richardson and Tom Smith. A direct test of the mixture of distributions hypothesis: Measuring the daily flow of information. *Journal of Financial and Quantitative Analysis*, 29(01):101–116, 1994.

[139] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[140] Stephen A Ross. Mutual fund separation in financial theory—the separating distributions. *Journal of Economic Theory*, 17(2):254–286, 1978.

[141] Stephen A Ross. *Neoclassical finance*. Princeton University Press, 2009.

[142] Mark Rubinstein. *Rubinstein on derivatives*. Risk Books, 1999.

[143] Doriana Ruffino and Jonathan Treussard. Derman and taleb's 'the illusions of dynamic replication': a comment. *Quantitative Finance*, 6(5):365–367, 2006.

[144] Harold Sackrowitz and Ester Samuel-Cahn. P values as random variables—expected p values. *The American Statistician*, 53(4):326–331, 1999.

[145] Gennady Samorodnitsky and Murad S Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[146] D Schleher. Generalized gram-charlier series with application to the sum of lognormal variates (corresp.). *IEEE Transactions on Information Theory*, 23(2):275–280, 1977.

[147] Jun Shao. *Mathematical Statistics*. Springer, 2003.

[148] SK Singh and GS Maddala. A function for size distribution of incomes: reply. *Econometrica*, 46(2), 1978.

[149] Didier Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization, and disorder: concepts and tools*. Springer, 2004.

[150] C.M. Sprenkle. *Warrant Prices as Indicators of Expectations and Preferences: in P. Cootner, ed., 1964, The Random Character of Stock Market Prices,*. MIT Press, Cambridge, Mass, 1964.

[151] AJ Stam. Regular variation of the tail of a subordinated probability distribution. *Advances in Applied Probability*, pages 308–327, 1973.

[152] Giitiro Suzuki. A consistent estimator for the mean deviation of the pearson type distribution. *Annals of the Institute of Statistical Mathematics*, 17(1):271–285, 1965.

[153] N N Taleb and R Douady. Mathematical definition, mapping, and detection of (anti) fragility. *Quantitative Finance*, 2013.

[154] Nassim N Taleb and Daniel G Goldstein. The problem is beyond psychology: The real world is more random than regression analyses. *International Journal of Forecasting*, 28(3):715–716, 2012.

[155] Nassim N Taleb and G Martin. The illusion of thin tails under aggregation (a reply to jack treynor). *Journal of Investment Management*, 2012.

[156] Nassim Nicholas Taleb. *Dynamic Hedging: Managing Vanilla and Exotic Options*. John Wiley & Sons (Wiley Series in Financial Engineering), 1997.

[157] Nassim Nicholas Taleb. *Incerto: Antifragile, The Black Swan , Fooled by Randomness, the Bed of Procrustes, Skin in the Game*. Random House and Penguin, 2001-2018.

[158] Nassim Nicholas Taleb. Black swans and the domains of statistics. *The American Statistician*, 61(3):198–200, 2007.

[159] Nassim Nicholas Taleb. Errors, robustness, and the fourth quadrant. *International Journal of Forecasting*, 25(4):744–759, 2009.

[160] Nassim Nicholas Taleb. Finiteness of variance is irrelevant in the practice of quantitative finance. *Complexity*, 14(3):66–76, 2009.

[161] Nassim Nicholas Taleb. *Antifragile: things that gain from disorder*. Random House and Penguin, 2012.

[162] Nassim Nicholas Taleb. Four points beginner risk managers should learn from jeff holman's mistakes in the discussion of antifragile. *arXiv preprint arXiv:1401.2524*, 2014.

[163] Nassim Nicholas Taleb. The meta-distribution of standard p-values. *arXiv preprint arXiv:1603.07532*, 2016.

[164] Nassim Nicholas Taleb. Stochastic tail exponent for asymmetric power laws. *arXiv preprint arXiv:1609.02369*, 2016.

[165] Nassim Nicholas Taleb. Election predictions as martingales: an arbitrage approach. *Quantitative Finance*, 18(1):1–5, 2018.

[166] Nassim Nicholas Taleb. How much data do you need? an operational, pre-asymptotic metric for fat-tailedness. *International Journal of Forecasting*, 2018.

[167] Nassim Nicholas Taleb. *Skin in the Game: Hidden Asymmetries in Daily Life*. Penguin (London) and Random House (N.Y.), 2018.

[168] Nassim Nicholas Taleb. *Technical Incerto Vol 1: The Statistical Consequences of Fat Tails, Papers and Commentaries*. Monograph, 2018.

[169] Nassim Nicholas Taleb, Elie Canetti, Tidiane Kinda, Elena Loukoianova, and Christian Schmieder. A new heuristic measure of fragility and tail risks: application to stress testing. *International Monetary Fund*, 2018.

[170] Nassim Nicholas Taleb and Raphael Douady. On the super-additivity and estimation biases of quantile contributions. *Physica A: Statistical Mechanics and its Applications*, 429:252–260, 2015.

[171] Nassim Nicholas Taleb and George A Martin. How to prevent other financial crises. *SAIS Review of International Affairs*, 32(1):49–60, 2012.

[172] Nassim Nicholas Taleb and Avital Pilpel. I problemi epistemologici del risk management. *Daniele Pace (a cura di)" Economia del rischio. Antologia di scritti su rischio e decisione economica", Giuffrè, Milano*, 2004.

[173] Nassim Nicholas Taleb and Constantine Sandis. The skin in the game heuristic for protection against tail events. *Review of Behavioral Economics*, 1:1–21, 2014.

[174] Jozef L Teugels. The class of subexponential distributions. *The Annals of Probability*, 3(6):1000–1011, 1975.

[175] Edward Thorp. A corrected derivation of the black-scholes option model. *Based on private conversation with Edward Thorp and a copy of a 7 page paper Thorp wrote around 1973, with disclaimer that I understood Ed. Thorp correctly.*, 1973.

[176] Edward O Thorp. Optimal gambling systems for favorable games. *Revue de l'Institut International de Statistique*, pages 273–293, 1969.

[177] Edward O Thorp. Understanding the kelly criterion. *The Kelly Capital Growth Investment Criterion: Theory and Practice', World Scientific Press, Singapore*, 2010.

[178] Edward O. Thorp and S. T. Kassouf. *Beat the Market*. New York: Random House, 1967.

[179] James Tobin. Liquidity preference as behavior towards risk. *The review of economic studies*, pages 65–86, 1958.

[180] Jack L Treynor. Insights-what can taleb learn from markowitz? *Journal of Investment Management*, 9(4):5, 2011.

[181] Constantino Tsallis, Celia Anteneodo, Lisa Borland, and Roberto Osorio. Nonextensive statistical mechanics and economics. *Physica A: Statistical Mechanics and its Applications*, 324(1):89–100, 2003.

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

[183] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[184] Willem Rutger van Zwet. *Convex transformations of random variables*, volume 7. Mathematisch centrum, 1964.

[185] SR Srinivasa Varadhan. *Large deviations and applications*, volume 46. SIAM, 1984.

[186] José A Villaseñor-Alva and Elizabeth González-Estrada. A bootstrap goodness of fit test for the generalized pareto distribution. *Computational Statistics & Data Analysis*, 53(11):3835–3841, 2009.

[187] Eric Weisstein. *Wolfram MathWorld*. Wolfram Research www.wolfram.com, 2017.

[188] Heath Windcliff and Phelim P Boyle. The 1/n pension investment puzzle. *North American Actuarial Journal*, 8(3):32–45, 2004.

[189] Yingying Xu, Zhuwu Wu, Long Jiang, and Xuefeng Song. A maximum entropy method for a robust portfolio problem. *Entropy*, 16(6):3401–3415, 2014.

[190] Yingying Yang, Shuhe Hu, and Tao Wu. The tail probability of the product of dependent random variables from max-domains of attraction. *Statistics & Probability Letters*, 81(12):1876–1882, 2011.

[191] IV Zaliapin, Yan Y Kagan, and Federic P Schoenberg. Approximating the distribution of pareto sums. *Pure and Applied geophysics*, 162(6-7):1187–1228, 2005.

[192] Rongxi Zhou, Ru Cai, and Guanqun Tong. Applications of entropy in finance: A review. *Entropy*, 15(11):4909–4931, 2013.

[193] VM Zolotarev. On a new viewpoint of limit theorems taking into account large deviationsr. *Selected Translations in Mathematical Statistics and Probability*, 9:153, 1971.

INDEX