PROBABILITY AND RISK IN THE REAL WORLD

MATHEMATICAL PARALLEL VERSION OF

THE INCERTO

I) ANTIFRAGILE, II) THE BLACK SWAN, III) THE BED OF PROCRUSTES & IV) FOOLED BY RANDOMNESS

BY

NASSIM NICHOLAS TALEB

2013

FREELY AVAILABLE

Contents

I Fat Tails 11

1 An Introduction to Fat Tails and Turkey Problems 13
1.1 Introduction: Fragility, not Statistics . . . 13
1.2 The Problem of (Enumerative) Induction . . . 14
1.3 Simple Risk Estimator . . . 14
1.4 Fat Tails, the Finite Moment Case . . . 16
1.5 A Simple Heuristic to Create Mildly Fat Tails . . . 18
1.6 Fattening of Tails Through the Approximation of a Skewed Distribution for the Variance . . . 20
1.7 Scalable and Nonscalable, A Deeper View of Fat Tails . . . 21
1.8 Subexponentials as a class of fat tailed (in at least one tail) distributions . . . 22
1.9 Different Approaches For Statistically Derived Estimators . . . 26
1.10 Economics Time Series Econometrics and Statistics Tend to imagine functions in $L^2 Space$ 29
1.11 Typical Manifestations of The Turkey Surprise . . . 30
1.12 Metrics for Functions Outside L^2 Space . . . 30

A Appendix: Special Cases of Fat Tails 33
A.1 Multimodality and Fat Tails, or the War and Peace Model . . . 33
A.1.1 A brief list of other situations where bimodality is encountered: . . . 34
A.2 Transition probabilites, or what can break will eventually break . . . 34

2 A Heuristic Hierarchy of Distributions For Inferential Asymmetries 37
2.1 Masquerade Example . . . 37
2.2 The Probabilistic Version of Absense of Evidence vs Evidence of Absence . . . 38
2.3 Via Negativa and One-Sided Arbitrage of Statistical Methods . . . 38
2.4 A Heuristic Hierarchy of Distributions in Term of Fat-Tailedness . . . 39

3 An Introduction to Higher Orders of Uncertainty 45
3.1 Metaprobability . . . 45
3.2 The Effect of Metaprobability on the Calibration of Power Laws . . . 46
3.3 The Effect of Metaprobability on Fat Tails . . . 47
3.4 Fukushima, Or How Errors Compound . . . 47

4 Payoff Skewness and Lack of Skin-in-the-Game 49
4.1 Transfer of Harm . . . 49
4.1.1 Simple assumptions of constant q and simple-condition stopping time . . . 50
4.1.2 Explaining why Skewed Distributions Conceal the Mean . . . 51

5 Large Numbers and Convergence in the Real World **53**
5.1 The Law of Large Numbers Under Fat Tails . . . 53
5.2 Preasymptotics and Central Limit in the Real World . . . 57
5.3 Using Log Cumulants to Observe Preasymptotics . . . 59
5.4 Illustration: Convergence of the Maximum of a Finite Variance Power Law . . . 62

B Where Standard Diversification Fails **63**

C Fat Tails and Random Matrices **65**

6 Some Misuses of Statistics in Social Science **67**
6.1 Attribute Substitution . . . 67
6.2 The Tails Sampling Property . . . 68
6.2.1 On the difference between the initial (generator) and the "recovered" distribution . . . 68
6.2.2 Case Study: Pinker (2011) Claims On The Stability of the Future Based on Past Data . . . 68
6.2.3 Claims Made From Power Laws . . . 69
6.3 A discussion of the Paretan 80/20 Rule . . . 69
6.3.1 Why the 80/20 Will Be Generally an Error: The Problem of In-Sample Calibration . . . 71
6.4 Survivorship Bias (Casanova) Property . . . 71
6.5 Left (Right) Tail Sample Insufficiency Under Negative (Positive) Skewness . . . 73
6.6 Why N=1 Can Be Very, Very Significant Statistically . . . 74
6.7 The Instability of Squared Variations in Regression Analysis . . . 74
6.7.1 Application to Economic Variables . . . 76
6.8 Statistical Testing of Differences Between Variables . . . 76
6.9 Studying the Statistical Properties of Binaries and Extending to Vanillas . . . 77
6.10 The Mother of All Turkey Problems: How Economics Time Series Econometrics and Statistics Don't Replicate . . . 77
6.10.1 Performance of Standard Parametric Risk Estimators, $f(x) = x^n$ (Norm $\mathcal{L}2$) . . . 77
6.10.2 Performance of Standard NonParametric Risk Estimators, f(x)= x or |x| (Norm $\mathcal{L}1$), A =(-∞, K] 79
6.11 A General Summary of The Problem of Reliance on Past Time Series . . . 80
6.12 Conclusion . . . 81

D On the Instability of Econometric Data **83**

7 On the Difference between Binary Prediction and True Exposure **85**
7.1 Binary vs Vanilla Predictions and Exposures . . . 85
7.2 A Semi-Technical Commentary on The Mathematical Differences . . . 88
7.3 The Applicability of Some Psychological Biases . . . 91

8 How Fat Tails Emerge From Recursive Epistemic Uncertainty **93**
8.1 Methods and Derivations . . . 93
8.1.1 Layering Uncertainties . . . 93
8.1.2 Higher order integrals in the Standard Gaussian Case . . . 93
8.1.3 Effect on Small Probabilities . . . 96
8.2 Regime 2: Cases of decaying parameters *a*(*n*) . . . 97
8.2.1 Regime 2-a;“bleed” of higher order error . . . 97
8.2.2 Regime 2-b; Second Method, a Non Multiplicative Error Rate . . . 97
8.3 Conclusion . . . 98

9 On the Difficulty of Risk Parametrization With Fat Tails **99**

9.1 Some Bad News Concerning power laws 99
9.2 Extreme Value Theory: Fuhgetaboudit 100
9.2.1 What is Extreme Value Theory? A Simplified Exposition 100
9.2.2 A Note. How does the Extreme Value Distribution emerge? 100
9.2.3 Extreme Values for Fat-Tailed Distribution 101
9.2.4 How Extreme Value Has a Severe Inverse Problem In the Real World 101
9.3 Using Power Laws Without Being Harmed by Mistakes 102

10 Brownian Motion in the Real World (Path Dependence and Fat Tails) 103
10.1 Path Dependence and History as Revelation of Antifragility 103
10.2 Brownian Motion in the Real World 103
10.3 tochastic Processes and Nonanticipating Strategies 104
10.4 0.4 Finite Variance not Necessary for Anything Ecological (incl. quant finance) 105

11 The Fourth Quadrant Mitigation (or "Solution") 107
11.1 Two types of Decisions 107

II Fragility and Nonlinear Exposure to Random Variables 109

12 Exposures and Nonlinear Transformations of Random Variables 111
12.1 The Conflation Problem: Exposures to x Confused With Knowledge About x 111
12.1.1 Exposure, not knowledge 111
12.1.2 *Limitations of knowledge* 111
12.1.3 *Bad news* 112
12.1.4 *The central point about what to understand* 112
12.1.5 *Fragility and Antifragility* 112
12.2 Transformations of Probability Distributions 112
12.2.1 Some Examples. 112
12.3 Application 1: Happiness $(f(x))$does not have the same statistical properties as wealth (x) 113
12.3.1 Case 1: The Kahneman Tversky Prospect theory, which is convex-concave 113
12.4 The Mistake of Using Regular Estimation Methods When the Payoff is Convex 115
12.4.1 Convexity and Explosive Payoffs 116

13 Generalized Dose-Response Curves and The Origin of Thin-Tails 119
13.1 Properties of the Inherited Probability Distribution 120
13.2 Conclusion and Remarks 122

14 Why The World Will Progressively Look Wierder to Us 123
14.1 How Noise Explodes Faster than Data 123
14.2 Derivations 124

15 Mapping Fragility and AntiFragility (Taleb Douady 2013) 129
15.1 Introduction 129
15.1.1 Fragility As Separate Risk From Psychological Preferences 130
15.1.2 Fragility and Model Error 132
15.1.3 Antifragility 132
15.2 Mathematical Expression of Fragility 134
15.2.1 Definition of Fragility: The *Intrinsic* Case 134

16 Why is the fragile nonlinear? **145**

16.1 Antifragility from Uneven Distribution (. . . . 146

17 The One Percent of the One Percent Depends on Inequality, not Wealth **149**

Introduction: Rigor in "The Real World" is a Different Kind of Rigor

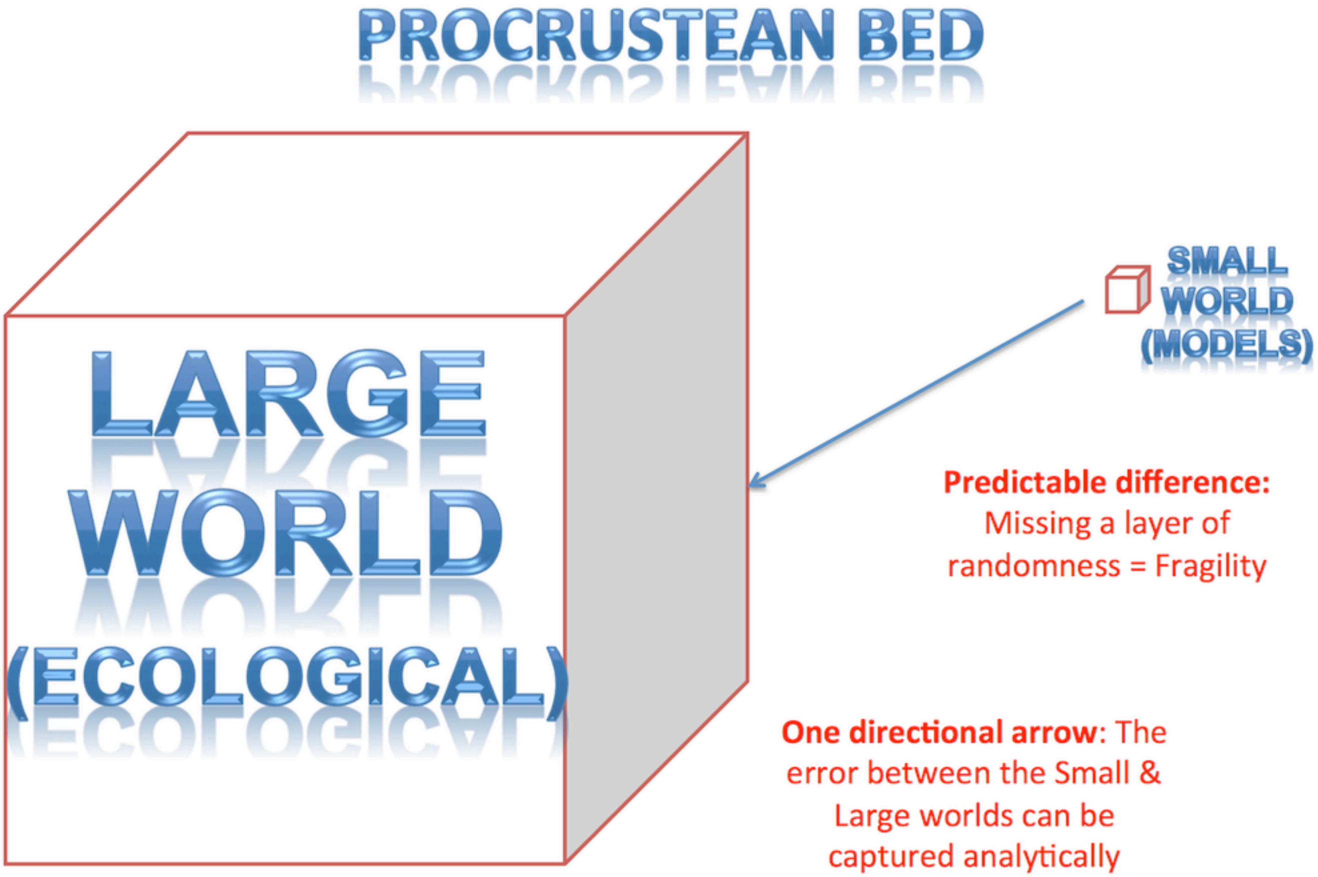


*Figure 1: The small World Large World Problem. In statistical domains assume **Small World= coin tosses** and **Large World = Real World**. Note that measure theory is not the small world, but large world, thanks to the degrees of freedom it confers.*

The problem of formal probability theory is that it necessarily covers narrower situations (small world Ω_S) than the real world (Ω_L), which produces Procrustean bed effects. $\Omega_S \subset \Omega_L$. The "academic" in the bad sense approach has been to assume that Ω_L is smaller rather than study the gap. The problems linked to incompleteness of models are largely in the form of preasymptotics and inverse problems.

Method: We cannot probe the Real World but we can get an idea (via perturbations) of relevant directions of the effects and difficulties coming from incompleteness, and make statements s.a. "incompleteness slows convergence to LLN by at least a factor of n$^{\alpha}$", or "increases the number of observations to make a certain statement by at least 2x".

So adding a layer of uncertainty to the representation in the form of model error, or metaprobability has a one-sided effect: expansion of Ω_S with following results:

i) **Fat tails**:
i-a)- Randomness at the level of the scale of the distribution generates fat tails. (Multi-level stochastic volatility).
i-b)- Model error in all its forms generates fat tails.
i-c) - Convexity of probability measures to uncertainty causes fat tails.
ii) Law of Large Numbers(weak): operates much more slowly, if ever at all. "P-values" are biased lower.
iii) Risk is larger than the conventional measures derived in Ω_S, particularly for payoffs in the tail.
iv) Allocations from optimal control and other theories (portfolio theory) have a higher variance than shown, hence increase risk.
v) The problem of induction is more acute.(epistemic opacity).
vi)The problem is more acute for convex payoffs, and simpler for concave ones.

Now i) $\Rightarrow$ ii) through vi).

The Difference Between Real World World and Models

Convex Heuristic

We give the reader a hint of the methodology and proposed approach with a semi-informal technical definition for now.

Definition 1. ***Rule***. *A rule is a decision-making heuristic that operates under a broad set of circumtances. Unlike a theorem, which depends on a specific (and closed) set of assumptions, it holds across a broad range of environments —which is precisely the point. In that sense it is stronger than a theorem for decision-making.*

Chapter x discusses robustness under perturbation or metamodels (or metaprobability). Here is the preview of the idea of convex heuristic, which in plain English, is at least robust to model uncertainty.

Definition 2. ***Convex Heuristic***. *In short it is required to not produce concave responses under parameter perturbation.*

Result of Chapter x Let $\{f_i\}$ be the family of possible functions, or "exposures" to x a random variable with probability measure $\lambda_{\sigma^-}(x)$, where σ^- is a parameter determining the scale (say, mean absolute deviation) on the left side of the distribution (below the mean). A rule is said "nonconcave" for payoff below K with respect to σ^- up to perturbation Δ if, taking the partial expected payoff

$$\mathbb{E}^K_{\sigma^-}(f_i) = \int_{-\infty}^{K} f_i(x)\, \mathrm{d}\lambda_{\sigma^-}(x),$$

f_i is deemed member of the family of convex heuristic $(x, K, \sigma^-, etc.)$:

$$\left\{ f_i : \frac{1}{2}\left(\mathbb{E}^K_{\sigma^- - \Delta}(f_i) + \mathbb{E}^K_{\sigma^- + \Delta}(f_i) \right) \geq \mathbb{E}^K_{\sigma^-}(f_i) \right\}$$

Note that we call these decision rules "convex" not necessarily because they have a convex payoff, but because their payoff is comparatively "more convex" (less concave) than otherwise. In that sense, finding protection is a convex act. The idea that makes life easy is that we can capture model uncertainty with simple tricks, namely the scale of the distribution.

A Class With an Absurd Name

This author currently teaching a class with the absurd title "risk management and decision - making in the real world", a title he has selected himself; this is a total absurdity since risk management and decision making should never have to justify being about the real world, and what' s worse, one should never be apologetic about it. In "real" disciplines, titles like "Safety in the Real World", "Biology and Medicine in the Real World" would be lunacies. But in social science all is possible as there is no exit from the gene pool for blunders, nothing to check the system, so skin in the game for researchers. You cannot blame the pilot of the plane or the brain surgeon for being "too practical", not philosophical enough; those who have done so have exited the gene pool. The same applies to decision making under uncertainty and incomplete information. The other absurdity in is the common separation of risk and decision making, as the latter cannot be treated in any way except under the constraint : in the real world.

And the real world is about incompleteness : incompleteness of understanding, representation, information, etc., what one does when one does not know what' s going on, or when there is a non - zero chance of not knowing what' s going on. It is based on focus on the unknown, not the production of mathematical certainties based on weak assumptions; rather measure the robustness of the exposure to the unknown, which can be done mathematically through metamodel (a model that examines the effectiveness and reliability of the model), what I call metaprobability, even if the meta - approach to the model is not strictly probabilistic.

This first volume presents a mathematical approach for dealing with errors in conventional risk models, taking the bulls ***t out of some, adding robustness, rigor and realism to others. For instance, if a "rigorously" derived model (say Markowitz mean variance, or Extreme Value Theory) gives a precise risk measure, but ignores the central fact that the parameters of the model don' t fall from the sky, but have some error rate in their estimation, then the model is not rigorous for risk management, decision making in the real world, or, for that matter, for anything (other than academic tenure). We need to add another layer of uncertainty, which invalidates some models (but not others). The mathematical rigor is therefore shifted from focus on asymptotic (but rather irrelevant) properties to making do with a certain set of incompleteness. Indeed there is a mathematical way to deal with incompletness. Adding disorder has a one-sided effect and we can deductively estimate its lower bound. For instance we know from Jensen's inequality that tail probabilities and risk measures are understimated in some class of models.

Fat Tails

The focus is squarely on "fat tails", since risks and harm lie principally in the high - impact events, The Black Swan and some statistical methods fail us there. The section ends with an identification of classes of exposures to these risks, the Fourth Quadrant idea, the class of decisions that do not lend themselves to modelization and need to be avoided. Modify your decisions. The reason decision making and risk management are insparable is that there are some exposure people should never take if the risk assessment is not reliable, something people understand in real life but not when modeling. About every rational person facing an plane ride with an unreliable risk model or a high degree of uncertainty about the safety of the aircraft would take a train instead; but the same person, in the absence of skin in the game, when working as "risk expert" would say : "well, I am using the best model we have" and use something not reliable, rather than be consistent with real-life decisions and subscribe to the straightforward principle : "let's only take those risks for which we have a reliable model".

Combination of Small World and Lack of Skin in the Game. The disease of formalism in the application of probability to real life by people who are not harmed by their mistakes can be illustrated as follows, with a very sad case study. One of the most

"cited" document in risk and quantitative methods is about "coherent measures of risk", which set strong principles on how to compute tail risk measures, such as the "value at risk" and other methods. Initially circulating in 1997, the measures of tail risk −while coherent− have proven to be underestimating risk at least 500 million times (sic). We have had a few blowups since, including Long Term Capital Management fiasco −and we had a few blowups before, but departments of mathematical probability were not informed of them. As I am writing these lines, it was announced that J.-P. Morgan made a loss that should have happened every ten billion years. The firms employing these "risk minds" behind the "seminal" paper blew up and ended up bailed out by the taxpayers. But we now now about a "coherent measure of risk". This would be the equivalent of risk managing an airplane flight by spending resources making sure the pilot *uses proper grammar* when communicating with the flight attendants, in order to "prevent incoherence". Clearly the problem, just as similar fancy *b***t* under the cover of the discipline of Extreme Value Theory is that tail events are very opaque computationally, and that such misplaced precision leads to confusion.
The "seminal" paper: Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical finance, 9(3), 203-228.*

Orthodoxy

Finally, someone recently asked me to give a talk at unorthodox statistics session of the American Statistical Association. I refused : the approach presented here is about as orthodox as possible, much of the bone of this author come precisely from enforcing rigorous standards of statistical inference on process. Risk (and decisions) require more rigor than other applications of statistical inference.

Measure Theory is not restrictive

In his wonderful textbook, Leo Breiman referred to probability as having two sides, the left side represented by his teacher, Michel Loève, which concerned itself with formalism and measure theory, and the right one which is typically associated with coin tosses and similar applications. Many have the illusion that the "real world" would be closer to the coin tosses. It is not: coin tosses are fake practice for probability theory, artificial setups in which people know the probability (what is called the **ludic fallacy** in *The Black Swan*. Measure theory, while formal, is liberating because it sets us free from these narrow structures. Its abstraction allows the expansion out of the small box, all the while remaining rigorous, in fact, at the highest possible level of rigor.

General Problems

The Black Swan Problem

Incomputability of Small Probalility: It is 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. And the smaller the probability, the larger the error, affecting events of high impact. The idea is to work with measures that are less sensitive to the issue (a statistical approch), or conceive exposures less affected by it (a decision theoric approach). Mathematically, the problem arises from the use of degenerate metaprobability.
In fact the central point is the 4^{th} quadrant where prevails both high-impact and non-measurability, where the max of the random variable determines most of the properties (which to repeat, has not computable probabilities).

	Problem	Description	Chapters/Sections
P 1	Preasymptotics, Incomplete Convergence	The real world is before the asymptote. This affects the applications (under fat tails) of the Law of Large Numbers and the Central Limit Theorem.	?
P2	Inverse Problems	a) The direction Model ⇒ Reality produces larger biases than Reality ⇒ Model b) Some models can be "arbitraged" in one direction, not the other .	1,?,?
P3	Conflation	a) The statistical properties of an exposure, f(x) are different from those of a r.v. x, with very significant effects under nonlinearities (when f(x) convex). b)Exposures and decisions can be modified, not probabilities.	1, 9
P4	Degenerate Metaprobability	Uncertainty about the probability distributions can be expressed as additional layer of uncertainty, or, simpler, errors, hence nested series of errors on errors. The Black Swan problem can be summarized as degenerate metaprobability.[1]	?,?

Definition 3. ***Arbitrage of Probability Measure.*** *A probability measure μ_A can be arbitraged if one can produce data fitting another probability measure μ_B and systematically fool the observer that it is μ_A based on his metrics in assessing the validity of the measure.*

We will rank probability measures along this arbitrage criterion.

Associated Specific "Black Swan Blindness" Errors (Applying Thin-Tailed Metrics to Fat Tailed Domains)

These are shockingly common, arising from mechanistic reliance on software or textbook items (or a culture of bad statistical insight). I skip the elementary "Pinker" error of mistaking journalistic fact - checking for scientific statistical "evidence" and focus on less obvious but equally dangerous ones.

1. **Overinference**: Making an inference from fat-tailed data assuming sample size allows claims (very common in social science). Chapter 3.
2. **Underinference**: Assuming $N{=}1$ is insufficient under large deviations. Chapters 1 and 3.
   (In other words both these errors lead to refusing true inference and accepting anecdote as "evidence")
3. Asymmetry: Fat-tailed probability distributions can masquerade as thin tailed ("great moderation", "long peace"), not the opposite.
4. The econometric (*very severe*) violation in using standard deviations and variances as a measure of dispersion without ascertaining the stability of the fourth moment (**??**.**??**) . This error alone allows us to discard everything in economics/econometrics using σ as irresponsible nonsense (with a narrow set of exceptions).
5. Making claims about "robust" statistics in the tails. Chapter 1.
6. Assuming that the errors in the estimation of x apply to *f(x)* (*very severe*).
7. Mistaking the properties of "Bets" and "digital predictions" for those of Vanilla exposures, with such things as "prediction markets".

Chapter 9.

8. Fitting tail exponents power laws in interpolative manner. Chapters 2, 6
9. Misuse of Kolmogorov-Smirnov and other methods for fitness of probability distribution. Chapter 1.
10. Calibration of small probabilities relying on sample size and not augmenting the total sample by a function of 1/p , where p is the probability to estimate.
11. Considering ArrowDebreu State Space as exhaustive rather than sum of known probabilities ≤ 1

	Principles	**Description**
$\mathcal{P}1$	Dutch Book	Probabilities need to add up to 1*
$\mathcal{P}2$	Asymmetry	Some errors are largely one sided.
$\mathcal{P}3$	Nonlinear Response	Fragility is more measurable than probability**
$\mathcal{P}4$	Conditional Precautionary Principle	Domain specific precautionary, based on fat tailedness of errors and asymmetry of payoff.
$\mathcal{P}5$	Decisions	Exposures can be modified, not probabilities.

* This and the corrollary that there is a non-zero probability of visible and known states spanned by the probability distribution adding up to <1 confers to probability theory, when used properly, a certain analytical robustness.

**The errors in measuring nonlinearity of responses are more robust and smaller than those in measuring responses. (Transfer theorems)

Definition 4. *Metaprobability: the two statements 1) "the probability of Rand Paul winning the election is 15.2%" and 2) the probability of getting n odds numbers in N throws of a fair die is x%" are different in the sense that the first statement has higher undertainty about its probability, and you know (with some probability) that it may change under an alternative analysis or over time.*

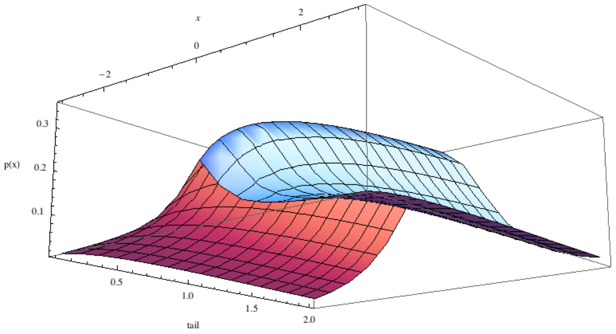


Figure 2: Metaprobability: we add another dimension to the probability distributions, as we consider the effect of a layer of uncertainty over the probabilities. It results in large effects in the tails, but, visually, these are identified through changes in the "peak" at the center of the distribution.

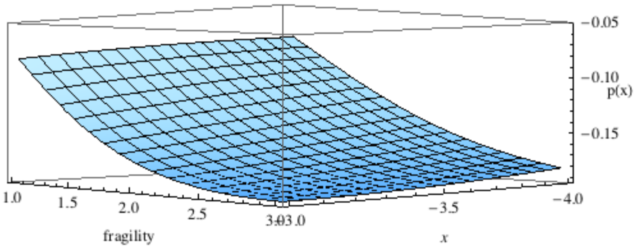


Figure 3: Fragility: Can be seen in the slope of the sensitivity of payoff across metadistributions

Part I

Fat Tails

1 An Introduction to Fat Tails and Turkey Problems

This is an introductory chapter outlining the turkey problem, showing its presence in data, explaining why an assessment of fragility is more potent than data-based methods of risk detection, introducing fat tails, and showing how fat tails cause turkey problems.

Figure 1.1: The risk of breaking of the coffee cup is not necessarily in the past time series of the variable; in fact surviving objects have to have had a "rosy" past.

1.1 Introduction: Fragility, not Statistics

Fragility (Volume 2) can be defined as an accelerating sensitivity to a harmful stressor: this response plots as a concave curve and mathematically culminates in more harm than benefit from the disorder cluster: (i) uncertainty, (ii) variability, (iii) imperfect, incomplete knowledge, (iv) chance, (v) chaos, (vi) volatility, (vii) disorder, (viii) entropy, (ix) time, (x) the unknown, (xi) randomness, (xii) turmoil, (xiii) stressor, (xiv) error, (xv) dispersion of outcomes, (xvi) unknowledge.

Antifragility is the opposite, producing a convex response that leads to more benefit than harm. We do not need to know the history and statistics of an item to measure its fragility or antifragility, or to be able to predict rare and random ('black swan') events. All we need is to be able to assess whether the item is accelerating towards harm or benefit.

The relation of fragility, convexity and sensitivity to disorder is thus mathematical and not derived from empirical data.

The problem with risk management is that "past" time series can be (and actually are) unreliable. Some finance journalist was commenting on my statement in *Antifragile* about our chronic inability to get the risk of a variable from the past with economic time series. "Where is he going to get the risk from since we cannot get it *from the past*? from the future?", he wrote. Not really, think about it: *from the present, the present state of the system.* This explains in a way why the detection of fragility is vastly more potent than that of risk -and much easier to do.

Asymmetry and Insufficiency of Past Data.

Our focus on fragility does not mean you can ignore the past history of an object for risk management, it is just accepting that the past is highly *insufficient.* The past is also *highly asymmetric.* There are instances (large deviations) for which the past reveals extremely valuable information about the risk of a process. Something that broke once before is breakable, but we cannot ascertain that what did not break is unbreakable. This asymmetry is extremely valuable with fat tails, as we can reject some theories, and get to the truth by means of *via negativa.*

This confusion about the nature of empiricism, or the difference between empiricism (rejection) and naive empiricism (anecdotal acceptance) is not just a problem with journalism. Naive inference from time series is incompatible with rigorous statistical inference; yet many workers with time series believe that it *is* statistical inference. One has to think of history as a sample path, just as one looks at a sample from a large population, and continuously keep in mind how representative the sample is of the large population. While analytically equivalent, it is psychologically hard to take the "outside view", given that we are all part of history, part of the sample so to speak.

General Principle To Avoid Imitative, Cosmetic (Job Market) Science:

From *Antifragile (2012):*
There is such a thing as nonnerdy applied mathematics: find a problem first, and p out the math that works for it (just as one acquires language), rather than study in a vacuum through theorems and artificial examples, then change reality to make it look like these examples.
The problem can be seen in the opposition between problems and inverse problems. To cite (Donald Geman), there are hundreds of theorems one can elaborate and prove, all of which may seem to find *some* application from the real world. But applying the idea of non-reversibility of the mechanism: there are very, very few theorems that can correspond to an exact selected problem. In the end this leaves us with a restrictive definition of what "rigor" means in mathematical treatments of the real world.

1.2 The Problem of (Enumerative) Induction

Turkey and Inverse Turkey (from the Glossary for *Antifragile*): The turkey is fed by the butcher for a thousand days, and every day the turkey pronounces with increased statistical confidence that the butcher "will never hurt it"−until Thanksgiving, which brings a Black Swan revision of belief for the turkey. Indeed not a good day to be a turkey. The inverse turkey error is the mirror confusion, not seeing opportunities− pronouncing that one has evidence that someone digging for gold or searching for cures will "never find" anything because he didn't find anything in the past. What we have just formulated is the philosophical problem of induction (more precisely of enumerative induction.) To this version of Bertrand Russel's chicken we add: mathematical difficulties, fat tails, and sucker problems.

1.3 Simple Risk Estimator

Let us define a risk estimator that we will work with throughout the book.
Definition 5. *Let X be, as of time T, a standard sequence of n+1 observations,* $X = (x_{t_0+i\Delta t})_{0\leq i\leq n}$ *(with* $x_t \in \mathbb{R}$*,* $i \in \mathbb{N}$*), as the discretely monitored history of a stochastic process* X_t *over the closed interval* $[t_0, T]$ *(with realizations at fixed interval* Δt *thus* $T = t_0 + n\Delta t$*). The empirical estimator* $M_T^X(A,f)$ *is defined as*

$$M_T^X(A,f) \equiv \frac{\sum_{i=0}^{n} \mathbf{1}_A f\left(x_{t_0+i\Delta t}\right)}{\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}'}} \tag{1.1}$$

where $\mathbf{1}_A$ $\mathcal{D} \to \{0,1\}$ is an indicator function taking values 1 if $x_t \in$ A and 0 otherwise, ($\mathcal{D}'$ subdomain of domain $\mathcal{D}$: $A \subseteq \mathcal{D}' \subset \mathcal{D}$) , and f is a function of x. For instance $f(x) = 1, f(x) = x$, and $f(x) = x^N$ correspond to the probability , the first moment, *and* N^{th} moment, respectively. A is the subset of the support of the distribution that is of concern for the estimation. Typically, $\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}} = n$.

Let us stay in dimension 1 for the next few chapters not to muddle things. Standard Estimators tend to be variations about $M_t^X(A,f)$ where f(x) =x and A is defined as the domain of the distribution of X, standard measures from x, such as moments of order z, etc., are calculated "as of period" T. Such measures might be useful for the knowledge of some properties, but remain insufficient for decision making as the decision-maker may be concerned for risk management purposes with the left tail (for distributions that are not entirely skewed, such as purely loss functions such as damage from earthquakes, terrorism, etc.), or any arbitrarily defined part of the distribution.

Standard Risk Estimators

Definition 6. *The empirical risk estimator S for the unconditional shortfall S below K is defined as, with $A = (-\infty, K)$, $f(x) = x$*

$$S \equiv \frac{\sum_{i=0}^{n} \mathbf{1}_A}{\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}'}} \tag{1.2}$$

An alternative method is to compute the conditional shortfall:

$$S' \equiv \mathbb{E}[M|X < K] = \frac{\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}'}}{\sum_{i=0}^{n} \mathbf{1}_A}$$

$$S' = \frac{\sum_{i=0}^{n} \mathbf{1}_A}{\sum_{i=0}^{n} \mathbf{1}_A}$$

One of the uses of the indicator function $\mathbf{1}_A$, for observations falling into a subsection A of the distribution, is that we can actually derive the past actuarial value of an option with X as an underlying struck as K as $M_T^X(A, x)$, with $A = (-\infty, K]$ for a put and $A = [K, \infty)$ for a call, with $f(x) = x$

Criterion 1. *The measure M is considered to be an estimator over interval [t- N Δt, T] if and only if it holds* in expectation *over a specific period $X_{T+i\Delta t}$ for a given i>0, that is across counterfactuals of the process, with a threshold ξ (a tolerated relative absolute divergence that can be a bias) so*

$$\frac{\mathbb{E}\left|M_T^X\left(A_z, 1\right) - M_{>T}^X\left(A_z, 1\right)\right|}{\left|M_T^X\left(A_z, 1\right)\right|} < \xi \tag{1.3}$$

when $M_T^X\left(A_z, 1\right)$ is computed; but while working with the opposite problem, that is, trying to guess the spread in the realizations of a stochastic process, when the process is known, but not the realizations, we will use $M_{>T}^X\left(A_z, 1\right)$ as a divisor.

In other words, the estimator as of some future time, should have some stability around the "true" value of the variable and stay below an upper bound on the tolerated bias.

We skip the notion of "variance" for an estimator and rely on absolute mean deviation so ξ can be the absolute value for the tolerated bias. And note that we use mean deviation as the equivalent of a "loss function"; except that with matters related to risk, the loss function is embedded in the subt A of the estimator.

This criterion is compatible with standard sampling theory. Actually, it is at the core of statistics. Let us rephrase:

Standard statistical theory doesn't allow claims on estimators made in a given set unless these are made on the basis that they can "generalize", that is, reproduce out of sample, into the part of the series that has not taken place (or not seen), i.e., for time series, for τ >t.

This should also apply in full force to the risk estimator. In fact we need more, much more vigilance with risks.

For convenience, we are taking some liberties with the notations, pending on context: $M_T^X(A, f)$ is held to be the estimator, or a conditional summation on data but for convenience, given that such estimator is sometimes called "empirical expectation", we will be also using the same symbol, namely with $M_{>T}^X(A, f)$ for the textit estimated variable for period $> T$ (to the right of T, as we will see, adapted to the filtration T). This will be done in cases M is the M-derived expectation operator $\mathbb{E}$ or $\mathbb{E}^P$ under real world probability measure $\mathbb{P}$ (taken here as a counting measure), that is, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a continuously increasing filtration $\mathcal{F}_t$, $\mathcal{F}_s \subset \mathcal{F}_t$ if s < t. the expectation operator (and other Lebesque measures) are adapted to the filtration $\mathcal{F}_T$ in the sense that the future is progressive and one takes a decision at a certain period $T + \Delta$t from information at period T, with an incompressible lag that we write as Δt −in the "real world", we will see in Chapter x there are more than one laging periods Δt, as one may need a lag to make a decision, and another for execution, so we necessarily need $> \Delta$t. The central idea of a *cadlag* process is that in the presence of discontinuities in an otherwise continuous stochastic process (or treated as continuous), we consider the right side, that is the first observation, and not the last.

1.4 Fat Tails, the Finite Moment Case

Fat tails are not about the incidence of low probability events, but the contributions of events away from the "center" of the distribution to the total properties.
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 $\frac{\left(M_T^X(A,x^n)\right)^{\frac{1}{n}}}{M_T^X(A,|x|)}$, $n > 1$, under the condition that the distribution has finite moments up to n, and the special case here n=2).
Simply, $x^n$is a weighting operator that assigns a weight, x^{n-1} 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.

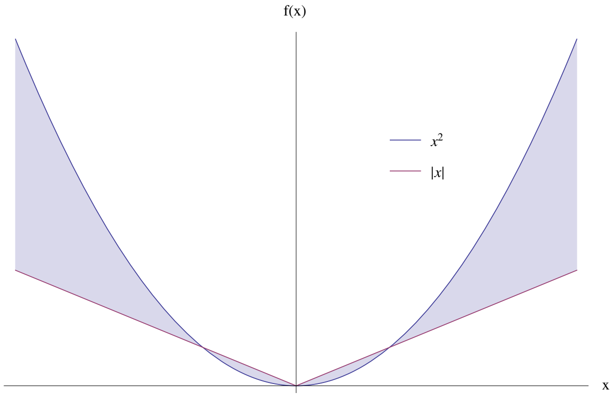


Figure 1.2: The difference between the two weighting functions increases for large values of x.

Proof: By Jensen's inequality under the counting measure.

Some formalism: L^p space

It is not just more rigorous, but more convenient to look at payoff in functional space, work with the space of functions having a certain integrability. Let Y be a measurable space with Lebesgue measure μ. The space $L^p$of f measurable functions on Y is defined as:

$$L^p(\mu) = \left\{ f : Y \to C \cup \infty : \left(\int_Y |f^p| \, d\mu \right)^{1/p} < \infty \right\}$$

The application of concern for our analysis is where the measure μ is a counting measure (on a countable set) and the function $f(y) \equiv y^p$, $p \geq 1$.
As a convention here, we write L^p for space, $\mathcal{L}^p$ for the norm in that space.
Let $X \equiv (x_i)_{i=1}^n$, The $\mathcal{L}^p$ Norm is defined (for our purpose) as, with p $\in \mathbb{N}$, $p \geq 1$):

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

The idea of dividing by n is to transform the norms into expectations,i.e., moments. For the Euclidian norm, $p = 2$.
The norm rises with higher values of p, as, with $a > 0$ [1],

$$\left(\frac{1}{n} \sum_{i=1}^n |x_i|^{p+a} \right)^{1/(p+a)} \geqslant \left(\frac{1}{n} \sum_{i=1}^n |x_i|^p \right)^{1/p}$$

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= $\{1, 1\}$. $\|X\|_1 = \|X\|_2 =$... $= \|X\|_n =$ 1. Perturbating while preserving $\|X\|_1$, $X = \left\{ \frac{1}{2}, \frac{3}{2} \right\}$ produces rising higher norms

$$\{\|X\|_n\}_{n=1}^5 = \left\{ 1, \frac{\sqrt{5}}{2}, \frac{\sqrt[3]{7}}{2^{2/3}}, \frac{\sqrt[4]{41}}{2}, \frac{\sqrt[5]{61}}{2^{4/5}} \right\}. \quad (1.4)$$

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

[1] An application of Hölder's inequality, $\left(\sum_{i=1}^n |x_i|^{p+a}\right)^{\frac{1}{a+p}} \geq \left(n^{\frac{1}{a+p} - \frac{1}{p}} \sum_{i=1}^n |x_i|^p\right)^{1/p}$

$$\{||X||_n\}_{n=1}^5 = \left\{1, \frac{5}{4}, \frac{\sqrt[3]{\frac{43}{2}}}{2}, \frac{\sqrt[4]{1201}}{4}, \frac{\sqrt[5]{2101}}{2 \times 2^{3/5}}\right\}. \tag{1.5}$$

So we can see it becomes rapidly explosive.

One property quite useful with power laws with infinite moment:

$$\|X\|_\infty = \sup \left(\frac{1}{n}|x_i|\right)_{i=1}^n \tag{1.6}$$

Gaussian Case

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

$$\frac{M_T^X\left(A, X^N\right)^{1/N}}{M_T^X\left(A,|X|\right)} = \frac{\pi^{\frac{N-1}{2N}}\left(2^{\frac{N}{2}-1}\left((-1)^N+1\right)\,\Gamma\left(\frac{N+1}{2}\right)\right)^{\frac{1}{N}}}{\sqrt{2}}$$

or, alternatively

$$\frac{M_T^X\left(A, X^N\right)}{} = 2^{\frac{1}{2}(N-3)}\left(1+(-1)^N\right)\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}-\frac{N}{2}}\Gamma\left(\frac{N+1}{2}\right) \tag{1.7}$$

where Γ(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:

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

hence

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

For a Gaussian the ratio $\sim$ 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.
As to the fourth moment, it equals $3\sqrt{\frac{\pi}{2}}\sigma^3$.
For a power law distribution with tail exponent α=3, say a Student T

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

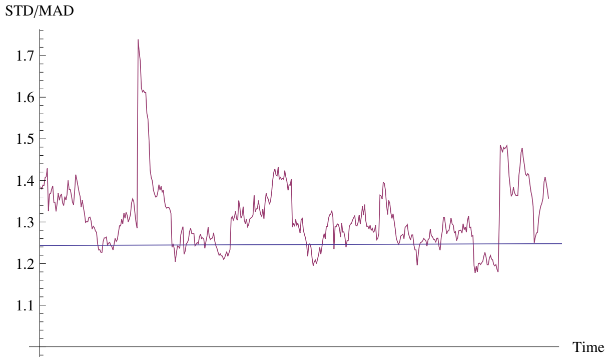


Figure 1.3: The Ratio Standard Deviation/Mean Deviation for the daily returns of the SP500 over the past 47 years, with a monthly window.

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 infinite means will always deliver a measurable mean in finite samples; but different samples will deliver completely different means. The next two figures illustrate the "drifting" effect of M a with increasing information.

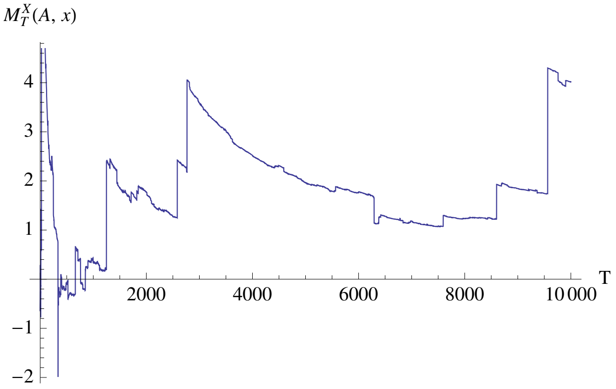


Figure 1.4: The mean of a series with Infinite mean (Cauchy).

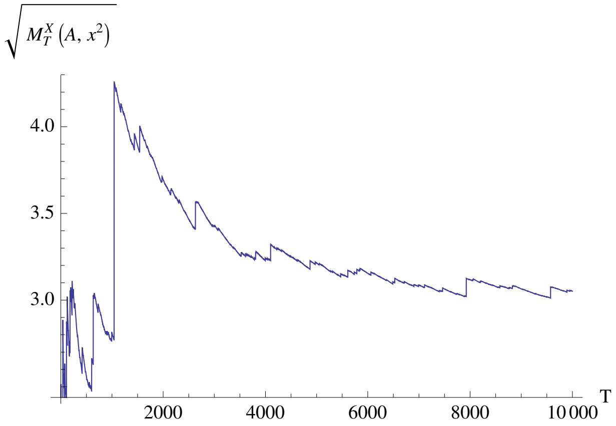


Figure 1.5: The standard deviation of a series with infinite variance (St(2)).

1.5 A Simple Heuristic to Create Mildly Fat Tails

Since higher moments increase under fat tails, as compared to lower ones, it should be possible so simply increase fat tails without increasing lower moments.

Note that the literature sometimes separates "Fat tails" from "Heavy tails", the first term being reserved for power laws, the second to subexponential distribution (on which, later). Fughtetaboutdit. We simply call "Fat Tails" something with a higher kurtosis than the Gaussian, even when kurtosis is not defined. The definition is functional as used by practioners of fat tails, that is, option traders and lends itself to the operation of "fattening the tails", as we will see in this section.

A Variance-preserving heuristic. Keep $\mathbb{E}\left(X^2\right)$ constant and increase $\mathbb{E}\left(X^4\right)$, by "stochasticizing" the variance of the distribution, since $<X^4>$ is itself analog to the variance of $<X^2>$ measured across samples ($\mathbb{E}\left(X^4\right)$ is the noncentral equivalent of $\mathbb{E}\left(\left(X^2-\mathbb{E}\left(X^2\right)\right)^2\right)$). Chapter x will do the "stochasticizing" in a more involved way.

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 $\mathrm{N}\left(0,\sigma\sqrt{1-a}\right)$ with probability $p=\frac{1}{2}$ and $\mathrm{N}\left(0,\sigma\sqrt{1+a}\right)$ with the remaining probability $\frac{1}{2}$, with $0 \leqslant \mathrm{a} < 1$.

The characteristic function is

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

Odd moments are nil. The second moment is preserved since

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

and the fourth moment

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

which puts the traditional kurtosis at $3\left(a^2+1\right)$. 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.

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 limited to getting some intuition about its effects. Section **??**.**??** will present a more involved technique.

As Figure **??**.**??** shows: fat tails are about higher peaks, a concentration of observations around the center of the distribution.

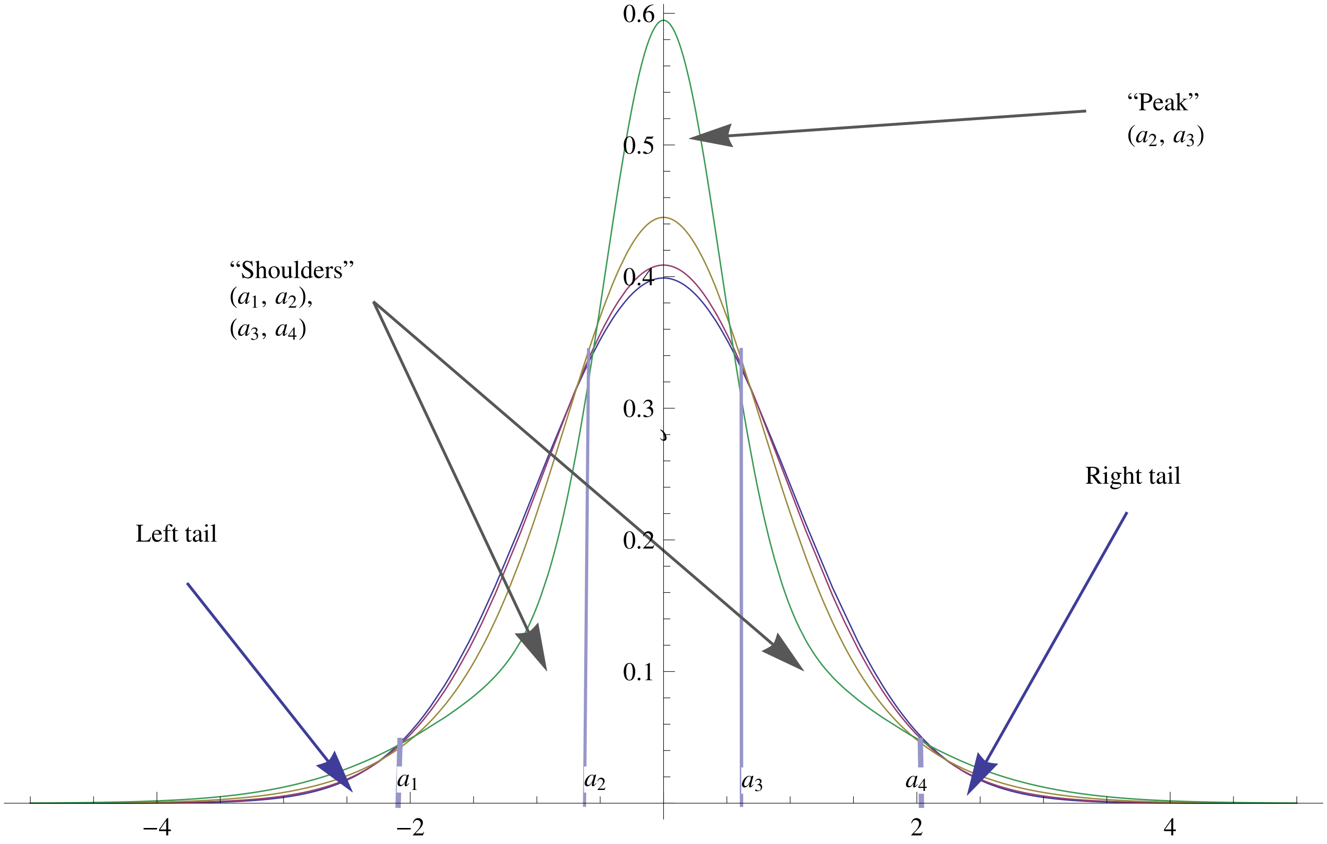


Figure 1.6: Fatter and Fatter Tails through perturbation of σ. The mixed distribution with values a=0,.25,.5, .75 . We can see crossovers a_1 through a_4. One can safely claim that the tails start at $a_4$on the right and $a_1$on the left.

The Crossovers and the Tunnel Effect

Notice in the figure x a series of crossover zones, invariant to a. Distributions called "bell shape" have a convex-concave-convex shape.

Let X be a random variable, the distribution of which p(x) is from a general class of all monomodal one-parameter continous pdfs p_σ with support $\mathcal{D} \subseteq \mathbb{R}$ and scale parameter σ.

1- If the variable is "two-tailed", that is, $\mathcal{D}$= (-∞,∞), where $p^\delta(x) \equiv \frac{p(x+\delta)+p(x-\delta)}{2}$

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 $x \in (a_2, a_3)$

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

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

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

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

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

which produces the following crossovers:

$$\{a_1, a_2, a_3, a_4\} =$$

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

In figure **??**, the crossovers for the intervals are numerically $\{-2.13\sigma, -.66\sigma, .66\sigma, 2.13\sigma\}$
As to a "cubic" symmetric power law(as we will see further down), the Student T Distribution with scale s and tail exponent 3

$$p(x) \equiv \frac{6\sqrt{3}}{\pi s \left(\frac{x^2}{s^2}+3\right)^2}$$

$$\{a_1, a_2, a_3, a_4\} =$$

$$\left\{-\sqrt{4-\sqrt{13}}s, -\sqrt{4+\sqrt{13}}s, \sqrt{4-\sqrt{13}}s, \sqrt{4+\sqrt{13}}s\right\}$$

2- 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.

$$\{a_1, a_2, a_3, a_4\} =$$

$$e^{\frac{1}{2}\left(2\mu-\sqrt{2}\sqrt{5\sigma^2-\sqrt{17}\sigma^2}\right)}, e^{\frac{1}{2}\left(2\mu-\sqrt{2}\sqrt{\sqrt{17}\sigma^2+5\sigma^2}\right)},$$

$$e^{\frac{1}{2}\left(2\mu+\sqrt{2}\sqrt{5\sigma^2-\sqrt{17}\sigma^2}\right)}, e^{\frac{1}{2}\left(2\mu+\sqrt{2}\sqrt{\sqrt{17}\sigma^2+5\sigma^2}\right)}$$

1.6 Fattening of Tails Through the Approximation of a Skewed Distribution for the Variance

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

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

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

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

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

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

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, 2006).

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, it can correspond to incomplete assessment of tail events.

Chapter x will show how tail events have large errors.

Why do we use Student T to simulate symmetric power laws? 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).

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 Chapt x, of concave-convex responses (sort of dampening of fragility and antifragility, bringing robustness, hence thinning tails).

1.7 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 Paretan tail, i.e., for x large enough, $P_{X>x} = Cx^{-\alpha}$ where α is the tail and C is a scaling constant.

k	$\mathbb{P}(X > k)^{-1}$	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\ k)}$	$\mathbb{P}(X > k)^{-1}$	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\ k)}$	$\mathbb{P}(X > k)^{-1}$	$\frac{\mathbb{P}(X>k)}{\mathbb{P}(X>2\ k)}$
	(Gaussian)	(Gaussian)	Student(3)	Student (3)	Pareto(2)	Pareto (2)
2	44	720	14.4	4.97443	8	4
4	31600.	5.1×10^{10}	71.4	6.87058	64	4
6	1.01×10^{9}	5.5×10^{23}	216	7.44787	216	4
8	1.61×10^{15}	9×10^{41}	491	7.67819	512	4
10	1.31×10^{23}	9×10^{65}	940	7.79053	1000	4
12	5.63×10^{32}	fughetaboudit	1610	7.85318	1730	4
14	1.28×10^{44}	fughetaboudit	2530	7.89152	2740	4
16	1.57×10^{57}	fughetaboudit	3770	7.91664	4100	4
18	1.03×10^{72}	fughetaboudit	5350	7.93397	5830	4
20	3.63×10^{88}	fughetaboudit	7320	7.94642	8000	4

Table 1.1: Scalability, comparing slowly varying functions to other distributions

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 $\lim_{x\to\infty}\frac{L(tx)}{Lx}$=1 for all constants $t > 0$.

For X large enough, $\frac{\text{logP}_{>x}}{\text{logx}}$ converges to a constant, the tail exponent -α. A scalable should produce the slope α in the tails on a log-log plot, as $x \to \infty$

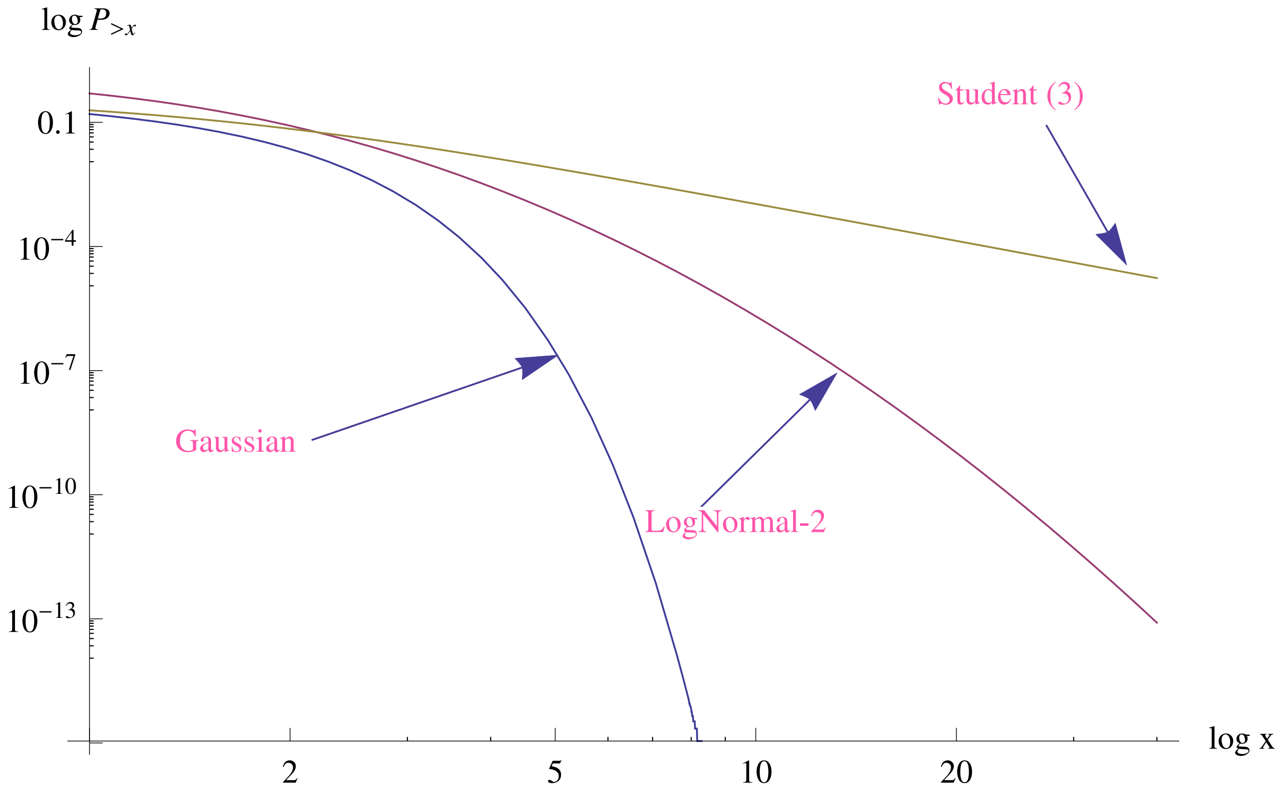


Figure 1.7: Three Distributions. As we hit the tails, the Student remains scalable while the Standard Lognormal shows an intermediate position.

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.

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 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 x.

1.8 Subexponentials as a class of fat tailed (in at least one tail) distributions

We introduced the category "true fat tails" as scalable power laws to differenciate it from the weaker one of fat tails as having higher kurtosis than a Gaussian.
Some use as a cut point infinite variance, but Chapter 2 will show it to be not useful, even misleading. Many finance researchers (Officer, 1972) and many private communications with *finance artists* reveal some kind of mental block in seeing the world polarized into finite/infinite variance.
Another useful distinction: Let $X = (x_i)_{2 \leq i \leq n}$ be i.i.d. random variables in $\mathbb{R}^+$, with cumulative distribution function F, by the Teugels (1975) and Pitman (1980) definition:

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

where F^2 is the convolution of$x$ with itself.
Note that X does not have to be limited to $\mathbb{R}^+$; we can split the variables in positive and negative domain for the analysis.

Example 1

Let $f^2(x)$ be the density of a once-convolved one-tailed Pareto distribution scaled at 1 with tail exponent α, where the density of the non-convolved distribution

$$f(x) = \alpha\, x^{-\alpha-1},$$

$x \geq 1$, $x \in [2, \infty)$,
which yields a closed-form density:
$f^2(x) =$

$$2\alpha^2 x^{-2\alpha-1}\left(B_{\frac{x-1}{x}}(-\alpha, 1-\alpha) - B_{\frac{1}{x}}(-\alpha, 1-\alpha)\right)$$

where $B_z(a,b)$ is the Incomplete Beta function, $B_z(a,b) \equiv \int_0^z t^{a-1}\,(1-t)^{b-1}\,\mathrm{d}t$

$$\left\{\frac{\int_K^\infty f^2(x,\alpha)\,\mathrm{d}x}{\int_K^\infty f(x,\alpha)\,\mathrm{d}x}\right\}\alpha = 1,2 =$$

$$\left\{\frac{2(K + \log(K-1))}{K}, \frac{2\left(\frac{K(K(K+3)-6)}{K-1} + 6\log(K-1)\right)}{K^2}\right\}$$

and, for $\alpha = 5$,

$$\frac{1}{2(K-1)^4 K^5}$$

$$K(K(K(K(K(K(K(K(4K+9)+24)+84)+504)-5250)+10920)$$
$$-8820)+2520)+2520(K-1)^4\log(K-1)$$

We know that the limit is 2 for all three cases, but it is important to observe the preasymptotics
As we can see in fig x, finite or nonfinite variance is of small importance for the effect in the tails.

Example 2

Case of the Gaussian. Since the Gaussian belongs to the family of the stable distribution (Chapter x), the convolution will produce a Gaussian of twice the variance. So taking a Gaussian, G $(0,1)$ for short (0 mean and unitary standard deviation), the densities of the convolution will be Gaussian $\left(0, \sqrt{2}\right)$, so the ratio

$$\frac{\int_K^\infty f^2(x)\,\mathrm{d}x}{\int_K^\infty f(x)\,\mathrm{d}x} = \frac{\operatorname{erfc}\left(\frac{K}{2}\right)}{\operatorname{erfc}\left(\frac{K}{\sqrt{2}}\right)}$$

will rapidly explode.

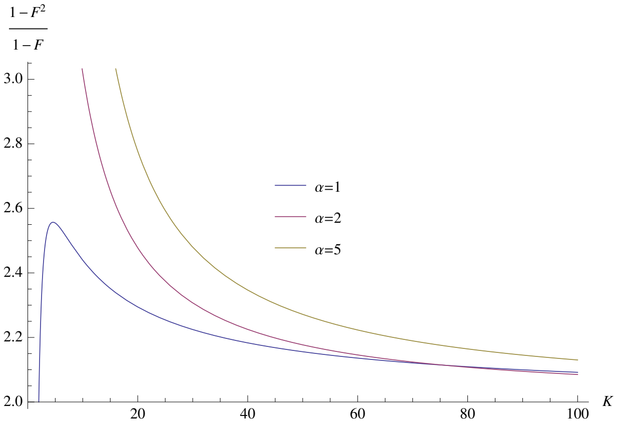


Figure 1.8: The ratio of the exceedance probabilities of a sum of two variables over a single one: power law

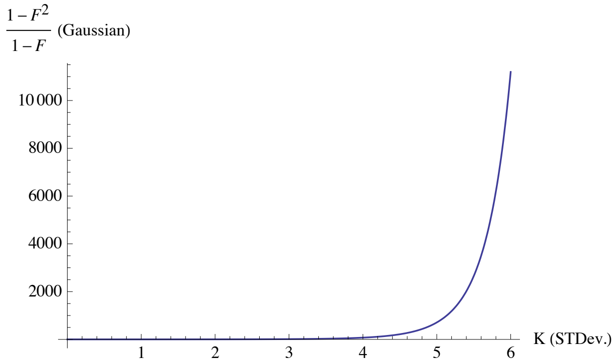


Figure 1.9: The ratio of the exceedance probabilities of a sum of two variables over a single one: Gaussian

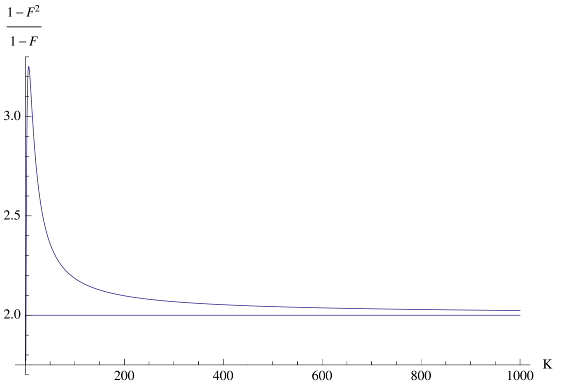


Figure 1.10: The ratio of the exceedance probabilities of a sum of two variables over a single one: Case of the Lognormal which in that respect behaves like a power law

Application: Two Real World Situations

We are randomly selecting two people, and the sum of their heights is 4.1 meters. What is the most likely combination? We are randomly selecting two people, and the sum of their assets, the total wealth is $30 million. What is the most likely breakdown?
Assume two variables X_1 and X_2 following an identical distribution, where f is the density function,

$$\begin{aligned} P\left[X_1+X_2=s\right] &= f^2(s) \\ &= \int f(y)\ f(s-y)\,\mathrm{d}y. \end{aligned}$$

The probability densities of joint events, with $0 \leq \beta < \frac{s}{2}$:

$$\bigcap \left(P\left(X_1=\frac{s}{2}+\beta\right), P\left(X_2=\frac{s}{2}-\beta\right)\right)$$
$$= P\left(X_1=\frac{s}{n}+\beta\right) \times P\left(X_2=\frac{s}{n}-\beta\right)$$

Let us work with the joint distribution for a given sum:

For a Gaussian, the product becomes

$$f\left(\frac{s}{n}+\beta\right) f\left(\frac{s}{n}-\beta\right) = \frac{e^{-\beta^2-\frac{s^2}{n^2}}}{2\pi}$$

For a Power law, say a Pareto distribution with α tail exponent, $f(x)$= $\alpha\ x^{-\alpha-1}x_{\min}^{\alpha}$where $x_{\min}$ is minimum value , $\frac{s}{2} \geq x_{\min}$, and $\beta \geq \frac{s}{2}-x_{\min}$

$$f\left(\beta+\frac{s}{n}\right) f\left(\beta-\frac{s}{n}\right) = \alpha^2 x_{\min}^{2\alpha}\left(\left(\beta-\frac{s}{2}\right)\left(\beta + \frac{s}{2}\right)\right)^{-\alpha-1}$$

Now the standard Lognormal belongs to the subexponential category, but just barely so (we used in the graph above Log Normal-2 as a designator for a distribution with the tail exceedance $\sim K e^{-\beta(\log(x)-\mu)^{\gamma}}$ where γ=2)

The product of two densities decreases with β for the Gaussian[2], and increases with the power law. For the Gaussian the maximal probability is obtained $\beta = 0$. For the power law, the larger the value of β, the better.

So the most likely combination is exactly 2.05 meters in the first example, and $x_{\min}$and $30 million $-x_{\min}$ in the second.

More General

More generally, distributions are called subexponential when the exceedance probability declines more slowly in the tails than the exponential.

a) $\lim_{x\to\infty} \frac{P_{X>\Sigma x}}{P_{X>x}} = n$, (Christyakov, 1964), which is equivalent to

b) $\lim_{x\to\infty} \frac{P_{X>\Sigma x}}{P(X>\max(x))} = 1$, (Embrecht and Goldie, 1980).

The sum is of the same order as the maximum (positive) value, another way of saying that the tails play a large role.

Clearly F has to have no exponential moment:

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

for all $\epsilon > 0$.
We can visualize the convergence of the integral at higher values of m: Figures **??** and **??** show the effect of $\mathbf{e}^{mx}\ f(x)$, that is, the product of the exponential moment m and the density of a continuous distributions $f(x)$ for large values of x.

[2] Technical comment: we illustrate some of the problems with continuous probability as follows. The sets 4.1 and $30\ 10^6$ have Lebesgue measures 0, so we work with densities and comparing densities implies Borel subsets of the space, that is, intervals (open or closed) ± a point. When we say "net worth is approximately 30 million", the lack of precision in the statement is offset by an equivalent one for the combinations of summands.

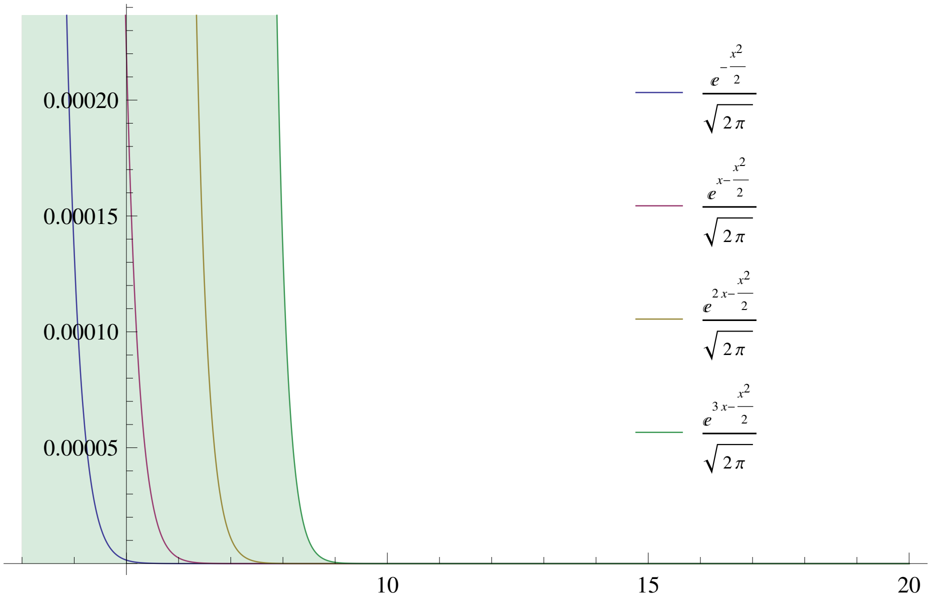


Figure 1.11: Multiplying the standard Gaussian density by e^{mx}, for $m = \{0, 1, 2, 3\}$.

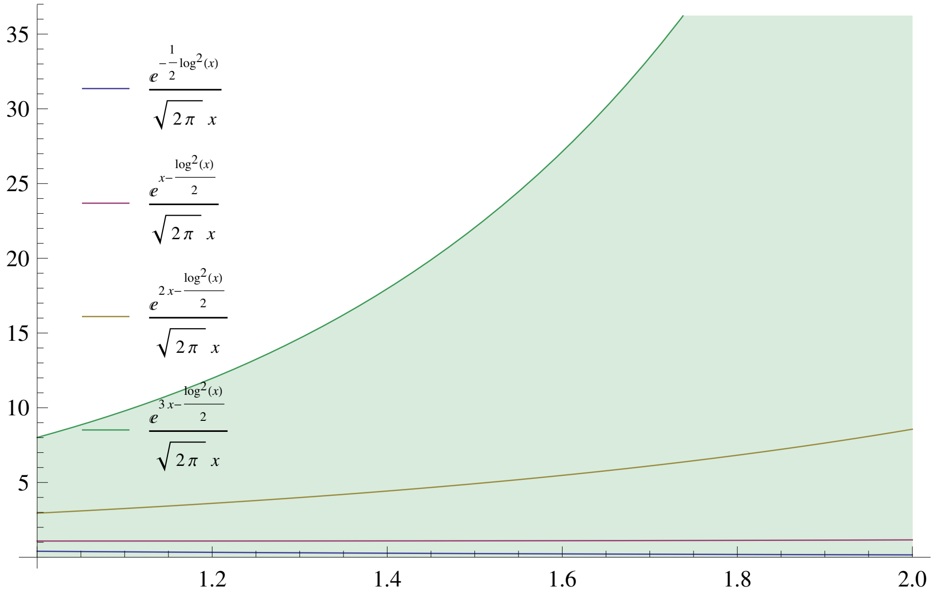


Figure 1.12: Multiplying the Lognormal (0,1) density by e^{mx}, for $m = \{0, 1, 2, 3\}$.

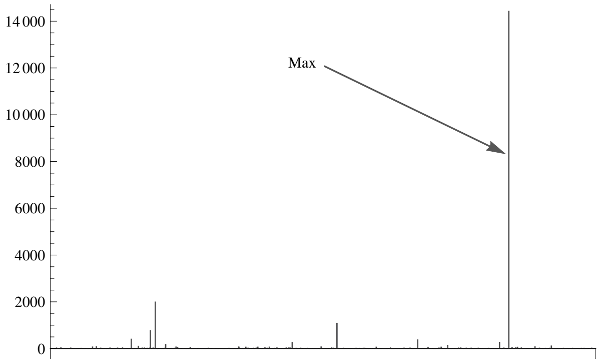


Figure 1.13: A time series of an extremely fat-tailed distribution. Given a long enough series, the contribution from the largest observation should represent the entire sum, dwarfing the rest.

1.9 Different Approaches For Statistically Derived Estimators

There are broadly two separate ways to go about estimators: nonparametric and parametric.

The nonparametric approach

it is based on observed raw frequencies derived from sample-size *n*. Roughly, it sets a subset of events *A* and $M_T^X(A,1)$ (i.e., *f(x) =1*), so we are dealing with the frequencies $\varphi(A) = \frac{1}{n}\sum_{i=0}^{n} 1_A$. Thus these estimates don't allow discussions on frequencies $\varphi < \frac{1}{n}$, at least not directly. Further the volatility of the estimator increases with lower frequencies. The error is a function of the frequency itself (or rather, the smaller of the frequency φ and *1-*φ). So if $\sum_{i=0}^{n} 1_A$=30 and $n = 1000$, only 3 out of 100 observations are expected to fall into the subset A, restricting the claims to too narrow a set of observations for us to be able to make a claim, even if the total sample $n = 1000$ is deemed satisfactory for other purposes. Some people introduce smoothing kernels between the various buckets corresponding to the various frequencies, but in essence the technique remains frequency-based. So if we nest subsets,$A_1 \subseteq A_2 \subseteq A$, the expected "volatility" (as we will see later in the chapter, we mean MAD, mean absolute deviation, not STD) of $M_T^X\left(A_z, f\right)$ will produce the following inequality:

$$\frac{E\left[\left|M_T^X\left(A_z, f\right) - M_{>T}^X\left(A_z, f\right)\right|\right]}{\left|M_T^X\left(A_z, f\right)\right|} \\ \leq \frac{E\left[\left|M_T^X\left(A_{<z}, f\right) - \left|M_{>T}^X\left(A_{<z}, f\right)\right|\right]}{\left|M_T^X\left(A_{<z}, f\right)\right|}$$

for all functions *f. (Proof via twinking of law of large numbers for sum of random variables).*

The parametric approach

it allows extrapolation but emprisons the representation into a specific off-the-shelf probability distribution (which can itself be composed of more sub-probability distributions); so M_T^X is an estimated parameter for use input into a distribution or model and the freedom left resides in differents values of the parameters.

Both methods make is difficult to deal with small frequencies. The nonparametric for obvious reasons of sample insufficiency in the tails, the parametric because small probabilities are very sensitive to parameter errors.

The Sampling Error for Convex Payoffs

This is the central problem of model error seen in consequences not in probability. The literature is used to discussing errors on probability which should not matter much for small probabilities. But it matters for payoffs, as *f* can depend on x. Let us see how the problem becomes very bad when we consider *f* and in the presence of fat tails. Simply, you are multiplying the error in probability by a large number, since fat tails imply that the probabilities *p(x)* do not decline fast enough for large values of x. Now the literature seem to have examined errors in probability, not errors in payoff.

Let $M_T^X\left(A_z, f\right)$ be the estimator of a function of x in the subset A_z= *(*δ_1 *,* δ_2 *)* of the support of the variable. Let $\xi(M_T^X\left(A_z, f\right))$ be the mean absolute error in the estimation of the probability in the small subset A_z= *(*δ_1 *,* δ_2 *)*, i.e.,

$$\xi\left(M_T^X\left(A_z, f\right)\right) \equiv \frac{\mathbb{E}\left|M_T^X\left(A_z, 1\right) - M_{>T}^X\left(A_z, 1\right)\right|}{M_T^X\left(A_z, 1\right)}$$

Assume *f(x)* is either linear or convex (but not concave) in the form *C+* Λ x^β, with both $\Lambda > 0$ and $\beta \geq 1$. Assume E[X], that is, $\mathbb{E}\left[M_{>T}^X\left(A_{\mathcal{D}}, f\right)\right] < \infty$, for $A_z \equiv A_{\mathcal{D}}$, a requirement that is not necessary for finite intervals.

Then the estimation error of $M_T^X\left(A_z, f\right)$ compounds the error in probability, thus giving us the lower bound in relation to ξ

$$\frac{\mathbb{E}\left[\left|M_T^X\left(A_z, f\right) - M_{>T}^X\left(A_z, f\right)\right|\right]}{M_T^X\left(A_z, f\right)} \\ \geq \left(\left|\delta_1 - \delta_2\right| \min\left(\left|\delta_2\right|, \left|\delta_1\right|\right)^{\beta-1} \\ + \min\left(\left|\delta_2\right|, \left|\delta_1\right|\right)^{\beta}\right) \frac{\mathbb{E}\left[\left|M_T^X\left(A_z, 1\right) - M_{>T}^X\left(A_z, 1\right)\right|\right]}{M_T^X\left(A_z, 1\right)}$$

Since $\frac{\mathbb{E}\left[M_{>T}^X(A_z,f)\right]}{\mathbb{E}\left[M_{>T}^X(A_z,1)\right]} = \frac{\int_{\delta_1}^{\delta_2} f(x)p(x)\,\mathrm{d}x}{\int_{\delta_1}^{\delta_2} p(x)\,\mathrm{d}x}$

and expanding $f(x)$, for a given n on both sides.

We can now generalize to the central inequality from convexity of payoff , which we shorten as *Convex Payoff Sampling Error Inequalities*, CPSEI:

Rule 1. *Under our conditions above, if for all* λ *∈(0,1) and* $f^{\{i,j\}}$*(x±Δ)* $\in A_z$, $\frac{(1-\lambda) f^i(x-\Delta)+\lambda f^i(x+\Delta)}{f^i(x)} \geq \frac{(1-\lambda) f^j(x-\Delta)+\lambda f^j(x+\Delta)}{f^j(x)}$*, (*$f^i$*is never less convex than* f^j*in interval* A_z *), then*

$$\xi\left(M_T^X(A_z, f^i)\right) \geq \xi\left(M_T^X(A_z, f^j)\right)$$

Rule 2. *Let* n_i *be the number of observations required for* $M_{>T}^X\left(A_{z_i}, f^i\right)$ *the estimator under* f^i *to get an equivalent expected mean absolute deviation as* $M_{>T}^X\left(A_{z_j}, f^j\right)$ *under* f^j *with observation size* n_j*, that is, for* ξ*(*$M_{T,n_i}^X\left(A_{z_i}, f^i\right.$*))=*$\xi$*(*$M_{T,n_j}^X\left(A_{z_j}, f^j\right.$*)), then*

$$n_i \geq n_j$$

This inequality turns into equality in the case of nonfinite first moment for the underlying distribution.

The proofs are obvious for distributions with finite second moment, using the speed of convergence of the sum of random variables expressed in mean deviations. We will not get to them until Chapter x on convergence and limit theorems but an example will follow in a few lines.

We will discuss the point further in Chapter x, in the presentation of the conflation problem.

For a sketch of the proof, just consider that the convex transformation of a probability distribution $p(x)$ produces a new distribution $f(x) \equiv \Lambda x^\beta$ with density $p_f(x) = \frac{\Lambda^{-1/\beta} x^{\frac{1-\beta}{\beta}} p\left(\left(\frac{x}{\Lambda}\right)^{1/\beta}\right)}{\beta}$ over its own adjusted domain, for which we find an increase in volatility, which requires a larger n to compensate, in order to maintain the same quality for the estimator.

Example

For a Gaussian distribution, the variance of the transformation becomes:

$$V\left(\Lambda x^\beta\right) = \frac{2^{\beta-2}\Lambda^2\sigma^{2\beta}}{\pi}\left(2\sqrt{\pi}\left((-1)^{2\beta}+1\right)\Gamma\left(\beta+\frac{1}{2}\right)\right.$$
$$\left.-\left((-1)^\beta+1\right)^2\Gamma\left(\frac{\beta+1}{2}\right)^2\right)$$

and to adjust the scale to be homogeneous degree 1, the variance of

$$V\left(x^\beta\right) = \frac{2^{\beta-2}\sigma^{2\beta}}{\pi}\left(2\sqrt{\pi}\left((-1)^{2\beta}+1\right)\Gamma\left(\beta+\frac{1}{2}\right)\right.$$
$$\left.-\left((-1)^\beta+1\right)^2\Gamma\left(\frac{\beta+1}{2}\right)^2\right)$$

For Λ=1, we get an idea of the increase in variance from convex transformations:

β	Variance $V(\beta)$	Kurtosis
1	σ^2	3
2	$2\ \sigma^4$	15
3	$15\ \sigma^6$	$\frac{231}{5}$
4	$96\ \sigma^8$	207
5	$945\ \sigma^{10}$	$\frac{46189}{63}$
6	$10170\ \sigma^{12}$	$\frac{38787711}{12769}$

Since the standard deviation drops at the rate $\sqrt{n}$ for non power laws, the number of $n(\beta)$, that is, the number of observations needed to incur the same error on the sample in standard deviation space will be $\frac{\sqrt{V(\beta)}}{\sqrt{n_1}} = \frac{\sqrt{V(1)}}{\sqrt{n}}$, hence $n_1 = 2$ n σ^2. But to equalize the errors in mean deviation space, since Kurtosis is higher than that of a Gaussian, we need to translate back into L^1 space, which is elementary in most cases.

For a Pareto Distribution with domain $v[x_{\min}^\beta, \infty)$,

$$V\left(\Lambda\ x^\beta\right) = \frac{\alpha\Lambda^2 x_{\min}^2}{(\alpha-2)(\alpha-1)^2}.$$

Using Log characteristic functions allows us to deal with the difference in sums and get the speed of convergence.[3]

Example illustrating the Convex Payoff Inequality

Let us compare the "true" theoretical value to random samples drawn from the Student T with 3 degrees of freedom, for $M_T^X\left(A, x^\beta\right)$, $A = (-\infty, -3]$, *n*=200, across *m* simulations $\left(> 10^5\right)$ by estimating $E\left|M_T^X\left(A, x^\beta\right) - M_{>T}^X\left(A, x^\beta\right)/M_T^X\left(A, x^\beta\right)\right|$ using

$$\xi = \frac{1}{m}\sum_{j=1}^{m}\left|\sum_{i=1}^{n}\frac{1_A\left(x_i^j\right)^\beta}{1_A}\right.$$
$$\left. - M_{>T}^X\left(A, x^\beta\right)/\sum_{i=1}^{n}\frac{1_A\left(x_i^j\right)^\beta}{1_A}\right| .$$

It produces the following table showing an explosive relative error ξ. We compare the effect to a Gausian with matching standard deviation, namely $\sqrt{3}$. The relative error becomes infinite as β approaches the tail exponent. We can see the difference between the Gaussian and the power law of finite second moment: both "sort of" resemble each others in many applications — but... not really.

β	$\xi_{\text{St}(3)}$	$\xi_{G(0,\sqrt{3})}$
1	0.17	0.05
$\frac{3}{2}$	0.32	0.08
2	0.62	0.11
$\frac{5}{2}$	1.62	0.13
3	$fuhgetaboudit$	0.18

Warning. Severe mistake (common in the economics literature)

One should never make a decision involving $M_T^X\left(A_{>z}, f\right)$ and basing it on calculations for $M_T^X\left(A_z, 1\right)$, especially when *f* is convex, as it violates CPSEI. Yet many papers make such a mistake. And as we saw under fat tails the problem is vastly more severe.

Utility Theory

Note that under a concave utility of negative states, decisions require a larger sample. By CPSEI the magnification of errors require larger number of observation. This is typically missed in the decision-science literature. But there is worse, as we see next.

Tail payoffs

: The author is disputing, in Taleb (2013), the results of a paper, Ilmanen (2013), on why tail probabilities are overvalued by the market: naively Ilmanen (2013) took the observed probabilities of large deviations,$f(x) = 1$ then made an inference for $f(x)$ an option payoff based on x, which can be extremely explosive (a error that can cause losses of several orders of magnitude the initial gain). Chapter x revisits the problem in the context of nonlinear transformations of random variables. The error on the estimator can be in the form of parameter mistake that inputs into the assumed probability distribution, say σ the standard deviation (Chapter x and discussion of metaprobability), or in the frequency estimation. Note now that if δ_1→-∞, we may have an infinite error on $M_T^X\left(A_z, f\right)$, the left-tail shortfall while, by definition, the error on probability is necessarily bounded.

If you assume in addition that the distribution *p(x)* is expected to have fat tails (of any of the kinds seen in **??**.**??** and **??**.**??**) , then the problem becomes more acute.

[3] The characteristic function of the transformation y= f(x) is

$$\frac{1}{32\sqrt{\pi}\,|t|^{7/4}\,\text{sgn}(t)}\;8\;\pi\;t\sqrt{|t|}\,{}_0\;\times\tilde{F}_1\left(;\frac{3}{4};-\frac{1}{4096\;|t|^2}\right)\left(\cos\left(\frac{\text{sgn}(t)+4\pi t}{32\;t}\right)+i\text{sgn}(t)\sin\left(\frac{\text{sgn}(t)+4\pi t}{32\;t}\right)\right)$$
$$-\pi t\,{}_0\;\tilde{F}_1\left(;\frac{5}{4};-\frac{1}{4096\,|t|^2}\right)\left(\cos\left(\frac{\text{sgn}(t)+12\pi t}{32\;t}\right)+i\;\text{sgn}(t)\;\sin\left(\frac{\text{sgn}(t)+12\pi t}{32\;t}\right)\right)-4e^{\frac{i}{32\;t}}\sqrt{it}\,|t|^{3/4}\,K_{\frac{1}{4}}\left(\frac{i}{32\;t}\right).$$

Now the mistake of estimating the properties of x, then making a decisions for a nonlinear function of it, $f(x)$, not realizing that the errors for $f(x)$ are different from those of x is extremely common. Naively, one needs a lot larger sample for $f(x)$ when $f(x)$ is convex than when $f(x) = x$. We will re-examine it along with the "conflation problem" in Chapter x.[4]

1.10 Economics Time Series Econometrics and Statistics Tend to imagine functions in L^2 Space

There is something Wrong With Econometrics, as Almost All Papers Don' t Replicate. Two reliability tests in Chapter x, 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. Like charlatans they keep spinning inconsistent *ad hoc* statements to explain failures.

We will show how, with economic variables one single observation in 10,000, that is, one single day in 40 years, can explain the bulk of the "kurtosis", a 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.

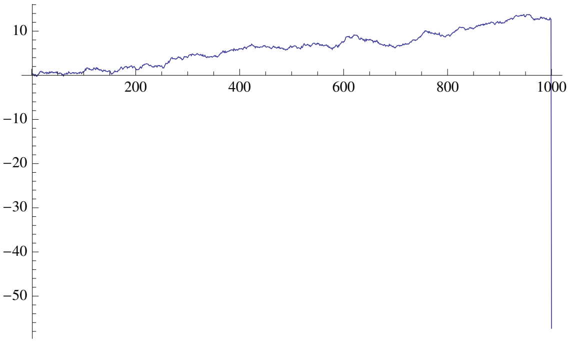


Figure 1.14: The Turkey Problem, where nothing in the past properties seems to indicate the possibility of the jump.

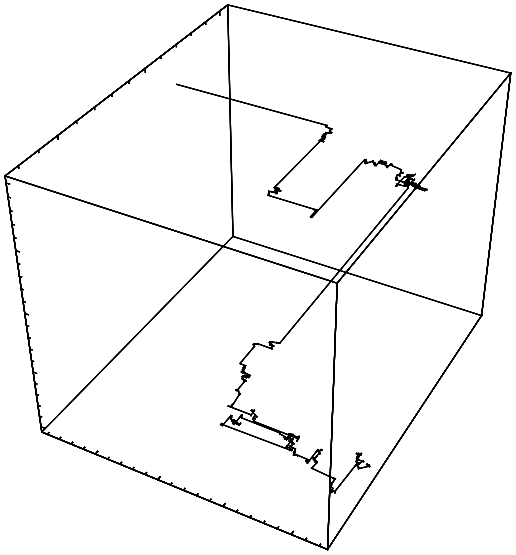


Figure 1.15: ***History moves by jumps****: A fat tailed historical process, in which events are distributed according to a power law that corresponds to the "80/20", with $\alpha \simeq 1.2$, the equivalent of a 3-D Brownian motion.*

[4] I thank Albert Wenger for corrections of mathematical typos in this section.

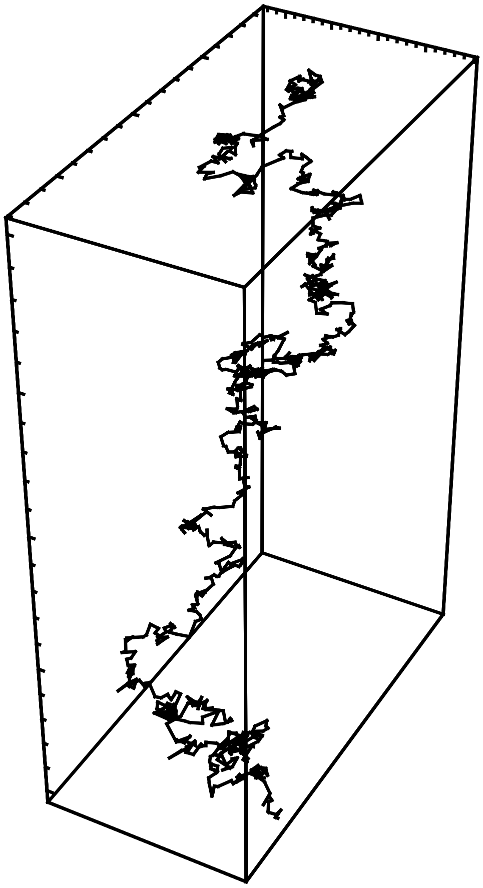


Figure 1.16: What the "fragilistas" have in mind: history as a thin-tailed process.

1.11 Typical Manifestations of The Turkey Surprise

Two critical (and lethal) mistakes, entailing mistaking inclusion in a class $\mathcal{D}_i$ for $\mathcal{D}_{<i}$ because of induced slowness in the convergence under the law of large numbers. We will see that in the hierarchy, scale (or variance) is swamped by tail deviations.

Great Moderation (Bernanke, 2006) consists in mistaking a two-tailed process with fat tails for a process with thin tails and low volatility.

Long Peace (Pinker, 2011) consists in mistaking a one-tailed process with fat tails for a process with thin tails and low volatility and low mean.

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

When I contacted social scientists I discovered that the familiarity with fat tails was pitifully small, highly inconsistent, and confused.

The Long Peace Mistake. Later, to my horror, I saw an amateurish book with an identical theory of great moderation produced by Steven Pinker with the same naive statistically derived discussions (>700 pages of them!). Except that it applied to security. The problem is that, unlike Bernanke, Pinker realized the process had fat tails, but did not realize the resulting errors in inference.

Chapter x will get into the details and what we can learn from it.

1.12 Metrics for Functions Outside L^2 Space

We can see from the data in Chapter 3 that the predictability of the Gaussian-style cumulants is low, the mean deviation of mean deviation is ~70% of the mean deviation of the standard deviation (in sample, but the effect is much worse in practice); working with squares is not a good estimator. Many have the illusion that we need variance: we don't, even in finance and economics (especially in finance and economics).

We propose different cumulants, that should exist whenever the mean exists. So we are not in the dark when we refuse standard deviation. It is just that these cumulants require more computer involvement and do not lend themselves easily to existing Platonic distributions.

And, unlike in the conventional Brownian Motion universe, they don't scale neatly.
Note finally that these measures are central since, to assess the quality of the estimation M_T^X, we are concerned with the expected mean error of the *empirical expectation,* here $E\left(\left|M_T^X\left(A_z,f\right)-M_{>T}^X\left(A_z,f\right)\right|\right)$, where z corresponds to the support of the distribution.

$$C_0 \equiv \frac{\sum_{i=1}^T x_i}{T}$$

(This is the simple case of $\mathbf{1}_A = \mathbf{1}_{\mathcal{D}}$; an alternative would be:

$C_0 \equiv \frac{1}{\sum_{i=1}^T \mathbf{1}_A} \sum_{i=1}^T x_i \mathbf{1}_A$ or $C_0 \equiv \frac{1}{\sum_{i=1}^T \mathcal{D}} \sum_{i=1}^T x_i \mathbf{1}_A$, depending on whether the function of concern for the fragility metric requires conditioning or not).

$$C_1 \equiv \frac{1}{T-1}\sum_{i=1}^T |x_i - C_0|$$

produces the Mean Deviation (but centered by the mean, the first moment).

$$C_2 \equiv \frac{1}{T-2}\sum_{i=1}^T ||x_i - Co| - C_1|$$

produces the mean deviation of the mean deviation. . . .

$$C_N \equiv \frac{1}{T-N-1}\sum_{i=1}^T |...|||x_i - Co| - C_1| - C_2|... - C_N|$$

Note the practical importance of C_1: under some conditions usually met, it measures the quality of the estimation $E\left[\left|M_T^X\left(A_z,f\right)-M_{>T}^X\left(A_z,f\right)\right|\right]$, since $M_{>T}^X\left(A_z,f\right)=C_0$. When discussing fragility, we will use a "tail cumulant", that is absolute deviations for 1_A covering a spccific tail.
The next table shows the theoretical first two cumulants for two symmetric distributions: a Gaussian, N (0,σ) and a symmetric Student T St$(0,s,\alpha)$ with mean 0, a scale parameter *s*, the PDF for x is

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

As to the PDF of the Pareto distribution, $p(x) = \alpha s^\alpha x^{-\alpha-1}$ for $x \geq s$ (and the mean will be necessarily positive).

Distr	**Mean**	$\mathbf{C_1}$	$\mathbf{C_2}$
Gaussian	0	$\sqrt{\frac{2}{\pi}}\sigma$	$2e^{-1/\pi}\sqrt{\frac{2}{\pi}}\left(1-e^{\frac{1}{\pi}}\text{erfc}\left(\frac{1}{\sqrt{\pi}}\right)\right)\sigma$
Pareto α	$\frac{\alpha s}{\alpha-1}$	$2(\alpha-1)^{\alpha-2}\alpha^{1-\alpha}s$	
ST α=3/2	0	$\frac{2\sqrt{\frac{6}{\pi}}s\Gamma\left(\frac{5}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}$	$\frac{8\sqrt{3}\Gamma\left(\frac{5}{4}\right)^2}{\pi^{3/2}}$
ST Square α=2	0	$\sqrt{2}s$	$s-\frac{s}{\sqrt{2}}$
ST Cubic α=3	0	$\frac{2\sqrt{3}s}{\pi}$	$\frac{8\sqrt{3}s\tan^{-1}\left(\frac{2}{\pi}\right)}{\pi^2}$

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

These cumulants will be useful in areas for which we do not have a good grasp of convergence of the sum of observations.

Chapter references

Borel, ÃĽmile Les probabilités dénombrables et leurs applications arithmétiques, Rendiconti del Circolo Matematico di Palermo, vol.27, no1, décembre 1909, p.247-271

Chistyakov., P. (1964) A theorem on sums of independent positive random variables. *Theory Prob Appl* 9, 640-648.

Doob, J. L. (1949) Heuristic approach to the Kolmogorov−−Smirnov theorems, *Ann. Math. Statistics*, 20 , pp. 393−−403.

Embrechts, P., & Goldie, C. M. (1980). *On closure and factorisation properties of subexponential and related distributions.* Katholieke Universiteit.

Embrechts, P., (1985) Subexponential distribution functions and their applications: a review. In Proc.7th Brasov Conf. Probability Theory,VNU Science Press, Utrecht,125-136.

Teugels, J. L. (1975). The class of subexponential distributions. *The Annals of Probability*, *3*(6), 1000-1011.

Pitman, E. J. G. (1980). Subexponential distribution functions. *J. Austral. Math. Soc. Ser.* A, 29(3), 337-347.

A Appendix: Special Cases of Fat Tails

A.1 Multimodality and Fat Tails, or the War and Peace Model

We noted in 1.x 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, and separate the potential outcomes wide enough, so that we get many modes, the kurtosis would drop. And if we associate different vairances 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.

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, 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, and more. Each state has its probability p_i.

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 ϕ(t) for the mixed distribution:

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

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

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

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

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

$$\begin{aligned} M_4 = {} & p_1 \left(6 m_1^2 \sigma_1^2 + m_1^4 + 3\sigma_1^4\right) \\ & + (1 - p_1) \left(6 m_2^2 \sigma_2^2 + m_2^4 + 3\sigma_2^4\right) \end{aligned}$$

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.

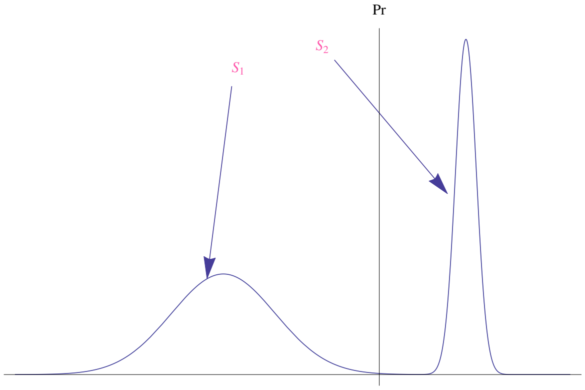


Figure A.1: The War and peace model. Kurtosis K=1.7, much lower than the Gaussian.

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(E(x)) the probabilitity densities of the expectation are smaller than at the different means of regimes, so $p(E(x)) < p\left(m_1\right)$ and $< p\left(m_2\right)$, but in the extreme case (bonds), p(E(x)) becomes increasingly small. The tail event is the realization around the mean.

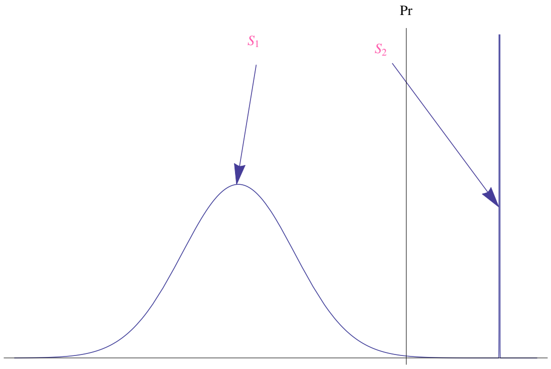


Figure A.2: The Bond payoff model. Absence of volatility, deterministic payoff in regime 2, mayhem in regime 1. Here the kurtosis K=2.5. Note that the coffee cup is a special case of both regimes 1 and 2 being degenerate.

In 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 will see in the next figure.

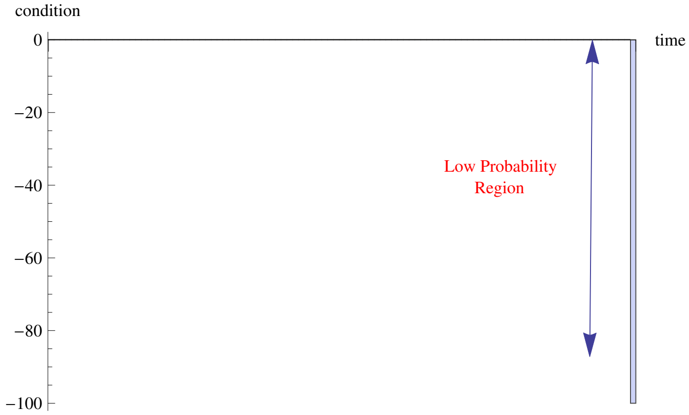


Figure A.3: The coffee cup cannot incur "small" harm; it is exposed to everything or nothing.

A.1.1 A brief list of other situations where bimodality is encountered:

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

A.2 Transition probabilites, or what can break will eventually 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, $P\left(S(t+\Delta \text{t})=s_j\right|S(t)=s_j))$

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

After n periods, that is, n steps,

$$P^n = \Bigg($$

$$\begin{array}{cc} \frac{(p_{1,1}-1)(p_{1,1}+p_{2,2}-1)^n+p_{2,2}-1}{p_{1,1}+p_{2,2}-2} & \frac{(1-p_{1,1})((p_{1,1}+p_{2,2}-1)^n-1)}{p_{1,1}+p_{2,2}-2} \\ \frac{(1-p_{2,2})((p_{1,1}+p_{2,2}-1)^n-1)}{p_{1,1}+p_{2,2}-2} & \frac{(p_{2,2}-1)(p_{1,1}+p_{2,2}-1)^n+p_{1,1}-1}{p_{1,1}+p_{2,2}-2} \end{array}$$

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 $\mathbf{1}^\mathsf{T}$ 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 $\pi = \begin{pmatrix} 1 & 1 \\ \frac{1-p_{1,1}}{1-p_{2,2}} & 1 \end{pmatrix}$

2 A Heuristic Hierarchy of Distributions For Inferential Asymmetries

This chapter explains in technical terms the "masquerade problem"discussed in *The Black Swan,* namely that one can be fooled by fat tails, not thin tails, as a fat tailed distribution can masquerade as a low-risk one, but not the reverse. Remarkably this point was missed, or never linked to the asymmetry between evidence of absence and absence of evidence.

Accordingly, one can make statements of the type "This is not Gaussian", or "this is not Poisson"(many people don't realize that Poisson are generally thin tails); but one cannot rule out Cauchy or other power laws. So this chapter puts some mathematical structure around the idea of which statements are permissible and which ones are not. (One can violate these statements but not from data analysis, only basing oneself on *a priori* statement of belonging to some probability distributions.)

Let us get deeper into the masquerade problem, as it concerns the problem of induction and Extremistan, and get to the next step. Simply, if a mechanism is fat tailed it can deliver large values; therefore the incidence of large deviations is possible, but *how* possible, *how often* these occur should occur, will be hard to know with any precision *beforehand.* This is similar to the water puddle problem: plenty of ice cubes could have generated it. As someone who goes from reality to possible explanatory models, I face a completely different spate of problems from those who do the opposite.

We said that fat tailed series can, in short episodes, masquerade as thin-tailed. At the worst, we don't know how long it would take to know. But we can have a pretty clear idea whether organically, because of the nature of the payoff, the "Black Swan"can hit on the left (losses) or on the right (profits). This point can be used in climatic analysis. Things that have worked for a long time are preferable—they are more likely to have reached their ergodic states.

We aim here at building a rigorous methodology for attaining statistical (and more general) knowledge by rejection, and cataloguing rejections, not addition. We can reject some class of statements concerning the fat-tailedness of the payoff, not others.

2.1 Masquerade Example

We construct the cases as switching between Gaussians with variances

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

with p $\in$ [0,1); a, b $\in$ (-1,1) and (to conserve the variance) $b = -a\frac{p}{1-p}$, which produces a Kurtosis $= \frac{3\left(\left(1-a^2\right)p-1\right)}{p-1}$ thus allowing polarized states and high kurtosis, with a condition that for a $>$ ($<$) 0, a $<$ ($>$)$\frac{1-p}{p}$.

Let us compare the two cases:

A) A switching process producing Kurtosis$= 10^7$ (using $p= 1/2000$, a sligtly below the upper bound a$=\frac{1-p}{p}-1$) to

B) The regular situation $p = 0$, a=1, the case of kurtosis $= 3$.

The two graphs in figures **??.??** and **??.??** show the realization of the processes A (to repeat, produced with the switching process) to a process B entirely Gaussian, both of the same variance.

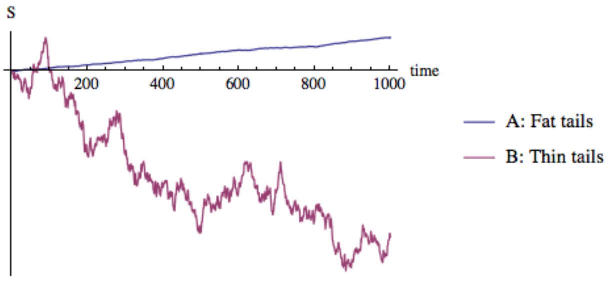


Figure 2.1: N=1000. Sample simulation. Both series have the exact same means and variances at the level of the generating process. Naive use of common metrics leads to the acceptance that the process A has thin tails.

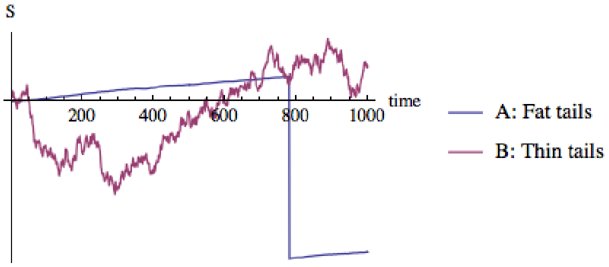


*Figure 2.2: N=1000. **Rejection:** Another simulation. there is 1/2 chance of seeing the real properties of A. We can now reject the hypothesis that the smoother process has thin tails.*

2.2 The Probabilistic Version of Absense of Evidence vs Evidence of Absence

Our concern is exposing some errors in probabilistic statements and statistical inference, in making inferences symmetric, when they are more likely to be false on one side than the other, or more harmful one side than another. Believe it or it, this pervades the entire literature.

Some people have the illusion that "because Kolmogorov-Smirnoff is nonparametric", it is therefore immune to the nature specific distribution under the test (perhaps from an accurate sentence in Feller (1971), vol 2 as we will see further down). The belief in Kolmogorov-Smirnoff is also built in the illusion that our concern is probability rather than expected payoff, or the associated problem of "confusing a binary for a vanilla", where by attribute substitution, one tests a certain variable in place of another, simpler one.

In other words, it is a severe mistake to treat epistemological inequalities as equalities. No matter what we do, we end up going back to the problem of induction, except that the world still exists and people unburdened with too many theories are still around. By making one-sided statements, or decisions, we have been immune to the muddle in statistical inference.

Remark on via negativa and the problem of induction

Test statistics are effective (and robust) at rejecting, but not at accepting, as a single large deviation allowed the rejection with extremely satisfactory margins (a near-infinitesimal P-Value). This illustrates the central epistemological difference between absence of evidence and evidence of absence.

2.3 Via Negativa and One-Sided Arbitrage of Statistical Methods

Via negativa

: In theology and philosophy, corresponds to the focus on what something is not, an indirect definition. In action, it is a recipe for what to avoid, what not to do− subtraction, not addition, say, in medicine. In epistemology: what to *not* accept, or accept as false. So a certain body of knowledge actually grows by rejection. (*Antifragile*, Glossary).

The proof and the derivations are based on climbing to a higher level of abstraction by focusing the discussion on a hierarchy of distributions based on fat-tailedness.

Remark: Test statistics can be arbitraged, or "fooled"in one direction, not the other.

Let us build a hierarchy of distributions based on tail events. But, first, a discussion of the link to the problem of induction.

From *The Black Swan* (Chapter 16 *)*: This author has learned a few tricks from experience dealing with power laws: whichever exponent one try to measure will be likely to be overestimated (recall that a lower exponent implies a smaller role for large deviations)−what you see is likely to be less Black Swannish than what you do not see. Let's say I generate a process that has an exponent of 1.7. You do not see what is inside the engine, only the data coming out. If I ask you what the exponent is, odds are that you will compute something like 2.4. You would do so even if you had a million data points. The reason is that it takes a long time for some fat tailed processes to reveal their properties, and you underestimate the severity of the shock. Sometimes a fat tailed distribution can make you believe that it is Gaussian, particularly when the process has mixtures. (Page 267, slightly edited).

2.4 A Heuristic Hierarchy of Distributions in Term of Fat-Tailedness

Let $\mathcal{D}_i$be a class of probability measures, $\mathcal{D}_i \subset \mathcal{D}_{>i}$ means in our terminology that a random event "in"$\mathcal{D}_i$ would necessarily "be in"$\mathcal{D}_j$, with $j > i$, and we can express it as follows. Let A_K be a one-tailed interval in $\mathbb{R}$, unbounded on one side K, s.a. $A_K^- = (-\infty, K\big]$ or $A_K^+ = \big[K, \infty)$, and $\mu(A)$ the probability measure on the interval, which corresponds to $\mu_i(A_K^-)$ the cumulative distribution function for K on the left, and $\mu_i(A_K^+) = 1 -$ the CDF (that is, the exceedance probability) on the right.

For continuous distributions, we can treat of the Radon-Nikodym derivatives for two measures $\frac{\partial \mu_i}{\partial \mu_j}$ over as the ratio of two probability with respect to a variable in A_K .

Definition 7. *We can define i) "acceptance"as being subject to a strictly positive probability of mistaking $\mathcal{D}_i$ for $\mathcal{D}_{<i}$ and ii) rejection as a claim that $\mathcal{D}_{>i}$. Likewise for what is called "confirmation"and "disconfirmation". Hence $\mathcal{D}_i \subset \mathcal{D}_j$ if either of these two conditions are satisfied:*

i) There exists a K_0 (called "in the negative tail") such that $\mu_j(A_{K_0}^-) > \mu_i(A_{K_0}^-)$ and $\mu_j(A_K^-) > \mu_i(A_K^-)$ for all K $< K_0$,

or

ii) There exists a K_0 ("in the positive tail") such that $\mu_j(A_{K_0}^+) > \mu_i(A_{K_0}^+)$ and $\mu_j(A_K^+) > \mu_i(A_K^+)$ for all K $>$ K_0

The derivations are as follows. Simply, the effect of the scale of the distribution (say, the variance in the finite second moment case) wanes in the tails. For the classes of distributions up to the Gaussian, the point is a no brainer because of compact support with 0 measure beyond a certain K. As as far as the Gaussian, there are two brands, one reached as a limit of, say, a sum of n Bernouilli variables, so the distribution will have compact support up to a multiple of n at infinity, that is, in finite processes (what we call the "real world"where things are finite). The second Gaussian category results from an approximation; it does not have compact support but because of the exponential decline in the tails, it will be dominated by power laws. To cite Adrien Douady, it has compact support for all practical purposes.

Case of Two Powerlaws

For powerlaws, let us consider the competing effects of scale, say σ (even in case of nonfinite variance), and α tail exponent, with $\alpha > 1$. Let the density be

$$P_{\alpha,\sigma}(x) = L(x)x^{-\alpha-1}$$

where $L(x)$ is a slowly varying function,

$$r_{\lambda,k}(x) \equiv \frac{P_{\lambda\alpha,k\ \sigma}(x)}{P_{\alpha,\sigma}(x)}$$

By only perturbating the scale, we increase the tail by a certain factor, since $\lim_{x\to\infty} r_{1,k}(x) = k^{\alpha}$, which can be significant. But by perturbating both and looking at the limit we get $\lim_{x\to\infty} r_{\lambda,k}(x) = \lambda\ k^{\alpha\lambda} \left(\frac{L}{x}\right)^{\alpha(-1+\lambda)}$, where L is now a constant, thus making the changes to α the tail exponent leading for large values of x.

Obviously, by symmetry, the same effect obtains in the left tail.

Rule 3. *When comparing two power laws, regardless of parametrization of the scale parameters for either distributions, the one with the lowest tail exponent will have higher density in the tails.*

Comparing Gaussian to Lognormal

Let us compare the Gaussian(μ, σ) to a Lognormal(m, s), in the right tail, and look at

how one dominates in the remote tails. There is no values of parameters σ and s such that the PDF of the Normal exceeds that of the Lognormal in the tails. Assume means of 0 for the Gaussian and the equivalent $e^{\frac{k^2 s^2}{2}}$ for the Lognormal with no loss of generality.
Simply, let us consider the the sign of the difference between the two densities, $\frac{\frac{e^{-\frac{\log^2(x)}{2k^2 s^2}}}{ksx} - \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma}}{\sqrt{2\pi}}$ by comparing the unscaled tail values of $\frac{e^{-\frac{\log^2(x)}{2k^2 s^2}}}{ksx}$ and $\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma}$. Taking logarithms of the ratio, $\delta(x) = \frac{x^2}{2\sigma^2} - \frac{\log^2(x)}{2k^2 s^2} - \log(ksx) + \log(\sigma)$, which is dominated by the first term x^2 as it is convex when the other terms are concave, so it will be > 0 for large values of x independently of parameters.

Rule 4. *Regardless of parametrization of the scale parameter (standard deviation) for either distribution, a lognormal will produce asymptotically higher tail densities in the positive domain than the Gaussian.*

Case of Mixture of Gaussians

Let us return to the example of the mixture distribution $N(0, \sigma)$ with probability $1 - p$ and $N(0, k\ \sigma)$ with the remaining probability p. The density of the second regime weighted by p becomes $p\frac{e^{-\frac{x^2}{2k^2\ \sigma^2}}}{k\sqrt{2\pi}\sigma}$. For large deviations of x, $\frac{p}{k}e^{-\frac{x^2}{2k^2}}$ is entirely dominated by k, so regardless of the probability $p > 0$, $k > 1$ sets the terms of the density.
In other words:
Rule 5. *Regardless of the mixture probabilities, when combining two Gaussians, the one with the higher standard deviations determines the density in the tails.*
Which brings us to the following epistemological classification:

	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 $\alpha < 2$, $\int e^{-tx} dF(x) = \infty$
$\mathcal{D}_8$	Infinite First Moment	Fuhgetaboutdit

Mixtures distributions entailing $\mathcal{D}_i$ and $\mathcal{D}_j$ are classified with the highest level of fat tails $\mathcal{D}_{\max(i,j)}$ regardless of the mixing. A mixture of Gaussians remains Gaussian for large deviations, even if the local properties can be confusing in small samples, except for the situation of infinite nesting of stochastic volatili-

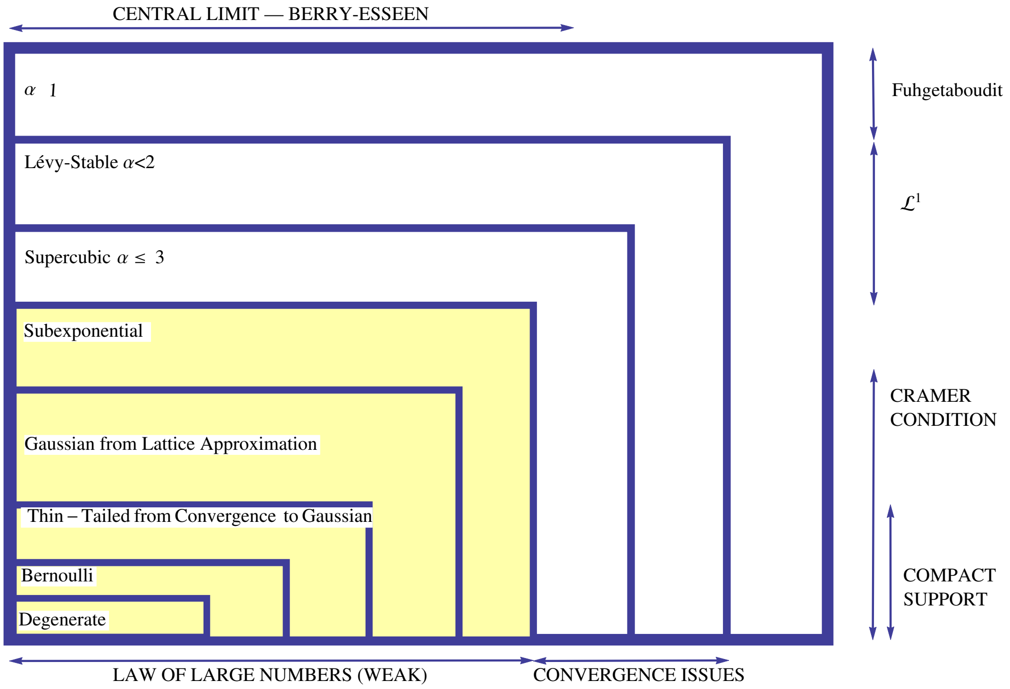

CENTRAL LIMIT — BERRY-ESSEEN
α 1
Lévy-Stable α<2
Supercubic $\alpha \leq 3$
Subexponential
Gaussian from Lattice Approximation
Thin – Tailed from Convergence to Gaussian
Bernoulli
Degenerate
Fuhgetaboudit
$\mathcal{L}^1$
CRAMER
CONDITION
COMPACT
SUPPORT
LAW OF LARGE NUMBERS (WEAK)
CONVERGENCE ISSUES

Figure 2.3: The tableau of Fat tails, along the various classifications for convergence purposes (i.e., convergence to the law of large numbers, etc.)

ties discussed in Chapter 6.

Now a few rapidly stated rules.

Rule 6. ***(General Decision Making Heuristic).*** *Rejection or acceptance of fitness to pre-specified probability distributions, based on suprema of distance between supposed probability distributions (say Kolmogorov Smirnoff and similar style) should only be able to "accept" the fatter tail one and "reject"the lower tail, i.e., based on the criterion $i > j$ based on the classification above.*

The point is discussed in **??.??** as we will start with an example "busting"a distribution concealing its properties.

Warning 1 : Always remember that one does not observe probability distributions, only realizations. (This even applies to the Extreme Value Laboratory of the Zurich ETH). Every probabilistic statement needs to be discounted by the probability of the parameter being away from the true one.

Warning 2 : Always remember that we do not live in probability space, but payoff space.

Rule 7. ***(Decision Mistakes)***. *Fatter tailed distributions are more likely to produce a lower in-sample variance (using empirical estimators) than a distribution of thinner tail of the same variance (in the finite variance case).*

For the derivation, recall that in **??.??** there in increase in observations in the "tunnel"(a_2, a_3) in response to increase in fat-tailedness.

How To Arbitrage Kolmogorov-Smirnov

Counterintuitively, when one raises the kurtosis, as in Figure **??**.**??** the time series looks "quieter". Simply, the storms are rare but deep. This leads to mistaken illusion of low volatility when in fact it is just high kurtosis, something that fooled people big-time with the story of the "great moderation"as risks were accumulating and nobody was realizing that fragility was increasing, like dynamite accumulating under the structure.

Kolmogorov - Smirnov, Shkmolgorov-Smirnoff

Remarkably, the fat tailed series passes general test of normality with better marks than the thin-tailed one, since it displays a lower variance. The problem discussed with with Avital Pilpel (Taleb and Pilpel, 2001, 2004, 2007) is that Kolmogorov-Smirnov and similar tests of normality are inherently self-referential.

These probability distributions are not directly observable, which makes any risk calculation suspicious since it hinges on knowledge about these distributions. Do we have enough data? If the distribution is, say, the traditional bell-shaped Gaussian, then yes, we may say that we have sufficient data. But if the distribution is not from such well-bred family, then we do not have enough data. But how do we know which distribution we have on our hands? Well, from the data itself .

If one needs a probability distribution to gauge knowledge about the future behavior of the distribution from its past results, and if, at the same time, one needs the past to derive a probability distribution in the first place, then we are facing a severe regress loop——a problem of self reference akin to that of Epimenides the Cretan saying whether the Cretans are liars or not liars. And this self-reference problem is only the beginning.

(Taleb and Pilpel, 2001, 2004)

Also,

From the Glossary in The Black Swan . *Statistical regress argument (or the problem of the circularity of statistics): We need data to discover a probability distribution. How do we know if we have enough? From the probability distribution. If it is a Gaussian, then a few points of data will suffice. How do we know it is a Gaussian? From the data. So we need the data to tell us what probability distribution to assume, and we need a probability distribution to tell us how much data we need. This causes a severe regress argument, which is somewhat shamelessly circumvented by resorting to the Gaussian and its kin.*

A comment on the Kolmogorov Statistic

It is key that the Kolmogorov-Smirnov test doesn't affect payoffs and higher moments, as it only focuses on probabilities. It is a severe problem because the approximation will not take large deviations into account, and doesn't make it useable for our purpose. But that's not the only problem. It is, as we mentioned, conditioned on sample size while claiming to be nonparametric.

Let us see how it works. Take the historical series and find the maximum point of divergence with *F(.)* the cumulative of the proposed distribution to test against:

$$D = \sup \left(\left(\left| \frac{1}{j} \sum_{i=1}^{J} X_{t_0+i\Delta \mathrm{t}} - F\left(X_{t_0+j\Delta \mathrm{t}}\right) \right| \right)_{j=1}^{n} \right)$$

where $n = \frac{T-t_0}{\Delta \mathrm{t}}$

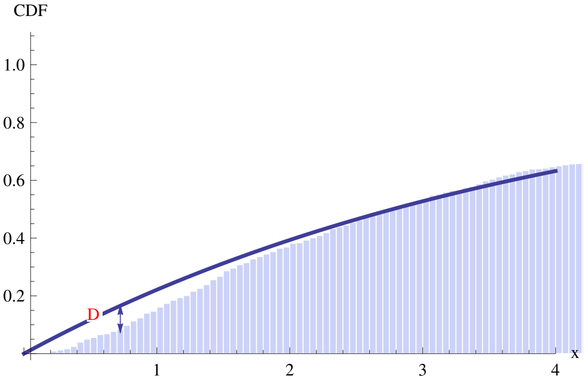


Figure 0.4 **The Kolmorov-Smirnov Gap.** D is the measure of the largest absolute divergence between the candidate and the target distribution.

We will get more technical in the discussion of convergence, take for now that the Kolmogorov statistic, that is, the distribution of *D*, is expressive of convergence, and should collapse with *n*.

The idea is that, by a Brownian Bridge argument (that is a process pinned on both sides, with intermediate steps subjected to double conditioning), $D_j = \left|\left(\frac{\sum_{i=1}^{J} X_{\Delta ti+t_0}}{j} - F\left(X_{\Delta tj+t_0}\right)\right)\right|$which is Uniformly distributed.

The probability of exceeding $D, P_{>D} = H\left(\sqrt{n}D\right)$, where H is the cumulative distribution function of the Kolmogorov-Smirnov distribution,

$$H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$$

We can see that the main idea reposes on a decay of $\sqrt{n}D$ with large values of n. So we can easily fool the testing by proposing distributions with a small probability of very large jump, where the probability of switch $\lesssim \frac{1}{n}$.

The mistake in misinterpreting Feller: the distribution of Dwill be uniform independently of the distribution under scrutiny, or the two distributions to be compared. But it does not mean that the test is immune to sample size$n$, that is, the possibility of jump with a probability an inverse function of n.

Table of the "fake"Gaussian when not busted

Let us run a more involved battery of statistical tests (but consider that it is a single run, one historical simulation).

Comparing the Fake and genuine Gaussians (Figure **??**.**??**) and subjecting them to a battery of tests. Note that some tests, such as the Jarque-Bera test, are more relevant to fat tailes as they include the payoffs:

		Statistic	P-Value
Fake	Anderson-Darling	0.406988	0.354835
	Cramér-von Mises	0.0624829	0.357839
	Jarque-Bera ALM	1.46412	0.472029
	Kolmogorov-Smirnov	0.0242912	0.167368
	Kuiper	0.0424013	0.110324
	Mardia Combined	1.46412	0.472029
	Mardia Kurtosis	−0.876786	0.380603
	Mardia Skewness	0.7466	0.387555
	Pearson χ^2	43.4276	0.041549
	Shapiro-Wilk	0.998193	0.372054
	Watson U^2	0.0607437	0.326458

		Statistic	P-Value
Genuine	Anderson-Darling	0.656362	0.0854403
	Cramér-von Mises	0.0931212	0.138087
	Jarque-Bera ALM	3.90387	0.136656
	Kolmogorov-Smirnov	0.023499	0.204809
	Kuiper	0.0410144	0.144466
	Mardia Combined	3.90387	0.136656
	Mardia Kurtosis	−1.83609	0.066344
	Mardia Skewness	0.620678	0.430795
	Pearson χ^2	33.7093	0.250061
	Shapiro-Wilk	0.997386	0.107481
	Watson U^2	0.0914161	0.116241

Table of the "fake" Gaussian when busted

And of course the fake Gaussian when caught. But recall that we have a small chance of observing the true distribution.

		Statistic	P-Value
Busted Fake (	Anderson-Darling	376.05	0.
	Cramér-von Mises	80.734	0.
	Jarque-Bera ALM	4.21×10^7	0.
	Kolmogorov-Smirnov	0.494547	0.
	Kuiper	0.967	0.
	Mardia Combined	4.21×10^7	0.
	Mardia Kurtosis	6430.	1.5×10^{-8979680}
	Mardia Skewness	166432.	1.07×10^{-36143}
	Pearson χ^2	30585.7	3.28×10^{-6596}
	Shapiro-Wilk	0.014	1.91×10^{-57}
	Watson U^2	80.58	0.

Use of the supremum of divergence

Note another manifestation of the error of ignoring the effect of the largest deviation. As we saw with Kolmogorov-Smirnoff and other rigorous methods in judging a probability distribution, one focuses on the maximum divergence, the supremum, as information. Another unused today but very potent technique, initially by Paul Levy (1924), called the concentration

function, also reposes on the use of a maximal distance:
From Petrov (1995):

$$Q_\lambda(X) \equiv \underset{x}{sup} P(x \leq X \leq x + \lambda)$$

for every $\lambda \geq 0$.
We will make use of it in discussion of the behavior of the sum of random variables and the law of large numbers.

Concluding Remarks: Mistaking Evidence for Anecdotes & The Reverse

Now some sad, very sad comments.

I emitted the following argument in a comment looking for maximal divergence: "Had a book proclaiming *The Long Peace* (on how violence has dropped) been published in $1913\frac{3}{4}$ it would carry similar arguments to those in Pinker's book", meaning that inability of an estimator period T to explain period $> t$, using the idea of maximum divergence. The author of the book complained that I was using "hindsight"to find the largest deviation, implying lack of rigor. This is a standard error in social science: data mining everywhere and not understanding the difference between meaningful disconfirmatory observation and anecdote.
We will revisit the problem upon discussing the "$N = 1$" fallacy (that is, the fallacy of thinking that $N = 1$ is systematically insufficient sample). Some social "scientists"(Herb Gintis, a representative mean imbecile) wrote about my approach to this problem, stating among other equally ignorant comments, something to the effect that "the plural of anecdotes is not data"(I have to deal with many mean social scientists). This elementary violation of the logic of inference from data is very common with social scientists as we will see in Chapter 3, as their life is based on mechanistic and primitive approaches to probability that miss the asymmetry. Yet, and here is the very, very sad part: *social science is the main consumer of statistical methods.*

3 An Introduction to Higher Orders of Uncertainty

3.1 Metaprobability

The Spectrum Between Uncertainty and Risk

There has been a bit of discussions about the distinction between "uncertainty" and "risk". We put the concepts on a spectrum, with one end of the spectrum "Knightian risk" not available for us mortals in the real world.

> When one assumes knowledge of a probability distribution, but has uncertainty attending the parameters, or when one has no knowledge of which probability distribution to consider, the situation is called "risk" in the Knightian sense (Knight, 1923). Such an animal does not exist in the real world. We find it preferable to talk about degrees of risk and degrees of uncertainty.

The Effect of Estimation Error, General Case

The idea of model error from missed uncertainty attending the parameters (another layer of randomness) is as follows.

Most estimations in economics (and elsewhere) take, as input, an average or expected parameter,

$$\bar{\alpha} = \int \alpha \ \phi(\alpha) \ d\alpha, \tag{3.1}$$

where α is ϕ distributed (deemed to be so a priori or from past samples), and regardles of the dispersion of α, build a probability distribution for x that relies on the mean estimated parameter, $p(X=x)= p\left(x \left| \bar{\alpha}\right.\right)$, rather than the more appropriate metaprobability adjusted probability for the density:

$$p(x) = \int \phi(\alpha) \, \mathrm{d}\alpha \tag{3.2}$$

In other words, if one is not certain about a parameter α, there is an inescapable layer of stochasticity; such stochasticity raises the expected (metaprobability-adjusted) probability if it is $< \frac{1}{2}$ and lowers it otherwise. The uncertainty is fundamentally epistemic, includes incertitude, in the sense of lack of certainty about the parameter.

The model bias becomes an equivalent of the Jensen gap (the difference between the two sides of Jensen's inequality), typically positive since probability is convex away from the center of the distribution. We get the bias ω_A from the differences in the steps in integration

$$\omega_A = \int \phi(\alpha) p(x|\alpha) \, \mathrm{d}\alpha - p(x| \int \alpha \phi(\alpha) \, \mathrm{d}\alpha)$$

With $f(x)$ a function , $f(x) = x$ for the mean, etc., we get the higher order bias $\omega_{A'}$

$$\begin{aligned} \omega_{A'} = & \int \left(\int \phi(\alpha) \ f(x) \ p(x|\alpha) \ \ d\alpha \right) \mathrm{d}x \\ & - \int f(x) \ p\left(x| \int \alpha \ \phi(\alpha) \, \mathrm{d}\alpha \right) \mathrm{d}x \end{aligned} \tag{3.3}$$

Now assume the distribution of α as discrete n states, with $\alpha = \{\alpha_i\}_{i=1}^n$ each with associated probability $\phi = \{\phi_i\}_{i=1}^n \sum_{i=1}^n \phi_i = 1$. Then 3.2 becomes

$$px = \phi_i \left(\sum_{i=1}^n p\left(x \left| \alpha_i \right.\right) \right) \tag{3.4}$$

So far this holds for α any parameter of any distribution.

3.2 The Effect of Metaprobability on the Calibration of Power Laws

In the presence of a layer of metaprobabilities (from uncertainty about the parameters), the asymptotic tail exponent for a powerlaw corresponds to the lowest *possible* tail exponent *regardless* of its probability. The problem explains "Black Swan" effects, i.e., why measurements tend to chronically underestimate tail contributions, rather than merely deliver imprecise but unbiased estimates.
When the perturbation affects the standard deviation of a Gaussian or similar nonpowerlaw tailed distribution, the end product is the weighted average of the probabilities. However, a powerlaw distribution with errors about the possible tail exponent will bear the asymptotic properties of the *lowest* exponent, not the average exponent.
Now assume p(X=x) a standard Pareto Distribution with α the tail exponent being estimated, $p(x|\alpha) = \alpha x^{-\alpha-1} x_{\min}^{\alpha}$, where $x_{\min}$is the lower bound for x,

$$p(x) = \sum_{i=1}^{n} \alpha_i x^{-\alpha_i - 1} x_{\min}^{\alpha_i} \phi_i$$

Taking it to the limit

$$\underset{x\to\infty}{\text{limit}}\; x^{\alpha^*+1} \sum_{i=1}^{n} \alpha_i x^{-\alpha_i - 1} x_{\min}^{\alpha_i} \phi_i = K$$

where K is a strictly positive constant and $\alpha^* = \underset{1\leq i\leq n}{min}\,\alpha_i$. In other words $\sum_{i=1}^{n} \alpha_i x^{-\alpha_i-1} x_{\min}^{\alpha_i} \phi_i$ is asymptotically equivalent to a constant times x^{α^*+1}. The lowest parameter in the space of all possibilities becomes the dominant parameter for the tail exponent.

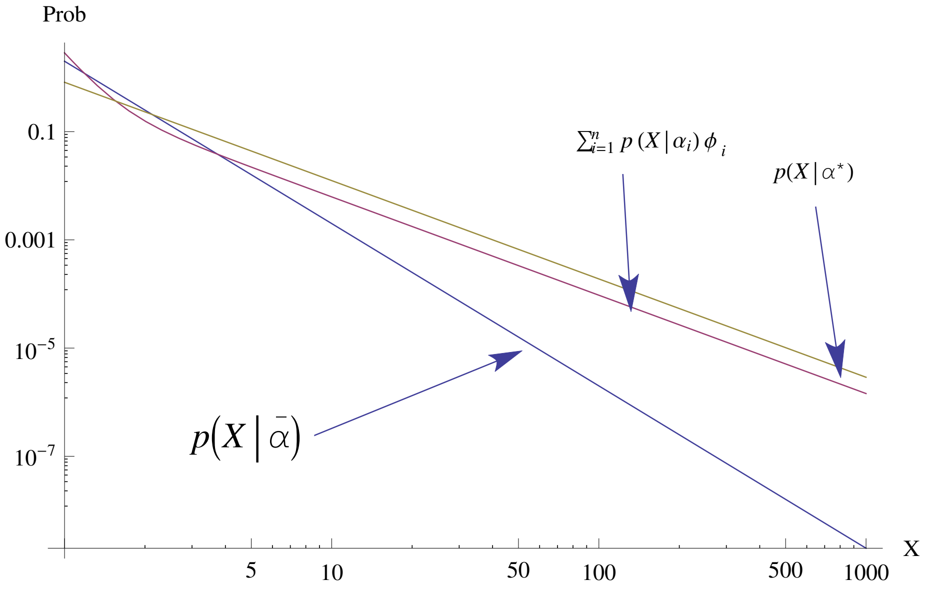


Figure 3.1: Log-log plot illustration of the asymptotic tail exponent with two states. The graphs shows the different situations, a) $p\left(x\,\middle|\,\bar{\alpha}\right)$ *b)* $\sum_{i=1}^{n} p\left(x\,|\alpha_i\right)\phi_i$ *and c)* $p\left(x\,|\alpha^*\right)$. *We can see how b) and c) converge*

The asymptotic Jensen Gap ω_A becomes $p\left(x\,|\alpha^*\right) - p(x|\bar{\alpha})$

Implications

Whenever we estimate the tail exponent from samples, we are likely to underestimate the thickness of the tails, an observation made about Monte Carlo

generated α-stable variates and the estimated results (the "Weron effect").
The higher the estimation variance, the lower the true exponent.
The asymptotic exponent is the lowest possible one. It does not even require estimation.
Metaprobabilistically, if one isn't sure about the probability distribution, and there is a probability that the variable is unbounded and "could be" powerlaw distributed, then it is powerlaw distributed, and of the lowest exponent.
The obvious conclusion is to in the presence of powerlaw tails, focus on changing payoffs to clip tail exposures to limit $\omega_{A'}$ and "robustify" tail exposures, making the computation problem go away.

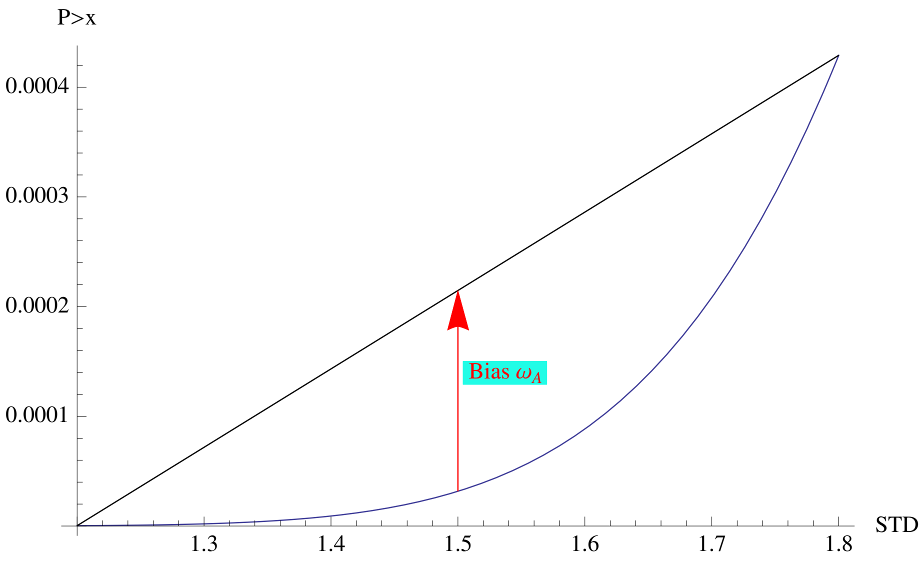


Figure 3.2: Illustration of the convexity bias for a Gaussian raising small probabilities: The plot shows the STD effect on P>x, and compares P>6 with a STD of 1.5 compared to P> 6 assuming a linear combination of 1.2 and 1.8 (here a(1)=1/5).

3.3 The Effect of Metaprobability on Fat Tails

Recall that the tail fattening methods **??.??** and **??.??** were based on randomizing the variance. Small probabilities rise precisely because they are convex to perturbations of the parameters (the scale) of the probability distribution.

3.4 Fukushima, Or How Errors Compound

"Risk management failed on several levels at Fukushima Daiichi. Both TEPCO and its captured regulator bear responsibility. First, highly tailored geophysical models predicted an infinitesimal chance of the region suffering an earthquake as powerful as the T014*d*hoku quake. This model uses historical seismic data to estimate the local frequency of earthquakes of various magnitudes; none of the quakes in the data was bigger than magnitude 8.0. Second, the plant's risk analysis did not consider the type of cascading, systemic failures that precipitated the meltdown. TEPCO never conceived of a situation in which the reactors shut down in response to an earthquake, and a tsunami topped the seawall, and the cooling pools inside the reactor buildings were overstuffed with spent fuel rods, and the main control room became too radioactive for workers to survive, and damage to local infrastructure delayed reinforcement, and hydrogen explosions breached the reactors' outer containment structures. Instead, TEPCO and its regulators addressed each of these risks independently and judged the plant safe to operate as is."Nick Werle, n+1, published by the n+1 Foundation, Brooklyn NY

4 Payoff Skewness and Lack of Skin-in-the-Game

This section will analyze the probabilistic mismatch or tail risks and returns in the presence of a principal-agent problem.

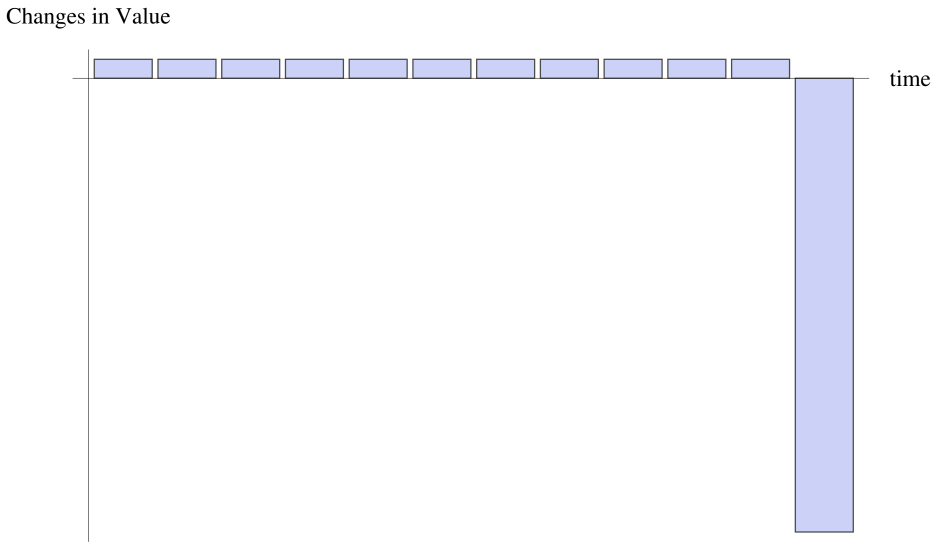


Figure 4.1: The most effective way to maximize the expected payoff to the agent at the expense of the principal.

4.1 Transfer of Harm

Rule 8. *If an agent has the upside of the payoff of the random variable, with no downside, and is judged solely on the basis of past performance, then the incentive is to hide risks in the left tail using a negatively skewed (or more generally, asymmetric) distribution for the performance. This can be generalized to any payoff for which one does not bear the full risks and negative consequences of one's actions.*

Let $P(K,M)$ be the payoff for the operator over *M* incentive periods

$$P(K,M) \equiv \gamma \sum_{i=1}^{M} q_{t+(i-1)\Delta \mathrm{t}} \left(x^j_{i+t\Delta \mathrm{t}} - K \right)^{+} \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau} \tag{4.1}$$

with $X^j = (x^j_{t+i\Delta \mathrm{t}})_{i=1}^{M} \in \mathbb{R}$, i.i.d. random variables representing the distribution of profits over a certain period $[t, t+i\Delta \mathrm{t}]$, $i \in \mathbb{N}$, $\Delta t \in \mathbb{R}^+$ and K is a "hurdle", $\tau = \inf\left\{ s : \left(\sum_{z \leq s} x_z \right) < x_{\min} \right\}$ is an indicator of stopping time when past performance conditions are not satisfied (namely, the condition of having a certain performance in a certain number of the previous years, otherwise the stream of payoffs terminates, the game ends and the number of positive incentives stops). The constant γ $\in$(0,1) is an "agent payoff", or compensation rate from the performance, which does not have to be monetary (as long as it can be quantified as "benefit"). The quantity $q_{t+(i-1)\Delta \mathrm{t}} \in$ [1,∞) indicates the size of the exposure at times *t+(i*-1 *)* Δt (because of an Ito lag, as the performance at period *s* is determined by *q* at a a strictly earlier period < *s*)

Let $\{f_j\}$ be the family of probability measures f_j of X^j , *j* $\in \mathbb{N}$. Each measure corresponds to certain mean/skewness characteristics, and we can split their properties in half on both sides of a "centrality" parameter K, as the "upper" and "lower" distributions. With some inconsequential abuse of notation we write $dF_j(x)$ as $f_j(x)dx$, so $F_j^{+}{=}\int_K^\infty f_j(x)\, dx$ and $F_j^{-}{=}\int_{-\infty}^K f_j(x)\, dx$, the "upper" and "lower" distributions, each corresponding to certain conditional expectation $\mathbb{E}_j^+ \equiv \frac{\int_K^\infty x f_j(x) dx}{\int_K^\infty f_j(x)\, dx}$ and $\mathbb{E}_j^- \equiv \frac{\int_{-\infty}^K x\ f_j(x) dx}{\int_{-\infty}^K f_j(x)\, dx}$.

Now define $\nu \in \mathbb{R}^+$as a K-centered nonparametric measure of asymmetry, $\nu_j \equiv \frac{F_j^-}{F_j^+}$, with values >1 for positive asymmetry, and <1 for negative ones. Intuitively, skewness has probabilities and expectations moving in opposite directions: the larger the negative payoff, the smaller the probability to compensate.

We do not assume a "fair game", that is, with unbounded returns *m* $\in$ (-∞,∞), $F_j^+\ \mathbb{E}_j^+ + F_j^-\ \mathbb{E}_j^- = m$, which we can write as

m^++m^-= m

4.1.1 Simple assumptions of constant *q* and simple-condition stopping time

Assume *q* constant, *q* =1 and simplify the stopping time condition as having no loss in the previous periods, τ =inf$\{(t+(i-1)\Delta \mathrm{t})\}$: $x_{\Delta \mathrm{t}(i-1)+t} < K\}$, which leads to

$$\mathbb{E}(P(K,M)) = \gamma\ \mathbb{E}_j^+ \times \mathbb{E}\left(\sum_{i=1}^M \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau}\right) \quad (4.2)$$

Since assuming independent and identically distributed agent's payoffs, the expectation at stopping time corresponds to the expectation of stopping time multiplied by the expected compensation to the agent γ $\mathbb{E}_j{}^+$. And $\mathbb{E}\left(\sum_{i=1}^M \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau}\right) = \left(\mathbb{E}\left(\sum_{i=1}^M \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau}\right) \wedge M\right)$.

The expectation of stopping time can be written as the probability of success under the condition of no previous loss:

$$\mathbb{E}\left(\sum_{i=1}^M \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau}\right) = \sum_{i=1}^M F_j^+\ \mathbf{1}_{x_{\Delta \mathbf{t}(i-1)+t}>K}$$

We can express the stopping time condition in terms of uninterrupted success runs. Let $\sum$ be the ordered set of consecutive success runs $\sum \equiv \{\{F\},\{\mathrm{SF}\},\{\mathrm{SSF}\},...,\{(M\ -\ 1)$ consecutive $S,F\}\}$, where *S* is success and *F* is failure over period Δt, with associated corresponding probabilities $\{(1-F_j^+)$, $F_j^+\left(1-F_j^+\right), F_j^{+2}\left(1-F_j^+\right), ..., F_j^{+M-1}\left(1-F_j^+\right)\}$,

$$\sum_{i=1}^M F_j^{+(i-1)}\left(1-F_j^+\right) = 1 - F_j^{+M} \simeq 1 \qquad (4.3)$$

For M large, since $F_j^+ \in$ (0,1) we can treat the previous as almost an equality, hence:

$$\sum_{i=1}^M \mathbf{1}_{t+(i-1)\Delta \mathrm{t}<\tau} = \sum_{i=1}^M (i{-}1)\ F_j^{+(i-1)}\left(1-F_j^+\right) = \frac{F_j^+}{1-F_j^+}$$

Finally, the expected payoff for the agent:

$$\mathbb{E}\left(P(K,M)\right) = \ \gamma\ \mathbb{E}_j^+ \frac{F_j^+}{1-F_j^+}$$

which increases by i) increasing $\mathbb{E}_j^+$, ii) minimizing the probability of the loss F_j^-, but, and that's the core point, even if i) and ii) take place at the expense of *m* the total expectation from the package.

Alarmingly, since $\mathbb{E}_j^+ = \frac{m-m^-}{F_j^+}$, the agent doesn't care about a degradation of the total expected return *m* if it comes from the left side of the distribution, m^-. Seen in skewness space, the expected agent payoff maximizes under the distribution *j* with the lowest value of ν_j (maximal negative asymmetry). The total expectation of the positive-incentive without-skin-in-the-game depends on negative skewness, not on *m*.

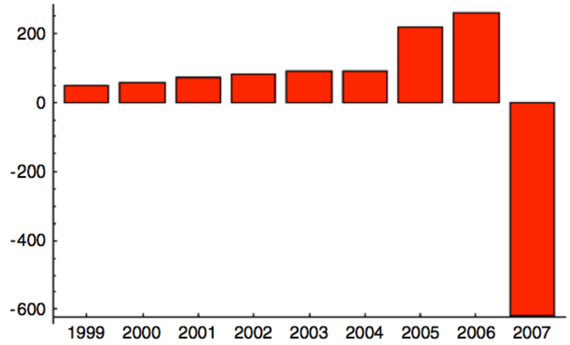


Figure 4.2: Indy Mac, a failed firm during the subprime crisis (from Taleb 2009). It is a representative of risks that keep increasing in the absence of losses, until explosive blowup.

Multiplicative q and the explosivity of blowups

Now, if there is a positive correlation between q and past performance, or survival length, then the effect become multiplicative. The negative payoff becomes explosive if the allocation q increases with visible profitability, as seen in Figure 2 with the story of IndyMac, whose risk kept growing until the blowup[1]. Consider that "successful" people get more attention, more funds, more promotion. Having "beaten the odds" imparts a certain credibility. In finance we often see fund managers experience a geometric explosion of funds under management after perceived "steady" returns. Forecasters with steady strings of successes become gods. And companies that have hidden risks tend to outperform others in small samples, their executives see higher compensation. so in place of a constant exposure q, consider a variable one:

$$q_{\Delta \mathbf{t}(i-1)+t} = q\ \omega(i)$$

where $\omega(i)$ is a multiplier that increases with time, and of course naturally collapses upon blowup.

Equation 4.1 becomes:

$$P(K,M) \equiv \gamma \sum_{i=1}^{M} q\, \omega(i) \left(x^{j}_{t+i\Delta \mathbf{t}} - K \right)^{+} \mathbf{1}_{t+(i-1)\Delta \mathbf{t} < \tau} \tag{4.4}$$

and the expectation, assuming the numbers of periods, M is large enough

$$\mathbb{E}(P(K,M)) = \gamma\ \mathbb{E}_j^{+}\ q\ \mathbb{E}\left(\sum_{i=1}^{M} \omega(i)\ \mathbf{1}_{\Delta \mathbf{t}(i-1)+t<\tau} \right) \tag{4.5}$$

Assuming the rate of conditional growth is a constant $r \in$ [0,∞) , and making the replacement ω(i)$\equiv e^{ri}$, we can call the last term in equation 4.5 the multiplier of the expected return to the agent:

$$\begin{aligned} &\mathbb{E}\left(\textstyle\sum_{i=1}^{M} e^{ir} \mathbf{1}_{\Delta \mathbf{t}(i-1)+t<\tau} \right) \\ &\quad = \textstyle\sum_{i=1}^{M} (i-1)\ F_j{}^{+} e^{ir} \mathbf{1}_{x_{\Delta \mathbf{t}(i-1)+t} > K} \end{aligned} \tag{4.6}$$

$$= \frac{\left(F^{+}-1\right)\left(\left(F^{+}\right)^{M}\left(M e^{(M+1)r} - F^{+}(M-1)e^{(M+2)r}\right) - F^{+} e^{2r}\right)}{\left(F^{+}e^{r}-1\right)^{2}} \tag{4.7}$$

We can get the table of sensitivities for the "multiplier" of the payoff:

	F=.6	0.7	0.8	0.9
r=0	1.5	2.32	3.72	5.47
0.1	2.57	4.8	10.07	19.59
0.2	4.93	12.05	34.55	86.53
0.3	11.09	38.15	147.57	445.59

Table 1 Multiplicative effect of skewness

4.1.2 Explaining why Skewed Distributions Conceal the Mean

Note that skewed distributions conceal their mean quite well, with $P(X < \mathbb{E}(x)) < \frac{1}{2}$ in the presence of negative skewness. And such effect increases with fat-tailedness. Consider a negatively skewed power law distribution, say the mirror image of a standard Pareto distribution, with maximum value $x_{\min}$, and domain $(-\infty, x_{\min}]$, with

[1] The following sad anecdote illustrate the problem with banks. It was announces that "JPMorgan Joins BofA With Perfect Trading Record in Quarter" (Dawn Kopecki and Hugh Son - Bloomberg News, May 9, 2013). Yet banks while "steady earners" go through long profitable periods followed by blowups; they end up losing back all cumulative profits in short episodes, just in 2008 they lost around 4.7 trillion U.S. dollars before government bailouts. The same took place in 1982-1983 and in the Savings and Loans crisis of 1991, see Taleb (2009).

[2] This discussion of a warped probabilistic incentive corresponds to what John Kay has called the "Taleb distribution", John Kay "A strategy for hedge funds and dangerous drivers", Financial Times, 16 January 2003.

exceedance probability $P(X > x) = \quad - x^{-\alpha} x_{\min}^{\alpha}$, and mean $-\frac{\alpha x_{\min}}{\alpha-1}$, with $\alpha > 1$, have a proportion of $1 - \frac{\alpha-1}{\alpha}$ of its realizations rosier than the true mean. Note that fat-tailedness increses at lower values of α. The popular "eighty-twenty", with tail exponent $\alpha = 1.15$, has > 90 percent of observations above the true mean[2].

Forecasters

We can see how forecasters who do not have skin in the game have the incentive of betting on the low-impact high probability event, and ignoring the lower probability ones, even if these are high impact. There is a confusion between "digital payoffs" $\int f_j(x)\, dx$ and full distribution, called "vanilla payoffs", $\int x f_j(x) dx$, see Taleb and Tetlock (2013)[3].

[3] Money managers do not have enough skin in the game unless they are so heavily invested in their funds that they can end up in a net negative form the event. The problem is that they are judged on frequency, not payoff, and tend to cluster together in packs to mitigate losses by making them look like "industry event". Many fund managers beat the odds by selling tails, say covered writes, by which one can increase the probability of gains but possibly lower the expectation. They also have the optionality of multi-time series; they can manage to hide losing funds in the event of failure. Many fund companies bury hundreds of losing funds away, in the "cemetery of history" (Taleb, 2007)
.

5 Large Numbers and Convergence in the Real World

The Law of Large Numbers and The Central Limit Theorem are the foundation of modern statistics: The behavior of the sum of random variables allows us to get to the asymptote and use handy asymptotic properties, that is, Platonic distributions. But the problem is that in the real world we never get to the asymptote, we just get "close". Some distributions get close quickly, others very slowly (even if they have finite variance). Recall from Chapter 1 that the quality of an estimator is tied to its replicability outside the set in which it was derived: this is the basis of the law of large numbers.

5.1 The Law of Large Numbers Under Fat Tails

How do you reach the limit?

The common interpretation of the weak law of large numbers is as follows.
By the weak law of large numbers, consider a sum of random variables X_1, X_2,..., X_N independent and identically distributed with finite mean *m*, that is E[X_i] $<\infty$, then $\frac{1}{N}\sum_{1\leq i\leq N} X_i$ converges to m **in probability**, as N $\rightarrow \infty$. But the problem of convergence in probability, as we will see later, is that it does not take place in the tails of the distribution (different parts of the distribution have different speeds). This point is quite central and will be examined later with a deeper mathematical discussions on limits in Chapter x. We limit it here to intuitive presentations of turkey surprises.
(Hint: we will need to look at the limit without the common route of Chebychev's inequality which requires E[X_i^2] $<\infty$. Chebychev's inequality and similar ones eliminate the probabilities of some tail events).
So long as there is a mean, observations should *at some point* reveal it.

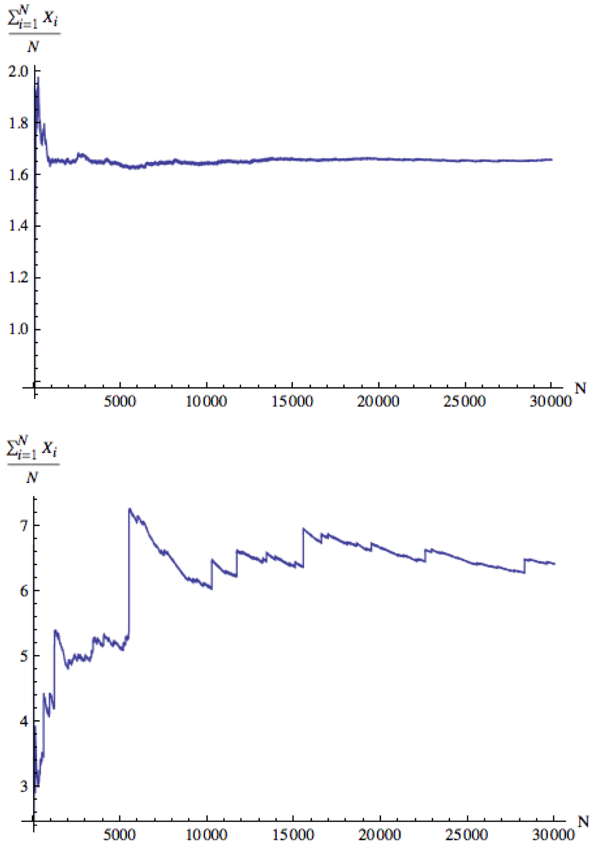


Figure 5.1: How thin tails (Gaussian) and fat tails (1< α ≤2) converge to the mean.

The law of iterated logarithms

For the "thin-tailed" conditions, we can see in Figure x how by the law of iterated logarithm, for x_i i.i.d. distributed with mean 0 and unitary variance, $\underset{n\to\infty}{lim\,sup}\,\frac{\sum_{i=1}^{n} x_i}{\sqrt{2nn \log\log}} = 1$ a.s. (and by symmetry $\underset{n\to\infty}{lim\,inf}$ $\frac{\sum_{i=1}^{n} x_i}{\sqrt{2nn \log\log}} =$ -1), thus giving us an acceptably narrow cone limiting the fluctuation of the sum.

Speed of convergence:

Let us examine the speed of convergence of the average $\frac{1}{N}\sum_{1\leq i\leq N} X_i$. For a Gaussian distribution (m, σ), the characteristic function for the convolution is $\varphi(t/N)^N {=} \left(e^{\frac{imt}{N} - \frac{s^2 t^2}{2N^2}}\right)^N$, which, derived twice at 0 yields $(-i)^2 \frac{\partial^2 c}{\partial t^2} - i \frac{\partial c}{\partial t} /.\, t \to 0$ which produces the standard deviation $\sigma(n) = \frac{\sigma(1)}{\sqrt{N}}$ so one can say that sum "converges" at a speed $\sqrt{N}$.
Another approach is by expanding φ and letting N go to infinity

$$\lim_{N\to\infty} \left(e^{\frac{imt}{N} - \frac{s^2 t^2}{2N^2}}\right)^N = e^{imt}$$

Now e^{imt} is the characteristic function of the degenerate distribution at *m,* with density $p(x) = \delta(m - x)$ *where* δ is the Dirac delta with values zero except at the point *m-x* . (Note that the strong law of large numbers imply as convergence takes place almost everywhere except for a set of probability 0; for that the same result should be obtained for all t).
But things are far more complicated with power laws. Let us repeat the exercise for a Pareto distribution with density $L^\alpha x^{-1-\alpha} \alpha$, x> L,

$$\varphi(t/N)^N = \alpha^N E_{\alpha+1} \left(-\frac{iLt}{N}\right)^N$$

where E is the exponential integral E; $E_n(z) = \int_1^\infty e^{-zt}/t^n dt$.
At the limit:

$$\lim_{N\to\infty} \varphi\left(\frac{t}{N}\right)^N = e^{\frac{\alpha}{\alpha-1} iLt}$$

which is degenerate Dirac at $\frac{\alpha}{\alpha-1}L$, and as we can see the limit only exists for α >1.
Setting $L = 1$ to scale, the standard deviation $\sigma_\alpha(N)$ for the *N*-average becomes, for α >2

$$\sigma_\alpha(N) = \frac{1}{N} \left(\alpha^N E_{\alpha+1}(0)^{N-2} \left(E_{\alpha-1}(0) E_{\alpha+1}(0) \right.\right.$$
$$\left.\left. + E_\alpha(0)^2 \left(-N\alpha^N E_{\alpha+1}(0)^N + N - 1\right)\right)\right)$$

Sucker Trap

After some tinkering, we get $\sigma_\alpha(N) = \frac{\sigma_\alpha(1)}{\sqrt{N}}$ as with the Gaussian, which is a sucker's trap. For we should be careful in interpreting $\sigma_\alpha(N)$, which will be very volatile since $\sigma_\alpha(1)$ is already very volatile and does not reveal itself easily in realizations of the process. In fact, let *p(.)* be the PDF of a Pareto distribution with mean *m*, standard deviation σ, minimum value *L* and exponent α, Δ_α the expected mean deviation of the variance for a given α will be $\Delta_\alpha{=}$ $\frac{1}{\sigma^2}\int_L^\infty \left(\left|(x-m)^2 - \sigma^2\right|\right) p(x) dx$

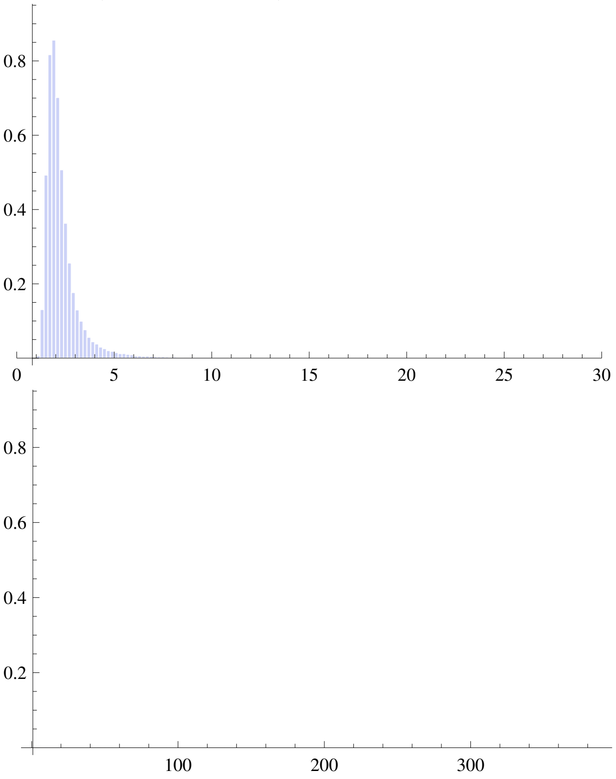


Figure 5.2: The distribution (histogram) of the standard deviation of the sum of N=100 α=13/6. The second graph shows the entire span of realizations.

Absence of Useful Theory:

As to situations, central situations, where 1< α <2, we are left hanging analytically (but we can do something about it in the next section). We will return to the problem in our treatment of the preasymptotics of the central limit theorem.

But we saw in **??.??** that the volatility of the mean is $\frac{\alpha}{\alpha-1}$ *s* and the mean deviation of the mean deviation, that is, the volatility of the volatility of mean is $2(\alpha-1)^{\alpha-2}\alpha^{1-\alpha}s$, where *s* is the scale of the distribution. As we get close to $\alpha = 1$ the mean becomes more and more volatile in realizations for a given scale. This is not trivial since we are not interested in the speed of convergence *per se* given a variance, rather the ability of a sample to deliver a meaningful estimate of some total properties.

Intuitively, the law of large numbers needs an infinite observations to converge at α=1. So, if it ever works, it would operate at a $>$20 times slower rate for an "observed" α of 1.15 than for an exponent of 3. To make up for measurement errors on the α, as a rough heuristic, just assume that one needs $>$ 400 times the observations. Indeed, 400 times! (The point of what we mean by "rate" will be revisited with the discussion of the Large Deviation Principle and the Cramer rate function in X.x; we need a bit more refinement of the idea of tail exposure for the sum of random variables).

Comparing N = 1 to N = 2 for a symmetric power law with 1< α ≤2 .

Let $\phi(t)$ be the characteristic function of the symmetric Student T with α degrees of freedom. After two-fold convolution of the average we get:

$$\phi(t/2)^2 = \frac{4^{1-\alpha}\alpha^{\alpha/2}\,|t|^\alpha\, K_{\frac{\alpha}{2}}\left(\frac{\sqrt{\alpha}|t|}{2}\right)^2}{\Gamma\left(\frac{\alpha}{2}\right)^2},$$

We can get an explicit density by inverse Fourier transformation of ϕ,

$$p_{2,\alpha}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(t/2)^{2-\text{i}tx}\text{dt},$$

which yields the following

$$p_{2,\alpha}(x) = \frac{\pi\ 2^{-4\alpha}\ \alpha^{5/2}\Gamma(2\alpha)\,{}_2F_1\left(\alpha+\frac{1}{2},\frac{\alpha+1}{2};\frac{\alpha+2}{2};-\frac{x^2}{\alpha}\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)^4}$$

where ${}_2F_1$is the hypergeometric function,${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty}(a)_k(b)_k/(c)_k\ z^k/\,k!$.

We can compare the twice-summed density to the initial one (with notation: p_n(x)= P($\sum_{i=1}^{N} x_i$=x))

$$p_{1,\alpha}(x) = \frac{\left(\frac{\alpha}{\alpha+x^2}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)}$$

From there, we see that in the Cauchy case (α=1) the sum conserves the density, so

$$p_{1,1}(x) = p_{2,1}(x) = \frac{1}{\pi\left(1+x^2\right)}$$

Let us use the ratio of mean deviations; since the mean is 0,

$$\mu(\alpha) \equiv \frac{\int |x| p_{2,\alpha}(x)dx}{\int |x| p_{1,\alpha}(x)dx}$$

$$\mu(\alpha) = \frac{\sqrt{\pi}\ 2^{1-\alpha}\ \Gamma\left(\alpha-\frac{1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)^2}$$

and

$$\lim_{\alpha\to\infty}\mu(\alpha) = \frac{1}{\sqrt{2}}$$

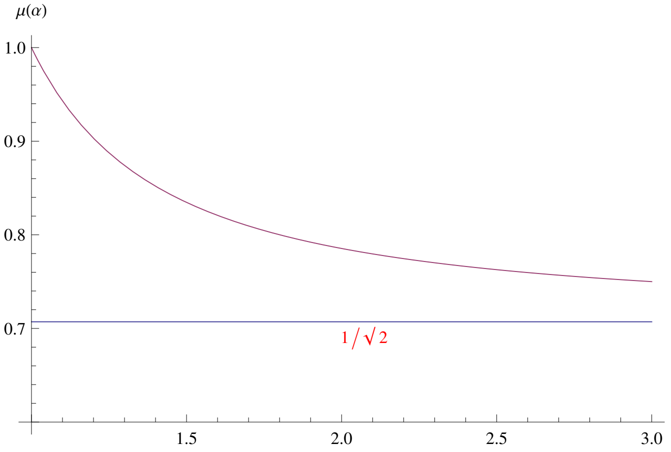


Figure 5.3: Preasymptotics of the ratio of mean deviations. But one should note that mean deviations themselves are extremely high in the neighborhood of ↓1. So we have a "sort of" double convergence to $\sqrt{n}$: convergence at higher n and convergence at higher α.

The double effect of summing fat tailed random variables: The summation of random variables performs two simultaneous actions, one, the "thinning" of the tails by the CLT for a finite variance distribution (or convergence to some basin of attraction for infinite variance classes); and the other, the lowering of the dispersion by the LLN.

Both effects are fast under thinner tails, and slow under fat tails. But there is a third effect: the dispersion of observations for *n*=1 is itself much higher under fat tails. Fatter tails for power laws come with higher expected mean deviation.

5.2 Preasymptotics and Central Limit in the Real World

The common mistake is to think that if we satisfy the criteria of convergence, that is, independence and ***finite variance***, that central limit is a given.Take the conventional formulation of the Central Limit Theorem [1]:

Let X_1 , X_2 ,... be a sequence of independent identically distributed random variables with mean *m* & variance σ^2 satisfying *m*$< \infty$ and $0 < \sigma^2 < \infty$, then

$$\frac{\sum_{i=1}^{N} X_i - Nm}{\sigma\sqrt{n}} \overset{D}{\rightarrow} N(0,1)\text{as } n \rightarrow \infty$$

Where $\overset{D}{\rightarrow}$ is converges "in distribution" and N(0,1) is the Gaussian with mean 0 and unit standard deviation.

Granted convergence "in distribution" is about the weakest form of convergence. Effectively we are dealing with a double problem.

The first, as uncovered by Jaynes, corresponds to the abuses of measure theory: Some properties that hold at infinity might not hold in all limiting processes .

There is a large difference between convergence a.s. (almost surely) and the weaker forms.

Jaynes 2003 (p.44):"The danger is that the present measure theory notation presupposes the infinite limit already accomplished, but contains no symbol indicating which limiting process was used (...) Any attempt to go directly to the limit can result in nonsense".

We accord with him on this point –along with his definition of probability as information incompleteness, about which later.

The second problem is that we do not have a "clean" limiting process –the process is itself idealized.

Now how should we look at the Central Limit Theorem? Let us see how we arrive to it assuming "independence".

The Kolmogorov-Lyapunov Approach and Convergence in the Body

The CLT works does not fill-in uniformily, but in a Gaussian way −−−indeed, disturbingly so. Simply, whatever your distribution (assuming one mode), your sample is going to be skewed to deliver more central observations, and fewer tail events. The consequence is that, under aggregation, the sum of these variables will converge "much" faster in theπ body of the distribution than in the tails. As N, the number of observations increases, the Gaussian zone should cover more grounds... but not in the "tails".

This quick note shows the intuition of the convergence and presents the difference between distributions.

Take the sum of of random independent variables X_i with *finite variance* under distribution φ(X). Assume 0 mean for simplicity (and symmetry, absence of skewness to simplify). A more useful formulation is the Kolmogorov or what we can call "Russian" approach of working with bounds:

$$P\left(-u \leq Z = \frac{\sum_{i=0}^{n} X_i}{\sqrt{n}\sigma} \leq u\right) = \frac{\int_{-u}^{u} e^{-\frac{Z^2}{2}}\, dZ}{\sqrt{2\pi}}$$

So the distribution is going to be:

$$\left(1 - \int_{-u}^{u} e^{-\frac{Z^2}{2}}\, dZ\right), \text{for} - u \leq z \leq u$$

inside the "tunnel" [-u,u] –the odds of falling inside the tunnel itself,

and

$$\int_{-\infty}^{u} Z\varphi'(N)dz + \int_{u}^{\infty} Z\varphi'(N)dz$$

outside the tunnel, in $\overline{[-u,u]}$,where $\varphi'(N)$ is the n-summed distribution of φ.

How $\varphi'(N)$ behaves is a bit interesting here –it is distribution dependent.

[1] Feller 1971, Vol. II

Before continuing, let us check the speed of convergence *per* distribution. It is quite interesting that we the ratio of observations in a given sub-segment of the distribution is in proportion to the expected frequency $\frac{N^u_{-u}}{N^\infty_{-\infty}}$ where N^u_{-u}, is the numbers of observations falling between *-u* and *u*. So the speed of convergence to the Gaussian will depend on $\frac{N^u_{-u}}{N^\infty_{-\infty}}$ as can be seen in the next two simulations.

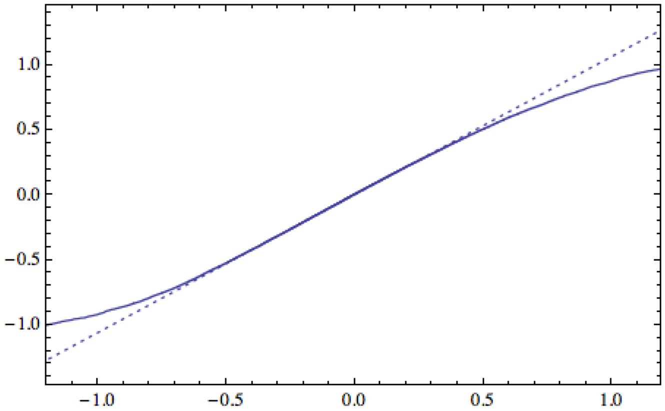


Figure 5.4: Q-Q Plot of N Sums of variables distributed according to the Student T with 3 degrees of freedom, N=50, compared to the Gaussian, rescaled into standard deviations. We see on both sides a higher incidence of tail events. 10^6*simulations*

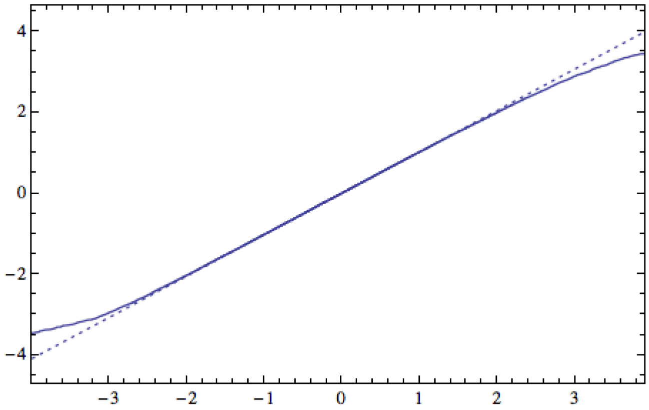


Figure 5.5: ***The Widening Center****. Q-Q Plot of variables distributed according to the Student T with 3 degrees of freedom compared to the Gaussian, rescaled into standard deviation, N=500. We see on both sides a higher incidence of tail events.* 10^7*simulations.*

To have an idea of the speed of the widening of the tunnel $(-u, u)$ under summation, consider the symmetric (0-centered) Student T with tail exponent $\alpha=$ 3, with density $\frac{2a^3}{\pi(a^2+x^2)^2}$, and variance a^2. For large "tail values" of x, $P(x) \to \frac{2a^3}{\pi x^4}$. Under summation of N variables, the tail $P(\Sigma x)$ will be $\frac{2Na^3}{\pi x^4}$. Now the center, by the Kolmogorov version of the central limit theorem, will have a variance of Na^2 in the center as well, hence

$$P(\Sigma\ x) = \frac{e^{-\frac{x^2}{2a^2 N}}}{\sqrt{2\pi}a\sqrt{N}}$$

Setting the point u where the crossover takes place,

$$\frac{e^{-\frac{x^2}{2aN}}}{\sqrt{2\pi}a\sqrt{N}} \simeq \frac{2Na^3}{\pi x^4},$$

hence $u^4 e^{-\frac{u^2}{2aN}} \simeq \frac{\sqrt{2}2a^3\sqrt{aN}N}{\sqrt{\pi}}$, which produces the solution

$$\pm u = \pm 2a\sqrt{N}\sqrt{-W\left(-\frac{1}{2N^{1/4}(2\pi)^{1/4}}\right)},$$

where W is the Lambert W function or *product log* which climbs very slowly[2], particularly if instead of considering the sum u we rescaled by $1/a\sqrt{N}$.

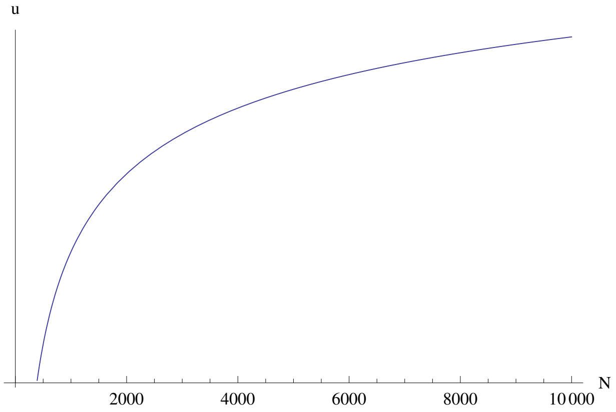


Figure 5.6: The behavior of the "tunnell" under summation

Note about the crossover

See the competing Nagaev brothers, s.a. S.V. Nagaev(1965,1970,1971,1973), and A.V. Nagaev(1969) etc. There are two sets of inequalities, one lower one

[2] Interestingly, among the authors on the paper on the Lambert W function figures Donald Knuth: Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., Knuth, D. E. (1996). On the LambertW function. Advances in Computational mathematics, 5(1), 329-359.

below which the sum is in regime 1 (thin-tailed behavior), an upper one for the fat tailed behavior, where the cumulative function for the sum behaves likes the maximum . By Nagaev (1965) For a regularly varying tail, where $\mathbb{E}\left(|X|^m\right) < \infty$ the minimum of the crossover should be to the left of $\sqrt{\left(\frac{m}{2}-1\right) N \log(N)}$ (normalizing for unit variance) for the right tail (and with the proper sign adjustment for the left tail).
So

$$\frac{\mathbb{P}_{> \sum^{N} X_i}}{\mathbb{P}_{> \frac{X}{\sqrt{N}}}} \to 1$$

for 0 $\leq$ x $\leq \sqrt{\left(\frac{m}{2}-1\right) N \log(N)}$

Generalizing for all exponents $>$ 2

More generally, using the reasoning for a broader set and getting the crossover for powelaws of all exponents:

$$\frac{\sqrt[4]{(\alpha-2)\alpha} e^{-\frac{\sqrt{\frac{\alpha-2}{\alpha}} x^2}{2 a N}}}{\sqrt{2\pi}\sqrt{a \alpha N}} \simeq \frac{a^\alpha \left(\frac{1}{x^2}\right)^{\frac{1+\alpha}{2}} \alpha^{\alpha/2}}{\text{Beta}\left[\frac{\alpha}{2}, \frac{1}{2}\ ,\right]}$$

since the standard deviation is $a\ \sqrt{\frac{\alpha}{-2+\alpha}}$

$$x \to \pm\ \sqrt{\pm \frac{a\ \alpha\ (\alpha+1)\ N\ W(\lambda)}{\sqrt{(\alpha-2)\ \alpha}}}$$

Where

$$\lambda = -\frac{(2\pi)^{\frac{1}{\alpha+1}} \sqrt{\frac{\alpha-2}{\alpha}} \left(\frac{\sqrt[4]{\alpha-2}\alpha^{-\frac{\alpha}{2}-\frac{1}{4}} a^{-\alpha-\frac{1}{2}} B\left(\frac{\alpha}{2}, \frac{1}{2}\right)}{\sqrt{N}}\right)^{-\frac{2}{\alpha+1}}}{a\ (\alpha+1)\ N}$$

5.3 Using Log Cumulants to Observe Preasymptotics

The normalized cumulant of order n, ${}_n$ is the derivative of the log of the characteristic function Φ which we convolute N times divided by the second cumulant (i,e., second moment).
This exercise show us how fast an aggregate of N-summed variables become Gaussian, looking at how quickly the 4th cumulant approaches 0. For instance the Poisson get there at a speed that depends inversely on Λ, that is, $1/(N^2\Lambda^3)$, while by contrast an exponential distribution reaches it at a slower rate at higher values of Λ since the cumulant is $(3!\,\Lambda^2)/N^2$.

Distribution	**Normal**(μ,σ)	**Poisson(**λ **)**	**Exponential(**λ **)**	$\mathbf{\Gamma(a,b)}$
PDF	$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$\frac{e^{-\lambda}\lambda^x}{x!}$	e^-x $\lambda\lambda$	$\frac{b^{-a}e^{-\frac{x}{b}}x^{a-1}}{\Gamma(a)}$
N-convoluted Log Characteristic	$N\log\left(e^{iz\mu-\frac{z^2\sigma^2}{2}}\right)$	$N\log\left(e^{\left(-1+e^{iz}\right)\lambda}\right)$	$N\log\left(\frac{\lambda}{\lambda-iz}\right)$	$N\log\left((1-ibz)^{-a}\right)$
2 nd **Cumulant**	1	1	1	1
3 rd	0	$\frac{1}{N\lambda}$	$\frac{2\lambda}{N}$	$\frac{2}{a\ b\ N}$
4 th	0	$\frac{1}{N^2\lambda^2}$	$\frac{3!\lambda^2}{N^2}$	$\frac{3!}{a^2\ b^2\ N^2}$
6 th	0	$\frac{1}{N^4\lambda^4}$	$\frac{5!\lambda^4}{N^4}$	$\frac{5!}{a^4b^4N^4}$
8 th	0	$\frac{1}{N^6\lambda^6}$	$\frac{7!\lambda^6}{N^6}$	$\frac{7!}{a^6b^6N^6}$
10 th	0	$\frac{1}{N^8\lambda^8}$	$\frac{9!\lambda^8}{N^8}$	$\frac{9!}{a^8b^8N^8}$

Table 5.1: Table of Normalized Cumulants -Speed of Convergence (Dividing by Σ^n where n is the order of the cumulant).

Distribution	Mixed Gaussians (Stoch Vol)	StudentT(3)	StudentT(4)
PDF	$p\frac{e^{-\frac{x^2}{2{\sigma_1}^2}}}{\sqrt{2\pi}\sigma_1}+(1-p)\frac{e^{-\frac{x^2}{2{\sigma_2}^2}}}{\sqrt{2\pi}\sigma_2}$	$\frac{6\sqrt{3}}{\pi(x^2+3)^2}$	$12\left(\frac{1}{x^2+4}\right)^{5/2}$
N-convoluted Log Characteristic	$N\log\left(pe^{-\frac{z^2{\sigma_1}^2}{2}}+(1-p)e^{-\frac{z^2{\sigma_2}^2}{2}}\right)$	$N\left(\log\left(\sqrt{3}\left\|z\right\|+1\right)-\sqrt{3}\left\|z\right\|\right)$	$N\log\left(2\left\|z\right\|^2K_2(2\left\|z\right\|)\right)$
2nd Cum	1	1	1
3 rd	0	Fuhgetaboudit	TK
4 th	$\frac{\left(3(1-p)p\left(\sigma_1^2-\sigma_2^2\right)^2\right)}{\left(N^2\left(p\sigma_1^2-(-1+p)\sigma_2^2\right)^3\right)}$	Fuhgetaboudit	Fuhgetaboudit
6 th	$\frac{\left(15(-1+p)p(-1+2p)\left(\sigma_1^2-\sigma_2^2\right)^3\right)}{\left(N^4\left(p\sigma_1^2-(-1+p)\sigma_2^2\right)^5\right)}$	Fuhgetaboudit	Fuhgetaboudit

Speed of Convergence of the Summed distribution using Edgeworth Expansions

A twinking of Feller (1971), Vol II by replacing the derivatives with our cumulants. Let $f_N(z)$ be the normalized sum of the i.i.d. distributed random variables $\Xi = \{\xi_i\}_{1<i\leq N}$ with variance σ^2 , $z \equiv \frac{\Sigma\xi_i - E(\Xi)}{\sigma}$ and $\phi_{0,\sigma}(z)$ the standard Gaussian with mean 0, then the convoluted sum approaches the Gaussian as follows assuming $\mathbb{E}\left(\Xi^p\right) < \infty$,i.e., the moments of Ξ of $\leq p$ exist:

$zf_N - z\phi_{0,\sigma}=$

$$(z\phi_{0,\sigma})\left(\sum_{s}^{p-2}\sum_{r}^{s}\frac{\sigma^s\left(zH_{2r+s}\right)\left(Y_{s,r}\left\{\frac{\kappa_k}{(k-1)k\sigma^{2k-2}}\right\}_{k=3}^{p}\right)}{\left(\sqrt{2}\sigma\right)\left(s!\,2^{r+\frac{s}{2}}\right)}\right.$$

$$\left.+1\right)$$

where $\kappa_k$is the cumulant of order k. $Y_{n,k}\left(x_1,\ldots,x_{-k+n+1}\right)$ is the partial Bell polynomial given by

$$Y_{n,k}\left(x_1,\ldots,x_{-k+n+1}\right) \equiv$$

$$\sum_{m_1=0}^{n}\cdots\sum_{m_n=0}^{n}\frac{n!}{\cdots m_1!\,m_n!}\times$$

$$\mathbf{1}_{[nm_n+m_1+2m_2+\cdots=n\wedge m_n+m_1+m_2+\cdots=k]}\prod_{s=1}^{n}\left(\frac{x_s}{s!}\right)^{m_s}$$

Notes on Levy Stability and the Generalized Cental Limit Theorem

Take for now that the distribution that concerves under summation (that is, stays the same) is said to be "stable". You add Gaussians and get Gaussians. But if you add binomials, you end up with a Gaussian, or, more accurately, "converge to the Gaussian basin of attraction". These distributions are not called "unstable" but they are.

There is a more general class of convergence. Just consider that the Cauchy variables converges to Cauchy, so the “stability’ has to apply to an entire class of distributions.

Although these lectures are not about mathematical techniques, but about the real world, it is worth developing some results converning stable distribution in order to prove some results relative to the effect of skewness and tails on the stability.

Let n be a positive integer, $n \geq$2 and $X_1, X_2, ..., X_n$ satisfy some measure of independence and are drawn from the same distribution,

i) there exist $c\ n \in\ \mathbb{R}^+$ and $d\ n \in\ \mathbb{R}^+$ such that

$$\sum_{i=1}^{n} X_i \overset{D}{=} c_n X + d_n$$

where $\overset{D}{=}$ means “equality” in distribution.

ii) or, equivalently, there exist sequence of i.i.d random variables $\{Y_i\}$, a real positive sequence $\{d_i\}$ and a real sequence $\{a_i\}$ such that

$$\frac{1}{d_n}\sum_{i=1}^{n} Y_i + a_n \overset{D}{\to} X$$

where$\overset{D}{\to}$ means convergence in distribution.
iii) or, equivalently,
The distribution of X has for characteristic function

$$\phi(t) = \begin{cases} \exp(i\mu t - \sigma\,|t|\,(1 + 2i\beta/\pi \mathsf{sgn}(t)\log(|t|))) & \alpha = 1 \\ \exp\left(i\mu t - |t\sigma|^{\alpha}\left(1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\mathsf{sgn}(t)\right)\right) & \alpha \neq 1 \end{cases}.$$

$\alpha \in$(0,2] $\sigma \in \mathbb{R}^{+}$, $\beta \in$[-1,1], $\mu \in \mathbb{R}$

Then if either of i), ii), iii) holds, X has the "alpha stable" distribution $\mathbf{S}(\alpha, \beta, \mu, \sigma)$, with β designating the symmetry, μ the centrality, and σ the scale.

Warning: perturbating the skewness of the Levy stable distribution by changing β without affecting the tail exponent is mean preserving, which we will see is unnatural: the transformation of random variables leads to effects on more than one characteristic of the distribution.

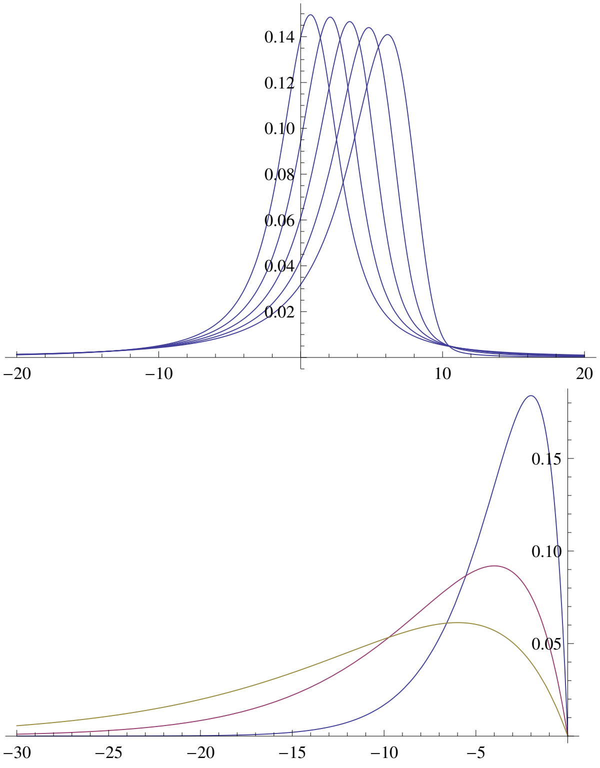


Figure 5.7: Disturbing the scale of the alpha stable and that of a more natural distribution, the gamma distribution. The alpha stable does not increase in risks! (risks for us in Chapter x is defined in thickening of the tails of the distribution). We will see later with "convexification" how it is rare to have an isolated perturbation of distribution without an increase in risks.

$\mathbf{S}(\alpha, \beta, \mu, \sigma)$represents the stable distribution S_{type} with index of stability α, skewness parameter β, location parameter μ, and scale parameter σ.

The Generalized Central Limit Theorem gives sequences a_n and b_n such that the distribution of the shifted and rescaled sum $Z_n = \left(\sum_i^n X_i - a_n\right)/b_n$ of n i.i.d. random variates X_i whose distribution function $F_X(x)$ has asymptotes $1 - cx^{-\mu}$ as x->$+\infty$ and $d(-x)^{-\mu}$ as x->$-\infty$ weakly converges to the stable distribution $S_1(\alpha, (c-d)/(c+d), 0, 1)$:

Note: Chebyshev's Inequality and upper bound on deviations under finite variance.

[To ADD MARKOV BOUNDS ⟶ CHEBYCHEV ⟶ CHERNOV BOUNDS.]

Even when the variance is finite, the bound is rather far. Consider Chebyshev's inequality:
$P(X > \alpha) \leq \frac{\sigma^2}{\alpha^2}$
$P(X > n\sigma) \leq \frac{1}{n^2}$
Which effectively accommodate power laws but puts a bound on the probability distribution of large deviations – but still significant.

The Effect of Finiteness of Variance

This table shows the inverse of the probability of exceeding a certain σ for the Gaussian and the lower on probability limit for any distribution with finite variance.

Deviation		
3	Gaussian	
$7. \times 10^2$	ChebyshevUpperBound	
9		
4	$3. \times 10^4$	16
5	$3. \times 10^6$	25
6	$1. \times 10^9$	36
7	$8. \times 10^{11}$	49
8	$2. \times 10^{15}$	64
9	$9. \times 10^{18}$	81
10	$1. \times 10^{23}$	100

5.4 Illustration: Convergence of the Maximum of a Finite Variance Power Law

The behavior of the maximum value as a percentage of a sum is much slower than we think, and doesn't make much difference on whether it is a finite variance, that is α >2 or not. (See comments in Mandelbrot & Taleb, 2011)
$\tau(N) \equiv E()$

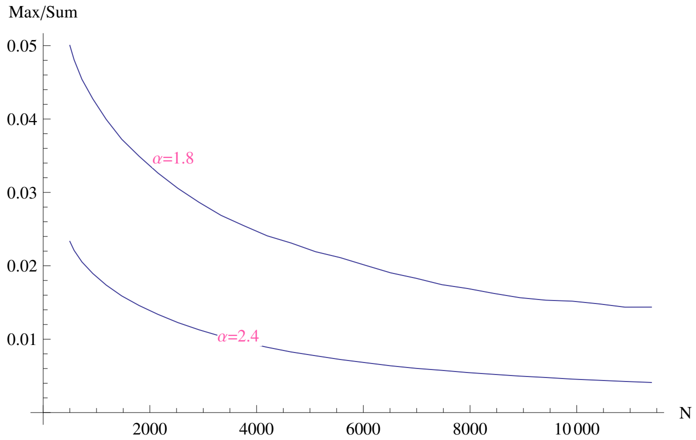


Figure 5.8: Pareto Distributions with different tail exponents α=1.8 and 2.5. The difference between the two (finite and infinite variance) is quantitative, not qualitative. impressive.

General References for the Chapter

P. LÃĺvy, 1937, "ThÃĺorie de l'addition des variables alÃĺatoires" , Gauthier-Villars
Gnedenko and Kolmogorov (1954, 1968 tr.), Limit Distributions for Sums of Independent Random Variables, Addison-Wesley
Prokhorov Y.V.(1959) Some remarks on the strong law of large numbers, Theor. Probab. Appl.
Prokhorov Y.V. (1959) One extremal problem of the theory of probability, Theory Probab. Appl., 1959, 4, No 2, 201 - 203.
Hoeffding W. (1963) Probability inequalities for sums of bounded random variables, J. Amer.Statist. Assoc.,
Petrov (1995 tr.), Limit Theorems of Probability Theory Sequences of Independent Random Variables, Clarendon Press
Blum, M. (1970) On the sums of independently distributed Pareto variates. SIAM J. Appl. Math. 19(1):191198
Nagaev S.V. , 1965, Some limit theorems for large deviations, Theory Probab. Appl.
Brennan, L. E., Reed, I. S., Sollfrey, W. (1968). A comparison of average likelihood and maximum likelihood ratio tests for detecting radar targets of unknown Doppler frequency. IEEE Trans. Info. Theor. IT-4:104110.
Ramsay, Colin,2006, The Distribution of Sums of Certain I.I.D. Pareto Variates. Communications in Statistics: Theory Methods.
Bennet G. Probability inequalities for the sum of independent random variables, J. Amer. Statist. Assoc., 1962, 57, No 297, 33-45.
Nagaev S.V. Large deviations of sums of independent random variables, Ann. Prob., 1979, 7, No 5, 745789.
A Discussion of Concentration functions
W. Doeblin, P. LÃĺvy, "Calcul des probabilitÃĺs. Sur les sommes de variables alÃĺatoires indÃĺpendantes Ãă dispersions bornÃĺes infÃĺrieurement" C.R. Acad. Sci. , 202 (1936) pp. 20272029[2]
W. Doeblin, "Sur les sommes d'un grand nombre de variables alÃĺatoires indÃĺpendantes" Bull. Sci. Math. , 63 (1939) pp. 2364[4]
Kolmogorov, "Sur les propriÃĺtÃĺs des fonctions de concentration de M. P. LÃĺvy" Ann. Inst. H. PoincarÃĺ , 16 (19581960) pp. 2734[5a]
B.A. Rogozin, "An estimate for concentration functions" Theory Probab. Appl. , 6 (1961) pp. 9496 Teoriya Veroyatnost. i Prilozhen. , 6 (1961) pp. 103105[5b]
B.A. Rogozin, "On the increase of dispersion of sums of independent random variables" Theory Probab. Appl. , 6 (1961) pp. 9799 Teoriya Veroyatnost. i Prilozhen. , 6 (1961) pp. 106108[6
H. Kesten, "A sharper form of the DoeblinLÃĺvyKolmogorovRogozin inequality for concentration functions" Math. Scand. , 25 (1969) pp. 133144
B.A. Rogozin, "An integral-type estimate for concentration functions of sums of independent random variables" Dokl. Akad. Nauk SSSR , 211 (1973) pp. 10671070 (In Russian)[8]V.V. Petrov, "Sums of independent random variables" , Springer (1975) (Translated from Russian)[9]
C.G. Esseen, "On the concentration function of a sum of independent random variables" Z. Wahrscheinlichkeitstheor. und Verw. Geb. , 9 (1968) pp. 290308
Rosenthal H.P. On the subspaces of Lp (p > 2) spanned by sequences of independent random variables// Israel J. Math., 1970, 8, 273303.
Nagaev S.V., Pinelis I.F. Some inequalities for the distributions of sums of independent random variables// Probab. Appl., 1977, 248-256, 22, No 2, 248 - 256.

B | Where Standard Diversification Fails

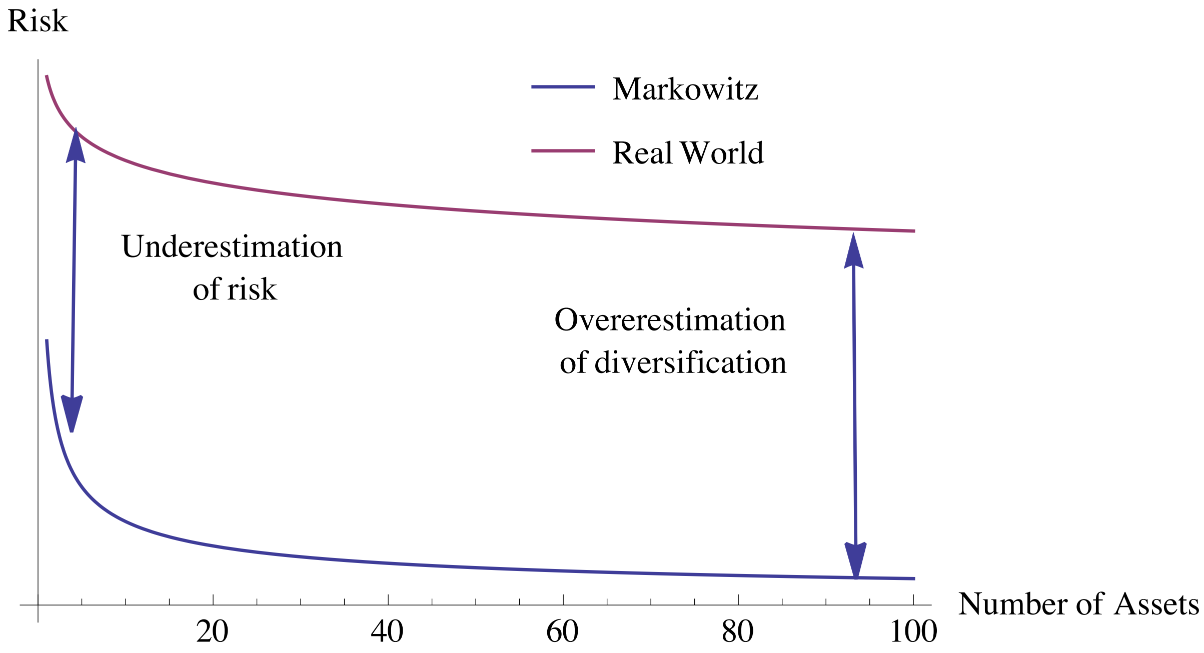


Figure B.1: The "diversification effect": difference between promised and delivered. Markowitz Mean Variance based portfolio construction will stand probably as the most empirically invalid theory ever used in modern times. If irresponsible charlatanism cannot describe this, what can?

C Fat Tails and Random Matrices

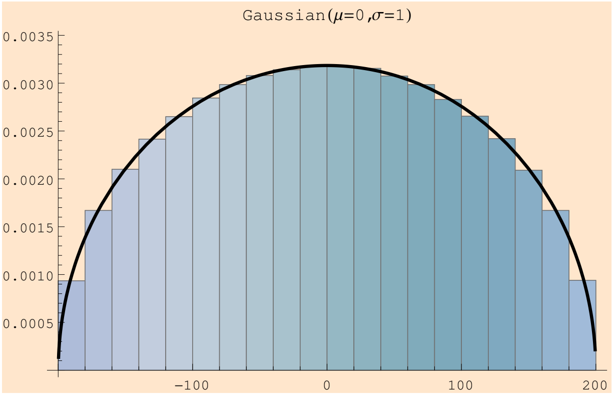


Table C.1: Gaussian

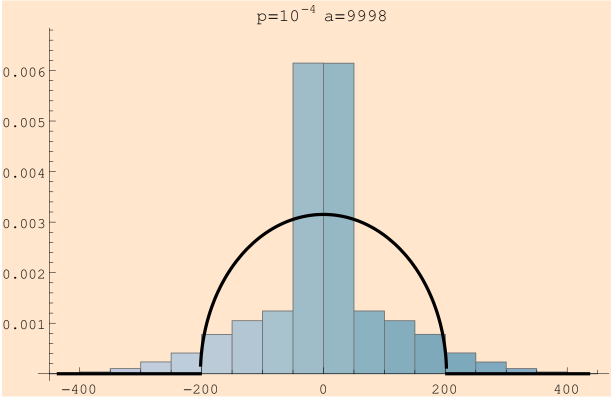


Table C.2: Standard Tail Fattening

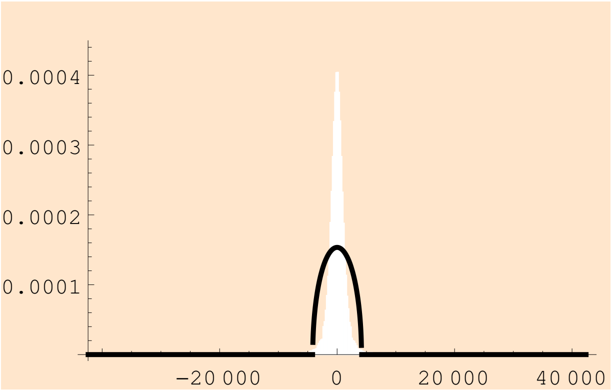


Table C.3: Student T $\frac{3}{2}$

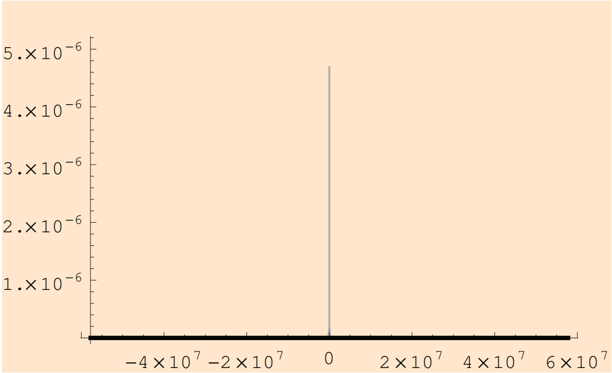


Table C.4: Cauchy

6 Some Misuses of Statistics in Social Science

6.1 Attribute Substitution

It occurs when an individual has to make a judgment (of a target attribute) that is complicated complex, and instead substitutes a more easily calculated one. There have been many papers (Kahneman and Tversky 1974, Hoggarth and Soyer, 2012) showing how statistical researchers overinterpret their own findings, as simplication leads to the *fooled by randomness* effect.

Goldstein and this author (Goldstein and Taleb 2007) showed how professional researchers and practitioners substitute $\|x\|_1$ for $\|x\|_2$). The common result is underestimating the randomness of the estimator M, in other words read too much into it. Standard deviation is ususally explained and interpreted as mean deviation. Simply, people find it easier to imagine that a variation of, say, (-5,+10,-4,-3, 5, 8) in temperature over successive day needs to be mentally estimated by squaring the numbers, averaging them, then taking square roots. Instead they just average the absolutes. But, what is key, they tend to do so while convincing themselves that they are using standard deviations.

There is worse. Mindless application of statistical techniques, without knowledge of the conditional nature of the claims are widespread. But mistakes are often elementary, like lectures by parrots repeating " *N* of 1" or " *p*", or "do you have evidence of?", etc. Many social scientists need to have a clear idea of the difference between science and journalism, or the one between rigorous empiricism and anecdotal statements. Science is not about making claims about a sample, but using a sample to make general claims and discuss properties that apply outside the sample.

Take *M'* (short for $M_T^X(A,f)$) the estimator we saw above from the realizations (a sample path) for some process, and *M** the "true" mean that would emanate from knowledge of the generating process for such variable. When someone announces: "The crime rate in NYC dropped between 2000 and 2010", the claim is limited *M'* the observed mean, not *M** the true mean, hence the claim can be deemed merely journalistic, not scientific, and journalists are there to report "facts" not theories. No scientific and causal statement should be made from *M'* on "why violence has dropped" unless one establishes a link to *M** the true mean. *M* cannot be deemed "evidence" by itself. Working with *M'* alone cannot be called "empiricism".

What we just saw is at the foundation of statistics (and, it looks like, science). Bayesians disagree on how *M'* converges to *M**, etc., never on this point. From his statements in a dispute with this author concerning his claims about the stability of modern times based on the mean casualy in the past (Pinker, 2011), Pinker seems to be aware that *M'* may have dropped over time (which is a straight equality) and sort of perhaps we might not be able to make claims on *M** which might not have really been dropping.

In some areas not involving time series, the differnce between *M'* and *M** is negligible. So I rapidly jot down a few rules before showing proofs and derivations (limiting *M'* to the arithmetic mean, that is, M'= $M_T^X((-\infty,\infty),x)$).

Note again that $\mathbb{E}$ is the expectation operator under "real-world" probability measure $\mathbb{P}$.

6.2 The Tails Sampling Property

E[| M'- M*|] increases in with fat-tailedness (the mean deviation of M* seen from the realizations in different samples of the same process). In other words, fat tails tend to mask the distributional properties. This is the immediate result of the problem of convergence by the law of large numbers.

6.2.1 On the difference between the initial (generator) and the "recovered" distribution

{Explanation of the method of generating data from a known distribution and comparing realized outcomes to expected ones}

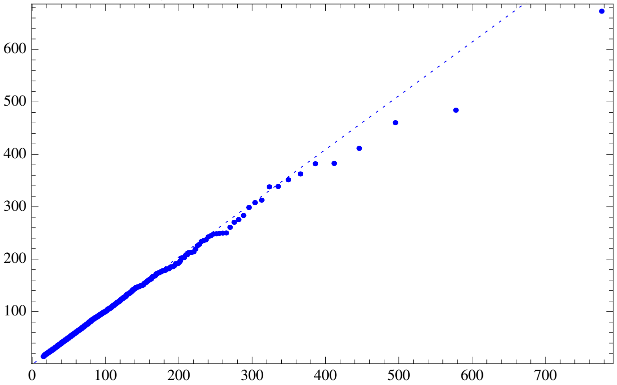


Figure 6.1: Q-Q plot" Fitting extreme value theory to data generated by its own process , the rest of course owing to sample insuficiency for extremely large values, a bias that typically causes the underestimation of tails, as the reader can see the points tending to fall to the right.

6.2.2 Case Study: Pinker (2011) Claims On The Stability of the Future Based on Past Data

When the generating process is power law with low exponent, plenty of confusion can take place.

For instance, Pinker(2011) claims that the generating process has a tail exponent ∼1.16 but made the mistake of drawing quantitative conclusions from it *about the mean from M'* and built *theories about drop in the risk* of violence that is contradicted by the data he was showing, since **fat tails plus negative skewness/asymmetry= hidden and underestimated risks of blowup**. His study is also missing the Casanova problem (next point) but let us focus on the error of being fooled by the mean of fat-tailed data.

The next two figures show the realizations of two subsamples, one before, and the other after the turkey problem, illustrating the inability of a set to naively deliver true probabilities through calm periods.

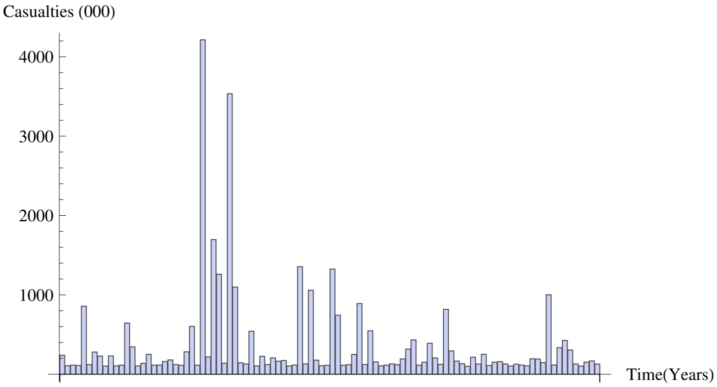


Figure 6.2: First 100 years (Sample Path):

A Monte Carlo generated realization of a process for casualties from violent conflict of the "80/20 or 80/02 style", that is tail exponent α= 1.15

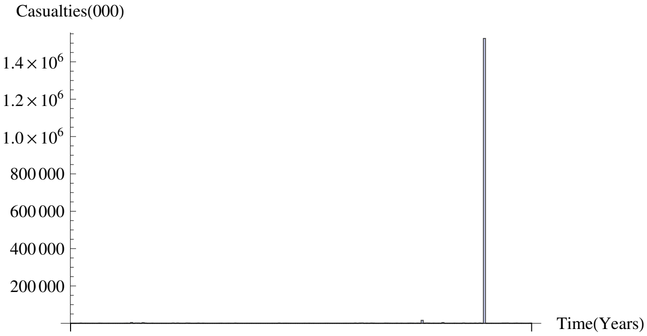


Figure 6.3: The Turkey Surprise: Now 200 years, the second 100 years dwarf the first; these are realizations of the exact same process, seen with a longer window and at a different scale.

The next simulations shows M1, the mean of casualties over the first 100 years across 10^4sample paths, and M2 the mean of casualties over the next 100 years.

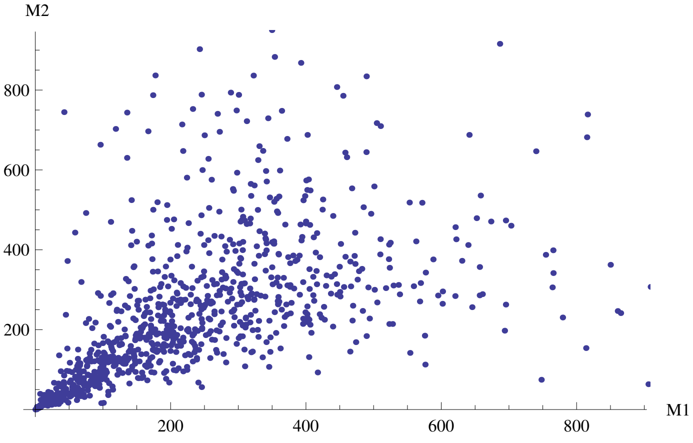


Figure 6.4: Does the past mean predict the future mean? Not so. M1 for 100 years,M2 for the next century. Seen at a narrow scale.

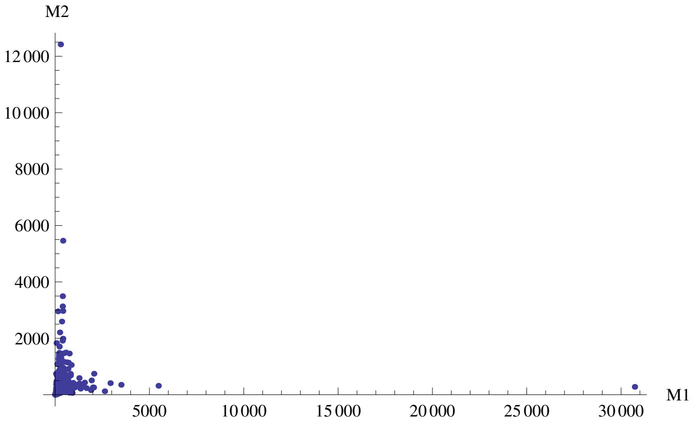


Figure 6.5: Does the past mean predict the future mean? Not so. M1 for 100 years,M2 for the next century. Seen at a wider scale.

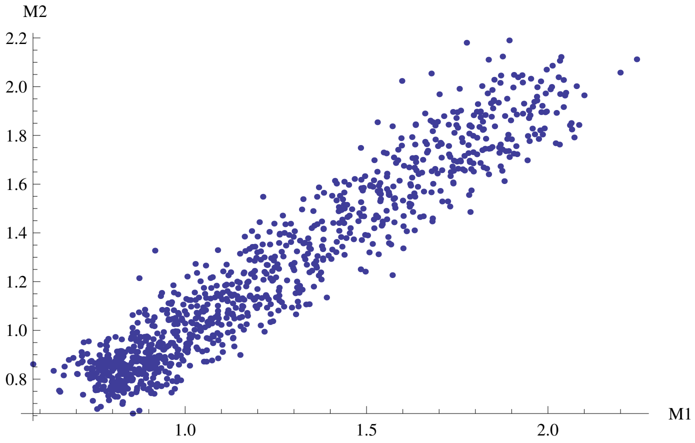


Figure 6.6: The same seen with a thin-tailed distribution.

So clearly it is a lunacy to try to read much into the mean of a power law with 1.15 exponent (and this is the mild case, where we *know* the exponent is 1.15. Typically we have an error rate, and the metaprobability discussion in Chapter x will show the exponent to be likely to be lower because of the possibility of error).

6.2.3 Claims Made From Power Laws

The Cederman graph – figure **??** shows exactly how *not* to make claims upon observing power laws.

6.3 A discussion of the Paretan 80/20 Rule

Next we will see how when one hears about the Paretan 80/20 "rule" (or, worse, "principle"), it is likely to underestimate the fat tails effect outside some narrow domains. It can be more like 95/20 or even 99.9999/.0001, or eventually 100/ϵ. Almost all economic reports applying power laws for "GINI" (Chapter x) or inequality miss the point. Even Pareto himself miscalibrated the rule.

As a heuristic, it is always best to assume underestimation of tail measurement. Recall that we are in a one-tailed situation, hence a likely underestimation of the mean.

Where does this 80/20 business come from?

Assume α the power law tail exponent, and an exceedant probability $P_{X>x} = x_{\min}\, x^{-\alpha}$, x $\in(x_{\min}, \infty)$. Simply, the top p of the population gets $S = p^{\frac{\alpha-1}{\alpha}}$ of the share of the total pie.

$$\alpha = \frac{\log(p)}{\log(p) - \log(S)}$$

which means that the exponent will be 1.161 for the 80/20 distribution.

Note that as α gets close to 1 the contribution explodes

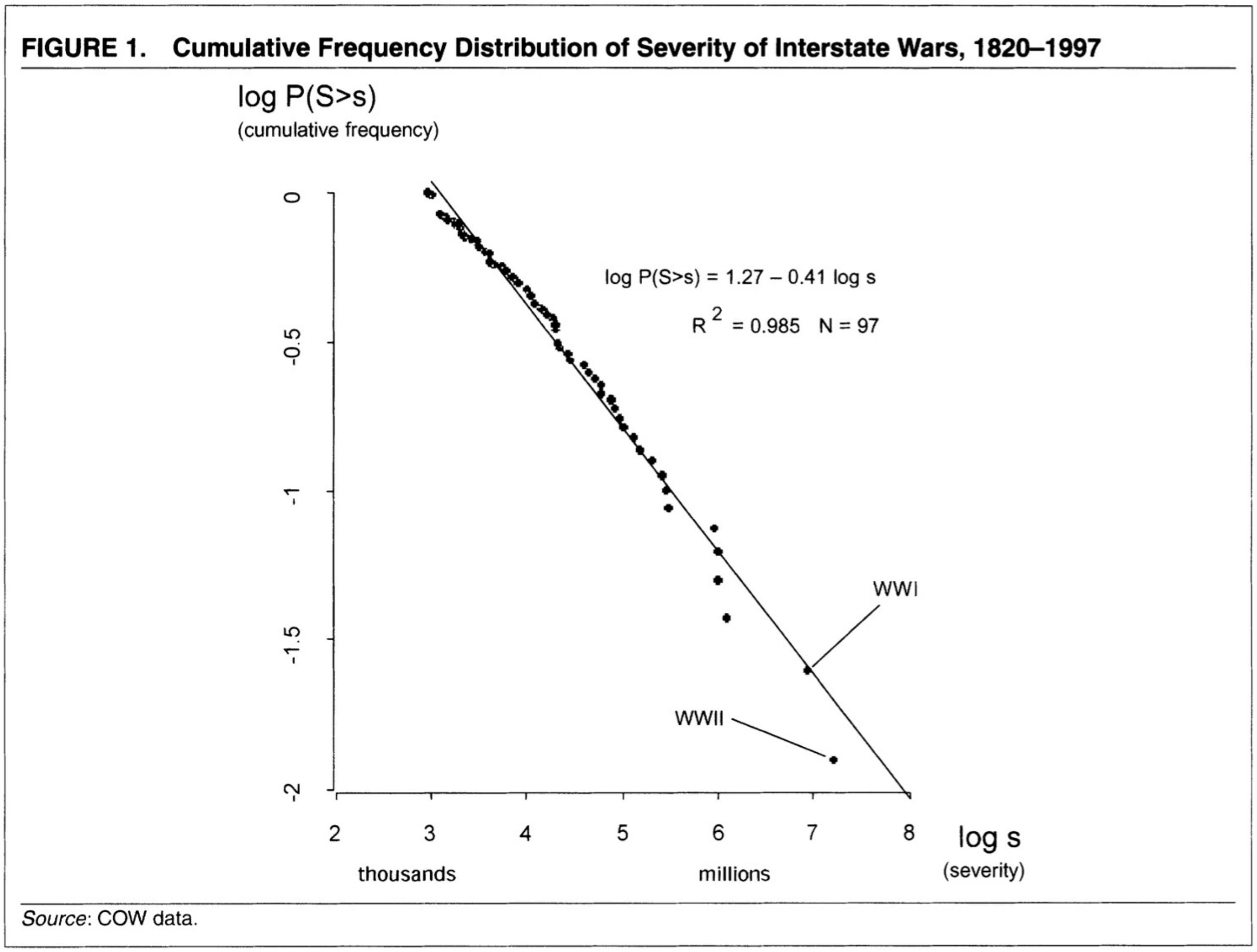


FIGURE 1. Cumulative Frequency Distribution of Severity of Interstate Wars, 1820–1997

Source: COW data.

Figure 6.7: Cederman 2003, used by Pinker. I wonder if I am dreaming or if the exponent α is really = .41. Chapters x and x show why such inference is centrally flawed, since low exponents do not allow claims on mean of the variableexcept to say that it is very, very high and not observable in finite samples. Also, in addition to wrong conclusions from the data, take for now that the regression fits the small deviations, not the large ones, and that the author overestimates our ability to figure out the asymptotic slope.

as it becomes close to infinite mean.

Derivation:

Start with the standard density $f(x) = x_{\min}^{\alpha} \alpha \, x^{-\alpha-1}$, $x \geq x_{\min}$.
1) The Share attributed above K, $K \geq x_{\min}$, becomes

$$\frac{\int_K^\infty x f(x)\, dx}{\int_{x_{\min}}^\infty x f(x)\, dx} = K^{1-\alpha}$$

2) The probability of exceeding K,

$$\int_K^\infty f(x) dx = K^{-\alpha}$$

3) Hence $K^{-\alpha}$ of the population contributes $K^{1-\alpha}$=$p^{\frac{\alpha-1}{\alpha}}$ of the result

6.3.1 Why the 80/20 Will Be Generally an Error: The Problem of In-Sample Calibration

Vilfredo Pareto figured out that 20% of the land in Italy was owned by 80% of the people, and the reverse. He later observed that 20 percent of the peapods in his garden yielded 80 percent of the peas that were harvested. He might have been right about the peas; but most certainly wrong about the land.

For fitting in-sample frequencies for a power law does not yield the proper "true" ratio since the sample is likely to be insufficient. One should fit a powerlaw using extrapolative, not interpolative techniques, such as methods based on Log-Log plotting or regressions. These latter methods are more informational, though with a few caveats as they can also suffer from sample insufficiency.

Data with infinite mean, $\alpha \leq$1, will masquerade as finite variance *in sample* and show about 80% contribution to the top 20% quantile. In fact you are expected to witness in finite samples a lower contribution of the top 20%/

Let us see. Generate *m* samples of α =1 data X_j=$\{x_{i,j}\}_{i=1}^n$, ordered $x_{i,j}\geq$ $x_{i-1,j}$, and examine the distribution of the top ν contribution Z_j^ν=$\frac{\sum_{i\leq \nu n} x_j}{\sum_{i\leq n} x_j}$, with $\nu \in$ (0,1).

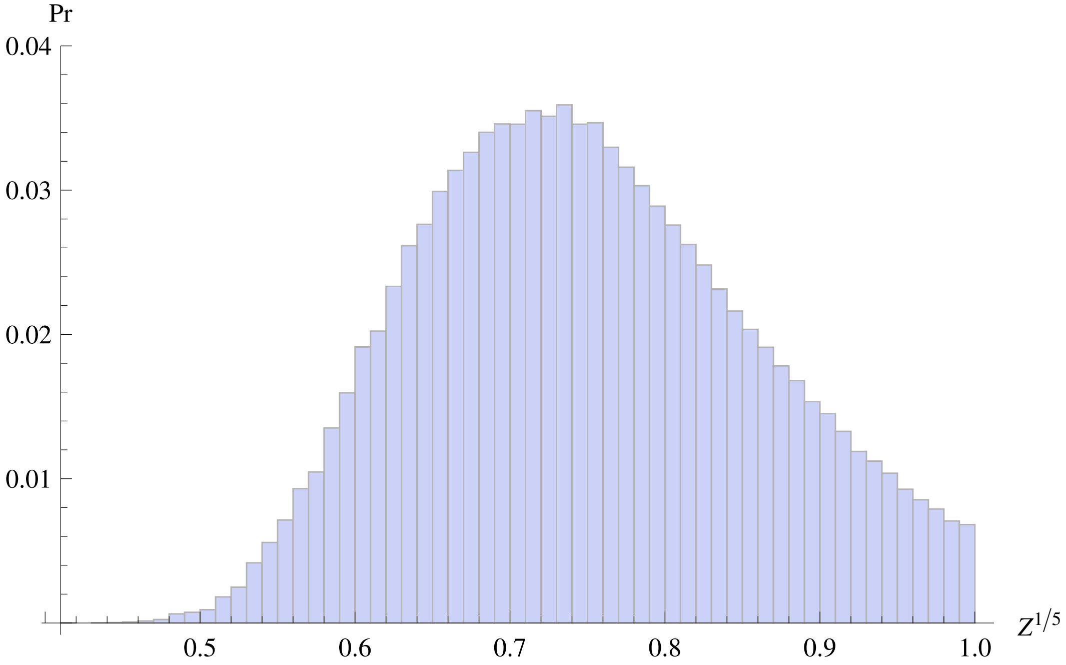


Figure 6.8: The difference betwen the generated (ex ante) and recovered (ex post) processes; ν=20/100, N= 10^7. Even when it should be .0001/100, we tend to watch an average of 75/20

6.4 Survivorship Bias (Casanova) Property

$mathbf E(M' - M*)$ increases under the presence of an absorbing barrier for the process. This is the Casanova effect, or fallacy of silent evidence see *The Black Swan,* Chapter 8. (**Fallacy of silent evidence**: Looking at history, we do not see the full story, only the rosier parts of the process, in the Glossary)

History is a single sample path we can model as a Brownian motion, or something similar with fat tails (say Levy flights). What we observe is one path among many "counterfactuals", or alternative histories. Let us call each one a "sample path", a succession of discretely observed states of the system between the initial state S_0 and S_T the present state.

Arithmetic process: We can model it as $S(t) = S(t-\Delta\mathrm{t}) + Z_{\Delta\mathrm{t}}$ where $Z_{\Delta\mathrm{t}}$ is noise drawn from any distribution.

Geometric process: We can model it as $S(t) = S(t-\Delta\mathrm{t})e^{W_t}$ typically $S(t-\Delta\mathrm{t})e^{\mu\Delta\mathrm{t}+s\sqrt{\Delta\mathrm{t}}Z_t}$ but W_t can be noise drawn from any distribution. Typically, $\log\left(\frac{S(t)}{S(t-i\Delta\mathrm{t})}\right)$ is treated as Gaussian, but we can use fatter tails. The convenience of the Gaussian is stochastic calculus and the ability to skip steps in the process, as S(t)=S(t-Δt)$e^{\mu\Delta\mathrm{t}+s\sqrt{\Delta\mathrm{t}}W_t}$, with W_t ∼N(0,1), works for all Δt, even allowing for a single period to summarize the total.

The Black Swan made the statement that history is more rosy than the "true" history, that is, the mean of the ensemble of all sample path.

Take an absorbing barrier H as a level that, when reached, leads to extinction, defined as becoming unobservable or unobserved at period *T*.

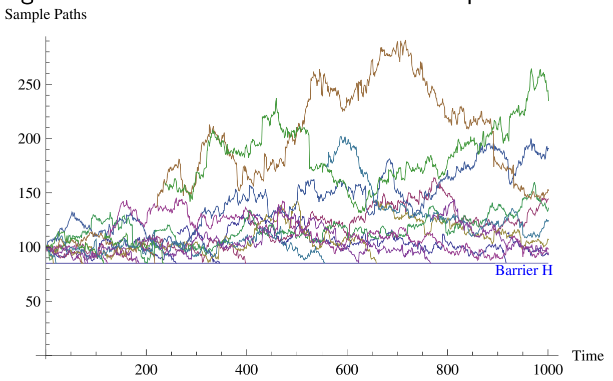


Table 6.1: Counterfactual historical paths subjected to an absorbing barrier.

When you observe history of a family of processes subjected to an absorbing barrier, i.e., you see the winners not the losers, there are biases. If the survival of the entity depends upon not hitting the barrier, then one cannot compute the probabilities along a certain sample path, without adjusting.

Begin

The "true" distribution is the one for all sample paths, the "observed" distribution is the one of the succession of points $\{S_{\mathrm{i}\Delta\mathrm{t}}\}_{i=1}^{T}$.

Bias in the measurement of the mean

In the presence of an absorbing barrier H "below", that is, lower than S_0, the "observed mean" ⩾ "true mean"

Bias in the measurement of the volatility

The "observed" variance (or mean deviation) ⩽ "true" variance

The first two results are well known (see Brown, Goetzman and Ross (1995)). What I will set to prove here is that fat-tailedness increases the bias.

First, let us pull out the "true" distribution using the reflection principle.

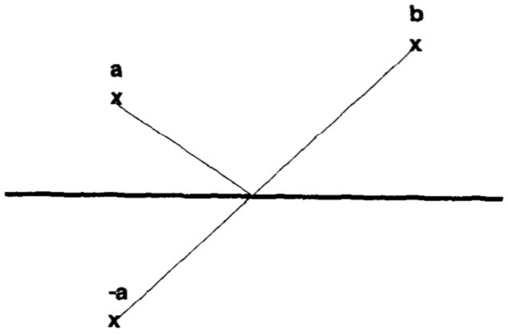


Figure 19.23 The reflection principle.

Table 6.2: ***The reflection principle*** *(graph from Taleb, 1997). The number of paths that go from point a to point b without hitting the barrier H is equivalent to the number of path from the point - a (equidistant to the barrier) to b.*

Thus if the barrier is *H* and we start at S_0then we have two distributions, one f(S), the other f(S-2(S_0-H))

By the reflection principle, the "observed" distribution $p(S)$ becomes:

$$p(S) = \begin{cases} f(S) - f\left(S - 2\left(S_0 - H\right)\right) & \text{if } S > H \\ 0 & \text{if } S < H \end{cases}$$

Simply, the nonobserved paths (the casualties "swallowed into the bowels of history") represent a mass of 1-$\int_H^\infty f(S) - f\left(S - 2\left(S_0 - H\right)\right) dS$ and, clearly, it is in this mass that all the hidden effects reside. We can prove that the missing mean is $\int_\infty^H S\left(f(S) - f\left(S - 2\left(S_0 - H\right)\right)\right) dS$ and perturbate $f(S)$ using the previously seen method to "fatten" the tail.

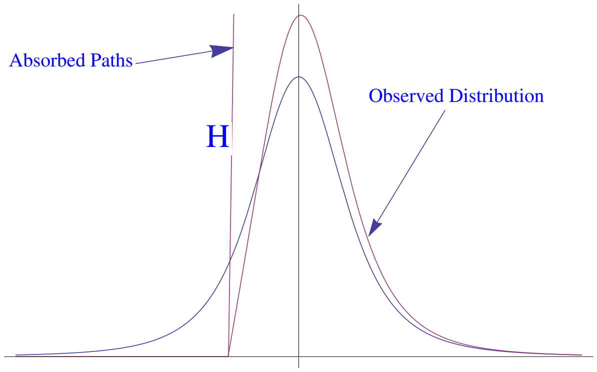


Table 6.3: If you don't take into account the sample paths that hit the barrier, the observed distribution seems more positive, and more stable, than the "true" one.

The interest aspect of the absorbing barrier (from below) is that it has the same effect as insufficient sampling of a left-skewed distribution under fat tails. The mean will look better than it really is.

6.5 Left (Right) Tail Sample Insufficiency Under Negative (Positive) Skewness

E[*M'*- *M**] increases (decreases) with negative (positive) skeweness of the true underying variable.

Some classes of payoff (those affected by Turkey problems) show better performance than "true" mean. Others (entrepreneurship) are plagued with in-sample underestimation of the mean. A naive measure of a sample mean, even without absorbing barrier, yields a higher oberved mean than "true" mean when the distribution is skewed to the left, and lower when the skewness is to the right.

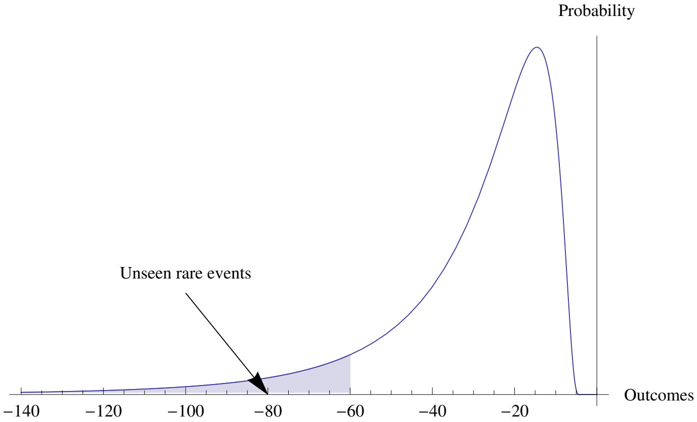


Figure 6.9: The left tail has fewer samples. The probability of an event falling below K in n samples is F(K), where F is the cumulative distribution.

This can be shown analytically, but a simulation works well.

To see how a distribution masks its mean because of sample insufficiency, take a skewed distribution with fat tails, say the standard Pareto Distribution we saw earlier.

The "true" mean is known to be *m*=$\frac{\alpha}{\alpha-1}$. Generate $\{X_{1,j},\ X_{2,j},\ ...,X_{N,j}\}$ random samples indexed by *j* as a designator of a certain history j. Measure $\mu_j = \frac{\sum_{i=1}^{N} X_{i,j}}{N}$. We end up with the sequence of various sample means $\{\mu_j\}_{j=1}^{T}$, which naturally should converge to M with both *N* and *T*. Next we calculate $\tilde{\mu}$ the median value of $\sum_{j=1}^{T} \frac{\mu_j}{M*T}$, such that P>$\tilde{\mu}$ $=\frac{1}{2}$ where, to repeat, M* is the theoretical mean we expect from the generating distribution.

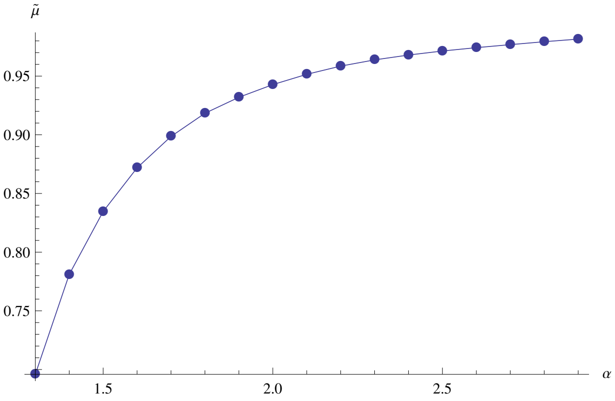


Figure 6.10: Median of $\sum_{j=1}^{T} \frac{\mu_j}{MT}$ in simulations (10^6 Monte Carlo runs). We can observe the underestimation of the mean of a skewed power law distribution as α exponent gets lower. Note that lower α imply fatter tails.

Entrepreneurship is penalized by right tail insufficiency making performance look worse than it is. Figures 0.1 and 0.2 can be seen in a symmetrical way, producing the exact opposite effect of negative skewness.

6.6 Why N=1 Can Be Very, Very Significant Statistically

The Power of Extreme Deviations: Under fat tails, large deviations from the mean are vastly more informational than small ones. They are not "anecdotal". (The last two properties corresponds to the black swan problem, inherently asymmetric).

We saw the point earlier (with the masquerade problem) in **??.??**. The gist is as follows, worth repeating and applying to this context.

A thin-tailed distribution is less likely to deliver a single large deviation than a fat tailed distribution a series of long calm periods. Now add negative skewness to the issue, which makes large deviations negative and small deviations positive, and a large *negative* deviation, under skewness, becomes extremely informational.

Mixing the arguments of **??.??** and **??.??** we get:

> Asymmetry in Inference: Under both negative skewness and fat tails, negative deviations from the mean are more informational than positive deviations.

6.7 The Instability of Squared Variations in Regression Analysis

Probing the limits of a standardized method by arbitrage. We can easily arbitrage a mechanistic method of analysis by generating data, the properties of which are known by us, which we call "true" properties, and comparing these "true" properties to the properties revealed by analyses, as well as the confidence of the analysis about its own results in the form of "p-values" or other masquerades.

This is no different from generating random noise and asking the "specialist" for an analysis of the charts, in order to test his knowledge, and, even more importantly, asking him to give us *a probability of his analysis being wrong*. Likewise, this is equivalent to providing a literary commentator with randomly generated giberish and asking him to provide comments. In this section we apply the technique to regression analyses, a great subject of abuse by the social scientists, particularly when ignoring the effects of fat tails.

In short, we saw the effect of fat tails on higher moments. We will start with 1) an extreme case of infinite mean (in which we know that the conventional regression analyses break down), then generalize to 2) situations with finite mean (but finite variance), then 3) finite variance but infinite higher moments. Note that except for case 3, these results are "sort of" standard in the econometrics literature, except that they are ignored away through tweaking of the assumptions.

Fooled by α=1

Assume the simplest possible regression model, as follows. Let y_i= β_0 + β_1 x_i + *s* z_i, with Y=$\{y_i\}_{1<i\leq n}$ the set of *n* dependent variables and X= $\{x_i\}_{1<i\leq n}$, the independent one; Y, X ϵ $\mathbb{R}$, i ϵ $\mathbb{N}$. The errors z_i are independent but drawn from a standard Cauchy (symmetric, with tail exponent α =1), multiplied by the amplitude or scale *s*; we will vary *s* across the thought experiment (recall that in the absence and variance and mean deviation we rely on *s* as a measure of dispersion). Since all moments are infinite, $\mathbb{E}[z_i^n] = \infty$ for all *n*≥1, we know *ex ante* that the noise is such that the "errors" or 'residuals" have infinite means and variances –but the problem is that in finite samples the property doesn't show. The sum of squares will be finite.

The next figure shows the effect of a very expected large deviation, as can be expected from a Cauchy jump.

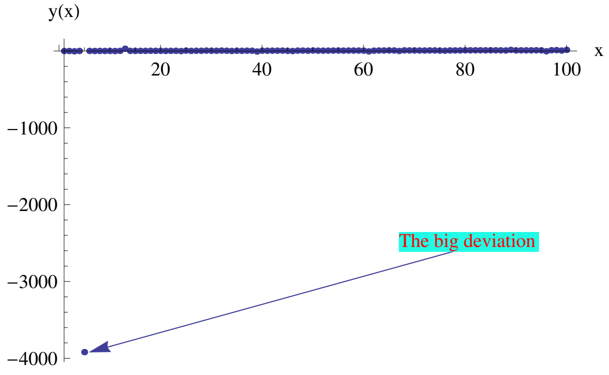


Figure 6.11: A sample regression path dominated by a large deviation. Most samples don't exhibit such deviation this, which is a problem. We know that with certainty (an application of the zero-one laws) that these deviations are certain as $n \to \infty$, so if one pick an arbitrarily large deviation, such number will be exceeded, with a result that can be illustrated as ***the sum of all variations will come from a single large deviation****.*

Next we generate T simulations (indexed by j) of n pairs $\{y_i, x_i\}_{1<i\leq n}$ for increasing values of x, thanks to Cauchy distributed variables variable $z^{\alpha}_{i,j}$ and multiplied $z^{\alpha}_{i,j}$ by the scaling constant s, leaving us with a set $\left\{\left\{\beta_0+\beta_1 x_i+s z^{\alpha}_{i,j}\right\}_{i=1}^{n}\right\}_{j=1}^{T}$. Using standard regression techniques of estimation we "regress" and obtain the standard equation $Y^{\text{est}}=\beta_0^{\text{est}}+X\beta_1^{\text{est}}$, where Y^{est} is the estimated Y, and E a vector of unexplained residuals E$\equiv\{\epsilon_{i,j}\}\equiv\left\{\left\{y^{\text{est}}_{i,j}-\beta_0^{\text{est}}-\beta_1^{\text{est}}x_{ij}\right\}_{i=1}^{n}\right\}_{j=1}^{T}$. We thus obtain T simulated values of $\rho\equiv\{\rho_j\}_{j=1}^{T}$, where ρ_j≡1- $\frac{\sum_{i=1}^{n}{\epsilon_{i,j}}^2}{\sum_{i=1}^{n}(y_{i,j}-\overline{y_j})^2}$, the R-square for a sample run j, where $\overline{y_j}=\frac{1}{n}\sum_{i=1}^{n}y_{i,j}$, in other words 1- (squared residuals) / (squared variations). We examine the distribution of the different realizations of ρ.

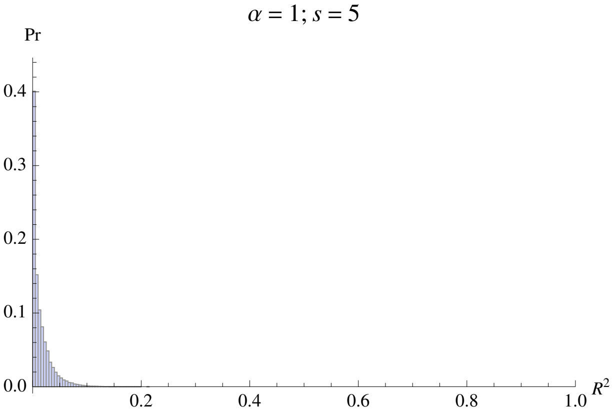


Figure 6.12: The histograms showing the distribution of R Squares; $T=10^6$ simulations.The "true" R-Square should be 0. High scale of noise.

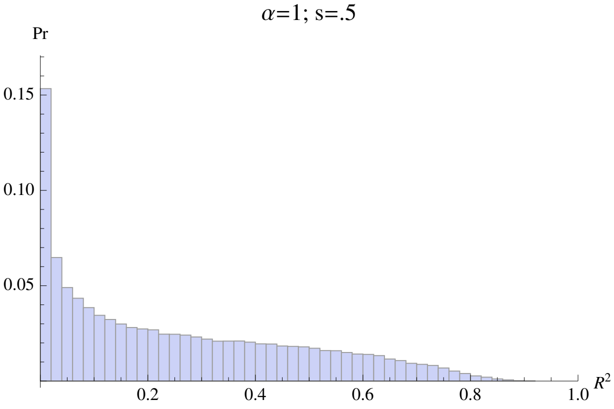


Figure 6.13: The histograms showing the distribution of R Squares; $T=10^6$ simulations.The "true" R-Square should be 0. Low scale of noise.

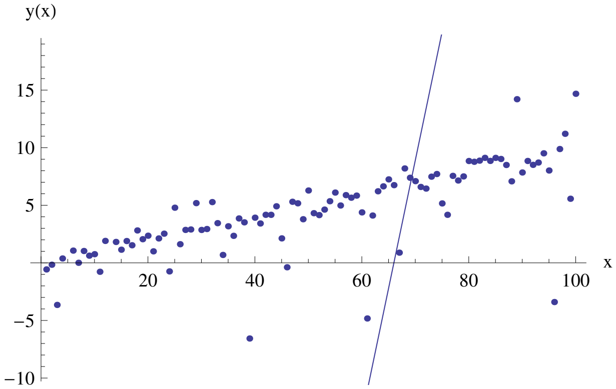


Figure 6.14: We can fit different regressions to the same story (which is no story). A regression that tries to accommodate the large deviation.

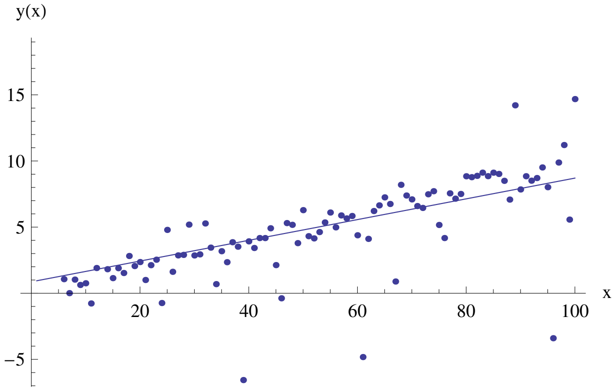


Figure 6.15: Missing the largest deviation (not necessarily voluntarily: the sample doesn't include the critical observation.

Arbitraging metrics

For a sample run which, typically, will not have a large deviation,
R-squared: 0.994813 (When the "true" R-squared would be 0)
The P-values are monstrously misleading.

	Estimate	Standard Error	t-Statistic	P-Value
1	4.99	0.417	11.976	$7.8 \times 10^{-}33$
x	0.10	0.00007224	1384.68	$9.3 \times 10^{-}11426$

6.7.1 Application to Economic Variables

We saw in **??.??** that kurtosis can be attributable to 1 in 10,000 observations (>50 years of data), meaning it is unrigorous to assume anything other than that the data has "infinite" kurtosis. The implication is that even if the squares exist, i.e., $\mathbb{E}[z_i^2] < \infty$, the distribution of z_i^2 has infinite variance, and is massively unstable. The "P-values" remain grossly miscomputed. The next graph shows the distribution of ρ across samples.

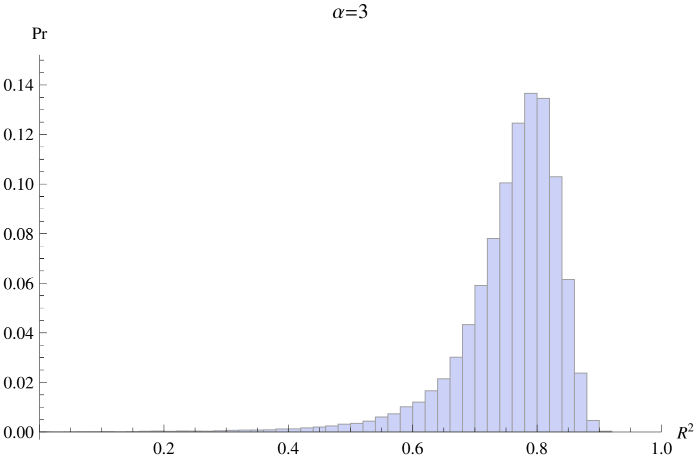


Figure 6.16: Finite variance but infinite kurtosis.

6.8 Statistical Testing of Differences Between Variables

A pervasive attribute substitution: Where X and Y are two random variables, the properties of X-Y, say the variance, probabilities, and higher order attributes are markedly different from the difference in properties. So $\mathbb{E}(X) - \mathbb{E}(Y) = \mathbb{E}(X) - \mathbb{E}(Y)$ but of course, $Var(X - Y) \neq Var(X) - Var(Y)$, etc. for higher norms. It means that P-values are different, and of course the coefficient of variation ("Sharpe"). Where σ is the Standard deviation of the variable (or sample):

$$\frac{\mathbb{E}(X-Y)}{\sigma(X-Y)} \neq \frac{\mathbb{E}X)}{\sigma(X)} - \frac{\mathbb{E}(Y))}{\sigma(Y)}$$

In *Fooled by Randomness* (2001):

> A far more acute problem relates to the outperformance, or the comparison, between two or more persons or entities. While we are certainly fooled by randomness when it comes to a single times series, the foolishness is compounded when it comes to the comparison between, say, two people, or a person and a benchmark. Why? Because both are random. Let us do the following simple thought experiment. Take two individuals, say, a person and his brother-in-law, launched through life. Assume equal odds for each of good and bad luck. Outcomes: lucky-lucky (no difference between them), unlucky-unlucky (again, no difference), lucky- unlucky (a large difference between them), unlucky-lucky (again, a large difference).

Ten years later (2011) it was found that 50% of neuroscience papers (peer-reviewed in "prestigious journals") that compared variables got it wrong.

> In theory, a comparison of two experimental effects requires a statistical test on their difference. In practice, this comparison is often based on an incorrect procedure involving two separate tests in which researchers conclude that effects differ when one effect is significant (P < 0.05) but the other is not (P > 0.05). We reviewed 513 behavioral, systems and cognitive neuroscience articles in five top-ranking journals (Science, Nature, Nature Neuroscience, Neuron and The Journal of Neuroscience) and found that 78 used the correct procedure and 79 used the incorrect procedure. An additional analysis suggests that incorrect analyses of interactions are even more common in cellular and molecular neuroscience.

In Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature neuroscience, 14(9), 1105-1107.

Fooled by Randomness was read by many professionals (to put it mildly); the mistake is still being made.

Ten years from now, they will still be making the mistake.

6.9 Studying the Statistical Properties of Binaries and Extending to Vanillas

See discussion in Chapter x.

6.10 The Mother of All Turkey Problems: How Economics Time Series Econometrics and Statistics Don't Replicate

(Debunking a Nasty Type of PseudoScience)

Something Wrong With Econometrics, as Almost All Papers Don't Replicate. The next two reliability tests, one about parametric methods the other about robust statistics, show that there is something wrong in econometric methods, fundamentally wrong, and that the methods are not dependable enough to be of use in anything remotely related to risky decisions.

6.10.1 Performance of Standard Parametric Risk Estimators, $f(x) = x^n$ (Norm $\mathcal{L}2$)

With economic variables one single observation in 10,000, that is, one single day in 40 years, can explain the bulk of the "kurtosis", a 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. The same problem is found with interest and exchange rates, commodities, and other variables. 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 Euclidian, or $\mathcal{L}^2$ 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.

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 Taleb (2009), using Log returns,

$$X_t \equiv \log \left(\frac{P(t)}{P(t - i\Delta\mathrm{t})} \right)$$

Take the measure $M_t^X \left((-\infty, \infty), X^4 \right)$ of the fourth noncentral moment

$$M_t^X \left((-\infty, \infty), X^4 \right) \equiv \frac{1}{n} \sum_{i=0}^{n} X_{t-i\Delta t}^4$$

and the n-sample maximum quartic observation Max$({X_{t-i\Delta\mathrm{t}}}^4)_{i=0}^n$. $Q(n)$ is the contribution of the maximum quartic variations over n samples.

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

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 ± .0028. Visibly we can see that the distribution 4^{th} moment has the property

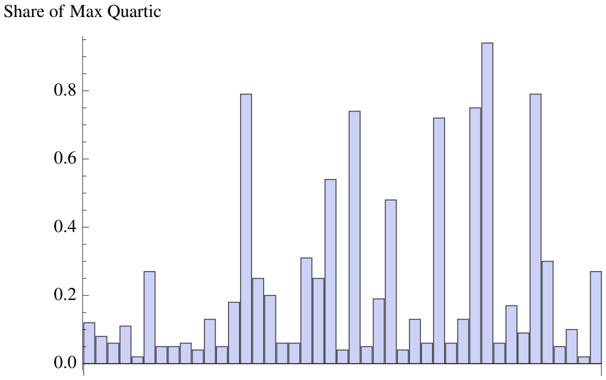
$$P\left(X > \max(x_i^4)_{i \leq 2 \leq n}\right) \approx P\left(X > \sum_{i=1}^n x_i^4\right)$$

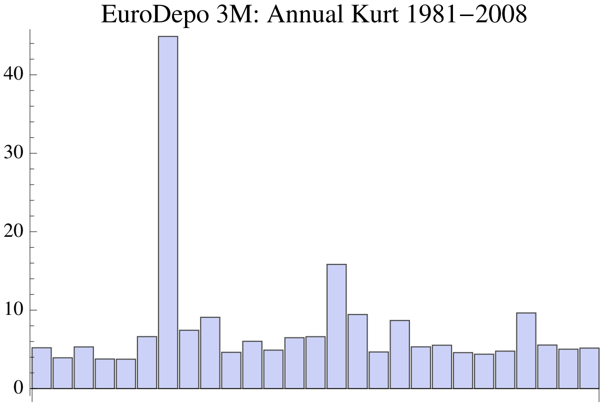
Recall that, naively, the fourth moment expresses the stability of the second moment. And the second moment expresses the stability of the measure across samples.

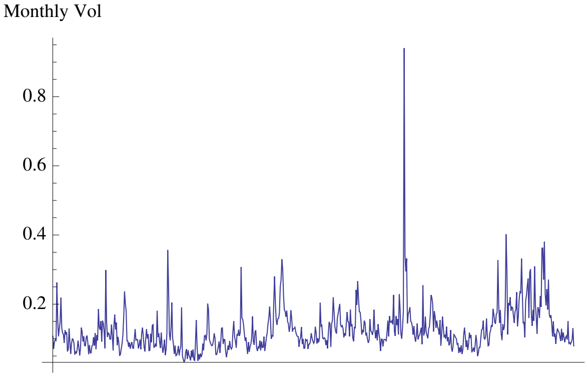
Security	**Max Q**	**Years**.
Silver	0.94	46.
SP500	0.79	56.
CrudeOil	0.79	26.
Short Sterling	0.75	17.
Heating Oil	0.74	31.
Nikkei	0.72	23.
FTSE	0.54	25.
JGB	0.48	24.
Eurodollar Depo 1M	0.31	19.
Sugar #11	0.3	48.
Yen	0.27	38.
Bovespa	0.27	16.
Eurodollar Depo 3M	0.25	28.
CT	0.25	48.
DAX	0.2	18.

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).







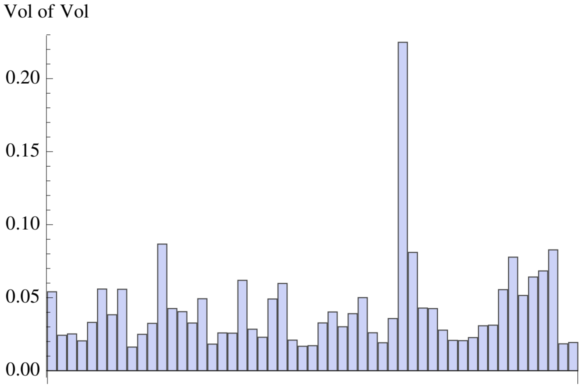


Figure 6.17: Montly delivered volatility in the SP500 (as measured by standard deviations). The only structure it seems to have comes from the fact that it is bounded at 0. This is standard.

Figure 6.18: Montly volatility of volatility from the same dataset, predictably unstable.

6.10.2 Performance of Standard Non-Parametric Risk Estimators, f(x)= x or |x| (Norm $\mathcal{L}1$), A =(-∞, K]

Does the past resemble the future in the tails? The following tests are nonparametric, that is entirely based on empirical probability distributions.

Figure 6.19: Comparing M[t-1, t] and M[t,t+1], where τ= 1year, 252 days, for macroeconomic data using extreme deviations, A= (-∞ ,-2 standard deviations (equivalent)], f(x) = x (replication of data from The Fourth Quadrant, Taleb, 2009)

Figure 6.20: The "regular" is predictive of the regular, that is mean deviation. Comparing M[t] and M[t+1 year] for macroeconomic data using regular deviations, A= (-∞ ,∞) , f(x)= |x|

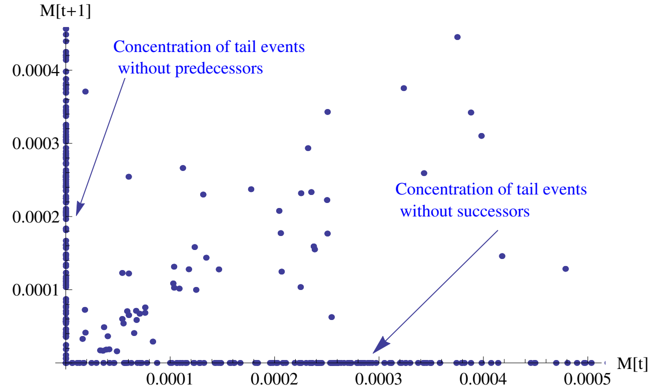


Figure 6.21: This are a lot worse for large deviations A= (-∞ ,-4 standard deviations (equivalent)], f(x) = x

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.

The Responses so far by members of the economics/econometrics establishment: "his books are too popular to merit attention", "nothing new" (sic), "egomaniac" (but I was told at the National Science Foundation that "egomaniac" does not apper to have a clear econometric significance). No answer as to why they still use STD, regressions, GARCH, value-at-risk and similar methods.

Peso problem: Note that many researchers invoke "outliers" or "peso problem" as acknowledging fat tails, yet ignore them analytically (outside of Poisson models that we will see are not possible to calibrate except after the fact). 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 unusable.

Peso problem attitude towards 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 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. The comment also shows the lack of common sense to look for robustness to these events).

Lack of Skin in the Game. Indeed one wonders why econometric methods can be used while being wrong, so shockingly wrong, how "University" researchers (adults) can partake of such a scam. Basically they 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,

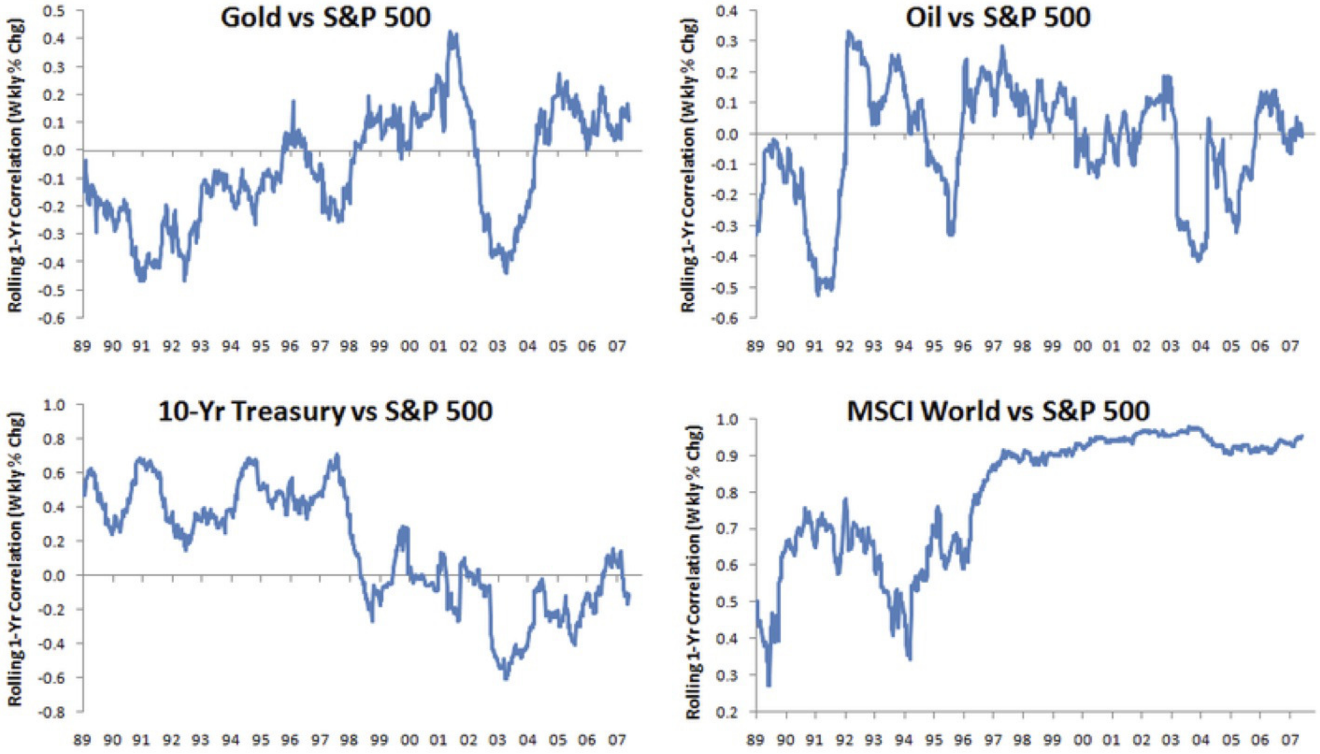


Figure 6.22: Correlations are also problematic, which flows from the instability of single variances and the effect of multiplication of the values of random variables.

"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 scam should continue perpetually so long as people are allowed to harm others with impunity. (More in Taleb and Sandis, 2013)

6.11 A General Summary of The Problem of Reliance on Past Time Series

The four aspects of what we will call the nonreplicability issue, particularly for mesures that are in the tails. These are briefly presented here and developed more technically throughout the book:

a- **Definition of statistical rigor (or Pinker Problem).** The idea that an estimator is not about fitness to past data, but related to how it can capture future realizations of a process seems absent from the discourse. Much of econometrics/risk management methods do not meet this simple point and the rigor required by orthodox, basic statistical theory.

b- **Statistical argument on the limit of knowledge of tail events**. Problems of replicability are acute for tail events. Tail events are impossible to price owing to the limitations from the size of the sample. Naively rare events have little data hence what estimator we may have is noisier.

c- **Mathematical argument about statistical decidability.** No probability without metaprobability. Metadistributions matter more with tail events, and with fat-tailed distributions.

1. The soft problem: we accept the probability distribution, but the imprecision in the calibration (or parameter errors) percolates in the tails.
2. The hard problem (Taleb and Pilpel, 2001, Taleb and Douady, 2009): We need to specify an *a priori* probability distribution from which we depend, or alternatively, propose a metadistribution with compact support.

3. Both problems are bridged in that a nested stochastization of standard deviation (or the scale of the parameters) for a Gaussian turn a thin-tailed distribution into a power law (and stochastization that includes the mean turns it into a jump-diffusion or mixed-Poisson).

   d- **Economic arguments**: The Friedman-Phelps and Lucas critiques, Goodhart's law. Acting on statistical information (a metric, a response) changes the statistical properties of some processes.

6.12 Conclusion

This chapter introduced the problem of "surprises" from the past of time series, and the invalidity of a certain class of estimators that seem to only work in-sample. Before examining more deeply the mathematical properties of fat-tails, let us look at some practical aspects.

D | On the Instability of Econometric Data

Table D.1: 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.

7 On the Difference between Binary Prediction and True Exposure

(With Implications For Forecasting Tournaments and Decision Making Research)

There are serious statistical differences between predictions, bets, and exposures that have a yes/no type of payoff, the "binaries", and those that have varying payoffs, which we call the "vanilla". Real world exposures tend to belong to the vanilla category, and are poorly captured by binaries. Yet much of the economics and decision making literature confuses the two. Vanilla exposures are sensitive to Black Swan effects, model errors, and prediction problems, while the binaries are largely immune to them. The binaries are mathematically tractable, while the vanilla are much less so. Hedging vanilla exposures with binary bets can be disastrous–and because of the human tendency to engage in attribute substitution when confronted by difficult questions,decision-makers and researchers often confuse the vanilla for the binary.

7.1 Binary vs Vanilla Predictions and Exposures

Binary: Binary predictions and exposures are about well defined discrete events, with yes/no types of answers, such as whether a person will win the election, a single individual will die, or a team will win a contest. We call them binary because the outcome is either 0 (the event does not take place) or 1 (the event took place), that is the set {0,1} or the set $\{a_L, a_H\}$, with $a_L < a_H$ any two discrete and exhaustive values for the outcomes. For instance, we cannot have five hundred people winning a presidential election. Or a single candidate running for an election has two exhaustive outcomes: win or lose.

Vanilla: "Vanilla" predictions and exposures, also known as natural random variables, correspond to situations in which the payoff is continuous and can take several values. The designation "vanilla" originates from definitions of financial contracts[1] ; it is fitting outside option trading because the exposures they designate are naturally occurring continuous variables, as opposed to the binary that which tend to involve abrupt institution-mandated discontinuities. The vanilla add a layer of complication: profits for companies or deaths due to terrorism or war can take many, many potential values. You can predict the company will be "profitable", but the profit could be $\$1 or 10 billion.$

There is a variety of exposures closer to the vanilla, namely bounded exposures that we can subsume mathematically into the binary category.

The main errors are as follows.

- Binaries always belong to the class of thin-tailed distributions, because of boundedness, while the vanillas don't. This means the law of large numbers operates very rapidly there. Extreme events wane rapidly in importance: for instance, as we will see further down in the discussion of the Chernoff bound, the probability of a series of 1000 bets to diverge more than 50% from the expected average is less than 1 in 10^{18}, while the vanilla can experience wilder fluctuations with a high probability, particularly in fat-tailed domains. Compar-

[1] The "vanilla" designation comes from option exposures that are open-ended as opposed to the binary ones that are called "exotic".

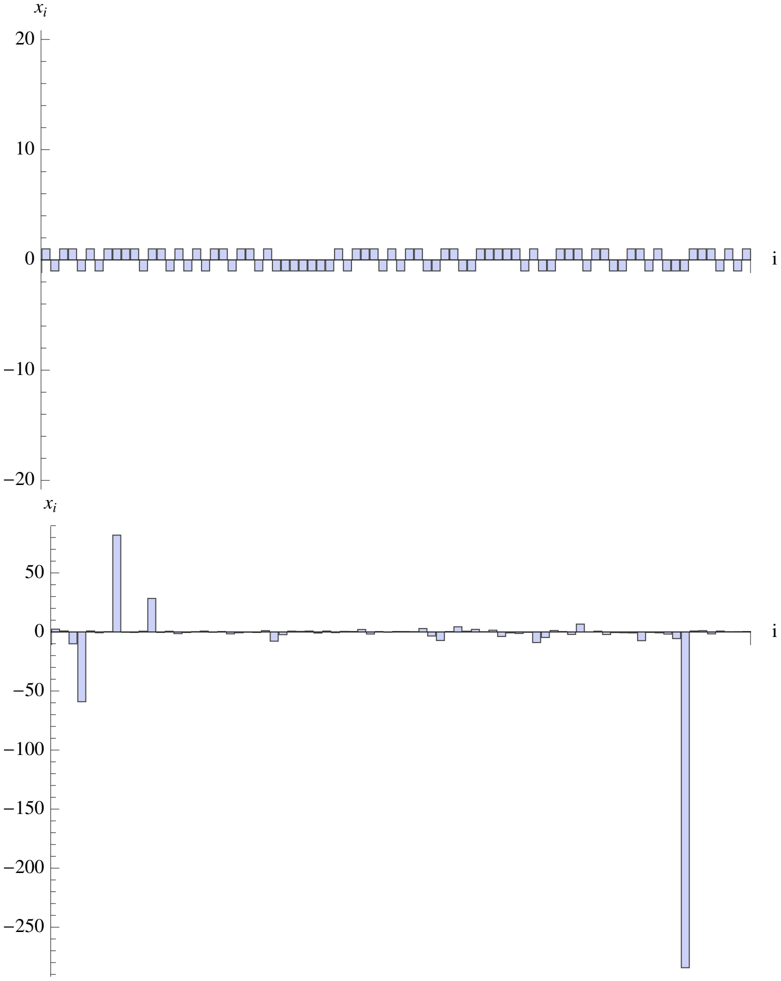


Figure 7.1: Comparing digital payoff (above) to the vanilla (below). The vertical payoff shows x_i $(x_1, x_2, ...)$ and the horizontal shows the index i= (1,2,...), as i can be time, or any other form of classification. We assume in the first case payoffs of {-1,1}, and open-ended (or with a very remote and unknown bounds) in the second.

ing one to another can be a lunacy.

- The research literature documents a certain class of biases, such as "dread risk" or "long shot bias", which is the overestimation of some classes of rare events, but derived from binary variables, then falls for the severe mathematical mitake of extending the result to vanilla exposures. If ecological exposures in the real world tends to have vanilla, not binary properties, then much of these results are invalid.

Let us return to the point that the variations of vanilla are not bounded, or have a remote boundary. Hence, the prediction of the vanilla is marred by Black Swan effects and need to be considered from such a viewpoint. For instance, a few prescient observers saw the potential for war among the Great Power of Europe in the early 20th century but virtually everyone missed the second dimension: that the war would wind up killing an unprecedented twenty million persons, setting the stage for both Soviet communism and German fascism and a war that would claim an additional 60 million, followed by a nuclear arms race from 1945 to the present, which might some day claim 600 million lives.

The Black Swan is Not About Probability But Payoff

In short, the vanilla has another dimension, the payoff, in addition to the probability, while the binary is limited to the probability. Ignoring this additional dimension is equivalent to living in a 3-D world but discussing it as if it were 2-D, promoting the illusion to all who will listen that such an analysis captures all worth capturing.

Now the Black Swan problem has been misunderstood. We are saying neither that there must be more volatility in our complexified world nor that there must be more outliers. Indeed, we may well have fewer such events but it has been shown that, under the mechanisms of "fat tails", their "impact" gets larger and larger and more and more unpredictable. The main cause is globalization and the spread of winner-take-all effects across variables (just think of the Google effect), as well as effect of the increased physical and electronic connectivity in the world, causing the weakening of "island effect" a well established fact in ecology by which isolated areas tend to have more varieties of species per square meter than larger ones. In addition, while physical events such as earthquakes and tsunamis may not have changed much in incidence and severity over the last 65 million years (when the dominant species on our planet, the dinosaurs, had a very bad day), their effect is compounded by interconnectivity.

So there are two points here.

Binary predictions are more tractable than exposures

First, binary predictions tend to work; we can learn to be pretty good at making them (at least on short timescales and with rapid accuracy feedback that teaches us how to distinguish signals from noise —all possible in forecasting tournaments as well as in electoral forecasting — see Silver, 2012). Further, these are mathematically tractable: your worst mistake is bounded, since probability is defined on the interval between 0 and 1. But the applications of these binaries tend to be restricted to manmade things, such as the world of games (the "ludic" domain).

It is important to note that, ironically, not only do Black Swan effects not impact the binaries, but they even make them more mathematically tractable, as will see further down.

Binary predictions are often taken as a substitute for vanilla ones

Second, most non-decision makers tend to confuse the binary and the vanilla. And well-intentioned efforts to improve performance in binary prediction tasks can have the unintended consequence of rendering us oblivious to catastrophic vanilla exposure.

The confusion can be traced to attribute substitution and the widespread tendency to replace difficult-to-answer questions with much-easier-to-answer ones. For instance, the extremely-difficult-to-answer question might be whether China and the USA are on an historical trajectory toward a rising-power/hegemon confrontation with the potential to claim far more lives than the most violent war thus far waged (say 10X more the 60M who died in World War II). The much-easier-binary-replacement questions —the sorts of questions likely to pop up in forecasting tournaments or prediction markets — might be whether the Chinese military kills more than 10 Vietnamese in the South China Sea or 10 Japanese in the East China Sea in the next 12 months or whether China publicly announces that it is restricting North Korean banking access to foreign currency in the next 6 months.

The nub of the conceptual confusion is that although predictions and payoffs are completely separate mathematically, both the general public and researchers are under constant attribute-substitution temptation of using answers to binary questions as substitutes for exposure to vanilla risks.

We often observe such attribute substitution in financial hedging strategies. For instance, Morgan Stanley correctly predicted the onset of a subprime crisis, but they had a binary hedge and ended up losing billions as the crisis ended up much deeper than predicted (*Bloomberg Magazine*, March 27, 2008).

Or, consider the performance of the best forecasters in geopolitical forecasting tournaments over the last 25 years (Tetlock, 2005; Tetlock & Mellers, 2011; Mellers et al, 2013). These forecasters may will be right when they say that the risk of a lethal confrontation claiming 10 or more lives in the East China Sea by the end of 2013 is only 0.04. They may be very "well calibrated" in the narrow technical sense that when they attach a

4% likelihood to events, those events occur only about 4% of the time. But framing a vanilla question as a binary question is dangerous because it masks exponentially escalating tail risks: the risks of a confrontation claiming not just 10 lives of 1000 or 1 million. No one has yet figured out how to design a forecasting tournament to assess the accuracy of probability judgments that range between .00000001% and 1% —and if someone ever did, it is unlikely that anyone would have the patience —or lifespan —to run the forecasting tournament for the necessary stretches of time (requiring us to think not just in terms of decades, centuries and millennia).

The deep ambiguity of objective probabilities at the extremes—and the inevitable instability in subjective probability estimates—can also create patterns of systematic mispricing of options. An option or option like payoff is not to be confused with a lottery, and the "lottery effect" or "long shot bias" often discussed in the economics literature that documents that agents overpay for these bets should not apply to the properties of actual options.

In *Fooled by Randomness*, the narrator is asked "do you predict that the market is going up or down?" "Up", he said, with confidence. Then the questioner got angry when he discovered that the narrator was short the market, i.e., would benefit from the market going down. The trader had a difficulty conveying the idea that someone could hold the belief that the market had a higher probability of going up, but that, should it go down, it would go down a lot. So the rational response was to be short.

This divorce between the binary (up is more likely) and the vanilla is very prevalent in real-world variables. Indeed we often see reports on how a certain financial institution "did not have a losing day in the entire quarter", only to see it going near-bust from a monstrously large trading loss. Likewise some predictors have an excellent record, except that following their advice would result in large losses, as they are rarely wrong, but when they miss their forecast, the results are devastating.

Remark:*More technically, for a heavy tailed distribution (defined as part of the subexponential family, see Taleb 2013), with at least one unbounded side to the random variable, the vanilla prediction record over a long series will be of the same order as the best or worst prediction, whichever in largest in absolute value, while no single outcome can change the record of the binary.*

Another way to put the point: to achieve the reputation of "Savior of Western civilization,"a politician such as Winston Churchill needed to be right on only one super-big question (such as the geopolitical intentions of the Nazis)– and it matters not how many smaller errors that politician made (e.g. Gallipoli, gold standard, autonomy for India). Churchill could have a terrible Brier score (binary accuracy) and a wonderful reputation (albeit one that still pivots on historical counterfactuals).

Finally, one of the authors wrote an entire book (Taleb, 1997) on the hedging and mathematical differences between binary and vanilla. When he was an option trader, he realized that binary options have nothing to do with vanilla options, economically and mathematically. Seventeen years later people are still making the mistake.

7.2 A Semi-Technical Commentary on The Mathematical Differences

Chernoff Bound

The binary is subjected to very tight bounds. Let $(X_i)_{1<i\leq n}$ bea sequence independent Bernouilli trials taking values in the set $\{0,1\}$, with $\mathbb{P}(X=1])=p$ and $\mathbb{P}(X=0)=1-p$, Take 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 $\delta>0$

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

and for $0<\delta\leq1$

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

Let us compute the probability of coin flips n of having 50% higher than the true mean, with p= $\frac{1}{2}$ and $\mu=$

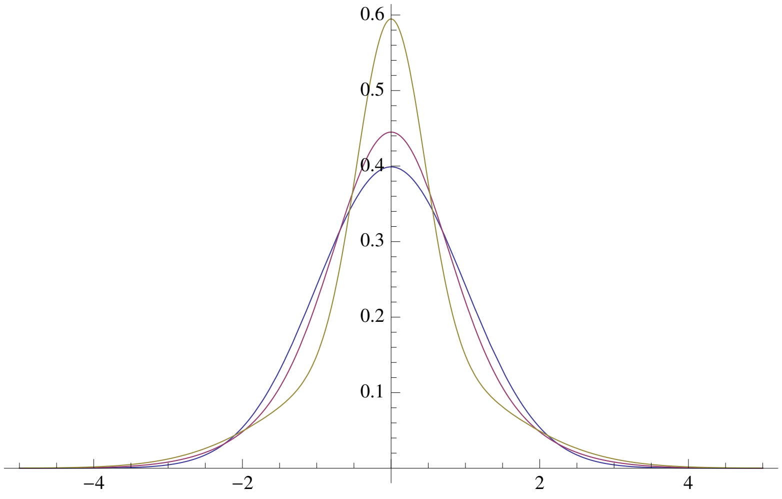


Figure 7.2: Fatter and fatter tails: different values for a. Note that higher peak implies a lower probability of leaving the ±1 σ tunnel

$\frac{n}{2}$: $\mathbb{P}\left(S \geq \left(\frac{3}{2}\right)\frac{n}{2}\right) \leq 2e^{-\frac{\mu\delta^2}{3}} = e^{-n/24}$
which for $n = 1000$ happens every 1 in 1.24×10^{18}.

Fatter tails lower the probability of remote events (the binary) and raise the value of the vanilla.

The following intuitive exercise will illustrate what happens when one conserves the variance of a distribution, but "fattens the tails" by increasing the kurtosis. The probability of a certain type of intermediate and large deviation drops, but their impact increases. Counterintuitively, the possibility of staying within a band increases.
Let x be a standard Gaussian random variable with mean 0 (with no loss of generality) and standard deviation σ. Let $P_{>1\sigma}$ be the probability of exceeding one standard deviation. $P_{>1\sigma}= 1 - \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}}\right)$, where erfc is the complementary error function, so $P_{>1\sigma} = P_{<1\sigma} \simeq$15.86% and the probability of staying within the "stability tunnel" between ± 1 σ is 1− $P_{>1\sigma}$− $P_{<1\sigma} \simeq 68.3\%$.
Let us fatten the tail in a variance-preserving manner, using the "barbell" standard method of linear combination of two Gaussians with two standard deviations separated by $\sigma\sqrt{1+a}$ and $\sigma\sqrt{1-a}$, a ∈(0,1), where a is the "vvol" (which is variance preserving, technically of no big effect here, as a standard deviation-preserving spreading gives the same qualitative result). Such a method leads to the immediate raising of the standard Kurtosis by $\left(1+a^2\right)$ since $\frac{\mathbb{E}\left(x^4\right)}{\mathbb{E}\left(x^2\right)^2} = 3\left(a^2+1\right)$, where $\mathbb{E}$ is the expectation operator.

$$\begin{aligned} P_{>1\sigma} &= P_{<1\sigma} \\ &= 1 - \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}\sqrt{1-a}}\right) \\ &\quad - \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}\sqrt{a+1}}\right) \end{aligned} \tag{7.1}$$

So then, for different values of a in Eq. 1 as we can see in Figure 2, the probability of staying inside 1 sigma rises, "rare" events become less frequent.
Note that this example was simplified for ease of argument. In fact the "tunnel" inside of which fat tailedness increases probabilities is between$-\sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma$ and $\sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma$ (even narrower than 1 σ in the example, as it numerically corresponds to the area between -.66 and .66), and the outer one is $\pm\sqrt{\frac{1}{2}\left(5+\sqrt{17}\right)}\sigma$, that is the area beyond ±2.13 σ.

The law of large numbers works better with the binary than the vanilla

Getting a bit more technical, the law of large numbers works much faster for the binary than the vanilla (for which it may never work, see Taleb, 2013). The more convex the payoff, the more observations one needs to make a reliable inference. The idea is as follows, as can be illustrated by an extreme example of very tractable binary and intractable vanilla.

Let x_t be the realization of the random variable *X* ∈ (-∞, ∞) at period *t*, which follows a Cauchy distribution with p.d.f. $f\left(x_t\right) \equiv \frac{1}{\pi((x_0-1)^2+1)}$. Let us set $x_0 = 0$ to simplify and make the exposure symmetric around 0. The Vanilla exposure maps to the variable x_t and has an expectation $\mathbb{E}\left(x_t\right) = \int_{-\infty}^{\infty} x_t f(x) dx$, which is undefined (i.e., will never converge to a fixed value). A bet at x_0 has a payoff mapped by as a Heaviside Theta Function $\theta_{>x_0}(x_t)$ paying 1 if $x_t > x_0$and 0 otherwise. The expectation of the payoff is simply $\mathbb{E}(\theta(x)) = \int_{-\infty}^{\infty} \theta_{>x_0}(x) f(x) dx = \int_{x_0}^{\infty} f(x) dx$, which is simply $P(x > 0)$. So long as a distribution exists, the binary exists and is Bernouilli distributed with probability of success and failure *p* and *1—p* respectively .

The irony is that the payoff of a bet on a Cauchy, admittedly the worst possible distribution to work with since it lacks both mean and variance, can be mapped by a Bernouilli distribution, about the most tractable of the distributions. In this case the Vanilla is the hardest thing to estimate, and the binary is the easiest thing to estimate.

Set $S_n = \frac{1}{n} \sum_{i=1}^{n} x_{t_i}$ the average payoff of a variety of vanilla bets x_{t_i}across periods t_i, and ${S^\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_{>x_0}\left(x_{t_i}\right)$. No matter how large *n*, $\lim_{n\to\infty} {S^\theta}_n$ has the same properties — the exact same probability distribution —as S_1. On the other hand $\lim_{n\to\infty} {S^\theta}_{n=p}$; further the presaymptotics of ${S^\theta}_n$ are tractable since it converges to $\frac{1}{2}$ rather quickly, and the standard deviations declines at speed $\sqrt{n}$, since $\sqrt{V\left({S^\theta}_n\right)} = \sqrt{\frac{V({S^\theta}_1)}{n}} = \sqrt{\frac{(1-p)p}{n}}$ (given that the moment generating function for the average is $M(z) = \left(p e^{z/n} - p + 1\right)^n$).

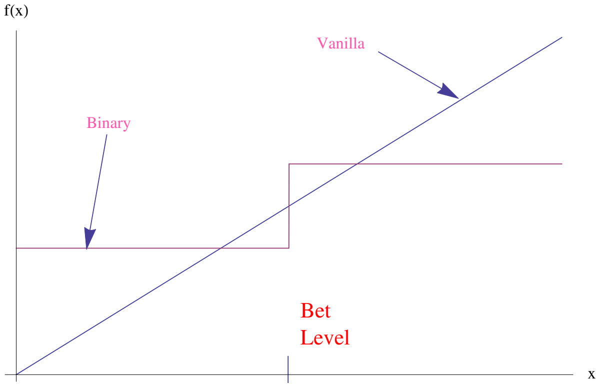


Figure 7.3: The different classes of payoff f(x) seen in relation to an event x. (When considering options, the vanilla can start at a given bet level, so the payoff would be continuous on one side, not the other).

The binary has necessarily a thin-tailed distribution, regardless of domain

More, generally, for the class of heavy tailed distributions, in a long time series, the sum is of the same order as the maximum, which cannot be the case for the binary:

$$\lim_{X\to\infty} \frac{P\left(X > \sum_{i=1}^{n} x_{t_i}\right)}{P\left(X > \max\left(x_{t_i}\right)_{i\leq 2\leq n}\right)} = 1 \qquad (7.2)$$

Compare this to the binary for which

$$\lim_{X\to\infty} P\left(X > \max\left(\theta(x_{t_i})\right)_{i\leq 2\leq n}\right) = 0 \qquad (7.3)$$

The binary is necessarily a thin-tailed distribution, regardless of domain.

We can assert the following:

- The sum of binaries converges at a speed faster or equal to that of the vanilla.
- The sum of binaries is never dominated by a single event, while that of the vanilla can be.

How is the binary more robust to model error?

In the more general case, the expected payoff of the vanilla is expressed as $\int_A x dF(x)$ (the unconditional shortfall) while that of the binary= $\int_{\mathrm{A}} dF(x)$, where A is the part of the support of interest for the exposure, typically A≡[K,∞), or (−∞,K]. Consider model error as perturbations in the parameters that determine the calculations of the probabilities. In the case of the vanilla, the perturbation's effect on the probability is multiplied by a larger value of x.

As an example, define a slighly more complicated vanilla than before, with option-like characteristics, $V(\alpha, K) \equiv \int_K^\infty x \; p_\alpha(x) dx$ and $B(\alpha, K) \equiv\int_K^\infty p_\alpha(x)\, dx$, where V is the expected payoff of vanilla, B is that of the binary, K is the "strike" equivalent for the bet level, and with x∈[1, ∞) let $p_\alpha(x)$ be the density of the Pareto distribution with minimum value 1 and tail exponent α, so $p_\alpha(x) \equiv \alpha x^{-\alpha-1}$.

Set the binary at .02, that is, a 2% probability of exceeding a certain number K, corresponds to an α=1.2275 and a K=24.2, so the binary is expressed as B(1.2, 24.2). Let us perturbate α, the tail exponent, to double the probability from .02 to .04. The result is $\frac{B(1.01,24.2)}{B(1.2,24.2)}$ = 2. The corresponding effect on the vanilla is $\frac{V(1.01,24.2)}{V(1.2,24.2)}$ = 37.4. In this case the vanilla was ~18 times more sensitive than the binary.

7.3 The Applicability of Some Psychological Biases

Without going through which papers identifying biases, Table 1 shows the effect of the error across domains. We are not saying that the bias does not exist; rather that, if the error is derived in a binary environment, or one with a capped payoff, it does not port outside the domain in which it was derived.

Table 7.1: True and False Biases

Bias	**Erroneous application**	**Justified applications**
Dread Risk	Comparing Terrorism to fall from ladders	Comparing risks of driving vs flying
General overestimation of small probabilities	Bounded bets in laboratory setting	Open-ended payoffs in fat-tailed domains
Long shot bias	Lotteries	Financial payoffs
Precautionary principle	Volcano eruptions	Climatic issues

Acknowledgments

Bruno Dupire, Raphael Douady, Daniel Kahneman, Barbara Mellers.

References

Chernoff, H. (1952), A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations, *Annals of Mathematic Statistics*, 23, 1952, pp. 493âĂŞ507.

Mellers, B. et al. (2013), How to win a geopolitical forecasting tournament: The power of teaming and training. Unpublished manuscript, Wharton School, University of Pennsylvania Team Good Judgment Lab.

Silver, Nate, 2012, *The Signal and the Noise.*

Taleb, N.N., 1997, *Dynamic Hedging: Managing Vanilla and Exotic Options*, Wiley

Taleb, N.N., 2001/2004, *Fooled by Randomness,* Random House

Taleb, N.N., 2013, *Probability and Risk in the Real World, Vol 1: Fat Tails*Freely Available Web Book, www.fooledbyrandomness.com

Tetlock, P.E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton: Princeton University Press.

Tetlock, P.E., Lebow, R.N., & Parker, G. (Eds.) (2006). *Unmaking the West: What-if scenarios that rewrite world history.* Ann Arbor, MI: University of Michigan Press.

Tetlock, P. E., & Mellers, B.A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. American Psychologist, 66(6), 542-554.

8 How Fat Tails Emerge From Recursive Epistemic Uncertainty

The Opposite of Central Limit

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, put add additional layers of uncertainty.

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.

8.1 Methods and Derivations

8.1.1 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 the uncertainty rate, and so forth. There is no reason to have certainty anywhere in the process.

8.1.2 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 ϕ(μ,σ,x) as the Gaussian PDF for value x with mean μ and standard deviation σ.

A 2^{nd}order stochastic standard deviation is the integral of ϕ across values of $\sigma \in \mathbb{R}^+$, under the measure $f\left(\bar{\sigma}, \sigma_1, \sigma\right)$, 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 $\overline{\sigma_1}$.

$$f(x)_1 = \int_0^\infty \phi(\mu, \sigma, x) f\left(\bar{\sigma}, \sigma_1, \sigma\right) \, \mathrm{d}\sigma$$

Generalizing to the N^{th} order, the density function *f(x)* becomes

$$f(x)_N = \int_0^\infty \dots \int_0^\infty \phi(\mu,\sigma,x) f\left(\bar{\sigma},\sigma_1,\sigma\right) f\left(\overline{\sigma_1},\sigma_2,\sigma_1\right) \dots f\left(\overline{\sigma_{N-1}},\sigma_N,\sigma_{N-1}\right) \mathrm{d}\sigma\, \mathrm{d}\sigma_1\, \mathrm{d}\sigma_2 \dots \mathrm{d}\sigma_N \qquad (8.1)$$

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 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 $\int_0^\infty \phi(\mu,\sigma,x) f\left(\bar{\sigma},\sigma_1,\sigma\right)\ mathrmd\sigma$ (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 $0 \leq a(1) < 1$, 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:

$$f(x)_1 = \frac{1}{2}\Big(\phi(\mu,\sigma\left(1+a(1)\right),x) + \phi(\mu,\sigma(1-a(1)),x)\Big) \qquad (8.2)$$

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± a(2)).

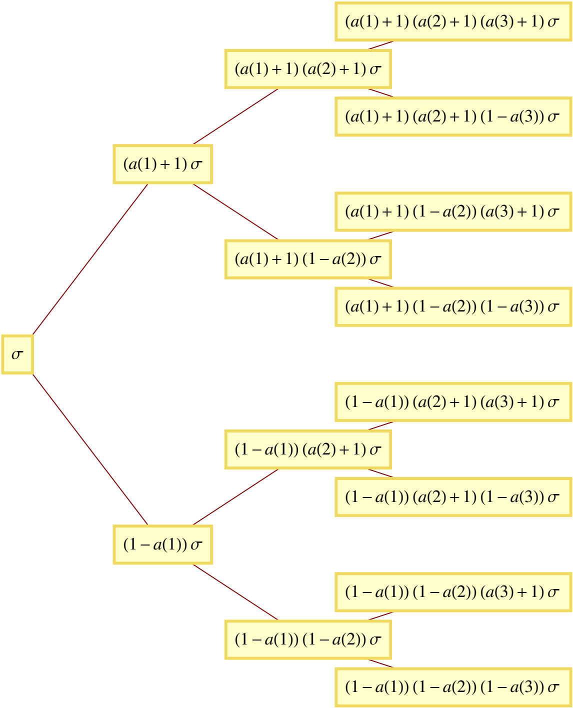

(a(1) + 1) (a(2) + 1) (a(3) + 1) σ
(a(1) + 1) (a(2) + 1) σ
(a(1) + 1) (a(2) + 1) (1 − a(3)) σ
(a(1) + 1) σ
(a(1) + 1) (1 − a(2)) (a(3) + 1) σ
(a(1) + 1) (1 − a(2)) σ
(a(1) + 1) (1 − a(2)) (1 − a(3)) σ
σ
(1 − a(1)) (a(2) + 1) (a(3) + 1) σ
(1 − a(1)) (a(2) + 1) σ
(1 − a(1)) (a(2) + 1) (1 − a(3)) σ
(1 − a(1)) σ
(1 − a(1)) (1 − a(2)) (a(3) + 1) σ
(1 − a(1)) (1 − a(2)) σ
(1 − a(1)) (1 − a(2)) (1 − a(3)) σ

Table 8.1: Three levels of error rates for σ following a multiplicative process

The second order stochastic standard deviation:

$$f(x)_2 = \frac{1}{4}\Bigg(\phi\bigg(\mu,\sigma(1+a(1)(1+a(2))),x\bigg)+ \phi\bigg(\mu,\sigma(1-a(1)(1+a(2))),x)+\phi(\mu,\sigma(1+a(1)(1-a(2)),x\bigg) + \phi\bigg(\mu,\sigma(1-a(1)(1-a(2))),x\bigg)\Bigg) \qquad (8.3)$$

and the N^{th} order:

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

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

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

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.

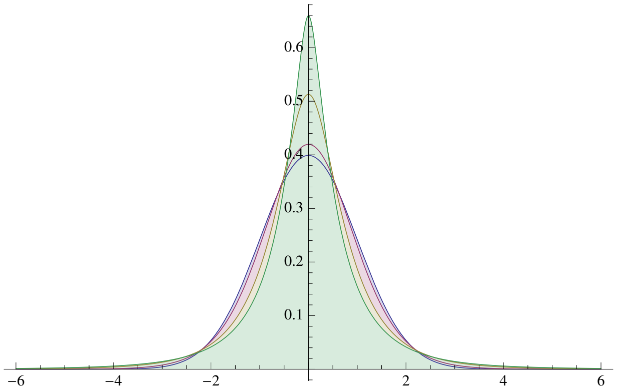


Figure 8.1: Thicker tails (higher peaks) for higher values of N; here $N=0,5,10,25,50$, all values of a=$\frac{1}{10}$

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*.

$$f(x|\mu,\sigma,M,N)=$$
$$2^{-N}\sum_{j=0}^{N}\binom{N}{j}\phi\left(\mu,\sigma(a+1)^j(1-a)^{N-j},x\right) \quad (8.4)$$

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 μ=0

$$\left(\begin{array}{cc} & \text{Moment} \\ 1 & 0 \\ 2 & \left(a^2+1\right)^N \sigma^2 \\ 3 & 0 \\ 4 & 3\left(a^4+6a^2+1\right)^N \sigma^4 \\ 5 & 0 \\ 6 & 15\left(a^6+15a^4+15a^2+1\right)^N \sigma^6 \\ 7 & 0 \\ 8 & 105\left(a^8+28a^6+70a^4+28a^2+1\right)^N \sigma^8 \end{array}\right)$$

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$\rightarrow \infty$. 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.

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.

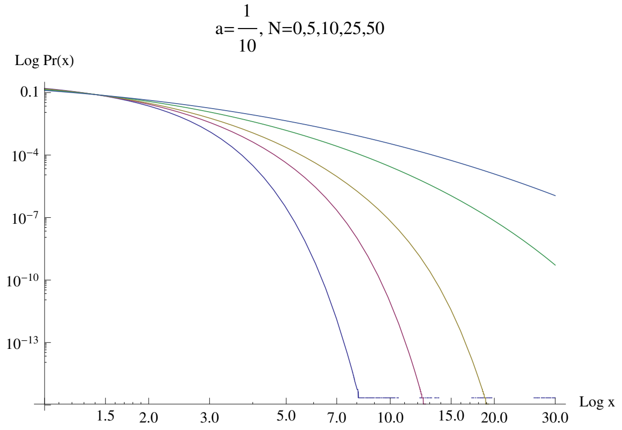


Figure 8.2: LogLog Plot of the probability of exceeding x showing power law-style flattening as N rises. Here all values of a= 1/10

The same effect takes place as a increases towards 1, as at the limit the tail exponent P>x approaches 1 but remains >1.

8.1.3 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:

$$P > K|N = \sum_{j=0}^{N} 2^{-N-1} \left(\begin{array}{c} N \\ j \end{array}\right) \text{erfc}\left(\frac{K}{\sqrt{2}\sigma(a+1)^j(1-a)^{N-j}}\right) \quad (8.5)$$

where erfc(.) is the complementary of the error function, 1-erf(.), 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 8.2: 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 8.3: 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}

8.2 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).

8.2.1 Regime 2-a;"bleed" of higher order error

Take a "bleed" of higher order errors at the rate λ, $0 \leq \lambda < 1$, such as *a(N)* = λ a(N-1), hence *a*(N) =λ^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:

$$M_3(N) = \left(a(1)^2+1\right)\sigma^2\prod_{i=1}^{N-1}\left(a(1)^2\lambda^{2i}+1\right) \quad (8.6)$$

We can reexpress 8.6 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:

$$M_4(N) = 3\sigma^4\prod_{i=0}^{N-1}\left(6a(1)^2\lambda^{2i}+a(1)^4\lambda^{4i}+1\right)$$

$$M_4(N) = 3\sigma^4\left(\left(2\sqrt{2}-3\right)a(1)^2;\lambda^2\right)_N \left(-\left(3+2\sqrt{2}\right)a(1)^2;\lambda^2\right)_N \quad (8.7)$$

$$\lim_{N\to\infty} M_4(N) = 3\sigma^4\left(\left(2\sqrt{2}-3\right)a(1)^2;\lambda^2\right)_\infty \left(-\left(3+2\sqrt{2}\right)a(1)^2;\lambda^2\right)_\infty \quad (8.8)$$

So the limiting second moment for λ=.9 and a(1)=.2 is just 1.28 σ^2, a significant but relatively benign convexity bias. The limiting fourth moment is just $9.88\sigma^4$, more than 3 times the Gaussian's (3 σ^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.

8.2.2 Regime 2-b; Second Method, a Non Multiplicative Error Rate

For N recursions,

$$\sigma(1 \pm (a(1)(1 \pm (a(2)(1 \pm a(3)(\;...)))$$

$$P(x,\mu,\sigma,N) = \frac{1}{L}\sum_{i=1}^{L} f\left(x,\mu,\sigma\left(1+\left(\mathbf{T}^N.\mathbf{A}^N\right)_i\right)\right.$$

$\left(\mathbf{M}^N.\mathbf{T}+1\right)_i^{\cdot\cdot}$ is the i^th component of the (N) dot product of T^N the matrix of Tuples in (xx) , L the length of the matrix, and A is the set of parameters

$$A^N = \left\{a^j\right\}_{j=1,\dots N}$$

So for instance, for *N*=3,T= {1, a, a^2,a^3}

$$\mathbf{A}^3\ \mathbf{T}^3 = \begin{pmatrix} a^3+a^2+a \\ -a^3+a^2+a \\ a^3-a^2+a \\ -a^3-a^2+a \\ a^3+a^2-a \\ -a^3+a^2-a \\ a^3-a^2-a \\ -a^3-a^2-a \end{pmatrix}$$

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.

8.3 Conclusion

Something Boring & something about epistemic opacity.

This part will be expanded

9 On the Difficulty of Risk Parametrization With Fat Tails

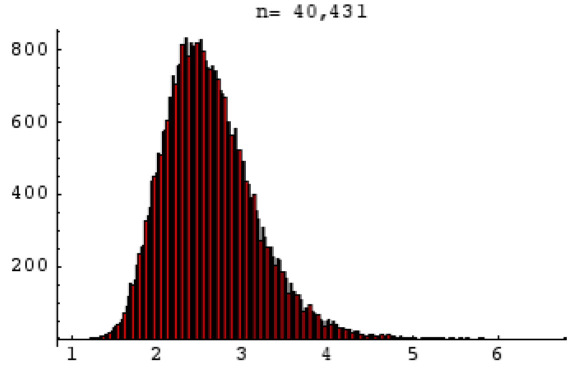
This chapter presents case studies around the point that, simply, some models depend quite a bit on small variations in parameters. The effect on the Gaussian is easy to gauge, and expected. But many believe in power laws as panacea. Even if I believed the r.v. was power law distributed, I still would not be able to make a statement on tail risks. Sorry, but that's life.

This chapter is illustrative; it will initially focus on nonmathematical limits to producing estimates of $M_T^X(A, f)$ when A is limited to the tail. We will see how things get worse when one is sampling and forecasting the maximum of a random variable.

9.1 Some Bad News Concerning power laws

We saw the shortcomings of parametric and nonparametric methods so far. What are left are power laws; they are a nice way to look at the world, but we can never really get to know the exponent α, for a spate of reasons we will see later (the concavity of the exponent to parameter uncertainty). Suffice for now to say that the same analysis on exponents yields a huge in-sample variance and that tail events are very sensitive to small changes in the exponent.

For instance, for a broad set of stocks over subsamples, using a standard estimation method (the Hill estimator), we get subsamples of securities. Simply, the variations are too large for a reliable computation of probabilities, which can vary by $>$ 2 orders of magnitude. And the effect on the mean of these probabilities is large since they are way out in the tails.



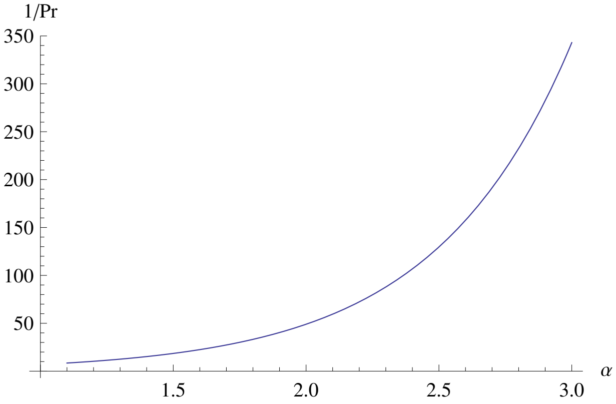


Table 9.1: The effect of small changes in tail exponent on a probability of exceeding a certain point. Here Pareto(L,α), probability of exceeding 7 L ranges from 1 in 10 to 1 in 350. For further in the tails the effect is more severe.

The way to see the response to small changes in tail exponent with probability: considering $P_{>K} \sim K^{-\alpha}$, the sensitivity to the tail exponent $\frac{\partial P_{>K}}{\partial \alpha} = -K^{-\alpha} \log(K)$.

Now the point that probabilities are sensitive to assumptions brings us back to the Black Swan problem. One might wonder, the change in probability might be large in percentage, but who cares, they may remain small. Perhaps, but in fat tailed domains, the event multiplying the probabilities is large. In life, it is not the probability that matters, but what one does with it, such as the expectation or other moments, and the contribution of the small probability to the total moments is large in power law domains.

For all powerlaws, when *K* is large, with $\alpha > 1$, the unconditional shortfall $S_+ = \int_K^\infty x \, \phi(x) dx$ and $S_- \int_{-\infty}^{-K} x \, \phi(x) dx$ approximate to $\frac{\alpha}{\alpha-1} K^{-\alpha+1}$ and -$\frac{\alpha}{\alpha-1} K^{-\alpha+1}$, which are extremely sensitive to α particularly at higher levels of *K*, $\frac{\partial S_+}{\partial \alpha} = -\frac{K^{1-\alpha}((\alpha-1)\alpha \log(K)+1)}{(\alpha-1)^2}$.

There is a deeper problem related to the effect of model error on the estimation of α, which compounds the problem, as α tends to be underestimated by Hill estimators and other methods, but let us leave it for now.

9.2 Extreme Value Theory: Fuhgetaboudit

We saw earlier how difficult it is to compute risks using power laws, owing to excessive model sensitivity. Let us apply this to the so-called Extreme Value Theory, EVT.

Extreme Value Theory has been considered a panacea for dealing with extreme events by some "risk modelers" . On paper it looks great. But only on paper. The problem is the calibration and parameter uncertainty – in the real world we don't know the parameters. The ranges in the probabilities generated we get are monstrous.

We start with a short presentation of the idea, followed by an exposition of the difficulty.

9.2.1 What is Extreme Value Theory? A Simplified Exposition

Let us proceed with simple examples.

Case 1, Thin Tailed Distribution

The Extremum of a Gaussian variable: Say we generate n Gaussian variables $(X_i)_{i=1}^n$ with mean 0 and unitary standard deviation, and take the highest value we find. We take the upper bound M_j for the n-size sample run j

$M_j = \text{Max}\,(X_{i,j})_{i=1}^n$

Assume we do so p times, to get p samples of maxima for the sequence M

$M = \left(\text{Max}\left\{X_{i,j}\right\}_{i=1}^n\right\}_{j=1}^p$

The next figure will plot a histogram of the result of both the simulation and .

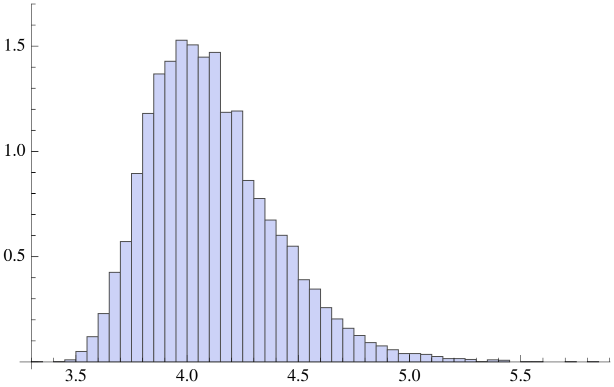


Figure 0.2. Taking p samples of Gaussian maxima; here $N = 30K$, $M = 10K$. We get the Mean of the maxima = 4.11159 Standard Deviation= 0.286938; Median = 4.07344

Let us now fit to the sample from the simulation to g, the density of an Extreme Value Distribution for x (or the Gumbel for the negative variable $-x$), with location and scale parameters α and β, respectively: $g(x; \alpha, \beta) = \frac{e^{\frac{\alpha-x}{\beta} - e^{\frac{\alpha-x}{\beta}}}}{\beta}$

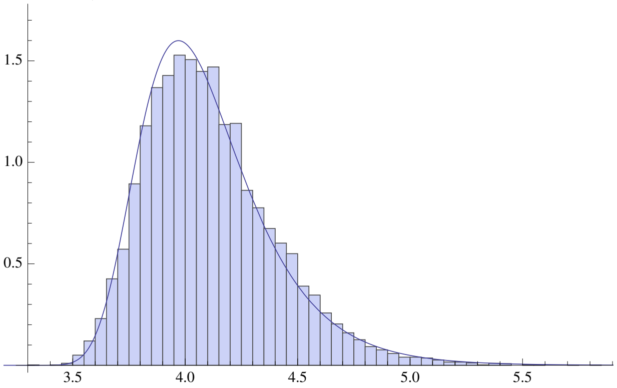


Figure 0.3. : Fitting an extreme value distribution (Gumbel for the maxima) α= 3.97904, β= 0.235239

9.2.2 A Note. How does the Extreme Value Distribution emerge?

Consider that the probability of exceeding the maximum corresponds to the rank statistics, that is the probability of all variables being below the observed sample. $P\left(X_1 < x, X_2 < x, ..., X_n < x\right) = 1 - ?\bigcap_{i=1}^{n} P\left(X_i\right) = F(x)^n$, where F is the cumulative Gaussian. Taking the first derivative of the cumulative distribution to get the density of the distribution of the maximum,

$p_n(x) \equiv \partial_x \left(F(x)^n\right) = -\frac{2^{\frac{1}{2}-n} n e^{-\frac{x^2}{2}} \left(\text{erf}\left(\frac{x}{\sqrt{2}}\right)+1\right)^{n-1}}{\sqrt{\pi}}$
Now we have norming constants $a_n$and b_n such that

$$G(x) \equiv P\left(\frac{M(n)-a_n}{b_n} > x\right)$$

.
But there is a basin of attraction condition for that. We need to find an $x_0 < \infty$ beyond which at the limit of n $\to \infty$, $x_0 = \sup\{$x $:F(x) < 1\}$

Derivations

$$(1 - P(X > a(n)x + b(n)))^N = G(x)$$

$$\exp(-NP(X > ax + b)) = G(x)$$

After some derivations[see below], $g(x) = \frac{e^{\frac{\alpha-x}{\beta} - e^{\frac{\alpha-x}{\beta}}}}{\beta}$, where
$\alpha = -\sqrt{2}\text{erfc}^{-1}\left(2 - \frac{2}{n}\right)$, where erfc^{-1}is the inverse error function, and
$\beta = \sqrt{2}\left(\text{erfc}^{-1}\left(2 - \frac{2}{n}\right) - \text{erfc}^{-1}\left(2 - \frac{2}{en}\right)\right)$
For $n = 30K$, $\{\alpha, \beta\} = \{3.98788, 0.231245\}$
The approximations become $\sqrt{2\log(n)} - \frac{\log(\log(n)) + \log(4\pi)}{2\sqrt{2\log(n)}}$ and $(2\log(n))^{-\frac{1}{2}}$ respectively + $o\left((\text{Log}n)^{-\frac{1}{2}}\right)$

9.2.3 Extreme Values for Fat-Tailed Distribution

Now let us generate, exactly as before, but change the distribution, with N random power law distributed variables X_i, with tail exponent α=3, generated from a Student T Distribution with 3 degrees of freedom. Again, we take the upper bound. This time it is not the Gumbel, but the Fréchet distribution that would fit the result, using −critically− the same α, Fréchet ϕ(x; α, β)=

$$\frac{\alpha e^{-\left(\frac{x}{\beta}\right)^{-\alpha}} \left(\frac{x}{\beta}\right)^{-\alpha-1}}{\beta},$$

for x>0

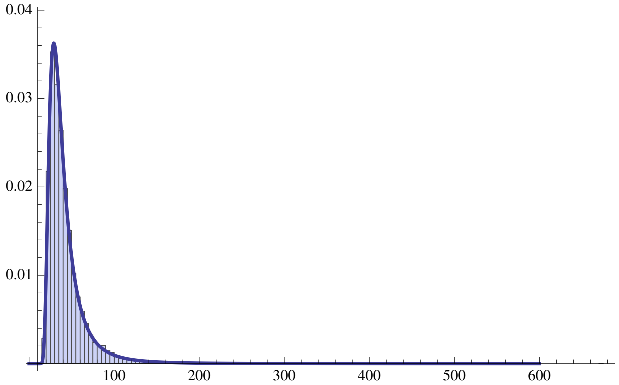


Figure 0.4. Fitting a Fréchet distribution to the Student T generated with α=3 degrees of freedom. The Frechet distribution α=3, β=32 fits up to higher values of E. But next two graphs shows the fit more closely.

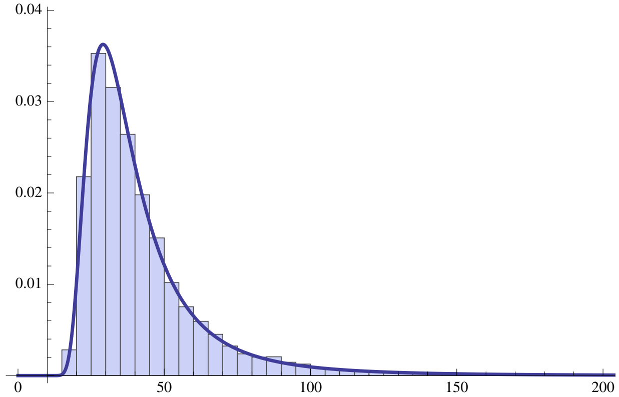


Figure 0.5. Seen more closely

9.2.4 How Extreme Value Has a Severe Inverse Problem In the Real World

In the previous case we start with the distribution, with the assumed parameters, then get the corresponding values, as these "risk modelers" do. In the real world, we don't quite know the calibration, the α of the distribution, assuming (generously) that we know the distribution. So here we go with the inverse problem. The next table illustrates the different calibrations of P_K the probabilities that the maximum exceeds a certain value K (as a multiple of β under different values of K and α.

α	$\frac{1}{P_{>3\beta}}$	$\frac{1}{P_{>10\beta}}$	$\frac{1}{P_{>20\beta}}$
1.	3.52773	10.5083	20.5042
1.25	4.46931	18.2875	42.7968
1.5	5.71218	32.1254	89.9437
1.75	7.3507	56.7356	189.649
2.	9.50926	100.501	400.5
2.25	12.3517	178.328	846.397
2.5	16.0938	316.728	1789.35
2.75	21.0196	562.841	3783.47
3.	27.5031	1000.5	8000.5
3.25	36.0363	1778.78	16918.4
3.5	47.2672	3162.78	35777.6
3.75	62.048	5623.91	75659.8
4.	81.501	10000.5	160000.
4.25	107.103	17783.3	338359.
4.5	140.797	31623.3	715542.
4.75	185.141	56234.6	1.51319×10^6
5.	243.5	100001.	3.2×10^6

Consider that the error in estimating the α of a distribution is quite large, often $> \frac{1}{2}$, and typically overstimated. So we can see that we get the probabilities mixed up > an order of magnitude.In other words the imprecision in the computation of the α compounds in the evaluation of the probabilities of extreme values.

9.3 Using Power Laws Without Being Harmed by Mistakes

We can use power laws in the "near tails" for information, not risk management. That is, not pushing outside the tails, staying within a part of the distribution for which errors are not compounded.

I was privileged to get access to a database with cumulative sales for editions in print that had at least one unit sold that particular week (that is, conditional of the specific edition being still in print). I fit a powerlaw with tail exponent $\alpha \simeq 1.3$ for the upper 10% of sales (graph), with N=30K. Using the Zipf variation for ranks of powerlaws, with r_x and r_y the ranks of book x and y, respectively, S_x and S_y the corresponding sales

$$\frac{S_x}{S_y} = \left(\frac{r_x}{r_y}\right)^{-\frac{1}{\alpha}}$$

So for example if the rank of x is 100 and y is 1000, x sells $\left(\frac{100}{1000}\right)^{-\frac{1}{1.3}}$ = 5.87 times what y sells.

Note this is only robust in deriving the sales of the lower ranking edition ($r_y > r_x$) because of inferential problems in the presence of fat-tails.

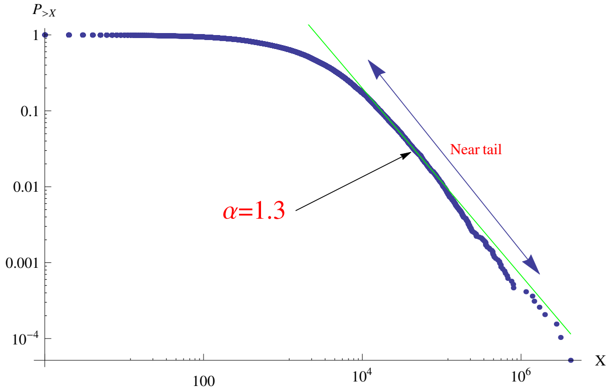


Figure 9.1: Log-Log Plot of the probability of exceeding X (book sales)

This works best for the top 10,000 books, but not quite the top 20 (because the tail is vastly more unstable). Further, the effective α for large deviations is lower than 1.3. But this method is robust as applied to rank within the "near tail".

10 | Brownian Motion in the Real World (Path Dependence and Fat Tails)

Most of the work concerning martingales and Brownian motion can be idealized to the point of lacking any match to reality, in spite of the sophisticated, rather complicated discussions. This section discusses the (consequential) differences.

10.1 Path Dependence and History as Revelation of Antifragility

Let us examine the non-Markov property of antifragility. Something that incurred hard times *but did not fall apart* is giving us information about its solidity, compared to something that has not been subjected to such stressors.

(The Markov Property for, say, a Brownian Motion $X_{N|\{X_1,X_2,\dots X_{N-1}\}}= X_{N|\{X_{N-1}\}}$, that is the last realization is the only one that matters. Now if we take fat tailed models, such as stochastic volatility processes, the properties of the system are Markov, but the history of the past realizations of the process matter in determining the present variance.)

Take *M* realizations of a Brownian Bridge process pinned at S_{t_0}= 100 and S_T= 120, sampled with N periods separated by Δt, with the sequence *S*, a collection of Brownian-looking paths with single realizations indexed by j ,

$$S_i^j = \left(\left(S^j_{\mathrm{i\Delta t}+t_0} \right)_{i=0}^{N} \right)_{j=1}^{M}$$

Take $m^* = \min_j \; min_i \S_i^j$ and $\left\{ j : \min S_i^j = m^* \right\}$

Take 1) the sample path with the most direct route (Path 1) defined as its lowest minimum , and 2) the one with the lowest minimum m^* (Path 2). The state of the system at period *T* depends heavily on whether the process S_T exceeds its minimum (Path 2), that is whether arrived there thanks to a steady decline, or rose first, then declined.

If the properties of the process depend on (S_T- m*), then there is path dependence. By properties of the process we mean the variance, projected variance in, say, stochastic volatility models, or similar matters.

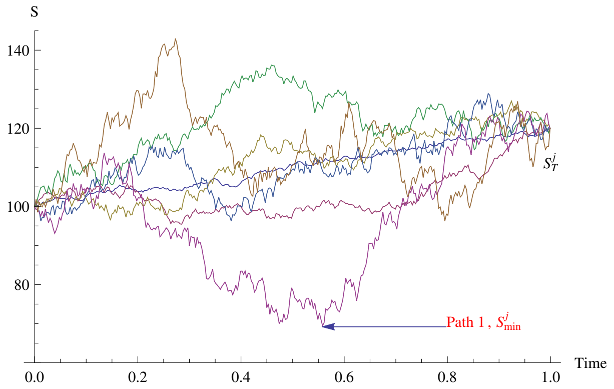


Figure 10.1: Brownian Bridge Pinned at 100 and 120, with multiple realizations $\{S_0^j,S_1^j\dots,S_T^j\}$*, each indexed by j ; the idea is to find the path j that satisfies the maximum distance* $D_j= \left| S_T - S^j_{\min} \right|$

10.2 Brownian Motion in the Real World

We mentioned in the discussion of the Casanova problem that stochastic calculus *requires* a certain class of distributions, such as the Gaussian. It is not as we expect because of the convenience of the smoothness in squares (finite Δx^2), rather because the distribution conserves across time scales. By central limit, a Gaussian remains a Gaussian under summation, that is sampling at longer time scales. But it also remains a Gaussian at shorter time scales. The foundation is infinite dividability.

The problems are as follows:

The results in the literature are subjected to the constaints that the Martingale **M** is member of the subset ($\mathbf{H}^2$) of square integrable martingales, $\sup_{t \leq T} \mathrm{E}[M^2] < \infty$

We know that the restriction does not work for lot or time series.

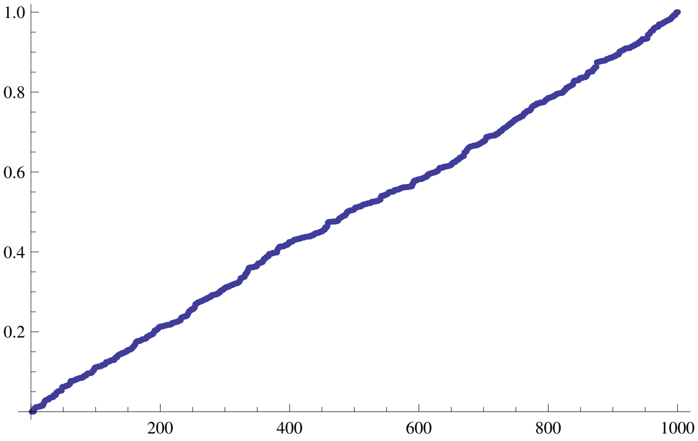
We know that, with θ an adapted process, without $\int_0^T \theta_s^2 \, ds < \infty$ we can't get most of the results of Ito's lemma.

Even with $\int_o^T dW^2 < \infty$, The situation is far from solved because of powerful, very powerful presamptotics.

Hint: Smoothness comes from $\int_o^T dW^2$ becoming linear to T at the continuous limit –Simply dt is too small in front of dW

Take the normalized (i.e. sum=1) cumulative variance (see Bouchaud & Potters), $\frac{\sum_{i=1}^{n}(W[i\Delta t]-W[(i-1)\Delta t])^2}{\sum_{i=1}^{T/\Delta t}(W[i\Delta t]-W[(i-1)\Delta t])^2}$.

Let us play with finite variance situations.



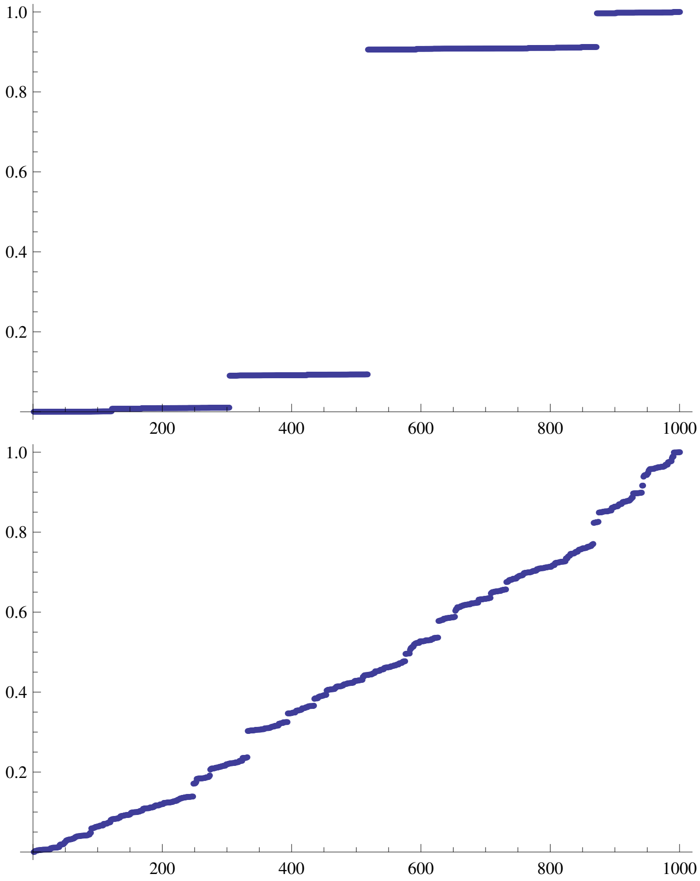
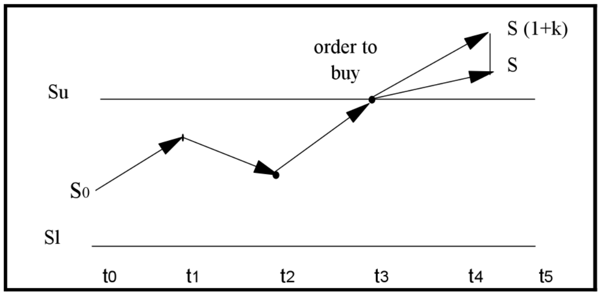


Figure 10.2: Ito's lemma in action. Three classes of processes with tail exonents: $\alpha = \infty$ (Gaussian), α = 1.16 (the 80/20) and, α = 3. Even finite variance does not lead to the smoothing of discontinuities except in the infinitesimal limit, another way to see failed asymptotes.

10.3 tochastic Processes and Nonanticipating Strategies

There is a difference between the Stratonovich and Ito's integration of a functional of a stochastic process. But there is another step missing in Ito: the gap between information and adjustment.



10.4 0.4 Finite Variance not Necessary for Anything Ecological (incl. quant finance)

11 The Fourth Quadrant Mitigation (or "Solution")

Let us return to $M[A, f(x)]$ of chapter 1. A quite significant result is that M[A,x^n] may not converge, in the case of, say power laws with exponent $\alpha < n$, but $M\left[A, x^m\right]$ where $m < n$, would converge. Well, where the integral $\int_{-\infty}^{\infty} f(x)p(x)\, dx$ does not exist, by "clipping tails", we can make the payoff integrable. There are two routes;
1) **Limiting f (turning an open payoff to a binary)**: when $f(x)$ is a constant as in a binary $\int_{-\infty}^{\infty} Kp(x)dx$ will necessarily converge if *p* is a probability distribution.
2) **Clipping tails:** (and this is the business we will deal with in *Antifragile*, Monograph 2), where the payoff is bounded, $A = [L, H]$, or the integral $\int_L^H f(x)p(x)dx$ will necessarily converge.

11.1 Two types of Decisions

M0 depends on the 0^{th} moment, that is, "Binary", or simple, i.e., as we saw, you just care if something is true or false. Very true or very false does not matter. Someone is either pregnant or not pregnant. A statement is "true" or "false" with some confidence interval. (I call these M0 as, more technically, they depend on the zeroth moment, namely just on probability of events, and not their magnitude — — —you just care about "raw" probability). A biological experiment in the laboratory or a bet with a friend about the outcome of a soccer game belong to this category.
M1$^+$Complex, depend on the 1^{st} or higher moments. You do not just care of the frequency— — —but of the impact as well, or, even more complex, some function of the impact. So there is another layer of uncertainty of impact. (I call these M1+, as they depend on higher moments of the distribution). When you invest you do not care how many times you make or lose, you care about the expectation: how many times you make or lose *times* the amount made or lost.
Two types of probability structures:
There are two classes of probability domains— — —very distinct qualitatively and quantitatively. The first, thin-tailed: Mediocristan", the second, thick tailed Extremistan:

Table 11.1: The Four Quadrants

	Simple payoffs	Complex payoffs
Distribution 1 ("thin tailed")	First Quadrant Extremely Safe	Second Quadrant: Safe
Distribution 2 (no or unknown characteristic scale)	Third Quadrant: Safe	Fourth Quadrant: Dangers

Conclusion

The 4th Quadrant is mitigated by changes in exposures. And exposures in the 4th quadrant can be to the negative or the positive, depending on if the domain subset A exposed on the left on on the right.

Table 11.2: Tableau of Decisions

Mo "True/False" *f(x)*=0	*M*1 Expectations LINEAR PAYOFF *f(x)*=1	*M*2+ NONLINEAR PAYOFF *f(x)* nonlinear(= x^2, x^3, etc.)
Medicine (health not epidemics)	Finance : nonleveraged Investment	Derivative payoffs
Psychology experiments	Insurance, measures of expected shortfall	Dynamically hedged portfolios
Bets (prediction markets)	General risk management	Leveraged portfolios (around the loss point)
Binary/Digital derivatives	Climate	Cubic payoffs (strips of out of the money options)
Life/Death	Economics (Policy)	Errors in analyses of volatility
	Security: Terrorism, Natural catastrophes	Calibration of nonlinear models
	Epidemics	Expectation weighted by nonlinear utility
	Casinos	Kurtosis-based positioning ("volatility trading")

Part II

Fragility and Nonlinear Exposure to Random Variables

12 Exposures and Nonlinear Transformations of Random Variables

12.1 The Conflation Problem: Exposures to x Confused With Knowledge About x

12.1.1 Exposure, not knowledge

.Take x a random or nonrandom variable, and *f(x)* the exposure, payoff, the effect of x on you, the end bottom line. (To be technical, x is higher dimensions, in $\mathfrak{R}^N$ but less assume for the sake of the examples in the introduction that it is a simple one-dimensional variable).

The disconnect. Practitioner and risk takers observe the following disconnect: people (nonpractitioners) talking x (with the implication that we practitioners should care about x in running our 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. 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.

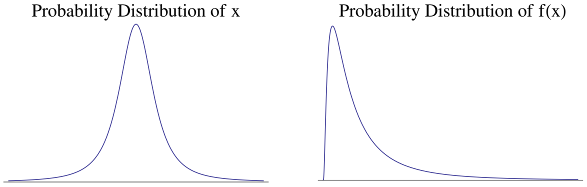
Examples

The variable x is 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 value function of Kahneman-Tversky, how these "affect" you. One can see that f(x) is vastly more stable or robust than x (it has thinner tails).

A convex and linear function of a variable x. 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(Mean of x), by Jensen's ineqality. 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 matter compared to the nonlinearity of f. Moral of the story: focus on f, which we can alter, rather than the measurement of the elusive properties of x.



There are infinite numbers of functions *F* depending on a unique variable x.

All utilities need to be embedded in *F*.

12.1.2 *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)*.

This seems naive, but people do, as something is lost in the translation.

The danger with the treatment of the Black Swan prob-

lem 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.
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 get a paper on "convex transformations of random variables", Van Zwet (1964).

12.1.3 *Bad news*

F is almost always nonlinear, often "S curved", that is convex-concave (for an increasing function).

12.1.4 *The central point about what to understand*

When *f(x)* is convex, say as in trial and error, or with an option, we do not need to understand x as much as our exposure to H. Simply the statistical properties of x are swamped by those of *H*. That's the point of *Antifragility* in which exposure is more important than the naive notion of "knowledge", that is, understanding x.

12.1.5 *Fragility and Antifragility*

When *f(x)* is concave (fragile), errors about x can translate into extreme negative values for F. When *f(x)* is convex, one is immune from negative variations.
The more nonlinear F the less the probabilities of x matter in the probability distribution of the final package *F*.
Most people confuse the probabilites of x with those of F. I am serious: the *entire* literature reposes largely on this mistake.
So, for now ignore discussions of x that do not have *F*. And, for Baal's sake, focus on *F*, not x.

12.2 Transformations of Probability Distributions

Say x follows a distribution p(x) and $z = f(x)$ follows a distribution g(z). Assume $g(z)$ continuous, increasing, and differentiable for now.
The density p at point r is defined by use of the integral

$$D(r) \equiv \int_{-\infty}^{r} p(x)dx$$

hence

$$\int_{-\infty}^{r} p(x)\, dx = \int_{-\infty}^{f(r)} g(z)\, dz$$

In differential form

$$g(z)dz = p(x)dx$$

since $x = f^{(-1)}(z)$, one obtains

$$g(z)dz = p\left(f^{(-1)}(z)\right) df^{(-1)}(z)$$

Now, the derivative of an inverse function

$$f^{(-1)}(z) = \frac{1}{f'\left(f^{-1}(z)\right)},$$

which provides the useful transformation heuristic:

$$g(z) = \frac{p\left(f^{(-1)}(z)\right)}{f'(u)|u = \left(f^{(-1)}(z)\right)} \tag{12.1}$$

In the event that g(z) is monotonic decreasing, then

$$g(z) = \frac{p\left(f^{(-1)}(z)\right)}{\left|f'(u)|u = \left(f^{(-1)}(z)\right)\right|}$$

Where f is convex (and continuous), $\frac{1}{2}(f(x - \Delta\text{x}) + f(\Delta\text{x} + x)) \geq f(x)$, concave if $\frac{1}{2}(f(x - \Delta\text{x}) + f(\Delta\text{x} + x)) \leq f(x)$. Let us simplify with sole condition, assuming f(.) twice differentiable, $\frac{\partial^2 f}{\partial x^2} \geq 0$ for all values of x in the convex case and <0 in the concave one.

12.2.1 Some Examples.

Squaring x: p(x) is a Gaussian(with mean 0, standard deviation 1) , f(x)= x^2

$$g(x)=\frac{e^{-\frac{x}{2}}}{2\sqrt{2\pi}\sqrt{x}},x$$

$\geqslant 0$

which corresponds to the Chi-square distribution with 1 degrees of freedom.

Exponentiating x :p(x) is a Gaussian(with mean μ, standard deviation σ)

$$g(x)=\frac{e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma x}$$

which is the lognormal distribution.

12.3 Application 1: Happiness ($f(x)$)does not have the same statistical properties as wealth (x)

There is a conflation of fat-tailedness of Wealth and Utility

12.3.1 Case 1: The Kahneman Tversky Prospect theory, which is convex-concave

$$v(x)=\begin{cases} x^a & x\geq 0 \\ -\lambda\ \left(-x^a\right) & x<0 \end{cases}$$

with a & λ calibrated a = 0.88 and λ = 2.25

For x (the changes in wealth) following a T distribution with tail exponent α,

$$f(x)=\frac{\left(\frac{\alpha}{\alpha+x^2}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)}$$

Where B is the Euler Beta function, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a{+}b) = \int_0^1 t^{a-1}(1{-}t)^{b-1}dt$; we get (skipping the details of *z= v(u)* and *f(u) du = z(x) dx*), the distribution z(x) of the utility of happiness v(x)

$$z(x|\alpha,a,\lambda)=\begin{cases} \frac{x^{\frac{1-a}{a}}\left(\frac{\alpha}{\alpha+x^{2/a}}\right)^{\frac{\alpha+1}{2}}}{a\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)} & x\geq 0 \\ \frac{\left(-\frac{x}{\lambda}\right)^{\frac{1-a}{a}}\left(\frac{\alpha}{\alpha+\left(-\frac{x}{\lambda}\right)^{2/a}}\right)^{\frac{\alpha+1}{2}}}{a\lambda\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)} & x<0 \end{cases}$$

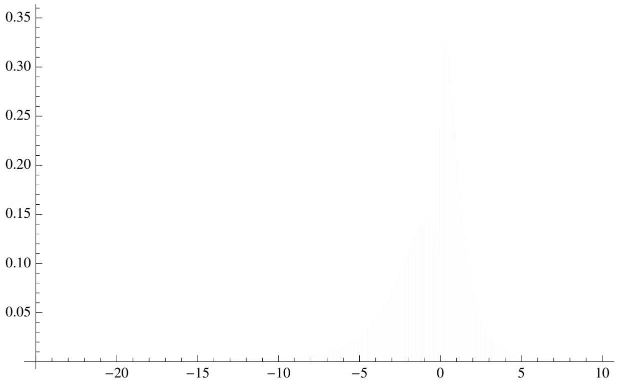


Figure 1: Simulation, first. The distribution of the utility of changes of wealth, when the changes in wealth follow a power law with tail exponent =2 (5 million Monte Carlo simulations).

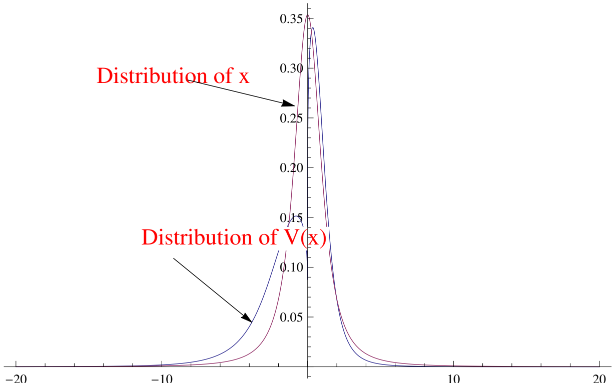


Figure 2: The graph in Figure 1 derived analytically

Fragility: as defined in the Taleb-Douady (2012) sense, on which later, i.e. tail sensitivity below K, v(x) is less “fragile” than x.

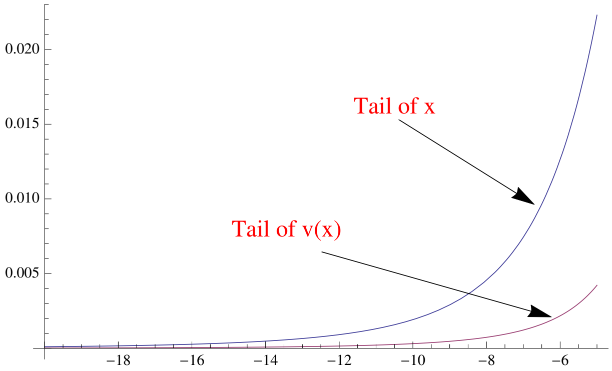


Figure 3: Left tail.

v(x) has thinner tails than x ⇔ more robust.

ASYMPTOTIC TAIL More technically the asymptotic tail for V(x) becomes $\frac{\alpha}{a}$ (i.e, for x and -x large, the exceedance probability for V, $P_{>x} \sim$ K $x^{-\frac{\alpha}{a}}$, with K a constant, or

$$z(x) \sim K x^{-\frac{\alpha}{a}-1}$$

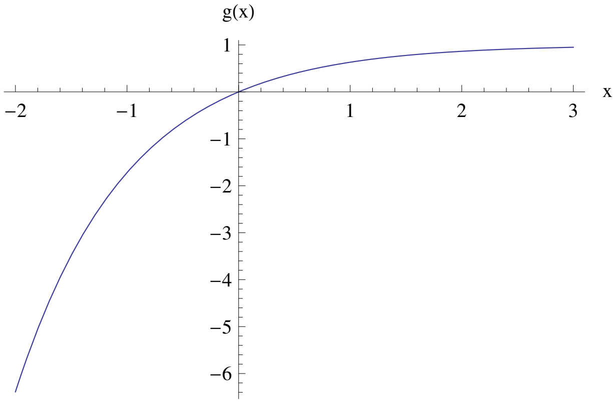
We can see that V(x) can easily have finite variance when x has an infinite one. The dampening of the tail has an increasingly consequential effect for lower values of α.

Case 2: Compare to the Monotone Concave of Classical Utility

Unlike the convex-concave shape in Kahneman Tversky, classical utility is monotone concave. This leads to plenty of absurdities, but the worst is the effect on the distribution of utility.

Granted one (K-T) deals with changes in wealth, the second is a function of wealth.

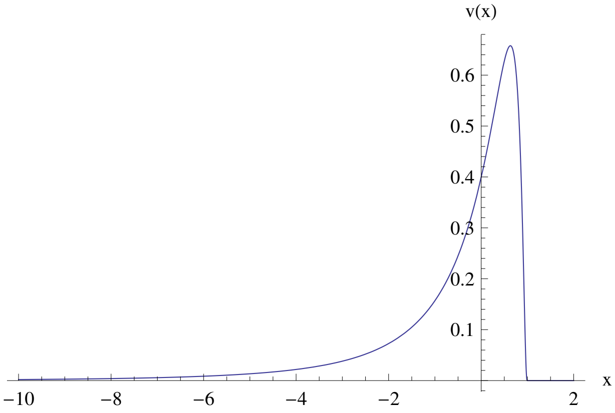
Take the standard concave utility function g(x)= 1- e^{-ax}. With a=1



Plot of 1- e^{-ax}

The distribution of v(x) will be

$$v(x) = -\frac{e^{-\frac{(\mu+\log(1-x))^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma(x-1)}$$



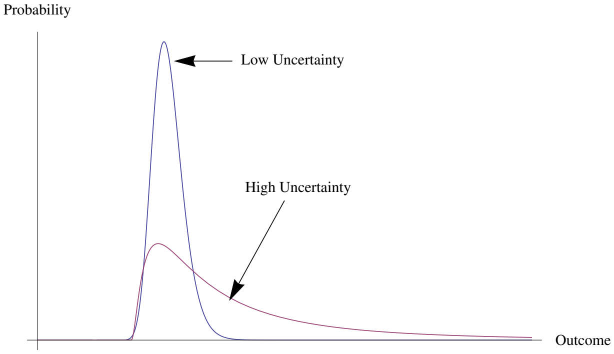
With such a distribution of utility it would be absurd to do anything.

The effect of convexity on the distribution of f(x)

Note the following property.

Distributions that are skewed have their mean dependent on the variance (when it exists), or on the scale. In other words, **more uncertainty raises the expectation.**

Demonstration 1:TK



Example: the Lognormal Distribution has a term $\frac{\sigma^2}{2}$ in its mean, linear to variance.

Example: the Exponential Distribution $1-e^{-x\lambda}$ $x \geq 0$ has the mean a concave function of the variance, that is, $\frac{1}{\lambda}$, the square root of its variance.

Example: the Pareto Distribution $L^\alpha x^{-1-\alpha}\alpha$ $x \geq L$, α >2 has the mean $\sqrt{\alpha-2}\sqrt{\alpha}$

× Standard Deviation, $\frac{\sqrt{\frac{\alpha}{\alpha-2}}L}{\alpha-1}$

12.4 The Mistake of Using Regular Estimation Methods When the Payoff is Convex

A simple way to see the point: the Ilmanen study assumes that one can derive strong conclusions from a single historical path not taking into account sensitivity to counterfactuals and completeness of sampling. It assumes that what one sees from a time series is the entire story.

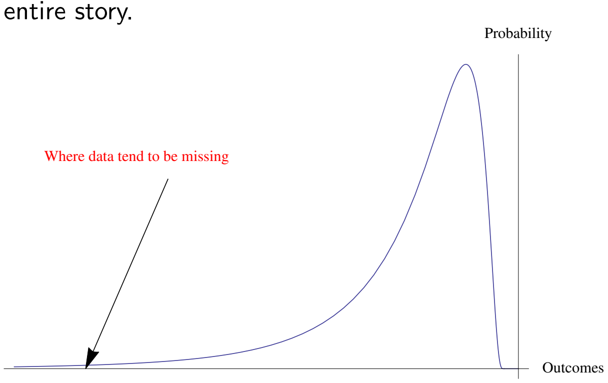


Figure 1: The Small Sample Effect and Naive Empiricism: When one looks at historical returns that are skewed to the left, most missing observations are in the left tails, causing an overestimation of the mean. The more skewed the payoff, and the thicker the left tail, the worst the gap between observed and true mean.

Now of concern for us is assessing the stub, or tail bias, that is, the difference between M and M*, or the potential contribution of tail events not seen in the window used for the analysis. When the payoff in the tails is powerful from convex responses, the stub becomes extremely large. So the rest of this note will go beyond the Ilmanen (2012) to explain the convexities of the payoffs in the tails and generalize to classical mistakes of testing strategies with explosive tail exposures on a finite simple historical sample. It will be based on the idea of metaprobability (or metamodel): by looking at effects of errors in models and representations. All one needs is an argument for a *very* small probability of a large payoff in the tail (devastating for the option seller) to reverse long shot arguments and make it uneconomic to sell a tail option. All it takes is a small model error to reverse the argument.

The Nonlineatities of Option Packages

There is a compounding effect of rarety of tail events and highly convex payoff when they happen, a convexity that is generally missed in the literature. To illustrate the point, we construct a "return on theta" (or return on time-decay) metric for a delta-neutral package of an option, seen at t_0 o given a deviation of magnitude $N\sigma_K$.

$$\Pi(N,K) \equiv \frac{1}{\theta_{S_0,t_0},\delta}\Big(O(S_0 e^{N\sigma_K\sqrt{\delta}},K,T-t_0,\sigma_K) - O\left(S_0,K,T-t_0-\delta,\sigma_K\right) - \Delta_{S_0,t_0}\left(1-S_0\right)e^{N\sigma_K\sqrt{\delta}}\Big), \tag{12.2}$$

where $0\left(S_0,K,T-t_0-\delta,\sigma_K\right)$is the European option price valued at time t_0 off an initial asset value S_0 , with a strike price K, a final expiration at time T, and priced using an "implied" standard deviation σ_K. The payoff of Π is the same whether O is a put or a call, owing to the delta-neutrality by hegding using a hedge ratio Δ_{S_0,t_0} (thanks to put-call parity, $\Delta_{S_0,t_0}$is negative if O is a call and positive otherwise). $\theta_{S_0,t_0}$is the discrete change in value of the option over a time increment δ (changes of value for an option in the absence of changes in any other variable). With the increment $\delta = 1/252$, this would be a single business day. We assumed interest rate are 0, with no loss of generality (it would be equivalent of expressing the problem under a risk-neutral measure). What 12.2 did is re-express the Fokker-Plank-Kolmogorov differential equation (Black Scholes), in discrete terms, away from the limit of δ →0. In the standard Black-Scholes World, the expectation of Π*(N,K)* should be zero, as N follows a Gaussian distribution with mean -$1/00082$ σ^2. But we are not about the Black Scholes world and we need to examine payoffs to potential distributions. The use of $\sigma_K$neutralizes the effect of "expensive" for the option as we will be using a multiple of σ_K as N standard deviations; if the option is priced at 15.87$\%$ volatility, then one standard deviation would correspond to a move of about 1$\%$, Exp[Sqrt[1/252]. 1587].

Clearly, for all K, Π[0,K]=-1 , Π[Sqrt[$2/\pi$],K]= 0 close to expiration (the break-even of the option without time premium, or when $T-t_0$= δ, takes place one mean deviation away), and Π[1,K]= 0.

12.4.1 Convexity and Explosive Payoffs

Of concern to us is the explosive nonlinearity in the tails. Let us examine the payoff of Π across many values of K= S_0 Exp[Λ σ_K $\sqrt{\delta}$], in other words how many "sigmas" away from the money the strike is positioned. A package about 20 σ out of the money , that is, Λ=20, the crash of 1987 would have returned 229,000 days of decay, compensating for $>$ 900 years of wasting premium waiting for the result. An equivalent reasoning could be made for subprime loans. From this we can assert that we need a minimum of 900 years of data to start pronouncing these options 20 standard deviations out-of-the money "expensive", in order to match the frequency that would deliver a payoff, and, more than 2000 years of data to make conservative claims. Clearly as we can see with Λ=0, the payoff is so linear that there is no hidden tail effect.

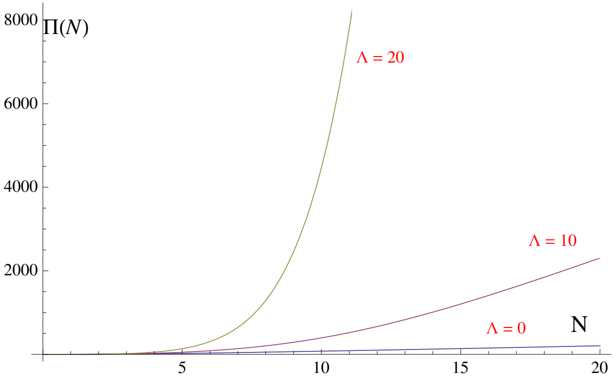


Figure 2: Returns for package Π*(N,K=* S_0Exp[Λ σ_K]) at values of Λ= 0,10,20 and N, the conditional "sigma" deviations.

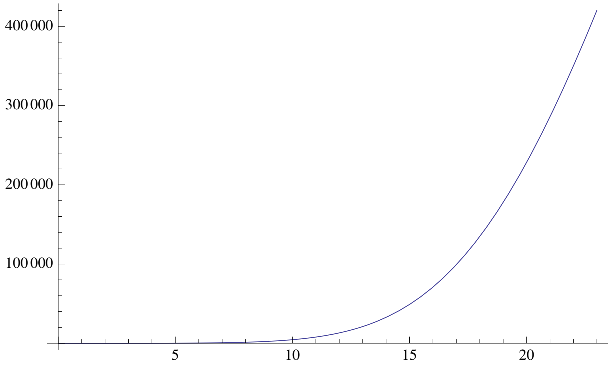


Figure 3: The extreme convexity of an extremely out of the money option, with Λ=20

Visibly the convexity is compounded by the fat-tailedness of the process: intuitively a convex transformation of a fat-tailed process, say a powerlaw, produces a powerlaw of considerably fatter tails. The Variance swap for instance results in $\frac{1}{2}$the tail exponent of the distribution of the underlying security, so it would have infinite variance with tail $\frac{3}{2}$ off the "cubic" exonent discussed in the literature (Gabaix et al,2003; Stanley et al, 2000) -and some out-of-the money options are more convex than variance swaps, producing tail equivalent of up to $\frac{1}{5}$ over a broad range of fluctuations.

For specific options there may not be an exact convex transformation. But we can get a Monte Carlo simulation illustrating the shape of the distribution and visually showing how sjewed it is.

Figure 4: In probability space. Histogram of the distribution of the returns Λ=10 using powerlaw returns for underlying distribution with α tail exponent =3.

Footnote 1: This convexity effect can be mitigated by some dynamic hedges, assuming no gaps but, because of "local time" for stochastic processes, in fact, some smaller deviations can carry the cost of larger ones: for a move of -10 sigmas followed by an upmove of 5 sigmas revision can end up costing a lot more than a mere -5 sigmas. Tail events can come from a volatile sample path snapping back and forth.

Fragility Heuristic and Nonlinear Exposure to Implied Volatility

Most of the losses from option portfolios tend to take place from the explosion of implied volatility, therefore acting as if the market had already experienced a tail event (say in 2008). The same result as Figure 3 can be seen for changes in implied volatility: an explosion of volatility by 5 × results in a 10 σ option gaining 270 × (the VIx went up $>$ 10 × during 2008). (In a well publicized debacle, the speculator Niederhoffer went bust because of explosive changes in implied volatility in his option portfolio, not from market movement; further, the options that bankrupted his fund ended up expiring worthless weeks later).

The Taleb and Douady (2012), Taleb Canetti et al (2012) fragility heuristic identifies convexity to significant parameters as a metric to assess fragility to model error or representation: by theorem, model error maps directly to nonlinearity of parameters. The heuristic corresponds to the perturbation of a parameter, say the scale of a probability distribution and looks at the effect of the expected shortfall; the same theorem asserts that the asymmetry between gain and losses (convexity) maps directly to the exposure to model error and to fragility. The exercise allows us to re-express the idea of convexity of payoff by ranking effects.

	$\times 2$	$\times 3$	$\times 4$
ATM	2	3	4
$\Lambda = 5$	5	10	16
$\Lambda = 10$	27	79	143
$\Lambda = 20$	7686	72741	208429

The Table presents differents results (in terms of multiples of option premia over intrinsic value) by multiplying implied volatility by 2, 3,4. An option 5 conditional standard deviations out of the money gains 16 times its value when implied volatility is multiplied by 4. Further out of the money options gain exponentially. Note the linearity of at-the-money options

Conclusion: The Asymmetry in Decision Making

To assert overpricing (or refute underpricing) of tail events expressed by convex instruments requires an extraordinary amount of "evidence", a much longer time series about the process and strong assumptions about temporal homogeneity. Out of the money options are so convex to events that a single crash (say every 50, 100, 200, even 900 years) could be sufficient to justify skepticism about selling *some* of them (or avoiding to sell them) –those whose convexity matches the frequency of the rare event. The further out in the tails, the less claims one can make about their "value", state of being "expensive', etc. One can make claims on "bounded" variables perhaps, not for the tails.

References

Ilmanen, Antti, 2012, "Do Financial Markets Reward Buying or Selling Insurance and Lottery Tickets?" Financial Analysts Journal, September/October, Vol. 68, No. 5 : 26 - 36.

Golec, Joseph, and Maurry Tamarkin. 1998. "Bettors Love Skewness, Not Risk, at the Horse Track." Journal of Political Economy, vol. 106, no. 1 (February) , 205-225.

Snowberg, Erik, and Justin Wolfers. 2010. "Explaining the Favorite - Longshot Bias : Is It Risk - Love or Misperceptions?" Working paper.

Taleb, N.N., 2004, "Bleed or Blowup? Why Do We Prefer Asymmetric Payoffs?" Journal of Behavioral Finance, vol. 5, no. 1.

13 | Generalized Dose-Response Curves and The Origin of Thin-Tails

The literature of heavy tails starts with a random walk and finds mechanisms that lead to fat tails under aggregation. We follow the inverse route and show how starting with fat tails we get to thin-tails from the probability distribution of the response to a random variable. We introduce a general dose-response curve show how the left amd right-boundedness of the reponse in natural things leads to thin-tails, even when the "underlying" variable of the exposure is fat-tailed.

The Origin of Thin Tails.

We have emprisoned the "statistical generator" of things on our planet into the random walk theory: the sum of i.i.d. variables eventually leads to a Gaussian, which is an appealing theory. Or, actually, even worse: at the origin lies a simpler Bernouilli binary generator with variations limited to the set {0,1}, normalized and scaled, under summation. Bernouilli, De Moivre, Galton, Bachelier: all used the mechanism, as illustrated by the Quincunx in which the binomial leads to the Gaussian. This has traditionally been the "generator" mechanism behind everything, from martingales to simple convergence theorems. Every standard textbook teaches the "naturalness" of the thus-obtained Gaussian.

In that sense, powerlaws are pathologies. Traditionally, researchers have tried to explain fat tailed distributions using the canonical random walk generator, but twinging it thanks to a series of mechanisms that start with an aggregation of random variables that does not lead to the central limit theorem, owing to lack of independence and the magnification of moves through some mechanism of contagion: preferential attachment, comparative advantage, or, alternatively, rescaling, and similar mechanisms.

But the random walk theory fails to accommodate some obvious phenomena.

First, many things move by jumps and discontinuities that cannot come from the random walk and the conventional Brownian motion, a theory that proved to be sticky (Mandelbrot, 1997).

Second, consider the distribution of the size of animals in nature, considered within-species. The height of humans follows (almost) a Normal Distribution but it is hard to find mechanism of random walk behind it (this is an observation imparted to the author by Yaneer Bar Yam).

Third, uncertainty and opacity lead to power laws, when a statistical mechanism has an error rate which in turn has an error rate, and thus, recursively (Taleb, 2011, 2013).

Our approach here is to assume that random variables, under absence of contraints, become power law-distributed. This is the default in the absence of boundedness or compactness. Then, the *response*, that is, a funtion of the random variable, considered in turn as an "inherited" random variable, will have different properties. If the response is bounded, then the dampening of the tails of the inherited distribution will lead it to bear the properties of the Gaussian, or the class of distributions possessing finite moments of all orders.

The Dose Response

Let $S^N(x)$: $\mathbb{R} \to [k_L, k_R]$ be a continuous function possessing derivatives $\left(S^N\right)^{(n)}(x)$ of all orders, expressed as an N-summed and scaled standard sigmoid functions:

$$S^N(x) \equiv \sum_{i=1}^{N} \frac{a_k}{1+\exp\left(-b_k x + c_k\right)}$$

(13.1)

where a_k, b_k, c_k are norming constants $\in \mathbb{R}$, satisfying:
i) S^N(-∞) $=k_L$
ii) $S^N(\infty)$ $=k_R$
and (equivalently for the first and last of the following conditions)
iii) $\frac{\partial^2 S^N}{\partial x^2} \geq$ 0 for $x \in$ (-∞, k_1) , $\frac{\partial^2 S^N}{\partial x^2} <$ 0 for $x \in$ (k_2, $k_{>2}$), and $\frac{\partial^2 S^N}{\partial x^2} \geq$ 0 for $x \in$ ($k_{>2}$, ∞), with $k_1 > k_2 \geq k_3 ... \geq k_N$.
The shapes at different calibrations are shown in Figure 1, in which we combined different values of N=2 $S^2\left(x, a_1, a_2, b_1, b_2, c_1, c_2\right)$, and the standard sigmoid $S^1\left(x, a_1, b_1, c_1\right)$, with a_1=1, b_1=1 and c_1=0. As we can see, unlike the common sigmoid, the asymptotic response can be lower than the maximum, as our curves are not monotonically increasing. The sigmoid shows benefits increasing rapidly (the convex phase), then increasing at a slower and slower rate until saturation. Our more general case starts by increasing, but the reponse can be actually negative beyond the saturation phase, though in a convex manner. Harm slows down and becomes "flat" when something is totally broken.

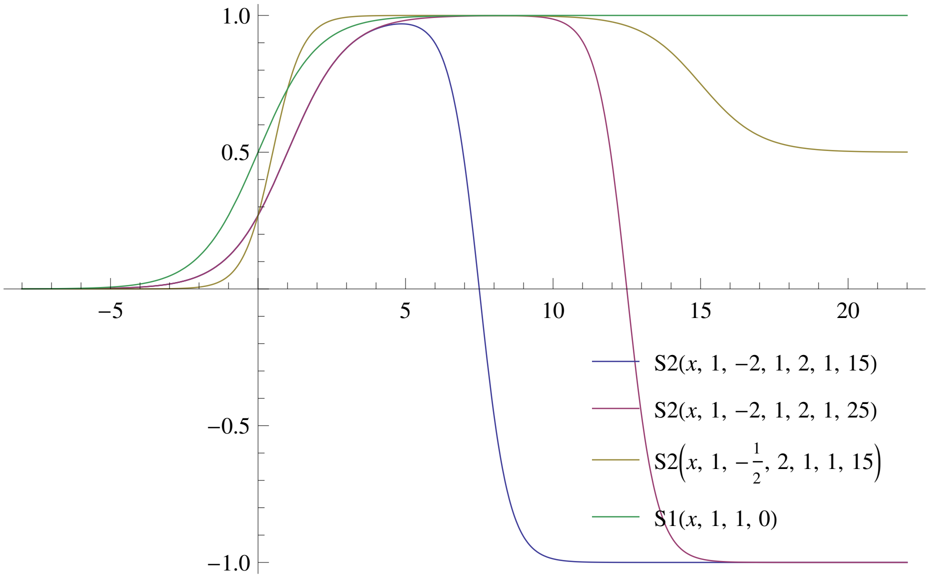


Figure 13.1: The Generalized Response Curve, $S^2\left(x, a_1, a_2, b_1, b_2, c_1, c_2\right)$, $S^1\left(x, a_1, b_1, c_1\right)$ *The convex part with positive first detivative has been designated as "antifragile"*

13.1 Properties of the Inherited Probability Distribution

Now let x be a random variable with distributed according to a general fat tailed distribution, with power laws at large negative and positive values, expressed (for clarity, without loss of generality) as a Student T Distribution with scale σ and exponent α, and support on the real line. Its domain $\mathcal{D}^f$= (∞, ∞), and density $f_{\sigma,\alpha}(x)$:

$$x f_{\sigma,\alpha} \equiv \frac{\left(\frac{\alpha}{\alpha+\frac{x^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\sigma B\left(\frac{\alpha}{2}, \frac{1}{2}\right)} \tag{13.2}$$

where $B(a,b) = \frac{(a\Gamma)(b\Gamma)}{\Gamma(a+b)} = \int_0^1 dt t^{a-1}(1-t)^{b-1}$
. The simulation effect of the convex-concave transformations of the terminal probability distribution is shown in Figure 2.

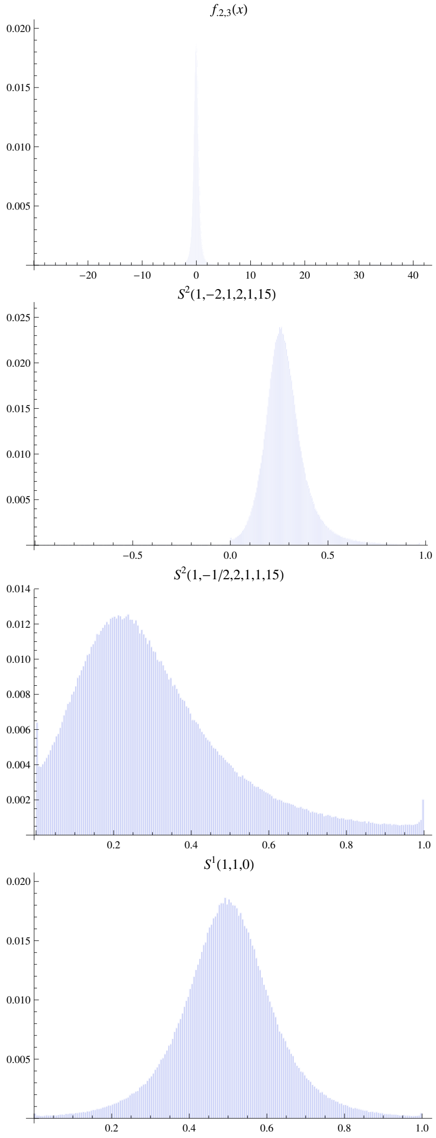


Figure 13.2: Histograms for the different inherited probability distributions (simulations ,$N = 10^6$)

And the Kurtosis of the inherited distributions drops at higher σ thanks to the boundedness of the payoff, making the truncation to the left and the right visible. Kurtosis for $f_{.2,3}$ is infinite, but in-sample will be extremely high, but, of course, finite. So we use it as a benchmark to see the drop from the calibration of the response curves.

Distribution	Kurtosis
$f_{.2,3}(x)$	86.3988
$S^2(1,-2,1,2,1,15)$	8.77458
$S^2(1,-1/2,2,1,1,15)$	4.08643
$S^1(1,1,0)$	4.20523

Case of the standard sigmoid, i.e., $N = 1$

$$S(x) \equiv \frac{a_1}{1+\exp(-b_1 x + c_1)}$$

(13.3)

g(x) is the inherited distribution, which can be shown to have a scaled domain $\mathcal{D}^g$= (k_L, k_R). It becomes

$$g(x) = \frac{a1\left(\frac{\alpha}{\alpha+\frac{\left(\log\left(\frac{x}{a1-x}\right)+c1\right)^2}{b1^2\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}b1\sigma x B\left(\frac{\alpha}{2},\frac{1}{2}\right)(a1-x)}$$

(13.4)

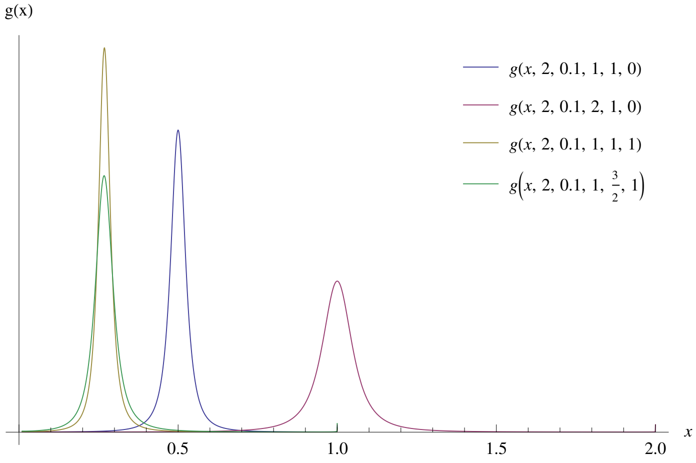


Figure 13.3: The different inherited probability distributions.

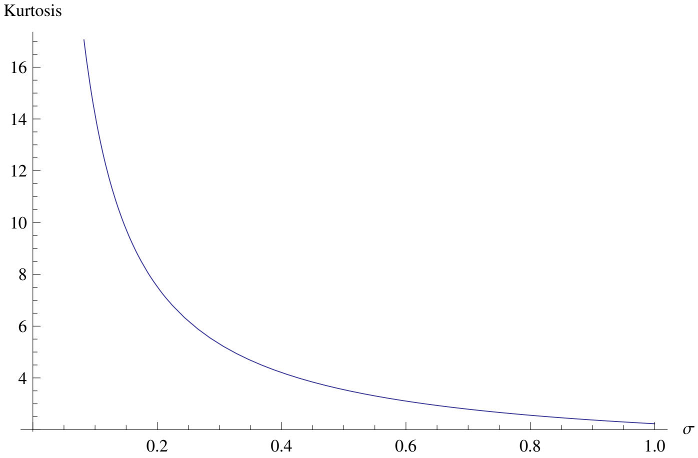


Figure 13.4: The Kurtosis of the standard drops along with the scale σ of the power law

Remark 1: The inherited distribution from S(x) will have a compact support regardless of the probability distribution of x.

13.2 Conclusion and Remarks

We showed the dose-response as the neglected origin of the thin-tailedness of observed distributions in nature. This approach to the dose-response curve is quite general, and can be used outside biology (say in the Kahneman-Tversky prospect theory, in which their version of the utility concept with respect to changes in wealth is concave on the left, hence bounded, and convex on the right.

14 Why The World Will Progressively Look Wierder to Us

The paradox is that increase in sample size *magnifies* the role of noise (or luck); it makes tail values even more extreme. There are some problems associated with big data.

14.1 How Noise Explodes Faster than Data

To the observer, every day will seem wierder than the previous one. It has always been absolutely silly to be exposed the news. Things are worse today thanks to the web.

Source	Effect
News	Wierder and wierder events reported on the front pages
Big Data	More spurious "statistical" relationships that eventually fail to replicate, with more accentuated effects and more statistical "significance" (sic)
Track Records	Greater performance for (temporary) "star" traders

We are getting more information, but with constant "consciouness", "desk space", or "visibility". Google News, Bloomberg News, etc. have space for, say, $<$100 items at any point in time. But there are millions of events every day. As the world is more connected, with the global dominating over the local, the number of sources of news is multiplying. But your consciousness remains limited. So we are experiencing a winner-take-all effect in information: like a large movie theatre with a small door.

Likewise we are getting more data. The size of the door is remaining constant, the theater is getting larger.

The winner-take-all effects in information space corresponds to more noise, less signal. In other words the spurious dominates.

Similarity with the Fooled by Randomness Bottleneck

This is similar to the idea that the more spurious returns dominate finance as the number of players get large, and swamp the more solid ones. Start with the idea (see Taleb 2001), that as a population of operators in a profession marked by a high degrees of randomness increases, the number of stellar results, and stellar for completely random reasons, gets larger. The "spurious tail" is therefore the number of persons who rise to the top for no reasons other than mere luck, with subsequent rationalizations, analyses, explanations, and attributions. The performance in the "spurious tail" is only a matter of number of participants, the base population of those who tried. Assuming a symmetric market, if one has for base population 1 million persons with zero skills and ability to predict starting Year 1, there should be 500K spurious winners Year 2, 250K Year 3, 125K Year 4, etc. One can easily see that the size of the winning population in, say, Year 10 depends on the size of the base population Year 1; doubling the initial population would double the straight winners. Injecting skills in the form of better-than-random abilities to predict does not change the story by much. (Note that this idea has been severely plagiarized by someone, about which a bit more soon).

Because of scalability, the top, say 300, managers get

the bulk of the allocations, with the lion's share going to the top 30. So it is obvious that the winner-take-all effect causes distortions: say there are m initial participants and the "top" k managers selected, the result will be $\frac{k}{m}$ managers in play. As the base population gets larger, that is, N increases linearly, we push into the tail probabilities.
Here read skills for information, noise for spurious performance, and translate the problem into information and news.

The paradox: This is quite paradoxical as we are accustomed to the opposite effect, namely that a large increases in sample size reduces the effect of sampling error; here the narrowness of M puts sampling error on steroids.

14.2 Derivations

Let $Z \equiv \left(z_i^j\right)_{1<j<m,1\leq i<n}$be a $(n \times m)$ sized population of variations, m population series and n data points per distribution, with $i, j \in \mathbb{N}$; assume "noise" or scale of the distribution $\sigma \in \mathbb{R}^+$, signal $\mu \geq$0 . Clearly σ can accommodate distributions with infinite variance, but we need the expectation to be finite. Assume i.i.d. for a start.

Cross Sectional ($n = 1$)

Special case n = 1: we are just considering news/data without historical attributes.
Let $F^{\leftarrow}$ be the generalized inverse distribution, or the quantile,

$$F^{\leftarrow}(w) = \inf\{t \in \mathbb{R} : F(t) \geq w\},$$

for all nondecreasing distribution functions $F(x) \equiv \mathbb{P}(X < x)$. For distributions without compact support, $w \in$ (0,1); otherwise $w \in [0, 1]$. In the case of continuous and increasing distributions, we can write F^{-1} instead.
The signal is in the expectaion, so $\mathbb{E}(z)$ is the signal, and σ the scale of the distribution determines the noise (which for a Gaussian corresponds to the standard deviation). Assume for now that all noises are drawn from the same distribution.
Assume constant probability the "threshold", ζ= $\frac{k}{m}$, where k is the size of the window of the arrival. Since we assume that k is constant, it matters greatly that the quantile covered shrinks with m.

Gaussian Noise

When we set ζ as the reachable noise. The quantile becomes:

$$F^{-1}(w) = \sqrt{2}\ \sigma\ \text{erfc}^{-1}(2w) + \mu,$$

where erfc^{-1}is the inverse complementary error function.
Of more concern is the survival function, $\Phi \equiv \overline{F(x)} \equiv \mathbb{P}(X > x)$, and its inverse Φ^{-1}

$$\Phi^{-1}{}_{\sigma,\mu}(\zeta) = -\sqrt{2}\sigma\text{erfc}^{-1}\left(2\frac{k}{m}\right) + \mu$$

Note that σ (noise) is multiplicative, when μ (signal) is additive.
As information increases, ζ becomes smaller, and Φ^{-1} moves away in standard deviations. But nothing yet by comparison with Fat tails.

Fat Tailed Noise

Now we take a Student T Distribution as a substitute to the Gaussian.

$$f(x) \equiv \frac{\left(\frac{\alpha}{\alpha+\frac{(x-\mu)^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\ \sigma\ B\left(\frac{\alpha}{2},\frac{1}{2}\right)} \tag{14.1}$$

Where we can get the inverse survival function.

$$\gamma^{-1}{}_{\sigma,\mu}(\zeta) = \mu + \sqrt{\alpha}\ \ \sigma\ \ \text{sgn}\,(1 - 2\ \zeta)\sqrt{\frac{1}{I^{-1}_{(1,(2\zeta-1)\text{sgn}(1-2\zeta))}\left(\frac{\alpha}{2},\frac{1}{2}\right)} - 1} \tag{14.2}$$

where I is the generalized regularized incomplete Beta function $I_{(z_0,z_1)}(a,b) = \frac{B_{(z_0,z_1)}(a,b)}{B(a,b)}$, and $B_z(a,b)$ the incomplete Beta function $B_z(a,b) = \int_0^z t^{a-1}(1-t)^{b-1}dt$. $B(a,b)$ is the Euler Beta function $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$.

As we can see in Figure 2, the explosion in the tails of noise, and noise only.

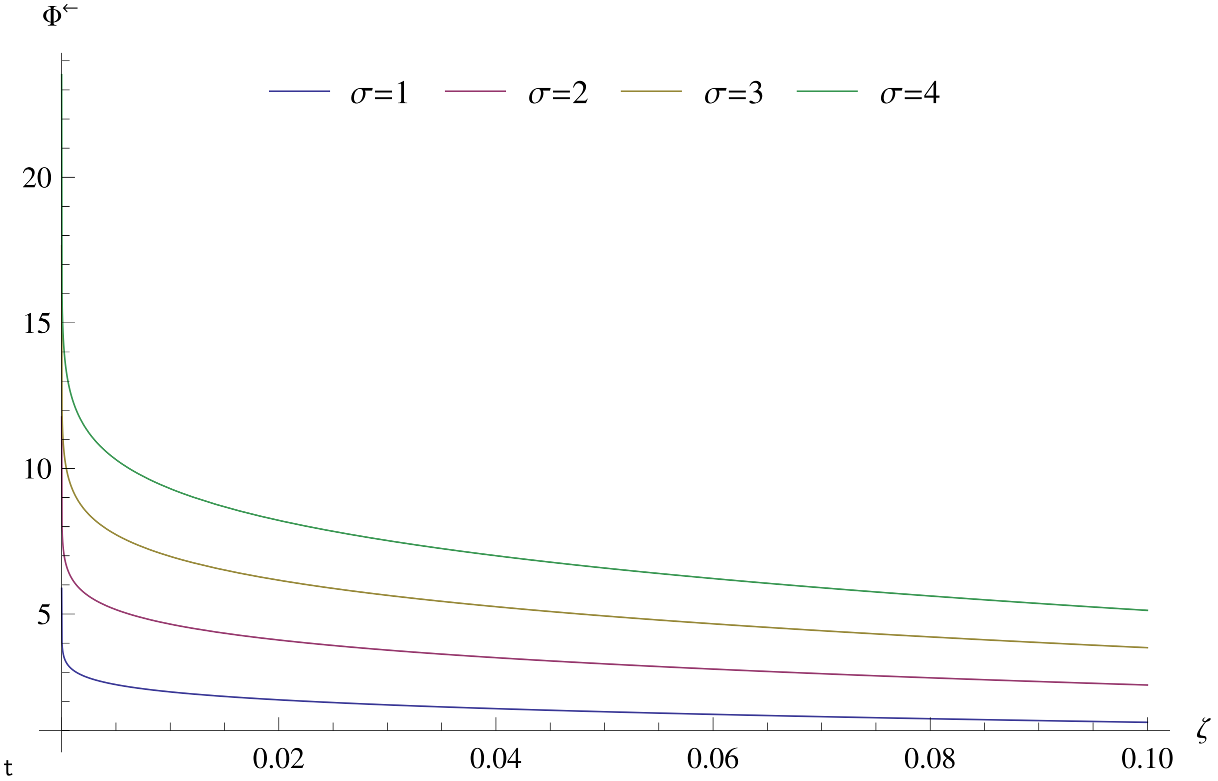


Figure 14.1: Gaussian, σ={1,2,3,4}

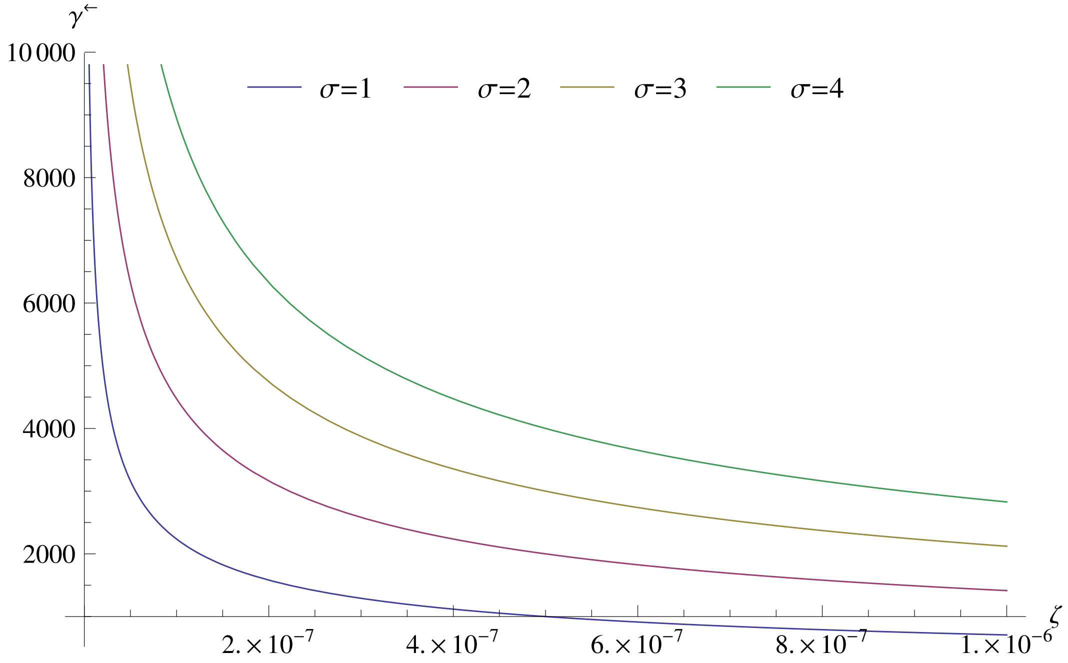


Figure 14.2: Power Law, σ={1,2,3,4}

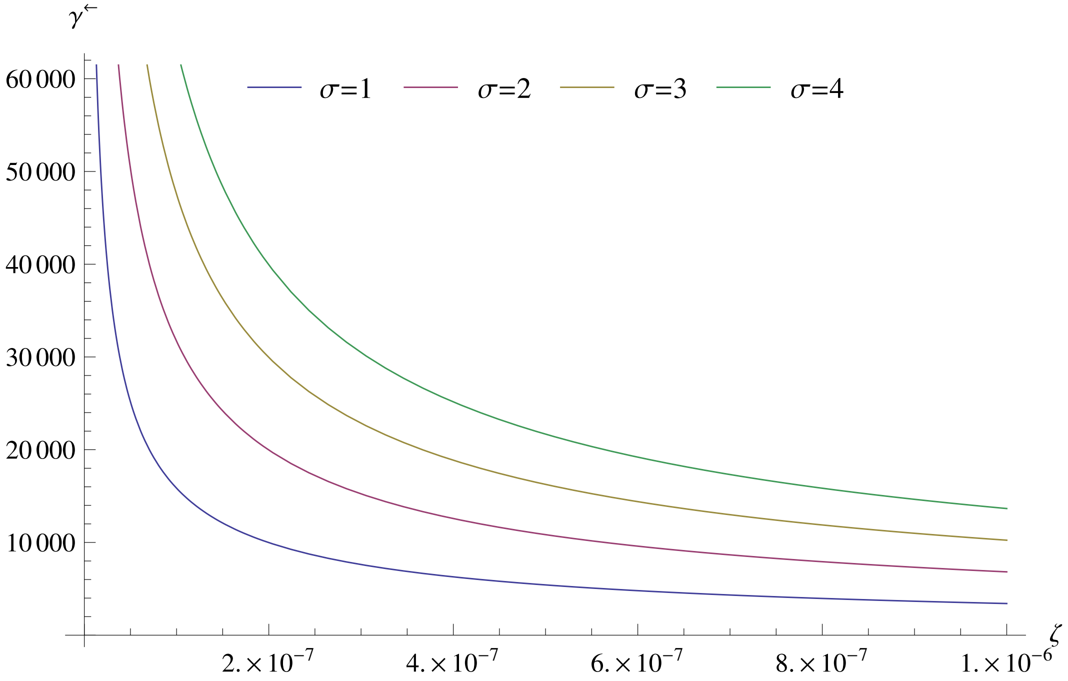


Figure 14.3: Alpha Stable Distribution

Fatter Tails: Alpha Stable Distribution

Part 2 of the discussion to come soon.