SILENT RISK

NASSIM NICHOLAS TALEB

Lectures on Probability, Fragility,
& Asymmetric Exposures

In which is provided a mathematical parallel version of the author's *Incerto*, with derivations, examples, theorems, & heuristics.

2015 (DRAFT)

This format is based on André Miede's *ClassicThesis*.

ABSTRACT

"Empirical evidence that the boat is safe", or how we tend to be fooled by silent risks. *Factum stultus cognoscit* (The fool only understand risks *after* the harm is done). Risk is both precautionary (fragility based) and evidentiary (statistical based); it is too serious a business to be left to mechanistic users of probability theory.
This figure encapsulates the scientific "nonsucker" approach to risk and probability. Courtesy George Nasr.

***

This book provides a mathematical framework for decision making and the analysis of (consequential) hidden risks, those tail events undetected or improperly detected by statistical machinery; and substitutes fragility as a more reliable measure of exposure. Model error is mapped as risk, even tail risk.[1]

Risks are seen in tail events rather than in the variations; this necessarily links them mathematically to an asymmetric response to intensity of shocks, convex or concave.

The difference between "models" and "the real world" ecologies lies largely in an additional layer of uncertainty that typically (because of the same asymmetric response by small probabilities to additional uncertainty) thickens the tails and invalidates *all* probabilistic tail risk measurements — models, by their very nature of reduction, are vulnerable to a chronic underestimation of the tails.

So tail events are not measurable; but the good news is that exposure to tail events is. In "Fat Tail Domains" (Extremistan), tail events are rarely present in past

1 This is a polite way to say No-BS approach to probability.

data: their statistical presence appears too late, and time series analysis is similar to sending troops after the battle. Hence the concept of fragility is introduced: is one vulnerable (i.e., asymmetric) to model error or model perturbation (seen as an additional layer of uncertainty)?

Part I looks at the consequences of fat tails, mostly in the form of slowness of convergence of measurements under the law of large number: some claims require 400 times more data than thought. Shows that much of the statistical techniques used in social sciences are either inconsistent or incompatible with probability theory. It also explores some errors in the social science literature about moments (confusion between probability and first moment, etc.)

Part II proposes a more realistic approach to risk measurement: fragility as nonlinear (concave) response, and explores nonlinearities and their statistical consequences. Risk management would consist in building structures that are not negatively asymmetric, that is both "robust" to both model error and tail events. Antifragility is a convex response to perturbations of a certain class of variables.

CONTENTS

Acknowledgments xv
Notes for Reviewers xviii

i FIGURING OUT PROBABILITY AND WHAT IT MEANS 1

1 WHAT IS PROBABILITY? WHAT IS EXPOSURE? 3
- 1.1 The Conflation of Events and Exposures 4
  - 1.1.1 Contract Theory 7
- 1.2 Payoff Classes $\mathfrak{P}_1$ through $\mathfrak{P}_4$ 10
  - 1.2.1 Atomic Payoff $\mathfrak{P}_1$ 12
  - 1.2.2 Binary Payoff Class $\mathfrak{P}_2$ 14
  - 1.2.3 Vanilla Payoff Class $\mathfrak{P}_3$, building blocks for regular exposures. 15
  - 1.2.4 Composite/Sigmoidal Payoff Class $\mathfrak{P}_4$ 18
- 1.3 Achieving Nonlinearity through $\mathfrak{P}_4$ 18
- 1.4 Main Errors in the Literature 18
- 1.5 The Applicability of Some Psychological Biases 19
- 1.6 Misfitness of Prediction Markets 19
  - 1.6.1 The Black Swan is Not About Probability But Payoff 19
  - 1.6.2 Chernoff Bound 21
  - 1.6.3 Fatter tails lower the probability of remote events (the binary) and raise the value of the vanilla. 21
  - 1.6.4 The law of large numbers works better with the binary than the variable 22
- 1.7 Finding Inconsistencies in Scholarly Treatments of Events 24
- 1.8 Metaprobability and the Payoff Kernel 26
- 1.9 Classification and Codification of Exposures 31
- 1.10 Numeraire definition 31
- 1.11 What is an Insurable Risk? 32
- 1.12 Ruin problems 32
- 1.13 Skepticism, Uncertainty, and Scale of a Distributon 32
- 1.14 Why Pascal Wager Has Nothing to Do With the Left Tail 32

2 THE "REAL WORLD" RIGOR PROJECT 33
- 2.1 A Course With an Absurd Title 35
- 2.2 Problems and Inverse Problems 38
- 2.3 Fragility, not Just Statistics, For Hidden Risks 47
- 2.4 Solution: The Convex Heuristic 48
  - 2.4.1 Convex Heuristics, Rationality, and Revelation of Preferences 49
- 2.5 Fragility and Model Error 50
  - 2.5.1 Why Engineering? 50
- 2.6 General Classification of Problems Related To Fat Tails 53

General Classification of Problems Related To Fat Tails 53

- 2.7 Closing the Introduction 55

A WHAT'S A CHARLATAN IN RISK AND PROBABILITY? 57

A.1 Charlatan 57

ii FAT TAILS: THE LLN UNDER REAL WORLD ECOLOGIES 63

3 FAT TAILS AND THE PROBLEM OF INDUCTION 65
3.1 The Problem of (Enumerative) Induction 65
3.2 Empirical Risk Estimators 65
3.3 Fat Tails, the Finite Moment Case 68
3.4 A Simple Heuristic to Create Mildly Fat Tails 71
3.5 The Body, The Shoulders, and The Tails 73
3.6 Fattening of Tails With Skewed Variance 76
3.7 Fat Tails in Higher Dimension 78
3.8 Scalable and Nonscalable, A Deeper View of Fat Tails 79
3.9 Subexponential as a class of fat tailed distributions 81
3.10 Joint Fat-Tailedness and Elliptical Distributions 87
3.11 Different Approaches For Statistical Estimators 90
3.12 Econometrics imagines functions in L^2 Space 94
3.13 Typical Manifestations of The Turkey Surprise 95
3.14 Metrics for Functions Outside L^2 Space 99
3.15 Using the Hilbert Transform 100
3.16 A Comment on Bayesian Methods in Risk Management 101

B SPECIAL CASES OF FAT TAILS 103
B.1 Multimodality and Fat Tails, or the War and Peace Model 103
B.2 Transition probabilites: what can break will break 105

C QUICK AND ROBUST MEASURE OF FAT TAILS 109
C.1 Introduction 109
C.2 First Metric, the Simple Estimator 109
C.3 Second Metric, the Ξ_2 estimator 111

4 HIERARCHY OF DISTRIBUTIONS FOR ASYMMETRIES 113
4.1 Permissible Empirical Statements 113
4.2 Masquerade Example 114
4.3 The Probabilistic Version of Absense of Evidence 115
4.4 Via Negativa and One-Sided Arbitrage of Statistical Methods 115
4.5 Hierarchy of Distributions in Term of Tails 116
4.6 How To Arbitrage Kolmogorov-Smirnov 120
4.7 Mistaking Evidence for Anecdotes & The Reverse 123

5 EFFECTS OF HIGHER ORDERS OF UNCERTAINTY 125
5.1 Meta-Probability Distribution 125
5.2 Metadistribution and the Calibration of Power Laws 127
5.3 The Effect of Metaprobability on Fat Tails 128
5.4 Fukushima, Or How Errors Compound 128
5.5 The Markowitz inconsistency 129
5.6 Psychological pseudo-biases under second layer of uncertainty. 131
5.6.1 The pathologization fallacy 131

6 LARGE NUMBERS AND CLT IN THE REAL WORLD 137
6.0.2 The "Pinker Problem" 137
6.1 The problem of Matching Errors 139
6.2 Generalizing Mean Deviation as Partial Expectation 141
6.3 Class of Stable Distributions 141

6.3.1 Results 142
6.3.2 Stochastic Alpha or Mixed Samples 144
6.4 Symmetric NonStable Distributions in the Subexponential Class 145
6.4.1 Symmetric Mixed Gaussians, Stochastic Mean 145
6.4.2 Half cubic Student T (Lévy Stable Basin) 146
6.4.3 Cubic Student T (Gaussian Basin) 147
6.5 Asymmetric NonStable Distributions in the Subexponetial Class 149
6.5.1 One-tailed Pareto Distributions 149
6.5.2 The Lognormal and Borderline Subexponential Class 149
6.6 Asymmetric Distributions in the Superexponential Class 149
6.6.1 Mixing Gaussian Distributions and Poisson Case 149
6.6.2 Skew Normal Distribution 149
6.6.3 Super-thin tailed distributions: Subgaussians 150
6.7 Acknowledgement 151
6.7.1 Derivations using explicit $\mathbb{E}(|X|)$ 151
6.7.2 Derivations using the Hilbert Transform and $\beta = 0$ 151
D IN PROGRESS DERIVATIONS FOR LLN ACROSS FAT TAILS 153
D.1 Comments on LLN and Bounds 153
D.1.1 Speed of Convergence for Simple Cases 153
D.1.2 Comparing N = 1 to N = 2 for a symmetric power law with 1< α ≤2. 155
D.2 Digression Into Inversion of Characteristic Function of NonStable Power Law 157
D.2.1 Integrable Characteristic Functions 158
D.3 Pulling the PDF of an n-Summed Student T 158
7 PREASYMPTOTICS AND CENTRAL LIMIT IN THE REAL WORLD 161
7.1 Using Log Cumulants to Observe Preasymptotics 167
7.2 Convergence of the Maximum of a Finite Variance Power Law 171
7.3 Sources and Further Readings 171
7.4 Convergence for Non-Lévy StablePower Laws Temporarily here) 172
E WHERE STANDARD DIVERSIFICATION FAILS 175
F FAT TAILS AND RANDOM MATRICES 179
8 SOME MISUSES OF STATISTICS IN SOCIAL SCIENCE 181
8.1 Mechanistic Statistical Statements 181
8.2 Attribute Substitution 181
8.3 The Tails Sampling Property 183
8.4 A discussion of the Paretan 80/20 Rule 186
8.5 Survivorship Bias (Casanova) Property 188
8.6 Left (Right) Tail Sample Insufficiency Under Negative (Positive) Skewness 190
8.7 Why N=1 Can Be Very, Very Significant Statistically 191
8.8 The Instability of Squared Variations in Regressions 191
8.9 Statistical Testing of Differences Between Variables 195
8.10 Studying the Statistical Properties of Binaries and Extending to Vanillas 196
8.11 Why Economics Time Series Don't Replicate 196

8.12 A General Summary of The Problem of Reliance on Past Time Series 201
8.13 Conclusion 202
G ON THE INSTABILITY OF ECONOMETRIC DATA 203
9 FAT TAILS FROM RECURSIVE UNCERTAINTY 205
9.1 Layering uncertainty 205
9.2 Regime 1 (Explosive): Case of a constant error parameter a 211
9.3 Convergence to Power Laws 212
9.4 Regime 1b: Preservation of Variance 214
9.5 Regime 2: Cases of decaying parameters a_n 214
9.6 Conclusion and Suggested Application 217
10 PARAMETRIZATION AND TAILS 219
10.1 Some Bad News Concerning power laws 219
10.2 Extreme Value Theory: Not a Panacea 220
10.3 Using Power Laws Without Being Harmed by Mistakes 224
H POISSON VS. POWER LAW TAILS 227
H.1 Beware The Poisson 227
H.2 Leave it to the Data 228
11 BROWNIAN MOTION IN THE REAL WORLD 231
11.1 Path Dependence and History as Revelation of Antifragility 231
11.2 SP and path dependence (incomplete) 232
11.3 Brownian Motion in the Real World 233
11.4 Stochastic Processes and Nonanticipating Strategies 235
11.5 Finite Variance not Necessary for Anything Ecological (incl. quant finance) 235
12 THE FOURTH QUADRANT "SOLUTION" 237
12.1 Two types of Decisions 237
13 RISK AND PRECAUTION 239
13.0.1 Stopping time equivalence 240
13.0.2 Sign of exposure 240
13.0.3 layering 240
13.1 What is the Precautionary Principle 240
13.2 Why Ruin is Serious Business 241
13.3 Skepticism and Precaution 241
13.4 Fallacious arguments in Risk Management 243
13.4.1 Crossing the road (the paralysis fallacy) 243
13.4.2 The Psychology of Risk and Thick Tailed Distributions 243
13.4.3 The Loch Ness fallacy 244
13.4.4 The fallacy of misusing the naturalistic fallacy 244
13.4.5 The "Butterfly in China" fallacy 245
13.4.6 The potato fallacy 245
13.4.7 The Russian roulette fallacy (the counterexamples in the risk domain) 246
13.4.8 The Carpenter Fallacy 246
13.4.9 The technological salvation fallacy 247
13.4.10 The pathologization fallacy 247
14 SKIN IN THE GAME AND RISK TAKING 249

14.1 Payoff Skewness and Lack of Skin-in-the-Game 254
iii (ANTI)FRAGILITY AND NONLINEAR RESPONSES TO RANDOM VARIABLES 261
15 EXPOSURES AS TRANSFORMED RANDOM VARIABLES 263
15.1 The Conflation Problem Redux: Exposures to x Confused With Knowledge About x 263
15.1.1 Limitations of knowledge 263
15.1.2 Bad news 264
15.1.3 The central point about what to understand 264
15.1.4 Fragility and Antifragility 264
15.2 Transformations of Probability Distributions 264
15.3 Application 1: Happiness ($f(x)$) is different from wealth (x) 266
15.4 The effect of convexity on the distribution of f(x) 268
15.5 Estimation Methods When the Payoff is Convex 269
15.5.1 Convexity and Explosive Payoffs 271
15.5.2 Conclusion: The Asymmetry in Decision Making 273
16 AN UNCERTAINTY APPROACH TO FRAGILITY 275
16.1 A Review of The General Notion of "Robustness" 275
16.2 Introduction 275
16.2.1 Intrinsic and Inherited Fragility: 278
16.2.2 Fragility As Separate Risk From Psychological Preferences 278
16.2.3 Avoidance of the Psychological 278
16.2.4 Beyond Jensen's Inequality 279
17 THE FRAGILITY THEOREMS 283
17.1 Tail Sensitivity to Uncertainty 283
17.1.1 Precise Expression of Fragility 286
17.2 Effect of Nonlinearity on Intrinsic Fragility 287
17.3 Fragility Drift 290
17.3.1 Second-order Fragility 291
17.4 Expressions of Robustness and Antifragility 291
17.4.1 Definition of Robustness 291
17.4.2 Antifragility 293
17.4.3 Remarks 294
17.4.4 Unconditionality of the shortfall measure ξ 295
18 APPLICATIONS TO MODEL ERROR 297
18.0.5 Example:Application to Budget Deficits 297
18.0.6 Model Error and Semi-Bias as Nonlinearity from Missed Stochasticity of Variables 298
18.1 Model Bias, Second Order Effects, and Fragility 299
19 THE FRAGILITY MEASUREMENT HEURISTICS 301
19.0.1 The Fragility/Model Error Detection Heuristic (detecting ω_A and ω_B when cogent) 301
19.1 Example 1 (Detecting Risk Not Shown By Stress Test) 301
19.2 The Heuristic applied to a stress testing 302
19.2.1 Further Applications Investigated in Next Chapters 302
19.3 Stress Tests 303
19.4 General Methodology 304
20 FRAGILITY AND ECONOMIC MODELS 305

20.1 The Markowitz Inconsistency 305
20.2 Application: Ricardian Model and Left Tail Exposure 305
20.2.1 Error and Probabilities 309
21 THE ORIGIN OF THIN-TAILS 311
21.1 Properties of the Inherited Probability Distribution 312
21.2 Conclusion and Remarks 315
22 SMALL IS BEAUTIFUL: RISK, SCALE AND CONCENTRATION 317
22.1 Introduction: The Tower of Babel 317
22.2 Unbounded Convexity Effects 320
22.3 A Richer Model: The Generalized Sigmoid 323
23 WHY IS THE FRAGILE NONLINEAR? 329
24 HOW THE WORLD WILL PROGRESSIVELY LOOK WEIRDER 331
24.1 How Noise Explodes Faster than Data 331
24.2 Derivations 333
25 THE CONVEXITY OF WEALTH TO INEQUALITY 337
25.1 The One Percent of the One Percent are Divorced from the Rest 337
26 NONLINEARITIES AND RISK IN MEDICINE 341
26.1 Antifragility from Uneven Distribution 341
27 AMERICAN OPTIONS AND HIDDEN CONVEXITY 345
27.1 This Note 345
27.2 The General Mathematical Results: Pricing Series of Hidden Options "Use One Lose All" 345
27.3 Misplaced Precision 347
27.4 The Pricing Equation 348
27.5 War Stories 348
27.6 The Stopping Time Problem 351
27.7 Expressing the various sub-options 351
27.8 Conclusion 351
Bibliography 353
Index 361

CHAPTER SUMMARIES

1 Probability defined –first things first. Why and how we cannot dissociate probability from decision. The notion of contract theory. Fallacies coming from verbalistic descriptions of probability. The difference between classes of payoffs with probabilistic consequences. Formal definition of metaprobability. 3

2 Outline of the book and project of the codification of Risk and decision theory as related to the real world (that is "no BS") in nonmathematical language (other chapters are mathematical). Introduces the main fallacies treated in the project. What can and should be mathematized. Presents the central principles of risk bearing. Introduces the idea of fragility as a response to volatility, the associated notion of convex heuristic, the problem of invisibility of the probability distribution and the spirit of the book. Explains why risk is in the tails not in the variations. Explains that the layering of random variables makes more ecological a view that is corresponds tot the "real world" and how layering of model errors generates fat tails. 33

3 Introducing mathematical formulations of fat tails. Shows how the problem of induction gets worse. Empirical risk estimator. Introduces different heuristics to "fatten" tails. Where do the tails start? Sampling error and convex payoffs. 65

4 Using the asymptotic Radon-Nikodym derivatives of probability measures, we construct a formal methodology to avoid the "masquerade problem" namely that standard "empirical" tests are not empirical at all and can be fooled by fat tails, though not by thin tails, as a fat tailed distribution (which requires a lot more data) can masquerade as a low-risk one, but not the reverse. Remarkably this point is the statistical version of the logical asymmetry between *evidence of absence* and *absence of evidence*. We put some refinement around the notion of "failure to reject", as it may misapply in some situations. We show how such tests as Kolmogorov Smirnoff, Anderson-Darling, Jarque-Bera, Mardia Kurtosis, and others can be gamed and how our ranking rectifies the problem. 113

5 The Spectrum Between Uncertainty and Risk. There has been a bit of discussions about the distinction between "uncertainty" and "risk". We believe in gradation of uncertainty at the level of the probability distribution itself (a "meta" or higher order of uncertainty.) One end of the spectrum, "Knightian risk", is not available for us mortals in the real world. We show how the effect on fat tails and on the calibration of tail exponents and reveal inconsistencies in models such as Markowitz or those used for intertemporal discounting (as many violations of "rationality" aren't violations 125

6 The Law of Large Numbers is the foundation of statistical knowledge –or, even (inductive) knowledge *tout court*. The behavior of the sum of random variables allows us to get to the asymptote and use handy asymptotic properties. However real life is more complicated. We cannot talk about LLN without figuring out the speed of convergence, which, when it is at $\sqrt{n}$, is only so asymptotically. Further, in some cases the LLN doesn't work at all. For very fat tailed, under the slightest parametric error, it will be more than 400 times slower than thought. 137

7 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). We examine how fat tailedness worsens the process. 161

8 We apply the results of the previous chapter on the slowness of the LLN and list misapplication of statistics in social science, almost all of them linked to misinterpretation of the effects of fat-tailedness (and often from lack of awareness of fat tails), and how by attribute substitution researchers can substitute one measure for another. Why for example, because of chronic small-sample effects, the 80/20 is milder in-sample (less fat-tailed) than in reality and why regression rarely works. 181

9 **Error about Errors.** Probabilistic representations require the inclusion of model (or representation) error (a probabilistic statement has to have an error rate), and, in the event of such treatment, one also needs to include second, third and higher order errors (about the methods used to compute the errors) and by a regress argument, to take the idea to its logical limit, one should be continuously reapplying the thinking all the way to its limit unless when one has a reason to stop, as a declared a priori that escapes quantitative and statistical method. We show how power laws emerge from nested errors on errors of the standard deviation for a Gaussian distribution. We also show under which regime regressed errors lead to non-power law fat-tailed distributions. 205

10 We present 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 one believed the r.v. was power law distributed, one still would not be able to make a precise statement on tail risks. Shows weaknesses of calibration of Extreme Value Theory. 219

11 Much of the work concerning martingales and Brownian motion has been idealized; we look for holes and pockets of mismatch to reality, with consequences. Infinite (or undefined) higher moments are not compatible with Ito calculus −outside the asymptote. Path dependence as a measure of fragility. 231

12 A less technical demarcation between Black Swan Domains and others. . . 237

13 We present the difference between ruin problems, particularly when systemic, and other risk management of a more computable nature. 239

14 Standard economic theory makes an allowance for the agency problem, but not the compounding of moral hazard in the presence of informational opacity, particularly in what concerns high-impact events in fat tailed domains (under slow convergence for the law of large numbers). Nor did it look at exposure as a filter that removes nefarious risk takers from the system so they stop harming others. (In the language of probability, skin in the game creates an absorbing state for the agent, not just the principal). But the ancients did; so did many aspects of moral philosophy. We propose a global and morally mandatory heuristic that anyone involved in an action which can possibly generate harm for others, even probabilistically, should be required to be exposed to some damage, regardless of context. While perhaps not sufficient, the heuristic is certainly necessary hence mandatory. It is supposed to counter voluntary and involuntary risk hiding − and risk transfer − in the tails. 249

15 Deeper into the conflation between a random variable and exposure to it. . 263

16 We provide a mathematical approach to fragility as negative sensitivity to a semi-measure of dispersion and volatility (a variant of negative or positive "vega") and examine the link to nonlinear effects. We link to the litterature on model "robustness" and show how we add nonlinearity to the conventional approaches. 275

17 Presents the fragility theorems and the transfer function between nonlinear response and the benefits and harm from increased uncertainty. 283

18 Presents the IMF fragility heuristics, particularly in the improvement of stress testing. 301

19 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 and right-boundedness of the reponse in natural things leads to thin-tails, even when the “underlying” variable of the exposure is fat-tailed. 311

20 We extract the effect of size on the degradation of the expectation of a random variable, from nonlinear response. The method is general and allows to show the "small is beautiful" or "decentralized is effective" or "a diverse ecology is safer" effect from a response to a stochastic stressor and prove stochastic diseconomies of scale and concentration (with as example the Irish potato famine and GMOs). We apply the methodology to environmental harm using standard sigmoid dose-response to show the need to split sources of pollution across independent 317

21 Explains why the fragilefragile is necessarily in the nonlinear. 329

22 Information is convex to noise. 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 and the increase of variables available for epidemiological and other "empirical" research. . . . 331

23 The one percent of the one percent has tail properties such that the tail wealth (expectation $\int_K^\infty x\, p(x)\, \mathrm{d}x$) depends far more on inequality than wealth. 337

24 Examines nonlinearities in medicine /iatrogenics as a risk management problem. 341
25 American Options have hidden optionalities. Using a European option as a baseline we heuristically add the difference. We also show how these hidden options are extremely significant compared to the focus of the research literature on insignificant marginal improvements of the pricing equations but in the Black Scholes World. 345

PREAMBLE/ NOTES ON THE TEXT

This author travelled two careers in the opposite of the usual directions:

> 1) **From risk taking to probability**: I came to deepening my studies of probability and did doctoral work during and *after* trading derivatives and volatility packages and maturing a certain bottom-up organic view of probability and probability distributions. The episode lasted for 21 years, interrupted in its middle for doctoral work. Indeed, volatility and derivatives (under the condition of skin in the game) are a great stepping stone into probability: much like driving a car at a speed of 600 mph (or even 6,000 mph) is a great way to understand its vulnerabilities.
>
> But this book goes beyond derivatives as it addresses probability problems in general, and only those that are generalizable,

and

> 2) **From practical essays (under the cover of "philosophical") to specialized work**: I only started publishing technical approaches (outside specialized option related matters) *after* publishing nontechnical "philosophical and practical" ones, though on the very same subject.

But the philosophical (or practical) essays and the technical derivations were written synchronously, not in sequence, largely in an idiosyncratic way, what the mathematician Marco Avellaneda called "private mathematical language", of which this is the translation – in fact the technical derivations for *The Black Swan*[110] and *Antifragile*[111] were started long before the essay form. So it took twenty years to mature the ideas and techniques of fragility and nonlinear response, the notion of probability as less rigorous than "exposure" for decision making, and the idea that "truth space" requires different types of logic than "consequence space", one built on asymmetries.

Risk takers view the world very differently from most academic users of probability and industry risk analysts, largely because the notion of "skin in the game" imposes a certain type of rigor and skepticism about we call in the next chapter cosmetic "job-market" science.

Risk is a serious business and it is high time that those who learned about it via risk-taking have something not "anecdotal" to say about the subject. In fact we will try to build a new maximally rigorous approach to it, one that incorporates practice.

ACKNOWLEDGMENTS

The text is not entirely that of the author. Four chapters contain recycled text written with collaborators in standalone articles: the late Benoit Mandelbrot (section of slowness of LLN under power laws, even with finite variance), Elie Canetti and the

stress-testing staff at the International Monetary Fund (for the heuristic to detect tail events), Phil Tetlock (binary vs variable for forecasting), Constantine Sandis (skin in the game) and Raphael Douady (mathematical mapping of fragility). But it is the latter paper that represents the biggest debt: as the central point of this book is convex response (or, more generally, nonlinear effects which subsume tail events), the latter paper is the result of 18 years of mulling that single idea, as an extention of *Dynamic Hedging* [108] applied outside the options domain, with 18 years of collaborative conversation with Raphael before the actual composition!

This book is in debt to three persons who left us. In addition to Benoit Mandelbrot, this author feels deep gratitude to the late David Freedman, for his encouragements to develop a rigorous model-error based, real-world approach to statistics, grounded in classical skeptical empiricism, and one that could circumvent the problem of induction: and the method was clear, of the type "don't use statistics where you can be a sucker" or "figure out where you can be the sucker". There was this "moment" in the air, when a group composed of the then unknown John Ioannidis, Stan Young, Philip Stark, and others got together –I was then an almost unpublished and argumentative "volatility" trader, something people couldn't quite understand unless expressed as "nonlinear payoffs", even then (*Dynamic Hedging* was unreadable to nonspecialists) and felt that getting David Freedman's attention was more of a burden than a blessing, as it meant a severe obligation.[2]

Indeed this exact book project was born from a 2012 Berkeley statistics department commencement lecture, given in his memory, with the message: "statistics is the most powerful weapon today, it comes with responsibility" (since numerical assessments increase risk taking) and the corrolary, directly taken from his legacy:

"Understand the model's errors before you understand the model".

leading to the theme of this book, that all one needs to do is figure out the answer to the following question:

Are you convex or concave to model errors?

Further, the Risk manager is the complement of the statistician:

The (conventional) statistician looks at the properties inside the confidence intervals, the risk analyst outside of them. (Figures 0.1 and 0.2

which is the reason statisticians have trouble with risk statements.

It was a very sad story to get a message from the statistical geophysicist Albert Tarantola linking to the electronic version of his book *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation* [112]. He had been maturing an idea on dealing with probability with his new work taking probability *ab ovo*. Tarantola had been piqued by the "masquerade" problem in *The Black Swan* presented in Chapter 4 and the notion that most risk methods "predict the irrelevant".

2 The late David Freedman was known to this author for his paper with Stark *What is the chance of an earthquake?*[44] but he really covered all manner of statistical mistakes, particular with the modeler's biases in his magisterial textbook [45].

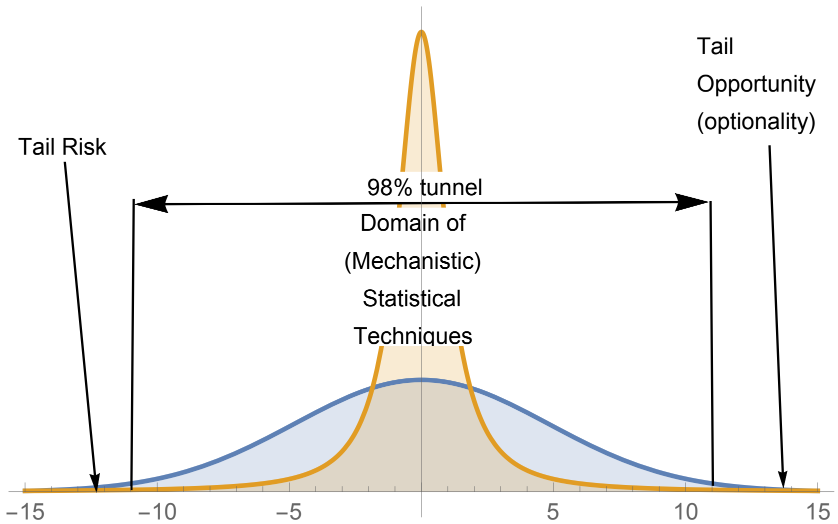


Figure 0.1: Conventional focus of statistical inference, which is why mechanistic uses cannot apply to tail risks. Scientific papers operate in that space and cannot be directly used to discuss risk. In addition confidence discussions are binary statements and do not address payoff (see the codifications in Chapter 1).(Note that the two distributions are calibrated to deliver exactly the same probability of staying in the tunnel $[-K, K]$, here 98%, with $K : \mathbb{P}(X < -K) = .01, \mathbb{P}(X > K) = .99$).

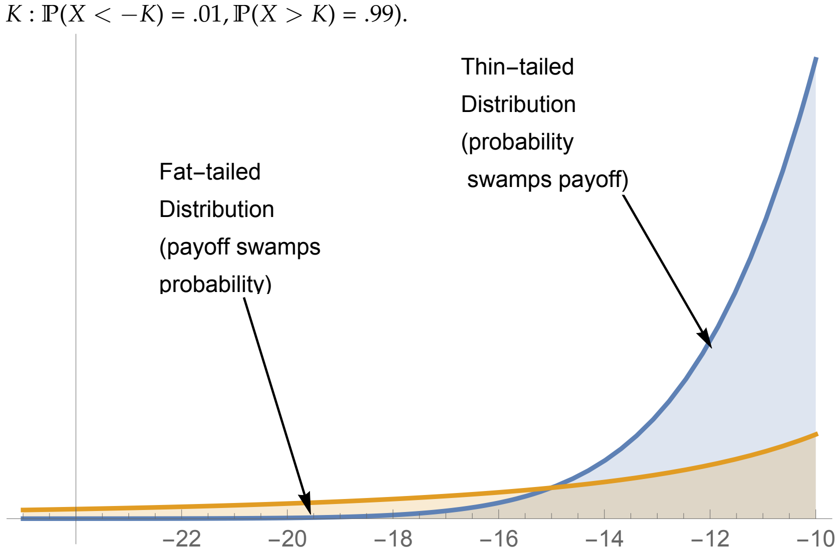


Figure 0.2: Zoom-in of the graph above, showing main difference between tail risks seen under thin (Mediocristan) and fat tails (Extremistan) and why the usual statistics texbook discussions on probabilities need to be refined for risk management .

Tragically, he passed away before the conference he was organizing took place, and while I ended up never meeting him, I felt mentored by his approach –along with the obligation to deliver technical results of the problem in its applications to risk management.

Sections of this text were presented in many places –as I said it took years to mature the point. Some of these chapters are adapted from lectures on hedging with Paul Wilmott and from my course "Risk Management in the Real World" at NYU which as I state in the next chapter is an absurd (but necessary) title. Outside of risk practitioners, in the first stage, I got invitations from statistical and mathematics departments initially to satisfy their curiosity about the exoticism of "outsider" and strange "volatility" trader or "quant" wild animal. But they soon got disappointed that the animal was not much of a wild animal but an orthodox statistician, actually overzealous about a nobullshit approach. I thank Wolfgang Härtle for, before this book was started in the following form, a full-day seminar at Humboldt University and Pantula Sastry for providing the inaugurating lecture of the International Year of Statistics at the National Science Foundation.

Carl Tony Fakhry has taken the thankless task of diligently rederiving every equation (at the time of writing he has just reached Chapter 3). I also thank Wenzhao Wu and Mian Wang for list of typos.

TO THE READER

The text can be read by (motivated) non-quants: everything mathematical in the text is accompanied with a "literary" commentary, so in many sections the math can be safely skipped. Its mission, to repeat, is to show a risk-taker perspective on risk management, integrated into the mathematical language, not to lecture on statistical concepts.

On the other hand, when it comes to math, it assumes a basic "quant level" advanced or heuristic knowledge of mathematical statistics, and is written as a monograph; it is closer to a longer research paper or old fashioned treatise. As I made sure there is little overlap with other books on the subject, I calibrated this text to the wonderful textbook by the late A. Papoulis *Probability, Random Variables, and Stochastic Processes* [84]: there is nothing basic discussed in this text that is not defined in Papoulis.

For more advanced, more mathematical, or deeper matters such as convergence theorems, the text provides definitions, but the reader is recommended to use Loeve's two volumes *Probability Theory* [70] and [71] for a measure theoretic approach, or Feller's two volumes, [40] and [39] and, for probability bounds, Petrov[86]. For extreme value theory, Embrecht et al [31] is irreplaceable.

STATUS/COMPLETION OF BOOK

This is a first draft for general discussion, not for equation-wise verification. There are still typos, errors and problems progressively discovered by readers thanks to the dissemination on the web. The bibliographical references are not uniform, they are in the process of being integrated into bibtex.

Note that there are redundancies that will be removed at the end of the composition.

August 2014 status: After completing most of the math sections, I started putting words and structure around the concepts, so I am progressively introducing "definitions", "remarks", and comments in plain English, inspired by both Bourbaki and legal codifications. The idea is to codify and explain all terms to clear up the usual confusions. As of August 2014, I am only 15% done.

Below is the list of the incomplete sections.

Incomplete Sections in Part I (mostly concerned with limitations of measurements of tail probabilities)

i A list of symbols.

ii Chapter 3 proposes a measure of fattailedness based on ratio of Norms for all(superexponential, subexponential, and powerlaws with tail exponent >2); it is more powerful than Kurtosis since we show it to be unstable in many domains. It lead us to a robust heuristic derivation of fat tails. We will add an Appendix comparing it to the Hill estimator.

iii An Appendix on the misfunctioning of maximum likelihood estimators (extension of the problem of Chapter 3).

iv In the chapter on pathologies of stochastic processes, a longer explanation of why a stochastic integral "in the real world" requires 3 periods not 2 with examples (event information for computation of exposureX_t $\rightarrow$ order $X_{t+\Delta}t$ $\rightarrow$ execution $X_{t+2\Delta t}$).

v The "Weron" effect of recovered α from estimates higher than true values.

vi A lengthier (and clearer) exposition of the variety of bounds: Markov–Chebychev–Lusin–Berhshtein–Lyapunov –Berry-Esseen – Chernoff bounds with tables.

vii A discussion of the Von Mises condition. A discussion of the Cramér condition. Connected: Why the research on large deviations remains outside fat-tailed domains.

viii A discussion of convergence (and nonconvergence) of random matrices to the Wigner semicirle, along with its importance with respect to Big Data

ix A section of pitfalls when deriving slopes for power laws, with situations where we tend to overestimate the exponent.

INCOMPLETE SECTIONS

(mostly concerned with building exposures and convexity of payoffs: What is and What is Not "Long Volatility")

i A discussion of gambler's ruin. The interest is the connection to tail events and fragility. "Ruin" is a better name because the idea of survival for an aggregate, such as probability of ecocide for the planet.

ii An exposition of the precautionary principle as a result of the fragility criterion.

iii A discussion of the "real option" literature showing connecting fragility to the negative of "real option".

iv A link between concavity and iatrogenic risks (modeled as short volatility).

v A concluding chapter.

Best Regards,
Nassim Nicholas Taleb
November 2014

Part I

FIGURING OUT PROBABILITY AND WHAT IT MEANS

1 WHAT IS PROBABILITY? WHAT IS EXPOSURE?

Chapter Summary **1: Probability defined –first things first. Why and how we cannot dissociate probability from decision. The notion of contract theory. Fallacies coming from verbalistic descriptions of probability. The difference between classes of payoffs with probabilistic consequences. Formal definition of metaprobability.**

The larger mission is as follows:

The project – both real-world and anti-anecdotal – is inspired of the many historical efforts and projects aimed to instil rigor in domains that grew organically in a confused way, by starting from the basics and expanding, Bourbaki-style, in a self-contained manner but aiming at maximal possible rigor. This would be a Bourbaki approach but completely aiming at putting the practical before the theoretical, a real-world rigor (as opposed to Bourbaki's scorn of the practical and useful, justified in a theoretical field like mathematics).

The "first thing" is not quite defining probability but rather formally mapping the pair probability and "event" under consideration, subsumed by the notion of probability –the two are inseparable. There has been a long tradition of attempts to define probability, with tons of discussions on what probability is, should be, can be, and cannot be. But, alas these discussions are at best minute Byzantine nuances, the importance of which has been magnified by the citation ring mechanism described in Chapter2; these discussions are "academic" in the worst sense of the word, dwarfed by the larger problem of:

- What is the random "event" under concern? Is it an "event" or something more complicated, like a distribution of outcomes with divergent desirability?
- How *should we use* "probability": probability is not an end product but an input in a larger integral transform, a payoff kernel.

We have done statistics for a century without a clear definition of probability (whether it is subjective, objective, or shmobjective plays little role in the equations of probability: all these probabilities end up adding up to 1 and following the same rules of calculus). But what matters significantly is the event of concern, which is not captured by the verbalistic approaches to probability.[1] Trying to define "what is fire" with academic precision is not something a firefighter should do, as interesting as it seems, given his higher priorities of figuring out the primary (nonacademic) variable, *what is (and what is not) burning*. Almost all these definitions of fire will end up burning the building in the same manner. People whithout skin in the game

1 In my basement I have shelves and shelves of treatises trying to define probability, from De Finetti, Keynes, von Mises, ... See Gillies for the latest approach. Compared to the major problems with metaprobability are mere footnote, as I am showing here by confining such discussion to a footnote.

(nonfirefighters) who spend time worrying about the composition of fire, but not its effect, would remain in the gene pool and divert scientific pursuit into interesting but inconsequential directions.[2]

For we truly quite don't know what we are talking about when we talk about probability. Often when we talk about probability, we are discussing something else –something far more fundamental.

1.1 THE CONFLATION OF EVENTS AND EXPOSURES

The problem can be best explained with this martial principle:

> The art of war teaches us to rely not on the likelihood of the enemy's not coming, but on our own readiness to receive him; not on the chance of his not attacking, but rather on the fact that we have made our position unassailable.

in Lao Tsu, *The Art of War*

Fallacy 1.1 (Verbalistic Expression of Probability, an Introduction to the problem)**.** *"Probability" is meaningless without an associated payoff function as its verbalistic expression doesn't necessarily match its mathematical one, the latter usually implicitly entailing a (contingent) payoff, except in rare cases where the payoff is "binary" (even then), a confusion that is prevalent in "research" on overestimation of rare events (Chapter* **??***). The probability distribution of the payoff, not that of the primary random variable being of concern, much of research on rare events in psychology and economics is actually invalidated by switching from the verbalistic to the mathematical-contractual definition.*

We skip the elements of measure theory for now in expressing random variables.

Take x a random or nonrandom variable (leave the exact definition of random variable and random event for later), and $f(x)$ the exposure, payoff, the effect of x on you, the end bottom line. 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 there has been a chronic confusion since Aristotle between x and $f(x)$. The mistake is at two level: one, simple confusion; second, a blind spot missing an elephant the decision-science literature, being aware the distinction and yet not realizing that action on $f(x)$ is easier than action on x.[3]

---

2 For an example of Byzantine concerns about probability so detailed and diverted from planet earth that they miss everything of relevance to risk, see the works of David Aldous on the central difference between "finite additivity" or "countable additivity", which can be classified as the hijacking of the most important discipline in the world, probability, by scholastic distinctions without (or with relatively minor) real-world difference.

3 Clearly $f(x)$ can be utility of x, or, even better, the combination; a utility of a function of x, $u(g(x))$, where u is utility and g a function. Utility theory has done some work focusing on the expectation of $\int f(x) \mathrm{d}P(x)$ where P is the probability. But there seems to have been a lack of focus on the *distribution* of the composite which, as we show in Chapter 15, would make standard concave and unbounded utility completely absurd for anyone to take risks under the slightest left-fat tailedness. It is as if utility theorists have been drowning too much in the axiomatic morass to consider what we can do about it in real life. Hence in this book our relation to utility will remain rather ambiguous except for specific discussions. As we will see with the idea of a contract, one can alter a payoff of x, not utility of x.
For option pricing the convenience of Black-Scholes approach has not been to show a pricing formula

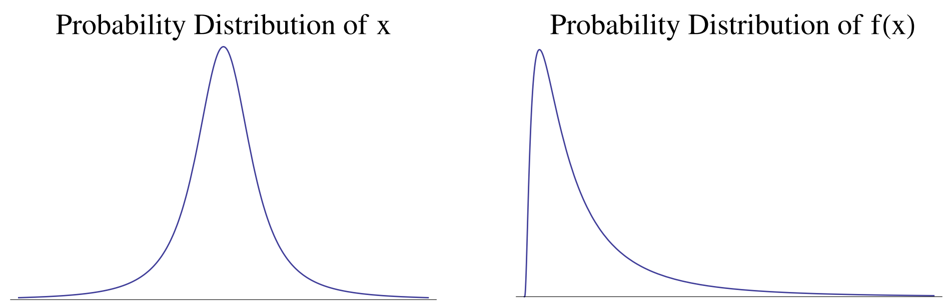


Figure 1.1: The conflation of x and $f(x)$: mistaking the statistical properties of the exposure to a variable for the variable itself. It is easier to modify exposure to get tractable properties than try to understand x. This is more general confusion of truth space and consequence space.

Mixed Convexities and Natural Systems

A more advanced point. In general, in nature, because $f(x)$ the response of entities and organisms to random events is generally thin-tailed while x can be fat-tailed, owing to $f(x)$ having the sigmoid "S" shape convex-concave (some type of floor below, progressive saturation above). This explains why the planet has not blown-up from tail events. And this also explains the difference (Chapter 21) between economic variables and natural ones, as economic variables can have the opposite effect of accelerated response at higher values of x (right-convex $f(x)$) hence a thickening of at least one of the tails.

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's well-being (which we assume is minimal).

The variable 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.

The variable 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).

I grew up under the rule that it is more reliable to modify $f(x)$ to the point where one can be satisfied with the reliability of the risk properties than try to understand the statistical properties of x, particularly under fat tails.[4]

Principle 1.1.
Risk management is less about understanding random events as much as what they can do to us.

–this has existed for a long time – but rather to exit the discussion on utility. But, as I have shown, we didn't even need Black-Scholes for that.

4 The reason decision making and risk management are inseparable is that there are some exposure people should never take if the risk assessment is not reliable, which, as we will see with the best map fallacy, is something people understand in real life but not when modeling.

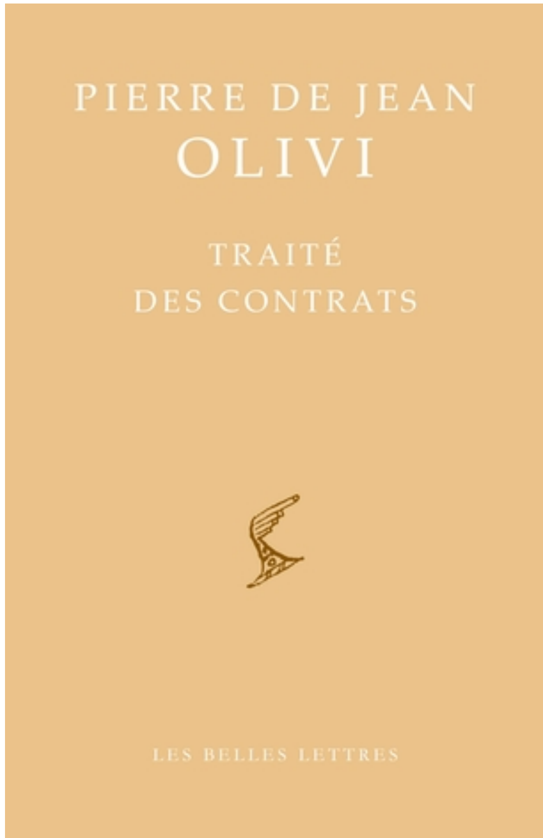


Figure 1.2: When you use the services of a lawyer for a contract, you are working on limiting or constructing $f(x)$ your exposure, where your risk and liability start and end. This 13th C. treatise by the legal and theological scholastic philosopher Pierre de Jean Olivi provide vastly more rigorous codification and deeper treatment of risk and probability than the subsequent mathematical ones grounded in the *narrower* ludic dimension (i.e., confined to games) by Fermat, Pascal, Huyguens, even De Finetti. Why? Because one can control exposure via contracts and structures rather than just narrowly defined knowledge of probability. Further, a ludic setup doesn't allow or perturbation of contractual agreements, as the terms are typically fixed.

The associated second central principle:

Principle 1.2 (Central Principle of (Probabilistic) Decision Making)**.**
It is more rigorous to take risks one understands than try to understand risks one is taking.

And the associated fallacy:

Definition 1.1 (The Best Map Fallacy)**.**
Unconditionally preferring a false map to no map at all. More technically, ignoring the fact that decision-making entails alterations in $f(x)$ in the absence of knowledge about x.

About every reasonable 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 a professor, professional manager or "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".

The best map is a violation of the central principle of risk management, Principle 1.2.

The fallacy is explained in *The Black Swan* [110]:

> I know few people who would board a plane heading for La Guardia airport in New York City with a pilot who was using a map of Atlanta's airport "because there is nothing else." People with a functioning brain would rather drive, take the train, or stay home. Yet once they get involved in economics, they prefer professionally to use a wrong measure, on the ground that "we have nothing else." The idea, well accepted by grandmothers, that one should pick a destination for which one has a good map, not travel and then find "the best" map, is foreign to PhDs in social science.

This is not a joke: the "give us something better" has been a recurring problem this author has had to deal with for a long time.

1.1.1 Contract Theory

The rigor of the 13th Century legal philosopher Pierre de Jean Olivi is as close to our ambition as that of Kolmogorov and Paul Lévy. It is a fact that stochastic concepts such as probability, contingency, risk, hazard, and harm found an extreme sophistication in philosophy and legal texts, from Cicero onwards, way before probability entered our vocabulary –and of course probability was made poorer by the mental gymnastics approach and the ludic version by Fermat-Pascal-Huygens-De Moivre ...

Remark 1.1 (Science v/s Contract Theory)**.**
Science is typically in binary space (that is, True/False) as defined below, not about exposure, while risk and decisions are necessarily in standard real-world full payoff space. Contract theory is in exposure space. Risk management is closer to the latter.

Remark 1.2 (Derivatives Theory)**.**
Option theory is mathematical contract theory.[5]

Remark 1.3.
A function of a random variable, s.a. exposure, needs to be treated as a separate random variable.

The point seems trivial but is not. Statisticians make the benign conflation of a random event ω for a random variable, which in most cases is just an abuse of notation. Much more severe –and common –is the conflation of a random variable for another one (the payoff).

Just consider how we define payoff of options, a combination of legal considerations and mathematical properties.

5 I thank Eric Briys for insights along these lines.

Definition 1.2 (Binary).
Binary statements, predictions and exposures are about well defined discrete events ω in probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with true/false, yes/no types of answers expressed as events in a specific probability space. The outcome random variable $X(\omega)$ is either 0 (the event does not take place or the statement is false) or 1 (the event took place or the statement is true), 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.

Example of binary: most scientific statements tested by "p-values", or most conversational nonquantitative "events" as whether a person will win the election, a single individual will die, or a team will win a contest.

Definition 1.3 (Standard, Real-World, Full Payoff, or "Vanilla" Space).
Statements, predictions and exposures, also known as natural random variables, correspond to situations in which the payoff is either continuous or can take several values. An event ω in probability space $(\Omega, \mathcal{F}, \mathbb{P})$ maps to random variable in $\mathbb{R}^1$, with $a_L < a_H \in \mathbb{R}$,

$$X(\omega) \in \textit{either } (a_L, a_H), [a_L, a_H), (a_L, a_H], \textit{or } [a_L, a_H],$$

where these intervals are Borel sets.

We start with a trivial error –trivial but frequently made.

Example 1.1 (Market Up or Down?).
In Fooled by Randomness (2001/2005) [106], the author was asked during a meeting which was more probable, that a given market would go higher or lower by the end of the month. "Higher", he said, insisting that it was "much more probable". But then it was revealed that he was making trades that benefit if that particular market went lower. The story wasn't retold for any paradox (too trivial) by as wonderment as to why people are surprised at all by such a story.

This of course, is most trivial for statisticians and professional quantitative traders but it appears to be frequently made since many performances are assessed on the frequency of profits not the expectation. (Yes, the market is more likely to go up, but should it go down it will fall much much more) and obvious when written down in probabilistic form, so for S_t the market price at period t there is nothing incompatible between probability and expectation having (sort of) opposite signs:

$$\operatorname{sgn}\left(\mathbb{P}(S_{t+1} > S_t) - \frac{1}{2}\right) = -\operatorname{sgn}(\mathbb{E}(S_{t+1}) - S_t)$$

(where $\mathbb{E}$ is the expectation). The left side of the equation expresses the "more likely" mathematically, and shows how trivial the divergence can be. This divergence in sign is possible once one takes into account a full distribution for S_{t+1}, which comes from having the mean much much lower than the median (under negative skewness of the distribution).

Beyond the trivial, this example illustrates the common confusion between a *bet* and an exposure. A bet is a binary outcome, an exposure has more nuanced results and depends on full distribution.

When we go deeper into the subject, many less obvious, or less known paradox-style problems occur. Simply, it is of the opinion of the author, that it is not rigorous to talk about "probability" as a final product, or even as a "foundation" of decisions.

Table 1: Four Classes

Class $\mathfrak{P}_i$	Name	Function notation	Fourier Transform of $\phi(\Psi^+)$: $\widehat{\phi}_1(t)$	$\mathbb{E}(\Psi)^+$
$\mathfrak{P}_1$	Atomic	Ψ_1	1	$p(x)$
$\mathfrak{P}_2$	Binary	Ψ_2^+, Ψ_2^-	$(1-\pi_K) + e^{it}\pi_K$	π_K
$\mathfrak{P}_3$	Vanilla	Ψ_3^+, Ψ_3^-	$(1-\pi_K)$ $+\int_K^\infty e^{it}\, \mathrm{dP}x$	$-K\,\pi_K+$ $\int_K^\infty x\, \mathrm{dP}x$
$\mathfrak{P}_{4a}$	Comp.	Ψ_4	$\prod \widehat{\phi}_i(t)$	$\Sigma\Omega_i\mathbb{E}(\Psi_i)$
$\mathfrak{P}_{4b}$	Gen. Sigm.			$\int \mathbb{E}(\Psi_i)\, \mathrm{d}\Omega$

The vanillas 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.

The conflation binary-vanilla is a mis-specification often made in probability, seen in as fundamental texts as in J.M. Keynes' approach to probability [64]. Such a conflation is almost always present in discussions of "prediction markets" and similar aberrations; it affects some results in research. It is even made in places by De Finetti in the assessment of what makes a good "probability appraiser"[22].[6]

> The central point here is that decision-making is not about being a good *probability appraiser* –life is not about probability as a standalone concept but something more complex in which probability only enters as a kernel, or integral transform.

The designation "vanilla" originates from definitions of financial contracts.[7]

Example 1.2 (Too much snow).
The owner of a ski resort in the Lebanon, deploring lack of snow, deposited at a shrine of the Virgin Mary a \$100 wishing for snow. Snow came, with such abundance, and avalanches, with people stuck in the cars, so the resort was forced to close, prompting the owner to quip "I should have only given \$25". What the owner did is discover the notion of nonlinear exposure under tail events.

Example 1.3 (Predicting the "Crisis" yet Blowing Up).
The financial firm Morgan Stanley correctly predicted the onset of a subprime crisis, but they misdefined the event they called "crisis"; they had a binary hedge (for small drop) and ended up losing billions as the crisis ended up much deeper than predicted.

As we will see, under fat tails, there is no such thing as a "typical event", and nonlinearity widens the difference between verbalistic and precisely contractual definitions.

6 The misuse comes from using the scoring rule of the following type:if a person gives a probability p for an event A, he is scored $(p-1)^2$ or p^2, according to whether A is subsequently found to be true or false. Consequences of A or the fact that there can be various versions of such event are, at best, an afterthought.

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

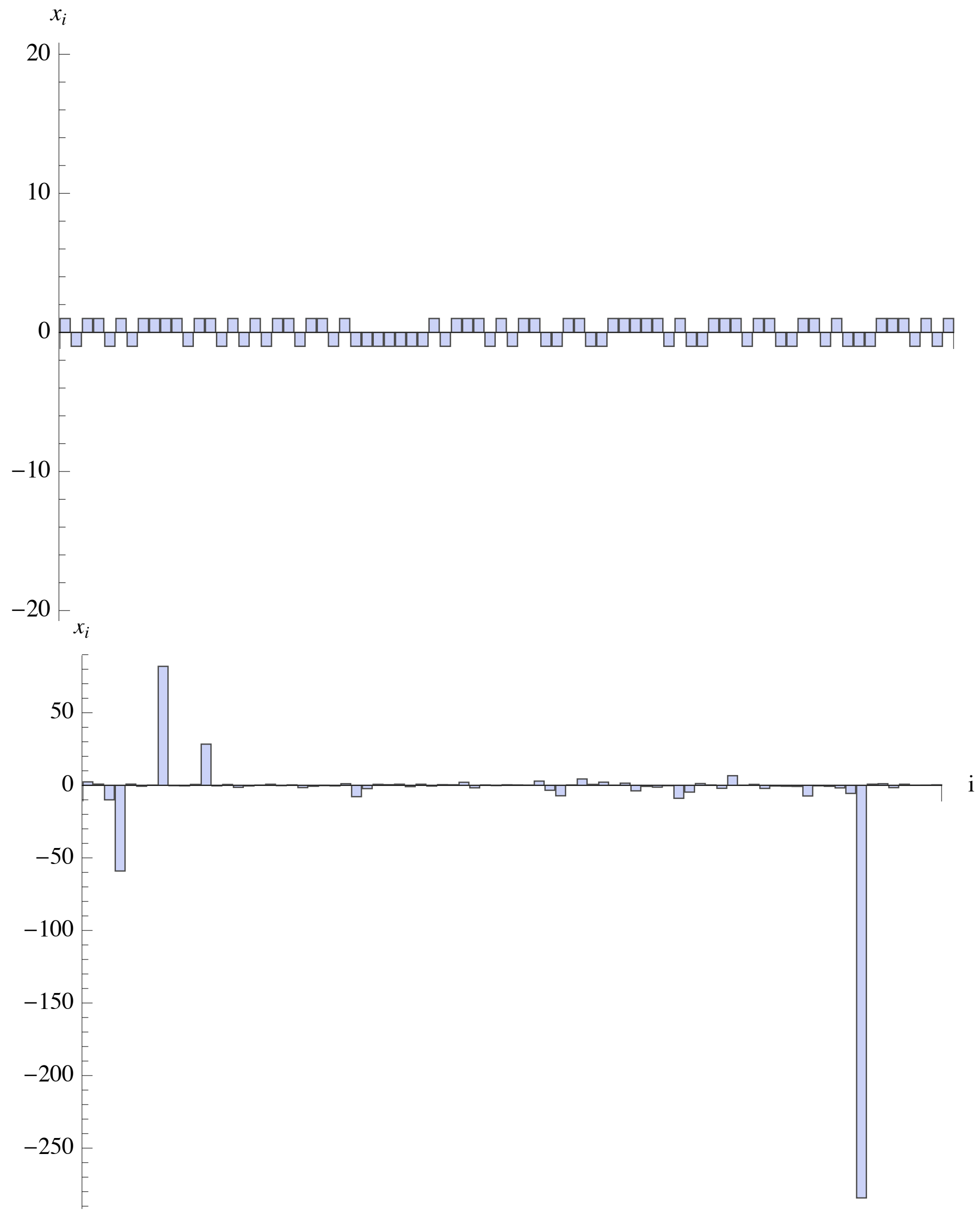


Figure 1.3: Comparing payoff in classes $\mathfrak{P}_2$ to those in $\mathfrak{P}_3$ (top), or binaries to the vanilla. The vertical payoff shows x_i, $(x_1, x_2, ...)$ and the horizontal shows the index *i= (1,2,...)*, as *i* can be time, or any other form of classification. We assume in the first case payoffs of {-1,1}, and open-ended (or with a very remote and unknown bounds) in the second.

1.2 PAYOFF CLASSES $\mathfrak{P}_1$ THROUGH $\mathfrak{P}_4$

Let $x \equiv x_T$ be a (non necessarily) Markovian continuous state variables observed at period T, $T \in \mathbb{R}^+$; x has support $\mathcal{D} = (\mathcal{D}^-, \mathcal{D}^+)$. The state variable is one-tailed or two-tailed, that is bounded on no more than one side, so either $\mathcal{D}^+ = \infty$ or $\mathcal{D}^- = -\infty$, or both.

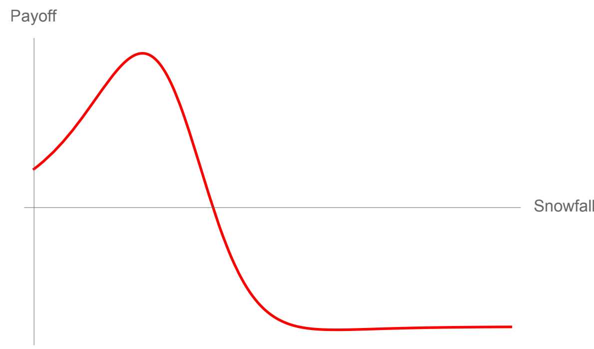


Figure 1.4: The graph shows the payofd to the ski resort as a function of snowfall. So the discrete variable "snow" (vs "no snow") is not a random event for our purpose. Note that such a payoff is built via a convex/concave combinations of vanillas.

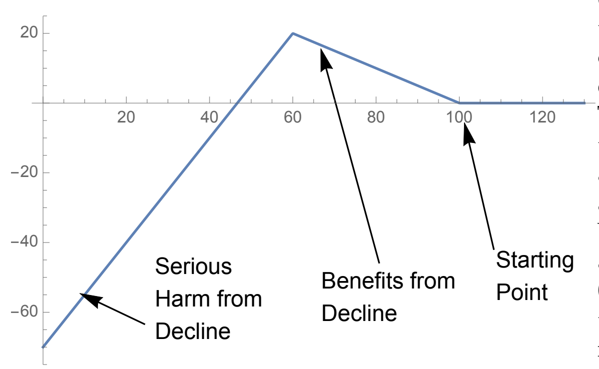


Figure 1.5: A confusing story: mistaking a decline for an "event". This shows the Morgan Stanley error of defining a crisis as a binary event; they aimed at profiting from a decline and ended up structuring their exposure in a way to blow up from it. This exposure is called in derivatives traders jargon a "Christmas Tree", achieved in with $\mathfrak{P}_4$ through an addition of the following contracts $\Psi_3^-(K)_{1\leq i\leq 3}$ and quantitities q_1, q_2 and q_3 such that $q_1 > 0$, $q_2, q_3 < 0$, and $q_1 < -q_2 < -q_3$, giving the toxic and highly nonlinear terminal payoff $\Psi_4 = q_1\Psi_3^-(K) + q_2\Psi_3^-(K - \Delta K) + q_3\Psi_3^-(K - k\Delta K)$, where $k > 1$. For convenience the figure shows K_2 triggered but not K_3 which kicks-in further in the tails.

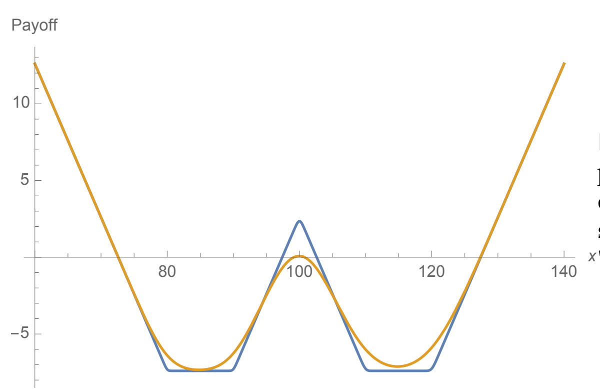


Figure 1.6: Even more confusing: exposure to events –in class $\mathfrak{P}_4$ –that escape straightforward verbalistic descriptions. Option traders call this a "butterfly exposure" in the jargon.

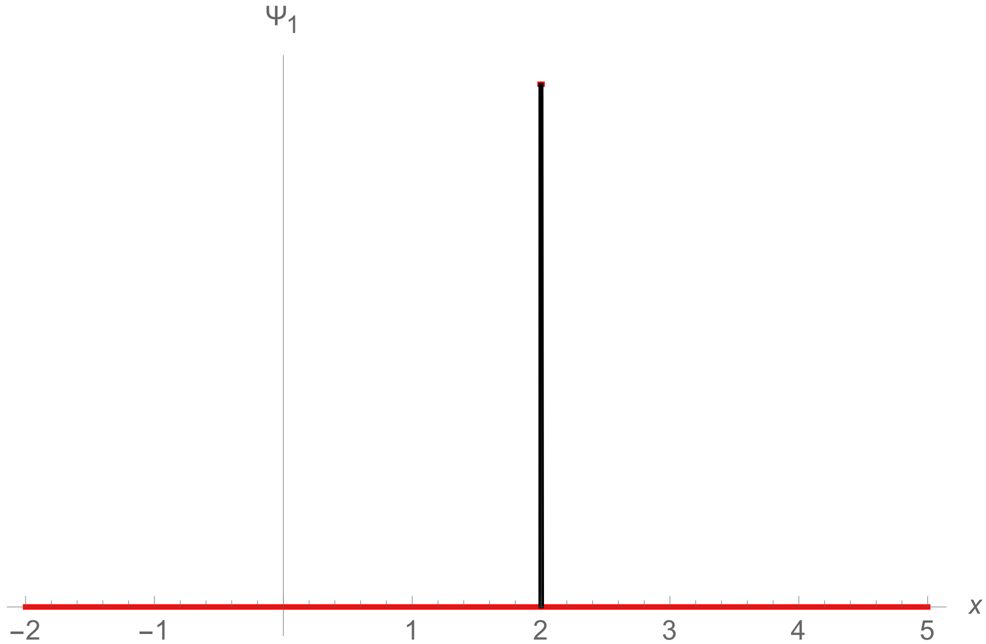


Figure 1.7: Payoff Class $\mathfrak{P}_1$

The "primitive" state variable x_t is continuously observed between discrete periods $T - \Delta t$ and T. The payoff or exposure function is $\Psi \mathbb{1}_{t>\tau}$ where $\tau = \{\inf(t) : x_t \notin A \, , t \leq T\}$, a stopping-time conditional discretization of the continuously sampled time.[8]

The "payoff kernel" Ψ at time T is a member of the exhaustive and mutually exclusive following 4 classes. We write its probability distribution $\phi(\Psi)$ and characteristic function $\widehat{\phi}(t)$ (the distributions of the payoff under the law of state variable x between $T - \Delta t$ and T, Ψ itself taken as a random variable) at T, and $p(x)$ the probability law for x at T.

Note that the various layers are obtained by integration over the state variable x over segments of the domain $\mathcal{D}$:

$$\Psi_i = \int \Psi_{i-1}(x) \, \mathrm{d}x$$

1.2.1 Atomic Payoff $\mathfrak{P}_1$

Definition 1.4 (Class $\mathfrak{P}_1$, or Arrow-Debreu State Variable)**.**
$\Psi \equiv \Psi_1(x, K)$, *which can be expressed as the Dirac Delta function:*

$$\Psi_1(x, K) = \delta(x - K)$$

where $\int_{K \in \mathcal{D}} \delta(x - K) \, \mathrm{d}x = 1$ *and* $\int_{K \notin \mathcal{D}} \delta(x - K) \, \mathrm{d}x = 0$ *otherwise.*

---

8 Without getting into details the stopping time does not have to be off the same primitive state variable x_t –even in dimention 1 –but can condition on any other state variable.

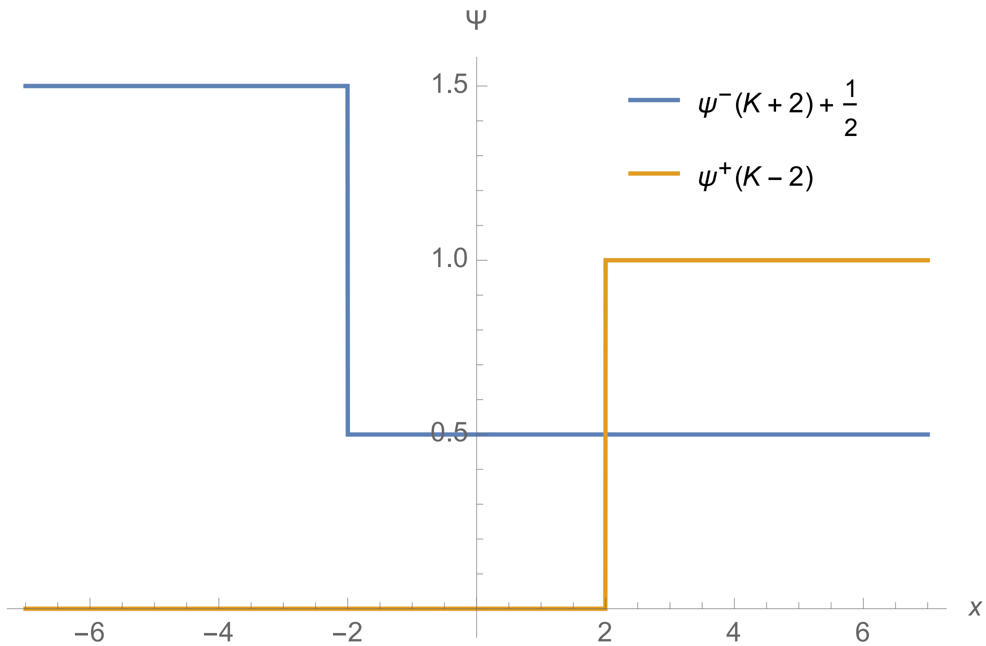


Figure 1.8: Payoff Class $\mathfrak{P}_2$

Remark 1.4 (Characteristic function invariance)**.**
The Characteristic function $\widehat{\phi}_1(t,K) = 1$ for all continuous probability distributions $p(x)$ of the primitive state variable x.

Proof. $\int_{\mathfrak{D}} e^{i\,t\,\delta(x-K)}p(x)\mathrm{d}(x) = \int_{\mathfrak{D}} p(x)\mathrm{d}(x) = 1$ when K is in the domain of integration. □

Remark 1.5.
The expectation of Ψ_1 maps to a probability density at K for all continuous probability distributions.

Proof. Consider that

$$\begin{aligned} i\frac{\partial}{\partial t}\widehat{\phi}_1(t,K) &= -i\frac{\partial}{\partial t}\int_{\mathfrak{D}} e^{(i\,t\,\delta(x-K))}p(x)\mathrm{d}x \\ &= \int_{\mathfrak{D}} e^{(i\,t\,\delta(x-K))}\delta(x-K)p(x)\mathrm{d}x \end{aligned} \tag{1.1}$$

Hence

$$\mathbb{E}(\Psi) = i\frac{\partial}{\partial t}\widehat{\phi}_1(t,K)|_{t=0} = p(K)$$

□

1.2.2 Binary Payoff Class $\mathfrak{P}_2$

Definition 1.5 ($\Psi \in \mathfrak{P}_2$, or Binary Payoffs)**.**
$\Psi \equiv \Psi_2(K)$ obtained by integration, so

$$\Psi_2^+(K) = \int_{\mathcal{D}^-}^{K} \Psi_1(x)\mathrm{d}x$$

which gives us, writing (for clarity) x for the state variable in the integrand and X for the observed one:

$$\Psi_2^+(X, K) = \begin{cases} 1 & \text{if } X \geq K; \\ 0 & \text{if } X < K. \end{cases}$$

and

$$\Psi_2^-(K) = \int_{K}^{\mathcal{D}^+} \Psi_1(x)\mathrm{d}x$$

giving us:

$$\Psi_2^-(X, K) = \begin{cases} 0 & \text{if } X > K; \\ 1 & \text{if } X \leq K. \end{cases}$$

which maps to the Heaviside θ function with known properties.

Remark 1.6.
The class $\mathfrak{P}_2$ is closed under affine transformation $a_H\Psi + a_L$, for all combinations $\{a_H, a_L : a_H x + a_L \in \mathfrak{D}\}$. This is true for affine transformations of all payoff functions in $\Psi_{\geq 2}$, the unit of payoff becoming $a_H + a_L$ and the lower (upper) bound a_L (a_H).

Proposition 1.1 (Binaries are Thin-Tailed)**.**
The probability distribution $\phi(\Psi_2)$, a "binary" payoff is a Bernouilli regardless of the underlying probability distribution over the state variable x.

Proof. First consider that Ψ_2^+ can be written as $\Psi_2^+(x) = \frac{1}{2}(1 + \operatorname{sgn}(x - K))$. Its characteristic function $\widehat{\phi}_2^+(t, K)$:

$$\begin{aligned} \widehat{\phi}_2^+(t, K) &= \int_{\mathcal{D}} \mathbf{e}^{\frac{1}{2} i\, t\, (1+\operatorname{sgn}(x-K))}\, p(x)\, \mathrm{d}x \\ &= \int_{<K} p(x)\, \mathrm{d}x + \int_{\geq K} \mathbf{e}^{i\, t} p(x)\, \mathrm{d}x \end{aligned} \tag{1.2}$$

So, with $\pi_K \equiv \mathbb{P}(X \geq K)$,

$$\widehat{\phi}_2^+(t, K) = (1 - \pi_K) + \mathbf{e}^{i\, t}\pi_K$$

which is the characteristic function of the Bernouilli distribution.

□

Note that we proved that Ψ_2 is subgaussian as defined in [61] regardless of $p(x)$ the probability distribution of the state variable, even if $p(x)$ has no moments.

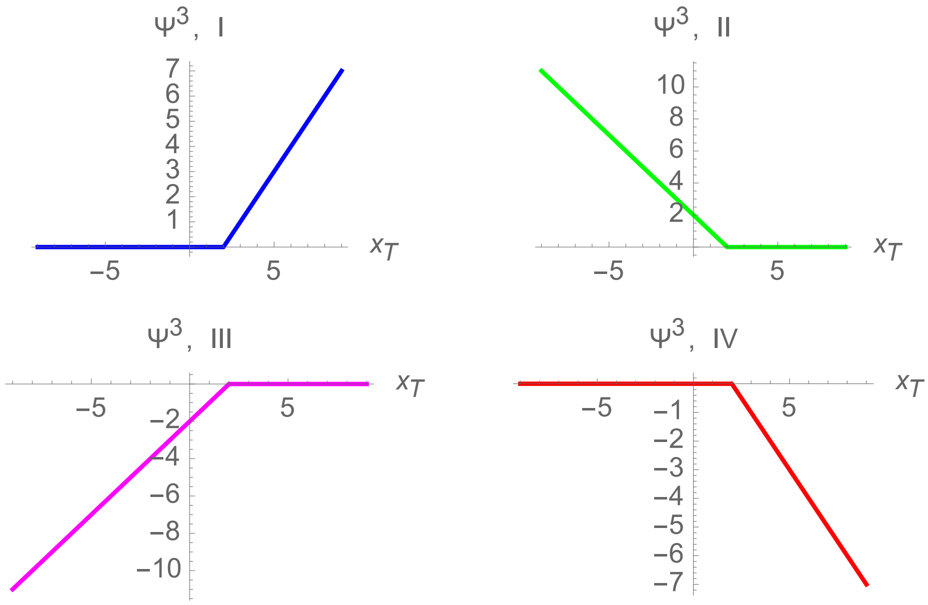


Figure 1.9: Payoff Class $\mathfrak{P}_3$

1.2.3 Vanilla Payoff Class $\mathfrak{P}_3$, building blocks for regular exposures.

Definition 1.6 ($\Psi \in \mathfrak{P}_3$, or Vanilla Payoff)**.**
$\Psi \equiv \Psi_3(X, K)$ *obtained by integration, so*

$$\Psi_3^{+}(X,K) = \int_{\mathcal{D}^-}^{X} \Psi_2(x-K)\mathrm{d}x$$

which gives us:

$$\Psi_3^{+}(X,K) = \begin{cases} X-K & \textit{if } X \geq K; \\ 0 & \textit{if } X < K. \end{cases}$$

and

$$\Psi_3^{-}(X,K) = \int_{X}^{\mathcal{D}^+} \Psi_2(x)\mathrm{d}x$$

giving us:

$$\Psi_3^{-}(X,K) = \begin{cases} K-X & \textit{if } X \leq K; \\ 0 & \textit{if } X > K. \end{cases}$$

Assume the support spans the real line. The characteristic function $\phi(t,K)$ can be expressed as:

$$\phi(t,K) = \int_{-\infty}^{\infty} p(X) e^{\frac{1}{2} it(X-K)(\mathrm{sgn}(X-k)+1)}\,\mathrm{d}X$$

which becomes

$$\phi(t,K) = (1-\pi_K) + e^{-itK} \int_{K}^{\infty} e^{itx} p(x)\mathrm{d}x \tag{1.3}$$

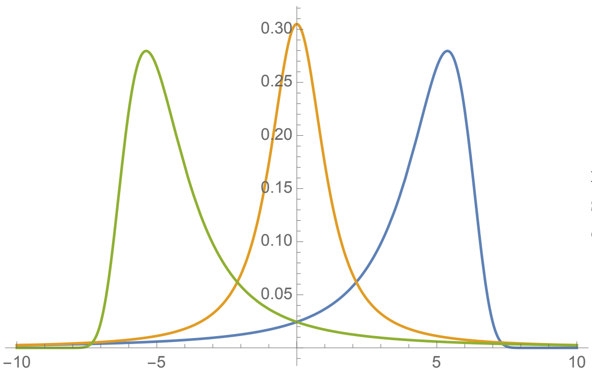


Figure 1.10: Stable Distributions: remarkably the three have exactly the same mean and mean deviation, but different β symmetry parameter.

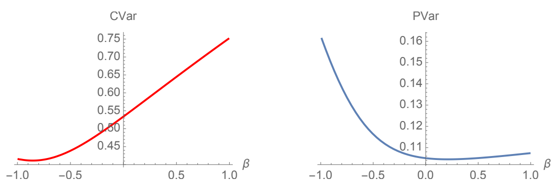


Figure 1.11: Stable Distribution. As we decrease skewness, with all other properties invariant, the CVar rises and the PVar (probability associated with VaR) declines.

Proposition 1.2 (Impossibility)**.**
It is possible to build a composite/sigmoidal payoff using the limit of sums of vanillas with strikes K, and $K + \Delta K$, but not possible to obtain vanillas using binaries.

Proof. The Fourier transform of the binary does not integrate into that of the vanilla as one need K struck at infinity. The sum requires open-ended payoffs on at least one side. □

For many distributions of the state variable the characteristic function allows explicit inversion (we can of course get numerical effects). Of some interest is the expectation that becomes:

$$\mathcal{E}(\Psi_3^+) = \int_K^\infty x\, p(x)\, dx - K\, \pi_K \tag{1.4}$$

which maps to common derivatives pricing such as the Bachelier approach[5] or it Lognormal generalizations popularized with [11].

As we can see Eq. 1.4 doesn't depend on the portion of the tail of the distribution below K. Of interest is the "stub" part of the pricing, which represents the difference between the vanilla and the binary of same strike K:

$$\Delta^+(K) \equiv E(\Psi_3^+ - K\Psi_2^+) = \int_K^\infty x\, p(x)\, \mathrm{d}x \tag{1.5}$$

The Δ contract has the convenience of sensitivity to fat tails (or other measures of uncertainty such as the scale of the distribution), as it extracts the "tail", segment of the distribution above (below) K.

The idea is to compare $\int_K^\infty x\, p(x)\, \mathrm{d}x$ and $\int_K^\infty p(x)\, \mathrm{d}x$ and see how they react in opposite directions to certain parameters that control the fatness of tails.

Remark 1.7 (Symmetry/Skewness Problem).
There exists a nondegenerate distribution $p^*(x)$ with $\mathbb{E}^{p^*}(X) = \mathbb{E}^{p}(X)$ and $\mathbb{E}^{p^*}(|X|^s) = \mathbb{E}^{p}(|X|^s)$ for $s \leq 2$ such that:

$$\begin{aligned} sgn\left(\int_K^\infty x\, p^*(x)\, \mathrm{d}x - \int_K^\infty x\, p(x)\, \mathrm{d}x\right) \\ = -sgn\left(\int_K^\infty p^*(x)\, \mathrm{d}x - \int_K^\infty p(x)\, \mathrm{d}x\right) \end{aligned} \tag{1.6}$$

Proof. The sketch of a proof is as follows. Just consider two "mirror" asymmetric distributions, p_1 and p_2, with equal left and right side expectations.

With $\mathbb{P}^+_{p_1} \equiv \int_0^\infty p_1(x)\, \mathrm{d}x$ and $\mathbb{P}^-_{p_2} \equiv \int_{-\infty}^0 p_2(x)\, \mathrm{d}x$, we assumed $\mathbb{P}^+_{p_1} = \mathbb{P}^-_{p_2}$. This is sufficient to have all moments the exact the same (should these exist) and all other attributes in L^1 as well: the distributions are identical except for the "mirror" of positive and negative values for attributes that are allowed to have a negative sign.

We write $\mathbb{E}^+_{p_1} \equiv \int_0^\infty x\, p_1(x)\, \mathrm{d}x$ and $\mathbb{E}^+_{p_1} \equiv -\int_{-\infty}^0 x\, p_2(x)\, \mathrm{d}x$. Since $\mathbb{E}^+_{p_1} = -\mathbb{E}^-_{p_2}$ we can observe that all changes in the expectation of the positive (negative) side of p_2 around the origin need to be offset by a change in the cumulative probability over the same domain in opposite sign. □

The argument is easily explored with discrete distributions or mixing Gaussians, but we can make it more general with the use of continuous non-mixed ones: the α-Stable offers the remarkable property of allowing changes in the symmetry parameter while retaining others (mean, scale, mean deviation) invariant, unlike other distribution such as the Skew-Normal distribution that have a skew parameter that affects the mean.[9]In addition to the skewness, the stable can also thus show us precisely how we can fatten the tails while preserving other properties.

Example 1.4 (Mirror Stable distributions).
Consider two mirror α-stable distributions as shown in Figure 1.11, $S_{\alpha,\beta,\mu,\sigma}$ with tail exponent $\alpha = \frac{3}{2}$ and $\beta = \pm 1$, centering at $\mu = 0$ to simplify;

$$p_1(x) = -\sqrt[3]{2}\; \frac{e^{\frac{(\mu-x)^3}{27\sigma^3}}\left(\frac{\sqrt[3]{3}(\mu-x)Ai\left(\frac{(\mu-x)^2}{3\ 2^{2/3}\sqrt[3]{3}\sigma^2}\right)}{\sigma} + 3\sqrt[3]{2}Ai'\left(\frac{(\mu-x)^2}{3\ 2^{2/3}\sqrt[3]{3}\sigma^2}\right)\right)}{3\ 3^{2/3}\sigma}$$

$$p_2(x) = -\sqrt[3]{2}\; \frac{e^{\frac{(\mu-x)^3}{27\sigma^3}}\left(\frac{\sqrt[3]{3}(\mu-x)Ai\left(\frac{(\mu-x)^2}{3\ 2^{2/3}\sqrt[3]{3}\sigma^2}\right)}{\sigma} + 3\sqrt[3]{2}Ai'\left(\frac{(x-\mu)^2}{3\ 2^{2/3}\sqrt[3]{3}\sigma^2}\right)\right)}{3\ 3^{2/3}\sigma}$$

9 For instance, the Skew-Normal $N(\mu, \sigma, \beta; x)$, where $\beta \in \mathbb{R}$ controls the skewness, with PDF $\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}\operatorname{erfc}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)}{\sqrt{2\pi}\sigma}$, has mean $\frac{\sqrt{\frac{2}{\pi}}\beta\sigma}{\sqrt{\beta^2+1}} + \mu$ and standard deviation $\sqrt{1 - \frac{2\beta^2}{\pi(\beta^2+1)}}\sigma$, meaning the manipulation of β leads to change in expectation and scale. The same applies to the mirrored Lognormal (where skewness and expectation depends on variance) and the Pareto Distribution (where the tail exponent controls the variance and the mean deviation if these exist.

$$\mathbb{E}^{+}_{p_1} = \frac{\sqrt[3]{2}\sigma}{\Gamma\left(\frac{2}{3}\right)},\ \mathbb{E}^{-}_{p_1} = -\frac{\sqrt[3]{2}\sigma}{\Gamma\left(\frac{2}{3}\right)}$$

$$\mathbb{E}^{+}_{p_2} = \frac{\sqrt[3]{2}\sigma}{\Gamma\left(\frac{2}{3}\right)},\ \mathbb{E}^{-}_{p_2} = -\frac{\sqrt[3]{2}\sigma}{\Gamma\left(\frac{2}{3}\right)}$$

$$\mathbb{P}^{+}_{p_1} = \frac{1}{3},\ \mathbb{P}^{+}_{p_1} = \frac{2}{3}$$

$$\mathbb{P}^{+}_{p_2} = \frac{2}{3},\ \mathbb{P}^{+}_{p_1} = \frac{1}{3}$$

Moving the $beta$ parameter which controls symmetry (and, only symmetry) to change the distribution have the effect of moving probabilities without altering expectations.

Stochastic Volatility Divergence Let s be the scale of the distribution with density $p_s(x)$. Consider the ratio of densities;

$$\exists \lambda : \forall K > \lambda, 0 < \delta < 1, \frac{1}{2}\frac{\left(p_{s-\delta s}(K) + p_{s+\delta s}(K)\right)}{p_s(K)} > 1$$

which is satisfied for continuous distributions with semi-concave densities.

> We will ferret out situations in which $\int_K^\infty x\ p(x)\,\mathrm{d}x$ (the "Cvar" or conditional value at risk) and $\int_K^\infty p(x)\,\mathrm{d}x$ (the Probability associated with "VaR" or value-at-risk) react to tail fattening situations in opposite manner.

1.2.4 Composite/Sigmoidal Payoff Class $\mathfrak{P}_4$

Definition 1.7 (P_4, or Composite Payoff)**.**
Pieced together sums of n payoffs weighted by Ω_j:

$$\Psi_4 = \sum_{j=1}^{n} \Omega_j^{+}\Phi^{+}_{i>1}(K_j) + \Omega_j^{-}\Phi^{-}_{i>1}(K_j)$$

This is the standard arithmetically decomposable composite payoff class, if we assume no conditions for stopping time –the ones encountered in regular exposures without utility taken into account, as a regular exposure can be expressed as the difference of two, more precisely $\Psi_2^{+}(K) - \Psi_2^{-}(K)\,, \forall K \in \mathcal{D}$.

Remark 1.8.
The class $\mathfrak{P}_4$ is closed under addition.

1.3 ACHIEVING NONLINEARITY THROUGH $\mathfrak{P}_4$

1.4 MAIN ERRORS IN THE LITERATURE

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 vanillas can experience wilder fluctuations with a high probability, particularly in fat-tailed domains. Comparing 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 vanillas exposures. If ecological exposures in the real world tends to have vanillas, not binary properties, then much of these results are invalid.

Let us return to the point that the variations of vanillas are not bounded. The consequence is that 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.

1.5 THE APPLICABILITY OF SOME PSYCHOLOGICAL BIASES

1.6 MISFITNESS OF PREDICTION MARKETS

1.6.1 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.

Two points.

Binary predictions are more tractable than standard ones 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

Table 2: True and False Biases in the Psychology Literature

Alleged Bias Derived in $\mathfrak{P}_2$	**Misspecified domain**	**Justified domain**
Dread Risk	Comparing Terrorism to fall from ladders	Comparing risks of driving vs flying
Overestimation of small probabilities	Open-ended payoffs in fat-tailed domains	Bounded bets in laboratory setting
Long shot bias	Convex financial payoffs	Lotteries

Table 3: Adequate and inadequade decision domains

Application	**Questionable domain**	**Justified domain**
Prediction markets	Revolutions	Elections
Prediction markets	"Crashes" in Natural Markets (Finance)	Sports
Forecasting	Judging by frequency in venture capital and other winner take all domains;	Judging by frequency in finite bets

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 standard 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.

Remark:*More technically, for a heavy tailed distribution (defined as part of the subexponential family), with at least one unbounded side to the random variable (one-tailedness), the variable 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.*

1.6.2 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 $0 < \delta \leq 1$

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

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

1.6.3 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 $\pm$ 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.

$$
\begin{aligned}
P_{>1\sigma} &= P_{<1\sigma} \\
&= 1 - \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}\sqrt{1-a}}\right) - \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2}\sqrt{a+1}}\right)
\end{aligned}
\tag{1.7}
$$

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 $\pm$2.13 σ.

1.6.4 The law of large numbers works better with the binary than the variable

Getting a bit more technical, the law of large numbers works much faster for the binary than the variable (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 variable.

Let x_t be the realization of the random variable $X \in$ (-∞, ∞) at period t, which follows a Cauchy distribution with p.d.f. $f\left(x_t\right) \equiv \frac{1}{\pi\left(\left(x_0-1\right)^2+1\right)}$. Let us set $x_0 = 0$ to simplify and make the exposure symmetric around 0. The variable 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 variable 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 variable 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

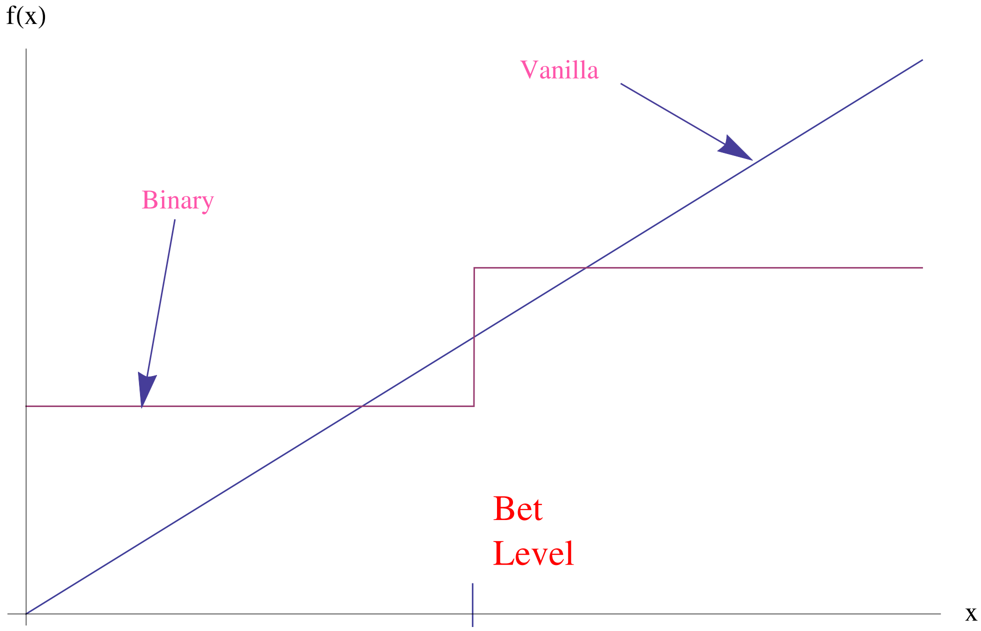


Figure 1.12: The different classes of payoff f(x) seen in relation to an event x. (When considering options, the vanilla can start at a given bet level, so the payoff would be continuous on one side, not the other).

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 $\sqrt{V\left(S^\theta{}_n\right)} = \sqrt{\frac{V\left(S^\theta{}_1\right)}{n}} = \sqrt{\frac{(1-p)p}{n}}$ (given that the moment generating function for the average is $M(z) = \left(pe^{z/n} - p + 1\right)^n$).

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:

$$\lim_{X\to\infty} \frac{P\left(X > \sum_{i=1}^n x_{t_i}\right)}{P\left(X > \max\left(x_{t_i}\right)_{i\leq 2\leq n}\right)} = 1 \tag{1.8}$$

Compare this to the binary for which

$$\lim_{X\to\infty} P\left(X > \max\left(\theta(x_{t_i})\right)_{i\leq 2\leq n}\right) = 0 \tag{1.9}$$

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 variable.

- The sum of binaries is never dominated by a single event, while that of the variable can be.

How is the binary more robust to model error?

In the more general case, the expected payoff of the variable is expressed as $\int_A x dF(x)$ (the unconditional shortfall) while that of the binary= $\int_A dF(x)$, 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 variable, the perturbation's effect on the probability is multiplied by a larger value of x.

As an example, define a slighly more complicated variable 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 variable, 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 variable is $\frac{V(1.01,24.2)}{V(1.2,24.2)} = 37.4$. In this case the variable was ∼18 times more sensitive than the binary.

—-

1.7 FINDING INCONSISTENCIES IN SCHOLARLY TREATMENTS OF EVENTS

Historians and Verbalistic definition of Events

Some people fancy being in binary space when they are in fact in vanilla payoff/exposure space.

Historians deal with events but risk being trapped in small narratives and webs of causation (another story). When this author made the statement that the nation-state was much more murderous than the collection of city-states and statelings that represented Italy, with the first war killing around 650, 000 people compared to previous event with casualties around two orders of magnitude lower, the reaction of historians was that no, there were *many more* wars in Italy before unification, with sieges, plots, and the kind of athmosphere one finds in Machiavelli. So the point

> "History is not a quantitative hence statistical statement. It is about events and trends".

Effectively the probability of war dropped, but the risk got bigger, yet historians insisted that their business is not probability. Their answer was of the sort "we deal with events defined as wars", hence they pretty much agreed that 2 wars is worse than a single one. But then plied with the question:

> Q1: Would you treat the second world war with the same "importance" as the Falkand Island war of 1982?

If the answer to the question is "of course not", then:

> Q2: Would you treat the second world war with less "importance" than the Falkand Island war of 1982 *plus* the Crimean war ?

Let us play the game. With Ω the event space, define $'' \succ ''$ as a binary relation such that event A is more "important" than B if $A, B \in \Omega$, A $\succ BorA \succsim B$ at least equal to B, then it looks like we can elicit from the historian that, "in general" (being careful about what the "general" means):

$$A \succsim B \text{ if } c(A) \geq c(B)$$

where $c : \Omega \to \mathbb{N}^+$ is *quantitative* measure of casualties, measured in number of death or similar metrics. Our questions Q1 and Q2 can establish monotonicity of the ordering relation.

We can assert that the historian is in fact not in binary space, even if he lives somewhat in the illusion that he is, otherwise it would lead to inconsistencies in the simplified ordering relation.[10]

We can continue the game for the satisfaction of certain axioms which would allow us to assert that in fact our judgment of historical events lines up to their risk, which, unavoidably, is quantitative. We can even adjust for the "severity" of events, where the binary relation is violated "except" when casualties are k greater, such that

$$\exists\, k \geq 1 : c(A) \geq k\, c(B) \Rightarrow {''}A \succsim B{''}$$

and still find out that, *for large events*, history while not being quantitative still depends on a quantitative ranking of severity. Given that we are in the tails business (that's what risk is about), history is in fact convincingly vanilla not binary.

"Exposure" (Hence Risk) Needs to be Far More Rigorous Than "Science"

People claiming a "scientific" approach to risk management needs to be very careful about what "science" means and how applicable it is for probabilistic decision making. Science consists in a body of rigorously verifiable (or, equivalently, falsifiable), replicable, and generalizable claims and statements –and those statements only, nothing that doesn't satisfy these constraints. Science scorns the particular. It never aimed at covering *all* manner of exposure management, and never about opaque matters. It is just a subset of our field of decision making. We need to survive by making decisions that do not satisfy scientific methodologies, and cannot wait a hundred years or so for these to be established–simply, extinction is an absorbing barrier. So phronetic approaches such as [41] or a broader class of matters we can call "wisdom" and precautionary actions are necessary. But not abiding by naive "evidentiary science", we embrace a larger set of human endeavors; it becomes necessary to build former protocols of decision akin to legal codes: rigorous, methodological, precise, adaptable, but certainly not standard "science" *per se*.

10 We could go deeper and express "fuzziness" about the importance of an event or a set of events as second-order effect similar to metaprobability modeling.

We will discuss the scientism later; for now consider a critical point. Textbook knowledge is largely about "True" and "False", which doesn't exactly map to payoff and exposure.

Parts have been solved in the paper

Let O be a family the one-dimensional payoff functions considered as of time t_0 over a certain horizon $t \in \mathbb{R}^+$, for:

A variable $X \in \mathfrak{D} = (\mathfrak{d}^-, \mathfrak{d}^+)$, with initial value x_{t_0} and value x_t at time of the payoff, upper bound $\mathfrak{d}^+ \geq 0$ and lower bound $\mathfrak{d}^- \leq \mathfrak{d}^+$

Let $\mathbb{1}_A$ be an indicator function, $\mathbb{1}_A \in \{1, -1\}$, q the size of the exposure, and P a constant(set at time t_0) (meant to represent the inital outlay, investment, or exposure).

We can define the kernel in many ways, depending on use and complexity of payoff.

The payoff kernel can be expressed as follows. With support $\mathfrak{D}$ and probability measure P which is is metaprobability adjusted:

$$\Psi(x_t, K) \equiv f(x_t, K)\, \mathrm{d}P_{t_0,t}(x_t, K)$$

With the expectation under discussion: $\int_{\mathfrak{D}} \Psi(x_t, K) \mathrm{d}P_{t_0,t}(x_t, K)$

1.8 METAPROBABILITY AND THE PAYOFF KERNEL

One must never accept a probability without probabilizing the source of the statement. In other words, if someone who is your sole source of information tells you "I am 100% certain", but you think that there is a 1% probability that the person is a liar, the probability must no longer be treated as 100% but 99% or so, perhaps even lower.[11] If you look at trust as an "epistemological notion" (Origgi, [83]), then the degree of trust maps directly to the metaprobability.

Risk, Uncertainty, and Layering: Historical Perspective

Principle 1.3 (The Necessity of Layering).
No probability without metaprobability. One cannot make a probabilistic statement without considering the probability of a statement being from an unreliable source, or subjected to measurement errors.

We can generalize to research giving certain conclusions in a dubious setup, like many "behavioral economics" results about, say, hyperbolic discounting (aside from the usual problem of misdefining contracts).

Definition 1.8 (Knightian Uncertainty).
It corresponds to a use of distribution with a degenerate metadistribution, i.e., fixed parameters devoid of stochasticity from estimation and model error.

11 I owe this to a long discussion with Paul Boghossian; it is remarkable how nonphilosophers have a rough time thinking of the meta-issue.

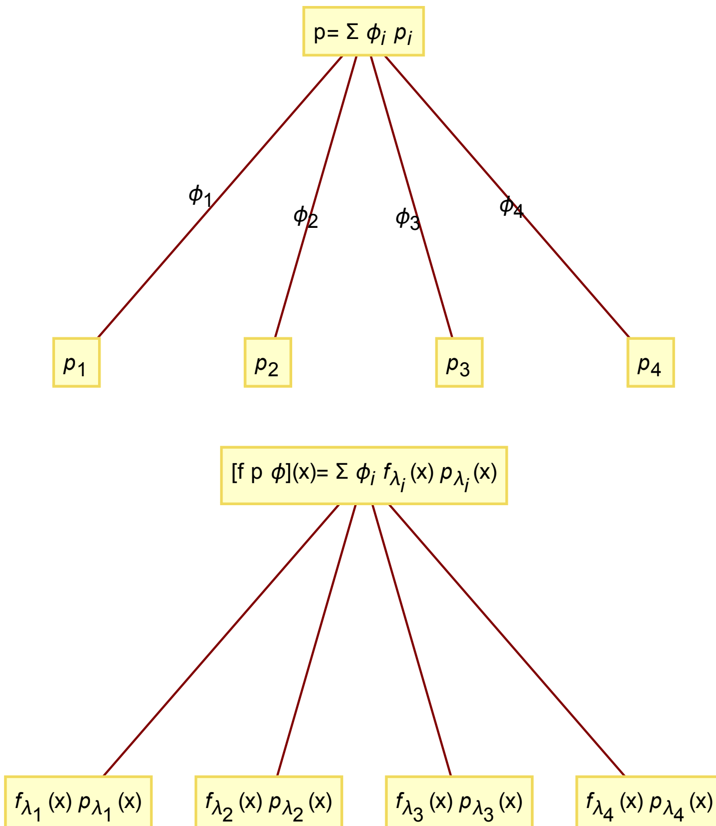


Figure 1.13: The idea of metaprobability Consider that uncertainty about probability can still get us a unique measure $\mathbb{P}$ equals the weighted average of the states ϕ_i, with $\Sigma\phi_i = 1$; however the nonlinearity of the response of the probability to λ requires every possible value of λ to be taken into account. Thus we can understand why under metaprobabilistic analysis small uncertainty about the probability in the extreme left tail can cause matters to blow up.

Remark 1.9 (A heuristic to spot incompetence)**.**
There is no such thing as "Knightian risk" in the real world, but gradations of computable risk. A good heuristic is to disqualify any adult who uses the idea of "Knightian risk" as incompetent.

We said that no probability without a payoff, and no probability without a metaprobability (at least), which produces a triplet 1) exposure, 2) probability, 3) metaprobability.

Definition 1.9 (**Metadistribution/Metaprobability**)**.**
Mounting any probability distribution on the probabilities or their distribution to examine sensitivity and higher order effects. It can be done:

a) Partially: By stochasticization of parameters (s.a. stochastic variance, stochastic tail exponents).

b) Globally: By stochasticization (subordination) of distributions.

Consider an ensemble of probability distributions, all identical except for the probability measures (in other words, same event space, same sigma-algebra, but different probability measures), that is $(\Omega, \mathcal{F}, \mathbb{P}_i)$. A variable $X \in \mathfrak{D} = (\mathfrak{d}^-, \mathfrak{d}^+)$, with upper bound $\mathfrak{d}^+ \geq 0$ and lower bound $\mathfrak{d}^- \leq \mathfrak{d}^+$

We associate a payoff (or decision) function f with a probability $\hat{p}$ of state x and a metaprobability weight ϕ. The atomic payoff Φ is an integral transform. If ϕ is discrete with states $D = \{1, 2, ..., n\}$, the constraint are that $\sum_{i \in D} \phi_i = 1$,and $0 \leq \phi_i \leq 1$. As to the probability p under concern, it can be discrete (mass function) or continuous(density) so let us pick the continuous case for the ease of exposition. The constaint on the probability is:

$$\forall i \in D, \int_{\mathfrak{D}} \hat{p}_{\lambda_i}(x)\, \mathrm{d}x = 1$$

$$\Psi_{p,f,\phi}(x) \equiv [pf\phi](x) \equiv \sum_{i \in D} f(x, \lambda_i) \phi_i \hat{p}_{\lambda_i}(x). \tag{1.10}$$

where λ is a hidden "driver" or parameter determining probability. The parameter λ could be the scale of a Gaussian (variance) or Levy-Stable distribution, but could also be the tail exponent.

In the simplified case of $< x\lambda >= 0$, i.e. when $\forall \lambda_i, f(x, \lambda_i) = f(x, \bar{\lambda})$ where $\bar{\lambda} = \sum_{\lambda_i \in D} \phi_i \lambda_i$, we can conveniently simplify 1.10 to:

$$\Psi_{p,f,\phi}(x) \equiv [pf\phi](x) \equiv f(x) \sum_{\lambda_i \in D} \phi_i \hat{p}_{\lambda_i}(x). \tag{1.11}$$

Equivalently, consider the continuous case $\phi(\lambda) : [0,1] \to [0,1]$:

$$\Psi_{p,f,\phi}(x) \equiv [pf\phi](x) \equiv \int_{\mathfrak{D}} f(x, \lambda) \phi(\lambda) \hat{p}_{\lambda}(x)\, \mathrm{d}\lambda. \tag{1.12}$$

which simplifies to:

$$\Psi_{p,f,\phi}(x) \equiv [pf\phi](x) \equiv f(x) \int_{\mathfrak{D}} \phi(\lambda) \hat{p}_{\lambda}(x)\, \mathrm{d}\lambda. \tag{1.13}$$

In the rest of the book we can consider the simplified case –outside of more serious cases of cross-dependence –and derive a metaprobality adjusted distribution, as a measure or a density on its own. Thus:

$$p(x) \equiv \int_{\mathfrak{D}} \phi(\lambda) \hat{p}_{\lambda}(x) \, \mathrm{d}\lambda$$

is treated as a probability density function with cumulative P.

More Complicated versions of parameter λ . The parameter in question can be multidimentional, which greatly complicates the integration delivering the kernel. However, note that we examine in Chapter 5 cases of the parameter λ driven by a parameter that has its own parametrized distribution, with the parametrization under concern too having a distribution, all the way to infinite regress for the "error on error" when the parameter is the scale of the distribution.

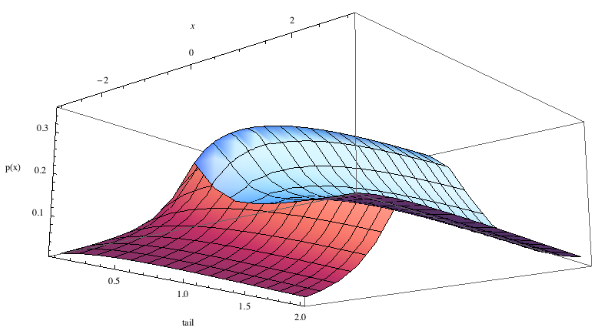


Figure 1.14: Metaprobability: we add another dimension to the probability distributions, as we consider the effect of a layer of uncertainty over the probabilities. It results in large effects in the tails, but, visually, these are identified through changes in the "peak" at the center of the distribution.

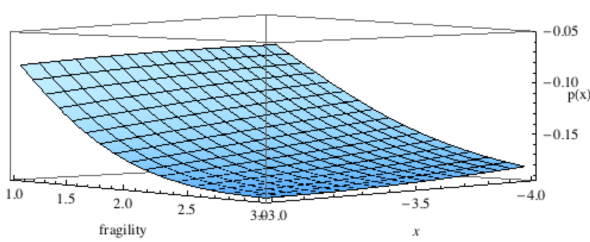


Figure 1.15: Fragility to Error (and Stressors): Can be seen in the slope of the sensitivity of payoff across metadistributions

subsectionHow to extract the distribution of the payoff

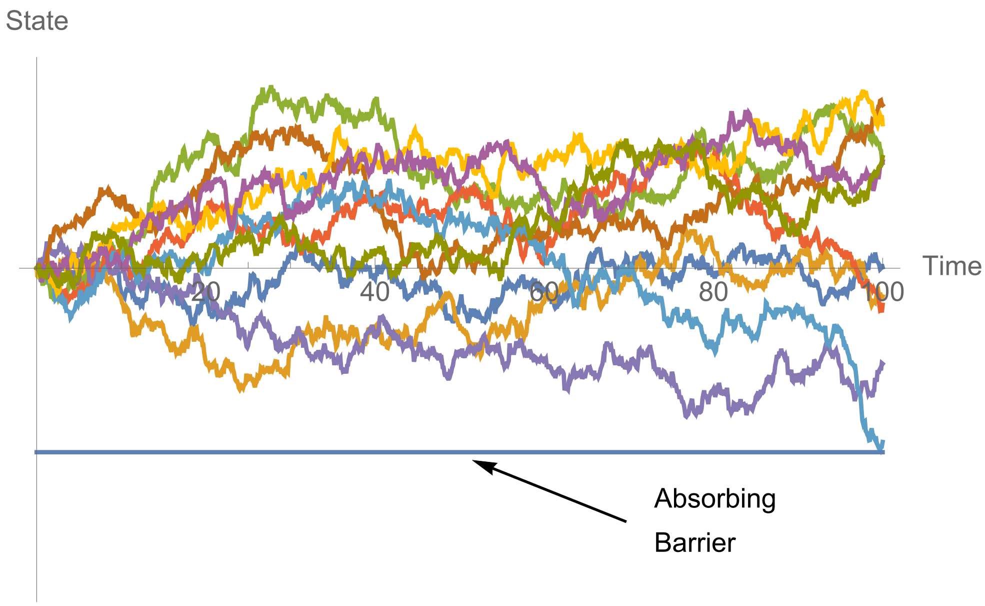


Figure 1.16: A variety of temporal states for a process subjected to an absorbing barrier. Once the absorbing barrier is hit, the process terminates, regardless of its future potential.

Note on Building a Smooth Payoff Function

$$\theta(x) = \lim_{k \to \infty} \frac{1}{2}(1 + \tanh kx) = \lim_{k \to \infty} \frac{1}{1 + e^{-2kx}}.$$

There are many other smooth, analytic approximations to the step function. Among the possibilities are:

$$H(x) = \lim_{k \to \infty} \left(\frac{1}{2} + \frac{1}{\pi} \arctan(kx) \right)$$

As to sign of x

$$\operatorname{sgn}(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin(ux)}{u} \, du,$$

which is useful for derivatives under the integral. Further, for more complex payoffs, in which the decision (payoff kernel) is the random variable, we can use the convolution theorem to prove that the Fourier transform of the product of two functions $f(x)$ and $g(x)$ is given as:

$$\mathcal{F}[f(x)g(x)] = \int_{-\infty}^{+\infty} F(\omega')G(\omega - \omega') \, d\omega',$$

the convolution of $F(\omega')G(\omega')$, where $F(\omega')$ and $G(\omega')$ are the Fourier transforms of $f(x)$ and $g(x)$ respectively.

These tools will be useful for working with payoffs analytically.

1.9 CLASSIFICATION AND CODIFICATION OF EXPOSURES

Definition 1.10 (Path dependent exposure).
A path dependent exposure has the payoff function depend on all the values of the underlying variable x between t_0 and a terminal value t.

Example 1.5.

Definition 1.11 (Contingent exposure).
When in **??**, $K > \mathfrak{d}^-$ *and* $K < \mathfrak{d}^+$.

Definition 1.12 (Time homogeneity of exposure).

Definition 1.13 (Absorbing barriers).
A special (and most common case) of path dependent exposure and critical for risk management.

Definition 1.14 (Decomposable payoff).

Definition 1.15 (Linear and nonlinear payoff).

Definition 1.16 (Quanto payoff).

Definition 1.17 (Asian payoff).

Definition 1.18 (Floating strike payoff).

Definition 1.19 (Multivariate scalar payoff).

1.10 NUMERAIRE DEFINITION

A critical problem with numeraire, in which the payoff is expressed, which is effectively problematic in many situations where the "barbell" (to be defined in section x) is implemented and something truly "risk-free" needs to be found. Well. only an invariant and unchanging metric is really risk-free.

Definition 1.20 (Numeraire related payoffs).

1.11 WHAT IS AN INSURABLE RISK?

1.12 RUIN PROBLEMS

1.13 SKEPTICISM, UNCERTAINTY, AND SCALE OF A DISTRIBUTON

1.14 WHY PASCAL WAGER HAS NOTHING TO DO WITH THE LEFT TAIL

2 THE "REAL WORLD" RIGOR PROJECT

Chapter Summary 2: **Outline of the book and project of the codification of Risk and decision theory as related to the real world (that is "no BS") in nonmathematical language (other chapters are mathematical). Introduces the main fallacies treated in the project. What can and should be mathematized. Presents the central principles of risk bearing. Introduces the idea of fragility as a response to volatility, the associated notion of convex heuristic, the problem of invisibility of the probability distribution and the spirit of the book. Explains why risk is in the tails not in the variations. Explains that the layering of random variables makes more ecological a view that is corresponds tot the "real world" and how layering of model errors generates fat tails.**

This chapter outlines the main ideas of the book; it can be read on its own as a summary of the project.

We start with *via negativa*, the definition of a negative, to show why fallacies matter (particularly that risk analysis and management are negative activities):

Definition 2.1 (Via Negativa)**.**
Consists in focusing on decision making by substraction, via the identification of errors. In theology and philosophy, it is 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.

Clearly, risk management is a *via negativa* endeavor, avoiding a certain class of adverse events.

Table 4: Via Negativa: Major Errors and Fallacies in This Book

Fallacy	Description	Section(s)
	Central Risk Fallacies	
Turkey Problem: Evidentiary fallacy	Requiring evidence of risk particularly in fat-tailed domains, violation of inferential asymmetries (evidence comes *after* risk).	Chapters 3, 6
Best Map Fallacy	Belief that a false map is unconditionally better than no map.	
Triffat Fallacy	Mistaking the inverse problem for the problem, finding the problem to fit the math.	

Table 4: (continued from previous page)

Fallacy	Description	Section(s)
Counter of Triffat Fallacy	Rejection of mathematical statements without showing mathematical flaw; rejection of mathematical rigor on grounds of failures in *some* domains or inverse problems.	
Knightian Risk Fallacy	Belief that probability is ever computable with 0 error rate, without having *any* model or parameter uncertainty.	
Convex Payoff Fallacy	Belief that loss function and required sample size in estimator for x is the same for $f(x)$ when f is convex.	Section 3.11
LLN Fallacy	Belief that LLN works naively with fat tails.	Chapter 6
Binary/Vanilla Conflation		
Crossing the Street Fallacy	Conflating systemic and local risk.	
Fallacy of Silent Evidence	Survivorship bias has large effects on small probabilities.	
CLT Error		
Fallacy of Silent Evidence	Survivorship bias has large effects on small probabilities.	
	Inferential Fallacies	
Froot Insurance fallacy/Pisano biotech fallacy (Harvard professors)	Making inference about mean in left/right skewed fat tailed domains by overestimating/underestimating it respectively owing to insufficience sample	

Table 4: (continued from previous page)

Fallacy	Description	Section(s)
Pinker Fallacy, 1 (another Harvard professor[1])	Mistaking fact-checking for statistical estimation.	
Pinker Fallacy, 2	Underestimating the tail risk and needed sample size for thick-tailed variables from inference from similar thin-tailed ones.	
The "n=1" Fallacy	Ignoring the effect of maximum divergence (Lévy, Kolmogorov) in disconfirmatory empiricism. (Counterfallacy is "n large" for confirmatory empiricism)	
The powerlaw fallacy	Rejecting powerlaw behavior from Log-Log plot or similar.	

2.1 A COURSE WITH AN ABSURD TITLE

This author is currently teaching a course 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, no 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, since risk taking requires reliability, hence our guiding principle in the next section.

Indeed something is completely broken in risk management.

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

1 Harvard University, because of the pressure to maintain a high status for a researcher in the academic community, which conflicts with genuine research, provides a gold mine for those of us searching for example of *fooled by randomness* effects.

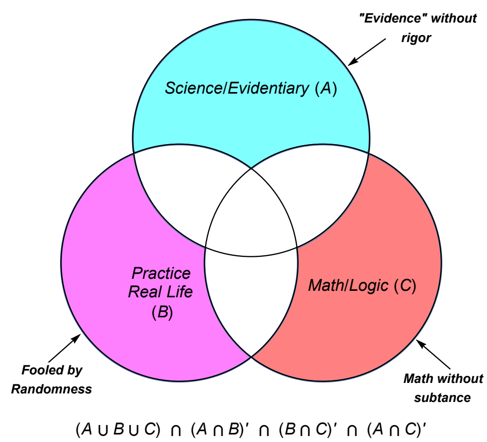


Figure 2.1: **Wrong! The Symmetric Difference of the Three Sets** The unhappy merger of theory and practice. Most academics and practitioners of risk and probability do not understand what **"intersection"** means. This explains why Wall Street "quants" blow up. It is hard trying to explain that yes, it is very mathematical but bringing what we call a math genius or acrobat won't do. It is jointly mathematical and practical.
"Math/Logic" includes probability theory, logic, philosophy.
"Practice" includes ancestral heuristics, inherited tricks and is largely convex, precautionary and **via negativa** .

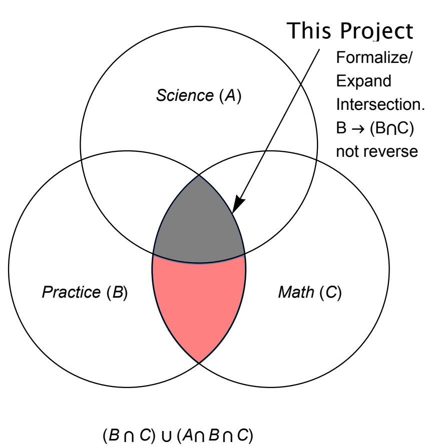


Figure 2.2: **The Right Way: Intersection is Not Sum** The rigorous way to formalize and teach probability and risk (though not to make decisions).
"Evidentiary" science is not robust enough in dealing with the unknown compared to heuristic decision-making. So this is about **what we can talk about in words/print and lecture about**, i.e., an explicit methodology.
The progress to "rigorify" practice consists in expanding the intersection by formalizing as much of **B** (i.e. learned rules of thumb) as possible.

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 by examining robustness to perturbation), what we call metaprobability, even if the meta-approach to the model is not strictly probabilistic.

Definition 2.2 (Rule)**.**
A rule in this project is a decision-making convex heuristic as defined in 2.4 page 48 that operates under a "broad set of circumtances" (that is, not concave under parameter perturbation as defined in Chapter 17). As illustrated in figures 2.1 and 2.2, a rule needs to lie outside the set $(A \cup B \cup C) \cap (A \cap B)' \cap (B \cap C)' \cap (A \cap C)'$ *(where ′ denotes the complement of the set).*

Unlike a theorem, which depends on a specific (and closed) set of assumptions, a rule holds across a broad range of environments – which is precisely the point. In that sense it is more rigorous than a theorem for decision-making, as it is in consequence space, concerning $f(x)$, not truth space, the properties of x as defined in 2.3.

Definition 2.3 (Evidentiary v/s Precautionary Approaches)**.**
(a) Evidentiary risk analysis consists in looking at properties of statistically derived empirical estimators as mapped in 3.2 page 66 and their loss functions as expressed in 3.4.

(b) Precautionary approaches are decisions based on absence of data, evidence, and clarity about the properties in the tails of the distribution. They consist in mapping using stability of the loss function under parametric perturbation or change in probability structure (fragility analysis) using methods defined in Chapter 17 (with summary in 2.4).

As shown in Table 5, in effect Evidentiary is narrowly probabilistic, while precautionary is metaprobabilistic (metaprobabilistic is defined in 1.9 on page 28).

Remark 2.1.
Tail risks and extreme deviations cannot be assessed solely by evidentiary methods, simply because of absence of rare events in past samples.

The point is deepened in Chapter 3

Figure 2.2 shows how and where mathematics imparts a necessary rigor in some places, at the intersection of theory and practice; and these are the areas we can discuss in this book. And the notion of intersection is not to be taken casually, owing to the inverse problem explained in section 2.2.

Principle 2.1 (Mathematics debunks mathematics)**.**
Mathematical "charlatanry" and fallacies in probabilities should be debunked using mathematics and mathematical arguments first.

Simply, the statement "everything cannot be mathematized", can be true, but "things that are falsely mathematized" can be detected from 1) assumptions, 2) richness of model, 3) inner sensitivity of model to parametric or distributional perturbations. For instance, we can show how "Value-at-Risk" fails by mathematical methods, using distibutional perturbations.

Table 5: The Difference Between Statistical/Evidentiary and Fragility-Based Risk Management

	Evidentiary Risk Management	**Analytical and Precautionary Risk Management**	
	Statistical/Actuarial Based	Model Based	**Fragility Based**
	Relies on past	Relies on theoretical model (with statistical backup/backtesting)	Relies on present attributes of object
Probabilistic?	Probabilistic	Probabilistic	Nonprobabilistic or indirectly probabilistic (only reasoning is probabilistic)
Typical Methods	Times series statistics, etc.	Use of estimated probability distribution Forecasting models	Detection of nonlinearity through heuristics
Expression	Variance, Value at Risk	Variance, Value at Risk, Tail exposure, (Shortfall)	Fragility Indicator, Heuristics
Characteristic	Dependence on both past sample and parameters	Dependence on parameters	Dependence on detection of second order effects
Performance	Erratic, Unreliable for tails	Erratic, Unreliable for tails	Robust, Focused on tails

2.2 PROBLEMS AND INVERSE PROBLEMS

Definition 2.4 (The inverse problem.)**.**
There are many more degrees of freedom (hence probability of making a mistake) when one goes from a model to the real world than when one goes from the real world to the model.

From *The Black Swan*, [110]

> Operation 1 (the melting ice cube): Imagine an ice cube and consider how it may melt over the next two hours while you play a few rounds of poker with your friends. Try to envision the shape of the resulting puddle.
>
> Operation 2 (where did the water come from?): Consider a puddle of water on the floor. Now try to reconstruct in your mind's eye the shape of the ice cube it may once have been. Note that the puddle may not have necessarily originated from an ice cube.

One can show probabilistically the misfitness of mathematics to many problems where it is used. It is much more rigorous and safer to start with a disease then look at the classes of drugs that can help (if any, or perhaps consider that no drug can be a potent alternative), than to start with a drug, then find some ailment that matches it, with the serious risk of mismatch. Believe it or not, the latter was the norm at the turn of last century, before the FDA got involved. People took drugs for the sake of taking drugs, particularly during the snake oil days.

What we are saying here is now accepted logic in healthcare but people don't get it when we change domains. In mathematics it is much better to start with a real problem, understand it well on its own terms, then go find a mathematical tool (if any, or use nothing as is often the best solution) than start with mathematical theorems then find some application to these. The difference (that between problem and inverse problem) is monstrous as the degrees of freedom are much narrower in the foreward than the backward equation, sort of). To cite Donald Geman (private communication), there are tens of thousands theorems one can elaborate and prove, all of which may seem to find *some* application in the real world, particularly if one looks hard (a process similar to what George Box calls "torturing" the data). 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. But people don't get that point there. The entire fields of mathematical economics and quantitative finance are based on that fabrication. Having a tool in your mind and looking for an application leads to the narrative fallacy.

> From *Antifragile* [111]:*There is such a thing as "real world" applied mathematics: find a problem first, and look for the mathematical methods that works for it (just as one acquires language), rather than study in a vacuum through theorems and artificial examples, then find some confirmatory representation of reality that makes it look like these examples.*

The point will be discussed in Chapter 8 in the context of statistical data mining.

> Nevertheless, once one got the math for it, stay with the math. Probabilistic problems can only be explained mathematically. We discovered that it is impossible to explain the difference thin tails/fat tails (Mediocristan/Extremistan) without mathematics. The same with the notion of "ruin".

This also explains why schooling is dangerous as it gives the illusion of the arrow theory $\rightarrow$ practice. Replace math with theory and you get an idea of what I call the *green lumber fallacy* in *Antifragile*.

An associated result is to ignore reality. Simply, risk management is about precautionary notes, cannot be separated from effect, the payoff, again, in the "real world", so the saying "this works in theory" but not in practice is nonsensical. And often people claim after a large blowup my model is right but there are "outliers" not realizing that we don't care about their model but the blowup.

Inverse Problem of Statistical Data

Principle 2.2 (Visibility of the Generator)**.**
In the real world one sees time series of events, not the generator of events, unless one is himself fabricating the data.

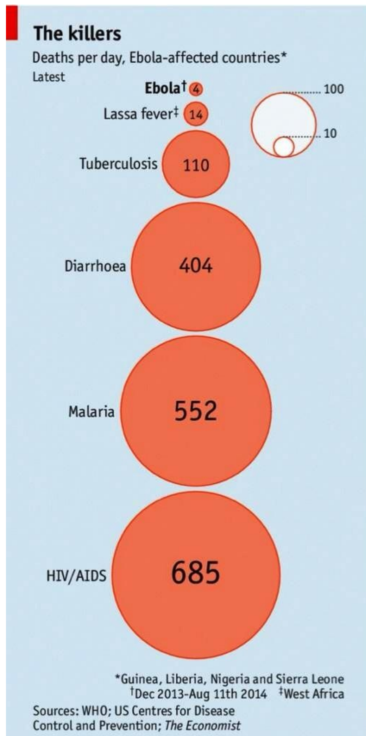


Figure 2.3: Naive evidentiary empiricism at work in treating potential epidemics as a scare. They compare number of past death, not taking into account acceleration (second order effects). While the other conditions were stable, ebola at the time was growing by 14 % a week.

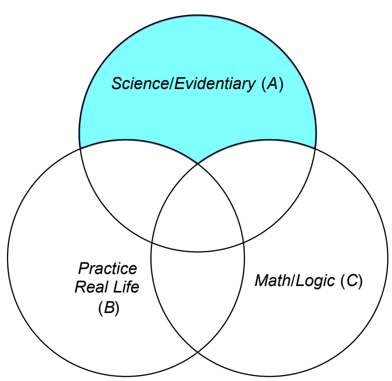


Figure 2.4: The way naive "empirical", say pro-GMOs science view nonevidentiary risk. In fact the real meaning of "empirical" is rigor in focusing on the unknown, hence the designation "skeptical empirical".
Empiricism requires logic (hence skepticism) but logic does not require empiricism.
The point becomes dicey when we look at mechanistic uses of statistics –parrotlike– and evidence by social scientists. One of the manifestation is the inability to think in nonevidentiary terms with the classical "where is the evidence?" mistake.

This section will illustrate the general methodology in detecting potential model error and provides a glimpse at rigorous "real world" decision-making.

The best way to figure out if someone is using an erroneous statistical technique is to apply such a technique on a dataset for which you have the answer. The best way to know the exact properties *ex ante* to generate it by Monte Carlo. So the technique throughout the book is to generate fat-tailed data, the properties of which we know with precision, and check how standard and mechanistic methods used by researchers and practitioners detect the *true* properties, then show the wedge between *observed* and *true* properties.

The focus will be, of course, on the effect of the law of large numbers.

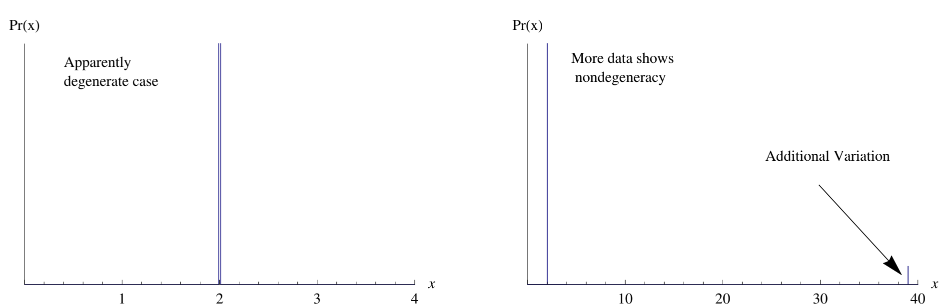


Figure 2.5: The Masquerade Problem (or Central Asymmetry in Inference). To the left, a degenerate random variable taking seemingly constant values, with a histogram producing a Dirac stick. One cannot rule out nondegeneracy. But the right plot exhibits more than one realization. Here one can rule out degeneracy. This central asymmetry can be generalized and put some rigor into statements like "failure to reject" as the notion of what is rejected needs to be refined. We produce rules in Chapter 4.

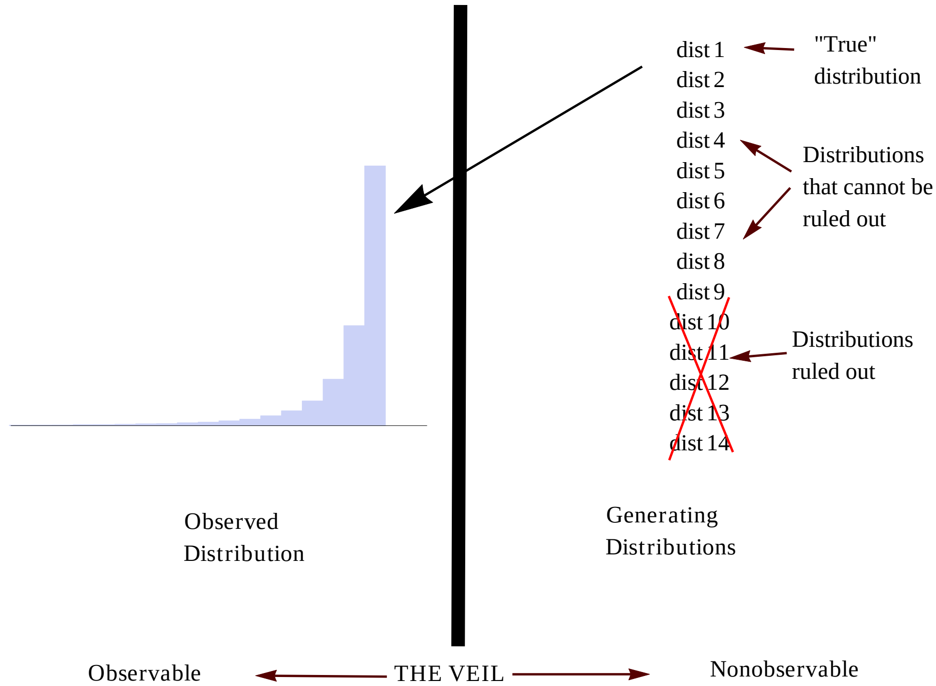


Figure 2.6: "The probabilistic veil". Taleb and Pilpel (2000,2004) cover the point from an epistemological standpoint with the "veil" thought experiment by which an observer is supplied with data (generated by someone with "perfect statistical information", that is, producing it from a generator of time series). The observer, not knowing the generating process, and basing his information on data *and data only*, would have to come up with an estimate of the statistical properties (probabilities, mean, variance, value-at-risk, etc.). Clearly, the observer having incomplete information about the generator, and no reliable theory about what the data corresponds to, will always make mistakes, but these mistakes have a certain pattern. This is the central problem of risk management.

The example in Figure 2.6 provides an idea of the methodolody, and Chapter 4 produces a formal "hierarchy" of statements that can be made by such an observer without violating a certain inferential rigor. For instance he can "reject" that the data is Gaussian, but not accept it as easily. And he can produce inequalities or "lower bound estimates" on, say, variance, never "estimates" in the standard sense since he has no idea about the generator and standard estimates require some associated statement about the generator.

Definition 2.5.
(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.

Chapter 4 will rank probability measures along this arbitrage criterion.

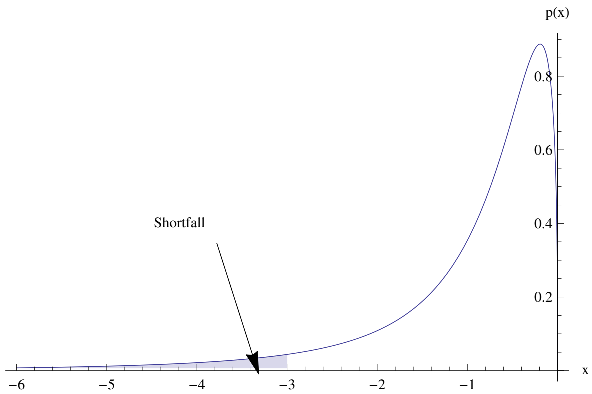


Figure 2.7: The "true" distribution as expected from the Monte Carlo generator

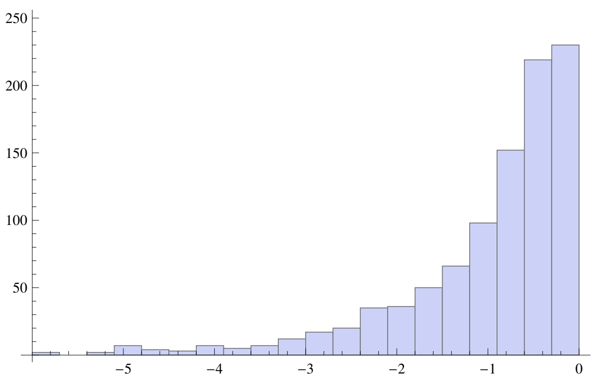


Figure 2.8: A typical realization, that is, an observed distribution for $N = 10^3$

Example of Finite Mean and Infinite Variance This example illustrates two biases: underestimation of the mean in the presence of skewed fat-tailed data, and illusion of finiteness of variance (sort of underestimation).

Let us say that x follows a version of Pareto Distribution with density $p(x)$,

$$p(x) = \begin{cases} \dfrac{\alpha k^{-1/\gamma}(-\mu-x)^{\frac{1}{\gamma}-1}\left(\left(\frac{k}{-\mu-x}\right)^{-1/\gamma}+1\right)^{-\alpha-1}}{\gamma} & \mu+x \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

By generating a Monte Carlo sample of size N with parameters $\alpha = 3/2, \mu = 1, k = 2$, and $\gamma = 3/4$ and sending it to a friendly researcher to ask him to derive the properties, we can easily gauge what can "fool" him. We generate M runs of N-sequence random variates $((x_i^j)_{i=1}^N)_{j=1}^M$

The expected "true" mean is:

$$\mathbb{E}(x) = \begin{cases} \frac{k\,\Gamma(\gamma+1)\Gamma(\alpha-\gamma)}{\Gamma(\alpha)} + \mu & \alpha > \gamma \\ \text{Indeterminate} & \text{otherwise} \end{cases}$$

and the "true" variance:

$$V(x) = \begin{cases} \frac{k^2\left(\Gamma(\alpha)\Gamma(2\gamma+1)\Gamma(\alpha-2\gamma)-\Gamma(\gamma+1)^2\Gamma(\alpha-\gamma)^2\right)}{\Gamma(\alpha)^2} & \alpha > 2\gamma \\ \text{Indeterminate} & \text{otherwise} \end{cases} \tag{2.2}$$

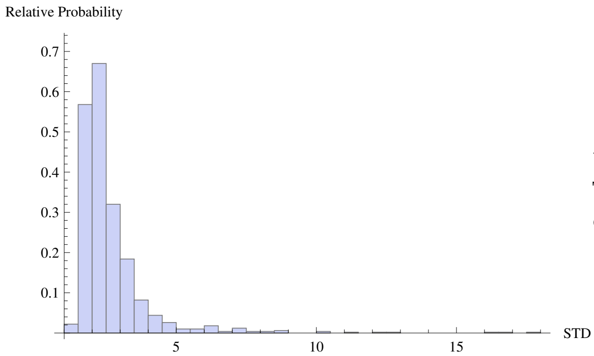


Figure 2.9: The Recovered Standard Deviation, which we insist, is infinite. This means that every run j would deliver a different average

which in our case is "infinite". Now a friendly researcher is likely to mistake the mean, since about $\tilde{6}0\%$ of the measurements will produce a higher value than the true mean, and, most certainly likely to mistake the variance (it is infinite and any finite number is a mistake).

Further, about 73% of observations fall *above* the true mean. The CDF= $1 - \left(\left(\frac{\Gamma(\gamma+1)\Gamma(\alpha-\gamma)}{\Gamma(\alpha)}\right)^{\frac{1}{\gamma}} + 1\right)^{-\alpha}$ where Γ is the Euler Gamma function $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, \mathrm{d}t$.

As to the expected shortfall, $S(K) \equiv \frac{\int_{-\infty}^{K} x\, p(x)\, dx}{\int_{-\infty}^{K} p(x)\, dx}$, close to 67% of the observations underestimate the "tail risk" below 1% and 99% for more severe risks. This exercise was a standard one but there are many more complicated distributions than the ones we played with.

Good News: Rules for Decision Theory

Table 6 provides a robust approach to the problem.

The good news is that the real world is about exposures, and exposures are asymmetric, leading us to focus on two aspects: 1) probability is about bounds, 2) the asymmetry leads to convexities in response, which is the focus of this text. Note that, thanks to inequalities and bounds (some tight, some less tight), the use of the classical theorems of probability theory can lead to classes of qualitative precautionary decisions that, ironically, do not rely on the computation of specific probabilities.

Table 6: General Rules of Risk Engineering

	Rules	Description
$\mathcal{R}1$	Dutch Book	Probabilities need to add up to 1* $-$ but cannot exceed 1
$\mathcal{R}1'$	Inequalities	It is more rigorous to work with probability **inequalities** and bounds than probabilistic estimates.
$\mathcal{R}2$	Asymmetry	Some errors have consequences that are largely, and clearly one sided.**
$\mathcal{R}3$	Nonlinear Response	Fragility is more measurable than probability***
$\mathcal{R}4$	Conditional Precautionary Principle	Domain specific precautionary, based on fat tailedness of errors and asymmetry of payoff.
$\mathcal{R}5$	Decisions	Exposures ($f(x)$)can be more reliably modified, instead of relying on computing probabilities of x.

* The Dutch book can be expressed, using the spirit of quantitative finance, as a no arbitrage situation, that is, no linear combination of payoffs can deliver a negative probability or one that exceeds 1. 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.

** Consider a plane ride. Disturbances are more likely to delay (or worsen) the flight than accelerate it or improve it. This is the concave case. The opposite is innovation and tinkering, the convex case.

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

Figure 2.10: The risk of breakage 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. Further, fragile objects are disproportionally more vulnerable to tail events than ordinary ones –by the concavity argument.

The Supreme Scientific Rigor of The Russian School of Probability One can believe in the rigor of mathematical statements about probability without falling into the trap of providing naive computations subjected to model error. If this author were to belong to a school of thought designated by a nationality, the

{Nationality} school of {discipline},

it would be the Russian school of probability.

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

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

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

[a] The way this connects to robustness, which we will formalize next section, is as follows. Is robust what does not change across perturbation of parameters of the probability distribution; this is the core of the idea in Part II with our focus on fragility and antifragility. The point is refined with concave or convex to such perturbations.

2.3 FRAGILITY, NOT JUST STATISTICS, FOR HIDDEN RISKS

Let us start with a sketch of the general solution to the problem of risk and probability, just to show that there is a solution (it will take an entire book to get there). The following section will outline both the problem and the methodology.

This reposes on the central idea that an assessment of fragility −and control of such fragility−is more ususeful, and more reliable,than probabilistic risk management and data-based methods of risk detection.

In a letter to *Nature* about the book *Antifragile*[111]: *Fragility* (the focus of Part III of this volume) 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.

Same with model errors –as we subject models to additional layers of uncertainty.

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 the statement in *Antifragile* about our chronic inability to get the risk of a variable from the past with economic time series, with associated overconfidence. "Where is he going to get the risk from since we cannot get it *from the past*? from the future?", he wrote. Not really, it is staring at us: *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. We can use the past to derive general statistical statements, of course, coupled with rigorous probabilistic inference but it is unwise to think that the data unconditionally yields precise probabilities, as we discuss next.

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 negative inference, *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. As we will see in Chapter x, it pervades social science and areas of science supported by statistical analyses. Yet 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 what Daniel Kahneman calls the "outside view", given that we are all part of history, part of the sample so to speak.

Let us now look at the point more formally, as the difference between an assessment of fragility and that of statistical knowledge can be mapped into the difference between x and $f(x)$

This will ease us into the "engineering" notion as opposed to other approaches to decision-making.

2.4 SOLUTION: THE CONVEX HEURISTIC

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

In his own discussion of the Borel-Cantelli lemma (the version popularly known as "monkeys on a typewriter")[13], Emile Borel explained that some events can be considered mathematically possible, but practically impossible. There exists a class of statements that are mathematically rigorous but practically nonsense, and vice versa.

If, in addition, one shifts from "truth space" to consequence space", in other words focus on (a function of) the payoff of events in addition to probability, rather than just their probability, then the ranking becomes even more acute and stark, shifting, as we will see, the discussion from probability to the richer one of fragility. In this book we will include costs of events as part of fragility, expressed as fragility under parameter perturbation. Chapter 5 discusses robustness under perturbation or metamodels (or metaprobability). But here is the preview of the idea of convex heuristic, which in plain English, is at least robust to model uncertainty.

Definition 2.6 (Convex Heuristic)**.**
In short what exposure is required to not produce concave responses under parameter perturbation.

Summary of a Convex Heuristic (from Chapter 17) Let $\{f_i\}$ be the family of possible functions, as "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 decision 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 heuristics $\mathscr{H}_{x,K,\sigma^-,\Delta,etc.}$:

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

Note that we call these decision rules "convex" in $\mathscr{H}$ not necessarily because they have a convex payoff, but also because, thanks to the introduction of payoff f, their payoff ends up comparatively "more convex" than otherwise. In that sense, finding protection is a convex act.

Outline of Properties (nonmathematical) of Convex Heuristics
Their aim is not to be "right" and avoid errors, but to ensure that errors remain small in consequences.

Definition 2.7.
A convex heuristic *is a decision rule with the following properties:*

- *Compactness: It is easy to remember, implement, use, and transmit.*
- *Consequences, not truth: It is about what it helps you do, not whether it is true or false. It should be judged not in "truth space" but in "consequence space."*
- *Antifragility: It is required to have a benefit when it is helpful larger than the loss when it is harmful. Thus it will eventually deliver gains from disorder.*
- *Robustness: It satisfies the fragility-based precautionary principle.*
- *Opacity: You do not need to understand how it works.*
- *Survivability of populations: Such a heuristic should not be judged solely on its intelligibility (how understandable it is), but on its survivability, or on a combination of intelligibility and survivability. Thus a long-surviving heuristic is less fragile than a newly emerging one. But ultimately it should never be assessed in its survival against other ideas, rather on the survival advantage it gave the populations who used it.*

The idea that makes life easy is that we can capture model uncertainty (and model error) with simple tricks, namely the scale of the distribution.

2.4.1 Convex Heuristics, Rationality, and Revelation of Preferences

One brilliant contribution by economists is the concept of "cheap talk", or the difference between "stated preferences" (what you say) and "revealed preferences" (those that can be inferred from actions). Actions are louder than words: what people say (in opinion polls or elsewhere) isn't as relevant, as individuals reveal their preferences with hard cash or, more generally costly action, or even more generally risky action (which, invariably, brings us to *skin in the game*). This is why opinion polls are considered largely entertainment. Further, the notion of "belief" is largely misunderstood.

Those who engage in actions that threaten their survival will eventually disappear, if their skin is in their game. Same with populations with the wrong heuristics.

Belief is deeply connected to probability (belief in the epistemic sense). Which means that violations of probability axioms and bounds can imply irrationality.

We showed here that the notion of "probability" raw is largely verbalistic and empty (probability maps to "degrees of belief" mathematically, is $\tilde{b}$elief), largely incomplete, more "binary" while revealed preferences via decisions is what matters (more technically probability is something deeply mathematical, useless on its own, an integral transform into something larger, and cannot be "summarized" in words). And decisions and decisions *only* can be a metric for "rationality"

Psychologists and journalistic types who make a living attacking "overreactions" and bubbles based on superficial assessments typically discuss "rationality" without getting what rationality means in its the decision-theoretic sense (the only definition that can be consistent, in terms of absence of violations of the standard axioms [CITE AXIOMS] and only from actions). But as we saw with convex heuristics the cause behind an action leading to survival is not necessarily apparent. Many seemingly irrational actions have led populations to survive. Dread risk and overreactions aren't just rational, but may be the only way to survive in some domains. [Cite Taleb and Read]

As an interesting application, one can even show that it is rational to "believe" in the supernatural if it leads to an increase in survival –*as a side effect*. [2]

This point matters a bit since "rational" in risk-taking needs to have a broader definition than "act according to model X-Y-Z" which can be incomplete. Hence the connection to metaprobability.

2.5 FRAGILITY AND MODEL ERROR

Crucially, we can gauge the nonlinear response to a parameter of a model using the same method and map "fragility to model error". For instance a small perturbation in the parameters entering the probability provides a one-sided increase of the likelihood of event (a convex response), then we can declare the model as unsafe (as with the assessments of Fukushima or the conventional Value-at-Risk models where small parameters variance more probabilities by 3 orders of magnitude). This method is fundamentally option-theoretic.

2.5.1 Why Engineering?

[Discussion of the problem- A personal record of the difference between measurement and working on reliability. The various debates.]

Risk is not Variations

On the common confustion between risk and variations. Risk is tail events, necessarily.

2 Many authors claim to be arbiters of "rationality" and, as we can see in Chapter on meta-distribution and the debunking of "pseudo-biases", accuse others of irrationality, but cannot come up with a coherent definition of rationality (unless model dependent, which means that a breakdown of a model or misspecification can justify actions otherwise deemed "irrational"); we can however certainly define irrationality in terms of violations of a certain set of axioms, so our definition is via negativa.

What Do Fat Tails Have to Do With This?

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. But they do so predictably. We end Part I 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 − in other words where x is so reliable that one needs an $f(x)$ that clips the left tail, hence allows for a computation of the potential shortfall. Again, to repat, it is more, much more rigorous to *modify your decisions*.

Fat Tails and Model Expansion

Next wee see how model uncertainty (or, within models, parameter uncertainty), or more generally, adding layers of randomness, cause fat tails.

Part I of this volume presents a mathematical approach for dealing with errors in conventional probability models 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.

So we may 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 because inapplicable) properties to making do with a certain set of incompleteness and preasymptotics. 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 can figure out from second order effects that tail probabilities and risk measures are understimated in some class of models.

Savage's Difference Between The Small and Large World

Ecologizing decision-making Luckily there is a profound literature on *satisficing* and various decision-making heuristics, starting with Herb Simon and continuing through various traditions delving into ecological rationality, [103], [48], [114]: in fact Leonard Savage's difference between small and large worlds will be the basis of Part I, which we can actually map mathematically.

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^{α}", 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:

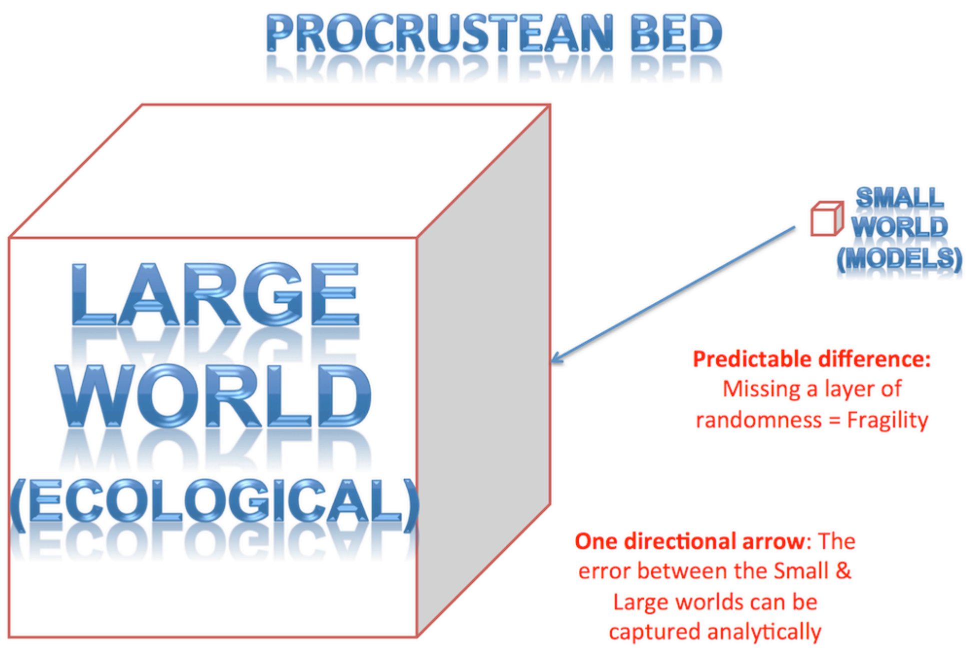


Figure 2.11: A Version of Savage's Small World/Large World Problem. In statistical domains assume **Small World= coin tosses** and **Large World = Real World**. Note that measure theory is not the small world, but large world, thanks to the degrees of freedom it confers.

> 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).

Risk (and decisions) require more rigor than other applications of statistical inference.

Coin tosses are not quite "real world" probability

In his wonderful textbook [15], 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*), and where bets are bounded, hence insensitive to problems of extreme fat tails. Ironically, measure theory, while formal, is less constraining and can set 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. Plenty of damage has been brought by the illusion that the coin toss model provides a "realistic" approach to the discipline, as we see in Chapter x, it leads to the random walk and the associated pathologies with a certain class of unbounded variables.

2.6 general classification of problems related to fat tails

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 for some classes of random variables *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).

Definition 2.8 (Degenerate Metaprobability)**.**
Indicates a single layer of stochasticity, such as a model with certain parameters.

Remark 2.2 (Knightian Risk)**.**
Degenerate metaprobability would be largely "Knightian Risk" when distribution under concern has a finite first moment.

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).We skip the elementary "Pinker" error of mistaking journalistic fact - checking for scientific statistical "evidence" and focus on less obvious but equally dangerous ones.

	Problem	Description	Chapters
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.	?
2	Inverse Problems	a) The direction Model $\Rightarrow$ Reality produces larger biases than Reality $\Rightarrow$ Model	
		b) Some models can be "arbitraged" in one direction, not the other .	1,?,?
3	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.[3]	

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 (*G*.*G*) . 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 3.
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 3.

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

2.7 CLOSING THE INTRODUCTION

We close the introduction with De Finetti's introduction to his course "On Probability":

> The course, with a deliberately generic title will deal with the conceptual and controversial questions on the subject of probability: questions which it is necessary to resolve, one way or another, so that the development of reasoning is not *reduced to a mere formalistic game of mathematical expressions* or to vacuous and simplistic pseudophilosophical statements or allegedly practical claims. (emph. mine.)

The next appendix deplores academic treatment of probability so we get it out of the way.

A | WHAT'S A CHARLATAN IN RISK AND PROBABILITY?

We start with a clean definition of charlatan. Many of us spend time fighting with charlatans; we need a cursory and useable definition that is both compatible with our probability business and the historical understanding of the snake-oil salesman.

A.1 CHARLATAN

Definition A.1 (Charlatan)**.**
In our context someone who meets at least two of the following. He

- *i- proposes complicated practical solutions to a problem that may or may not exist or has a practical simpler and less costly alternative*
- *ii- favors unconditional via positiva over via negativa*
- *iii- has small or no offsetting exposure to iatrogenics*[1]*, in a way to incur no or minimal harm should the solution to the problem be worse than doing nothing*
- *iv- avoids nonlucrative or noncareerenhancing solutions*
- *v- does not take professional, reputational or financial risks for his opinion*
- *vi- in assessments of small probability, tends to produce a number rather than a lower bound*
- *vii- tries to make his audience confuse "absence of evidence" for "evidence of absence" with small probability events.*

Definition A.2 (Skeptic/Empiricist)**.**
The skeptical empiricist is traditionally (contrary to casual descriptions) someone who puts a high burden on empirical data and focuses on the nonevidentiary unknown, the exact opposite to the naive empiricist.

Remark A.1 (Charlatan vs Skeptic)**.**
A charlatan is the exact opposite of the skeptic or skeptical empiricist.

Remark A.2 (Iatrogenics)**.**
Our definition of charlatan isn't about what he knows, but the ratio of iatrogenics in the consequences of his actions.

1 Iatrogenics is harm caused by the healer

The GMO Problem For instance we can spot a difficulty by the insistence by some "scientists" on the introduction of genetically modified "golden rice" with added vitamins as a complicated solution to some nutritional deficiency to "solve the problem" when simpler solutions (with less potential side effects) are on hand are available.

The charlatan can be highly credentialed (with publications in *Econometrica*) or merely one of those risk management consultants. A mathematician may never be a charlatan when doing math, but becomes one automatically when proposing models with potential iatrogenics and imposing them uncritically to reality. We will see how charlatans operate by implicit collusion using citation rings.

Citation Rings and Cosmetic Job Market Science

Citation rings are how charlatans can operate as a group. All members in citations rings are not necessarily charlatans, but group charlatans need citation rings.

How I came about citation rings? At a certain university a fellow was being evaluated for tenure. Having no means to gauge his impact on the profession and the quality of his research, they checked how many "top publications" he had. Now, pray, what does constitute a "top publication"? It turned out that the ranking is exclusively based on the citations *the journal* gets. So people can form of group according to the Latin expression *asinus asimum fricat* (donkeys rubbing donkeys), cite each other, and call themselves a discipline of triangularly vetted experts.

Detecting a "clique" in network theory is how terrorist cells and groups tend to be identified by the agencies.

Now what if the fellow got citations on his own? The administrators didn't know how to handle it.

Looking into the system revealed quite a bit of arbitrage-style abuse by operators.

Subdiscipline of Bullshittology I am being polite here. I truly believe that a scary share of current discussions of risk management and probability by *nonrisktakers* fall into the category called obscurantist, partaking of the "bullshitology" discussed in Elster: "There is a less polite word for obscurantism: bullshit. Within Anglo-American philosophy there is in fact a minor sub-discipline that one might call bullshittology." [29]. The problem is that, because of nonlinearities with risk, minor bullshit can lead to catastrophic consequences, just imagine a bullshitter piloting a plane. My angle is that the bullshit-cleaner in the risk domain is skin-in-the-game, which eliminates those with poor understanding of risk.

Definition A.3 (Higher order self-referential system)**.**
A_i references $A_{j\neq i}$, A_j references $A_{z\neq j}$, $\cdots$, A_z references A_i.

Definition A.4 (Academic Citation Ring)**.**
A legal higher-order self-referential collection of operators who more or less "anonymously"

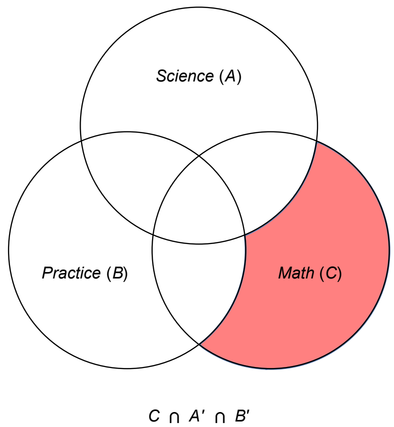


Figure A.1: The Triffat Fallacy, or the way academic decision theory and mathematical statistics view decision, probability, and risk.

peer-review and cite each other, directly, triangularly, or in a network manner, constituting a clique in a larger network, thus creating so-called academic impact ("highly cited") for themselves or their journals.

Citation rings become illegal when operators use fake identities; they are otherwise legal no matter how circular the system.

The mark of such system is engagement in incremental science in a given direction, calling each other's results "innovative". Example of dangerous citation ring: Markowitz mean-variance, GARCH, Value-At-Risk and more general risk management, some traditions of behavioral economics.

Definition A.5 (Job Market Science)**.**
A paper that follows recipes and tricks to attain higher ranking in a certain community. It seems a "contribution" but it is explained by connection to other parts which are triangularly self-referential ; it is otherwise substance-free.

Take GARCH methods (Rob Engle [35]): we know that, in practice, GARCH is totally useless to predict volatility; it is an academic PR machine. And, analytically, it is unsound under the structure of fat tails in the markets, as we will see in Chapter 3 and section 8.11 But the "Nobel" plus an active citation ring deems it a "successful" theory.

It is clear that, with rare exceptions articles published *Econometrica* are either substance-free or pure distortion (use of variance as measure of variability).

How do we break away from substance-free statistical science? Skin in the game, of course, reestablishes contact with reality, with details in Chapter 14 . The central idea is that survival matters in risk, people not truly exposed to harm can continue operating permanently.

The Triffat Fallacy

An illustration of our nighmare for risk management –and an explanation of why we can't accept current methods in economics for anything to do with the real world – is as follows. From *Antifragile*[111]:

> Modern members of the discipline of decision theory, alas, travel a one- way road from theory to practice. They characteristically gravitate to the most complicated but most inapplicable problems, calling the process "doing science."

> There is an anecdote about one Professor Triffat (I am changing the name because the story might be apocryphal, though from what I have witnessed, it is very characteristic). He is one of the highly cited academics of the field of decision theory, wrote the main textbook and helped develop something grand and useless called "rational decision making," loaded with grand and useless axioms and shmaxioms, grand and even more useless probabilities and shmobabilities. Triffat, then at Columbia University, was agonizing over the decision to accept an appointment at Harvard –many people who talk about risk can spend their lives without encountering more difficult risk taking than this type of decision. A colleague suggested he use some of his Very Highly Respected and Grandly Honored and Decorated academic techniques with something like "maximum expected utility," as, he told him, "you always write about this." Triffat angrily responded, "Come on, this is serious!"

Definition A.6 (The Triffat Fallacy)**.**
Consists in confusing the problem and the inverse problem, going from theory to practice, at the intersection $C \cap A' \cap B'$ according to definitions in A.1.

There has been a lot of trivial commentary, a recurring critique of theoretical risk management, (with the person feeling that he has just discovered it): things are "too mathematical", "mathematics does not work in the real world", or lists of what does or does not constitute "mathematical charlatanry".[2] But little or nothing seems to be done to figure out *where* math works and is needed; where standard methods ascribed to science, whether evidentiary (statistics) or analytical (mathematics/logic) do not apply in Risk management and decision making under opacity –since one doesn't have the whole story– except as constraints.

2 It has been fashionable to invoke the vague idea of mathematical "charlatanry" in the history of economics, first with Alfred Marshall famous procedure "(1) Use mathematics as shorthand language, rather than as an engine of inquiry. (2) Keep to them till you have done. (3) Translate into English. (4) Then illustrate by examples that are important in real life (5) Burn the mathematics. (6) If you can't succeed in 4, burn 3. This I do often.". Similarly, J.M. Keynes: "(...)we will endeavour to discredit the mathematical charlatanry by which, for a hundred years past, the basis of theoretical statistics have been greatly undermined", in *A Treatise On Probability* [64]. As one can see, these types of warnings proved ineffectual owing to citation rings. So our tack is different, largely constrained by the idea of skin in the game that would bring things to the missing link of reality.

Pseudo-Rigor and Lack of skin in the game

The disease of pseudo-rigor 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 rules 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 we are 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 *"science"* 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.[a]

[a] The "seminal" paper: Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999), Coherent measures of risk. [4]

Part II

FAT TAILS: THE LLN UNDER REAL WORLD ECOLOGIES

3 FAT TAILS AND THE PROBLEM OF INDUCTION

Chapter Summary 3: **Introducing mathematical formulations of fat tails. Shows how the problem of induction gets worse. Empirical risk estimator. Introduces different heuristics to "fatten" tails. Where do the tails start? Sampling error and convex payoffs.**

3.1 THE PROBLEM OF (ENUMERATIVE) INDUCTION

Turkey and Inverse Turkey (from the Glossary in *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.

3.2 EMPIRICAL RISK ESTIMATORS

Let us define an empirical risk estimator that we will work with throughout the book. We start with a partial first moment.

Definition 3.1.
(Estimator) Let X be, as of time T, a standard sequence of n+1 observations, $X = \left(x_{t_0+i\Delta t}\right)_{0\leq i\leq n}$ *(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* $T = t_0 + n\Delta t$*).* [1]

The empirical estimator $M_T^X(A, f)$ *is defined as*

$$M_T^X(A, f) \equiv \frac{\sum_{i=0}^{n} \mathbf{1}_A f\left(x_{t_0+i\Delta t}\right)}{\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}'}} \tag{3.1}$$

1 It is not necessary that Δt follows strictly calendar time for high frequency observations, as calendar time does not necessarily correspond to transaction time or economic time, so by a procedure used in option trading called "transactional time" or "economic time", the observation frequency might need to be rescaled in a certain fashion to increase sampling at some windows over others — a procedure not dissimilar to seasonal adjustment, though more rigorous mathematically. What matters is that, if there is scaling of Δt, the scaling function needs to be fixed and deterministic. But this problem is mostly present in high frequency. The author thanks Robert Frey for the discussion.

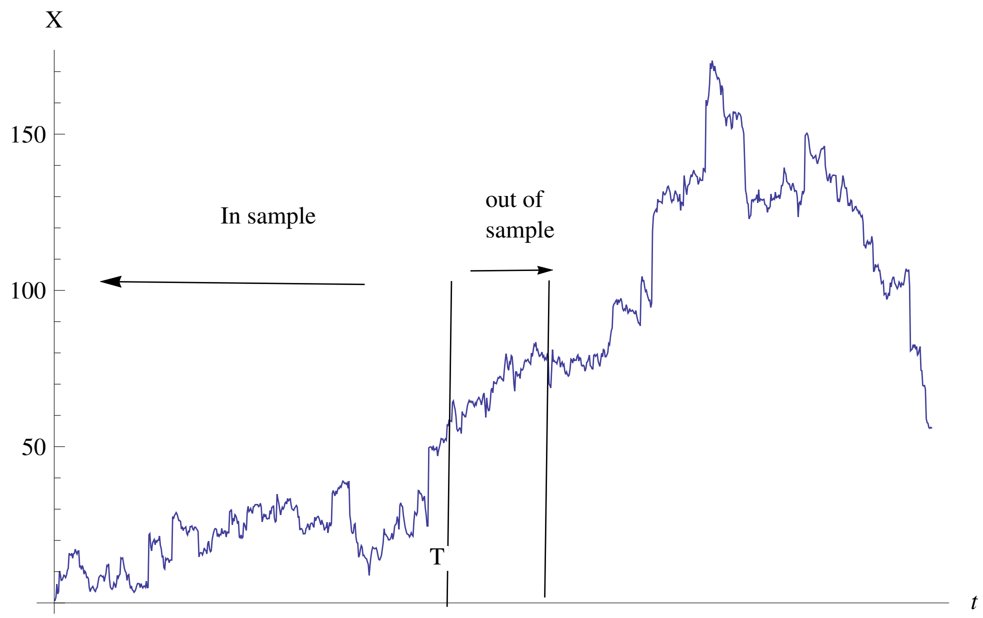


Figure 3.1: A rolling window: to estimate the errors of an estimator,it is not rigorous to compute in-sample properties of estimators, but compare properties obtained at T with prediction in a window outside of it. Maximum likelihood estimators should have their variance (or other more real-world metric of dispersion) estimated outside the window.

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, $\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}} = n$, the counting measure.

Let us stay in dimension 1 for now 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 3.2.
(Shortfall Empirical Estimator) The empirical risk estimator S for the unconditional shortfall S below K is defined as, with $A = (-\infty, K)$, $f(x) = x$

$$S \equiv \frac{\sum_{i=0}^{n} x\, \mathbf{1}_A}{\sum_{i=0}^{n} \mathbf{1}_{\mathcal{D}'}} \tag{3.2}$$

An alternative method is to compute the conditional shortfall:

$$S' \equiv \mathbb{E}[M|X < K] = \frac{\sum_{i=0}^{n} x\mathbf{1}_A}{\sum_{i=0}^{n} \mathbf{1}_A} \tag{3.3}$$

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 $f(x) = x - K$ or $K - x$.

Criterion 3.1.
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 ϵ (a tolerated relative absolute divergence; removing the absolute sign reveals the bias) so*

$$\xi(M_T^X(A_z, f)) = \frac{\mathbb{E}\left|M_T^X(A_z, f) - M_{>T}^X(A_z, f)\right|}{\left|M_T^X(A_z, f)\right|} < \epsilon \tag{3.4}$$

when $M_T^X(A_z, f)$ 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 $M_{>T}^X(A_z, 1)$ 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 use the loss function $\xi(.) = |.|$ measuring mean absolute deviations to accommodate functions and exposures and that do not have finite second moment, even if the process has such moments. Another reason is that in the real world gains and losses are in straight numerical deviations.[a]
>
> a Using absolute deviations would sound more robust than squared deviations, particularly for fat-tailed domains; it seems that the resistance comes, among other things, from the absence of derivability at 0.

So 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 subset A of the estimator.

This criterion makes our risk estimator 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 (Ω, $\mathcal{F}$, $\mathbb{P}$), 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.

3.3 FAT TAILS, THE FINITE MOMENT CASE

Some harmless formalism: L^p space. Let's look at payoff in functional space, to 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:

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

with $p \geq 1$. The application of concern for our analysis in this section is where the measure μ is a counting measure (on a countable set).

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.[2] 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).

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

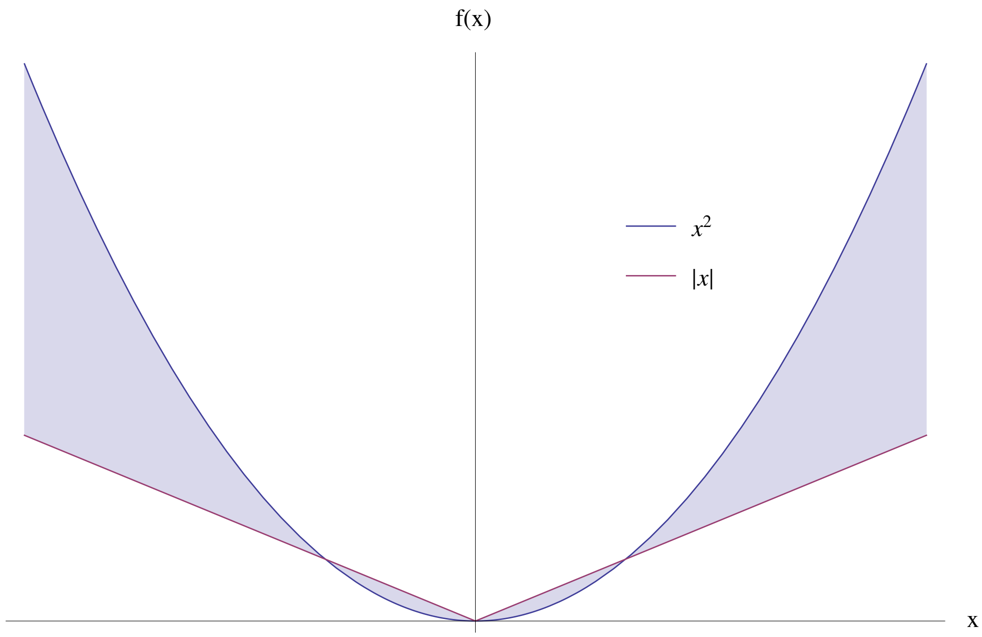


Figure 3.2: The difference between the two weighting functions increases for large values of x.

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.[3]

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

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 $a > 0$.[4],

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

3 TK Adding an appendix "Quick and Robust Estimates of Fatness of Tails When Higher Moments Don't Exist" showing how the ratios STD/MAD (finite second moment) and MAD(MAD)/STD (finite first moment) provide robust estimates and outperform the Hill estimator for symmetric power laws.

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

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

$$\{\|X\|_n\}_{n=1}^5 = \left\{1, \frac{\sqrt{5}}{2}, \frac{\sqrt[3]{7}}{2^{2/3}}, \frac{\sqrt[4]{41}}{2}, \frac{\sqrt[5]{61}}{2^{4/5}}\right\}. \tag{3.5}$$

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

$$\{||X||_n\}_{n=1}^5 = \left\{1, \frac{5}{4}, \frac{\sqrt[3]{\frac{43}{2}}}{2}, \frac{\sqrt[4]{1201}}{4}, \frac{\sqrt[5]{2101}}{2 \times 2^{3/5}}\right\}. \tag{3.6}$$

So we can see it becomes rapidly explosive.

One property quite useful with power laws with infinite moment:

$$\|X\|_\infty = \sup \left(\frac{1}{n} |x_i|\right)_{i=1}^n \tag{3.7}$$

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

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

or, alternatively

$$\frac{M_T^X\left(A, X^N\right)}{M_T^X(A, |X|)} = 2^{\frac{1}{2}(N-3)} \left(1 + (-1)^N\right) \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}-\frac{N}{2}} \Gamma\left(\frac{N+1}{2}\right) \tag{3.8}$$

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 ~ 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,

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

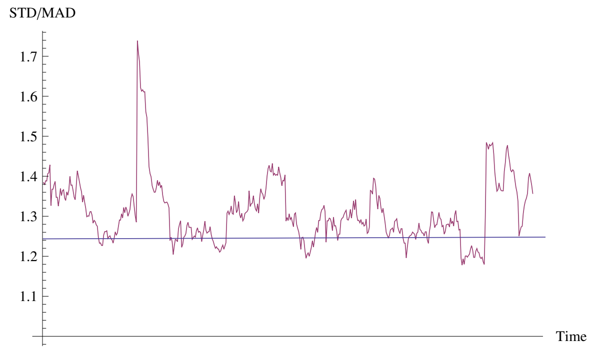


Figure 3.3: The Ratio Standard Deviation/Mean Deviation for the daily returns of the SP500 over the past 47 years, with a monthly window.

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

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

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. Figures 3.4 and 3.5 illustrate the "drifting" effect of M a with increasing information.

3.4 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

What is a "Tail Event"?

There seems to be a confusion about the definition of a "tail event", as it has different meanings in different disciplines. The three are only vaguely related.
1) In statistics: an event of low probability.
2) Here: an event of low probability but worth discussing, hence has to have some large consequence.
3) In measure and probability theory: Let $(X_i)_{i=1}^n$ be a n sequence of realizations (that is, roughly speaking a random variables–function of "event"). The tail sigma algebra of the sequence is $\mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(X_{n+1}, X_{n+2}, \ldots)$ and an event $\in \mathscr{T}$ is a tail event. So here it means a specific event extending infinitely into the future, or mathematically speaking the limiting behavior of sequence of random variables.

So when we discuss the Borel-Cantelli lemma or the zero-one law that the probability of a tail event happening infinitely often is 1 or0, it is the latter that is meant.

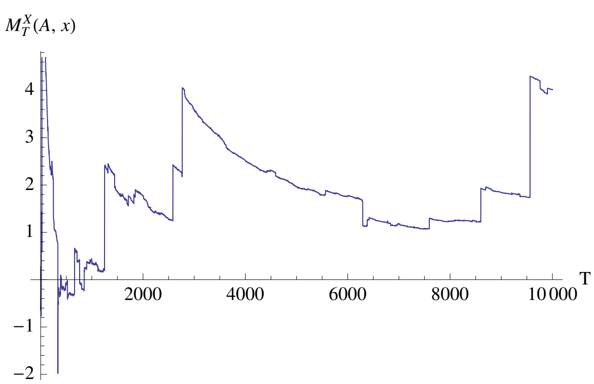


Figure 3.4: The mean of a series with Infinite mean (Cauchy).

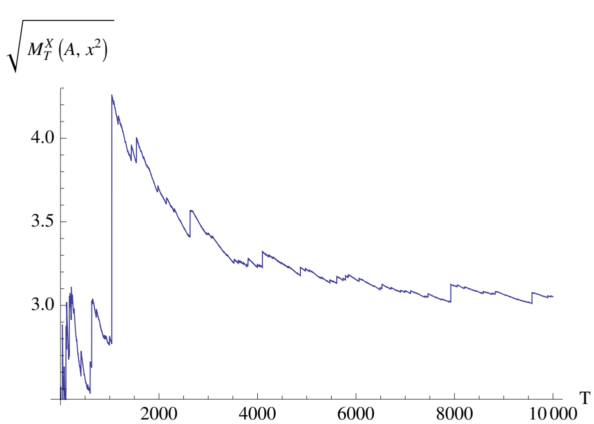


Figure 3.5: The standard deviation of a series with infinite variance (St(2)).

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

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

The characteristic function is

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

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 3.6 will present a more involved technique.

As Figure 3.6 shows: fat tails are about higher peaks, a concentration of observations around the center of the distribution.

3.5 THE BODY, THE SHOULDERS, AND THE TAILS

We assume tails start at the level of convexity of the segment of the probability distribution to the scale of the distribution.

The Crossovers and Tunnel Effect.

Notice in Figure 3.6 a series of crossover zones, invariant to a. Distributions called "bell shape" have a convex-concave-convex shape (or quasi-concave shape).

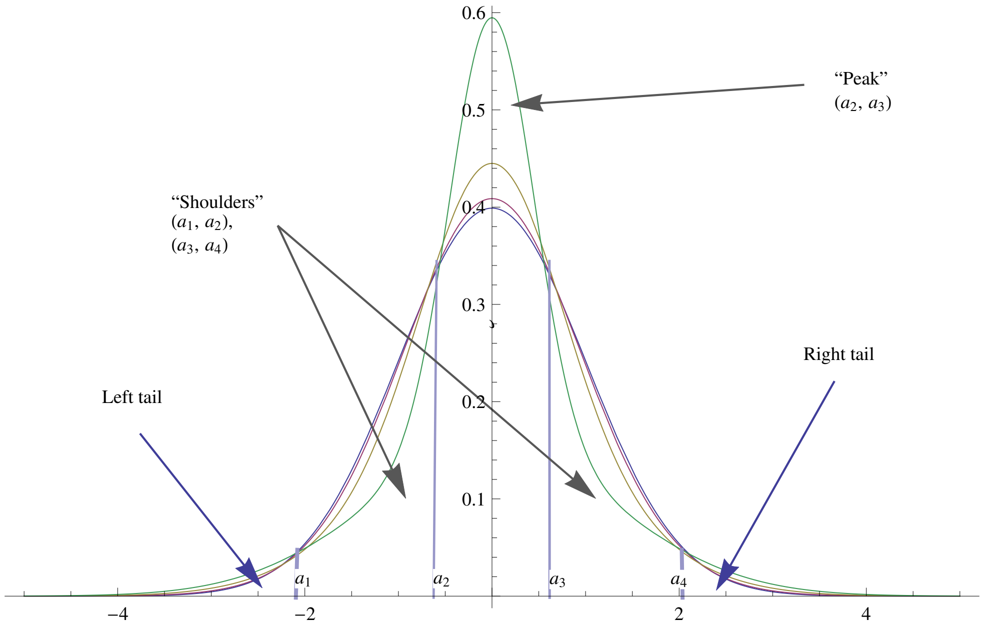


Figure 3.6: Fatter and Fatter Tails through perturbation of σ. The mixed distribution with values for the stochastic volatility coefficient a: $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. We can see crossovers a_1 through a_4. The "tails" proper start at a_4 on the right and $a_1$on the left.

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

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

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

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

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

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

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

In Summary, Where Does the Tail Start?

For a general class of symmetric distributions with power laws, the tail starts at: $\pm \frac{\sqrt{\frac{5\alpha+\sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}$, with α infinite in the stochastic volatility Gaussian case and s the standard deviation. The "tail" is located between around 2 and 3 standard deviations. This flows from our definition: which part of the distribution is convex to errors in the estimation of the scale.
But in practice, because historical measurements of STD will be biased lower because of small sample effects (as we repeat fat tails accentuate small sample effects), the deviations will be > 2-3 STDs.

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

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

which produces the following crossovers:

$$\{a_1, a_2, a_3, a_4\} = \left\{\mu - \sqrt{\frac{1}{2}\left(5+\sqrt{17}\right)}\sigma, \mu - \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma, \mu + \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma, \mu + \sqrt{\frac{1}{2}\left(5+\sqrt{17}\right)}\sigma\right\} \tag{3.9}$$

In figure 3.6, the crossovers for the intervals are numerically $\{-2.13\sigma, -.66\sigma, .66\sigma, 2.13\sigma\}$.
As to a symmetric power law(as we will see further down), the Student T Distribution with scale s and tail exponent α:

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

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

[EXPLAIN B[] is BETA HERE OR IN A TABLE OF SYMBOLS]
When the Student is "cubic", that is, $\alpha = 3$:

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

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

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

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

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

3.6 FATTENING OF TAILS WITH SKEWED VARIANCE

We can improve on the fat-tail heuristic in 3.4, (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 well the effect on probabilities of a lognormal weighting for the characteristic function

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

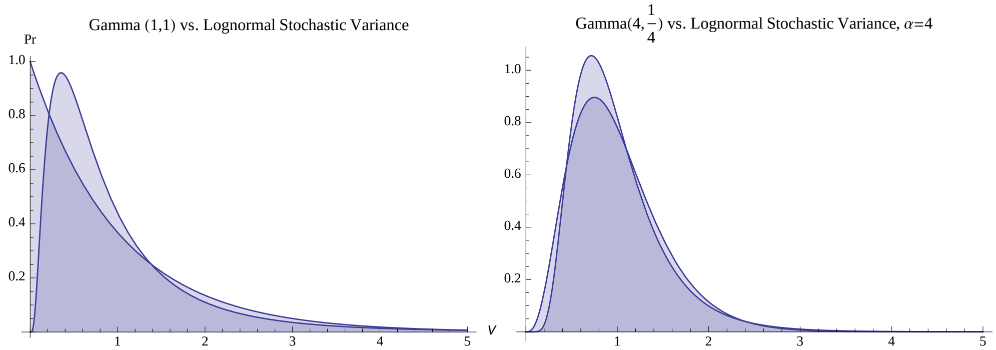


Figure 3.7: Stochastic Variance: Gamma distribution and Lognormal of same mean and variance.

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). But the absence of a closed-form distribution can be remedied as follows.

Gamma Variance A shortcut for a full lognormal distribution without the narrow scope of heuristic is to use Gamma Variance. Assume that the variance of the Gaussian follows a gamma distribution.

$$\Gamma_\alpha(v) = \frac{v^{\alpha-1}\left(\frac{V}{\alpha}\right)^{-\alpha} e^{-\frac{\alpha v}{V}}}{\Gamma(\alpha)}$$

with mean V and standard deviation $\frac{V^2}{\alpha}$. Figure 3.7 shows the matching to a lognormal with same first two moments as we get the lognormal with mean and standard deviation, respectively, $\left\{\frac{1}{2}\log\left(\frac{\alpha V^3}{\alpha V+1}\right)\right.$ and $\sqrt{-\log\left(\frac{\alpha V}{\alpha V+1}\right)}$. The final distribution becomes (once again, assuming, without loss, a mean of 0):

$$f_{\alpha,V}(x) = \int_0^\infty \frac{e^{-\frac{x^2}{2v}}}{\sqrt{2\pi}\sqrt{v}}\Gamma_\alpha(v)\mathrm{d}v$$

allora:

$$f_{\alpha,V}(x) = \frac{2^{\frac{3}{4}-\frac{\alpha}{2}}\left(\frac{V}{\alpha}\right)^{-\alpha}\left(\frac{\alpha}{V}\right)^{\frac{1}{4}-\frac{\alpha}{2}}\left(\frac{1}{x^2}\right)^{\frac{1}{4}-\frac{\alpha}{2}} K_{\frac{1}{2}-\alpha}\left(\frac{\sqrt{2}\sqrt{\frac{\alpha}{V}}}{\sqrt{\frac{1}{x^2}}}\right)}{\sqrt{\pi}\Gamma(\alpha)} \tag{3.10}$$

Chapter x will show how tail events have large errors.

Why do we use Student T to simulate symmetric power laws? For convenience, only for convenience. It is not that we *believe* that the generating process is Student T. Simply, the center of the distribution does not matter much for the properties involved in certain classes of decision making. The lower the exponent, the less the center plays a role. The higher the exponent, the more the student T resembles the Gaussian, and the more justified its use will be accordingly. More advanced

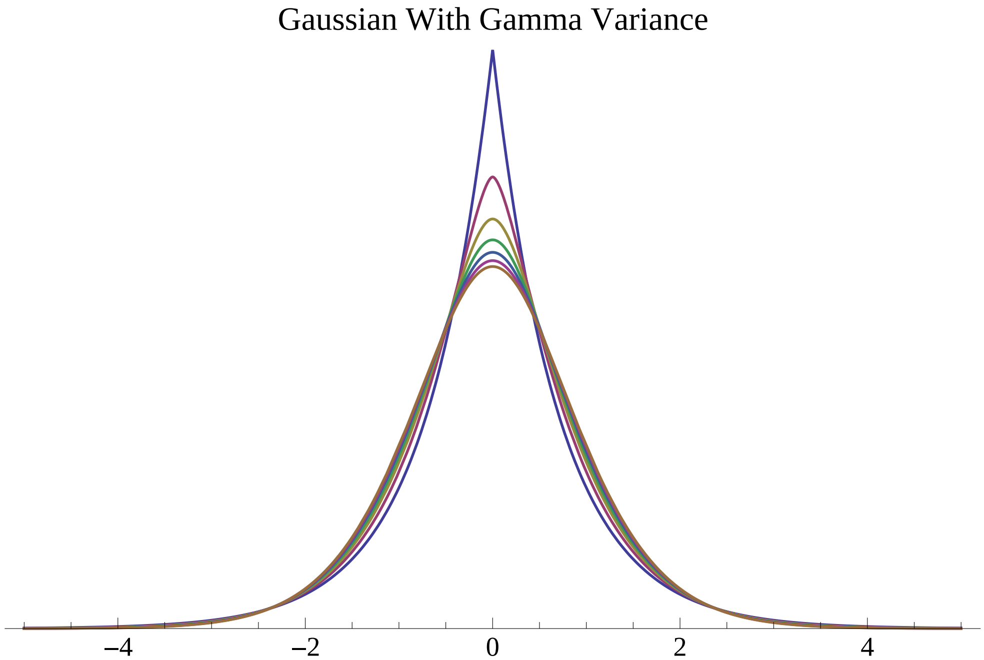


Figure 3.8: Stochastic Variance using Gamma distribution by perturbating α in equation 3.10.

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).

3.7 FAT TAILS IN HIGHER DIMENSION

$\overrightarrow{X} = (X_1, X_2, \ldots, X_m)$ the vector of random variables. Consider the joint probability distribution $f(x_1, \ldots, x_m)$. We denote the m-variate multivariate Normal distribution by $N(0, \Sigma)$, with mean vector $\overrightarrow{\mu}$, variance-covariance matrix Σ, and joint pdf,

$$f\left(\overrightarrow{x}\right) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\left(\overrightarrow{x}-\overrightarrow{\mu}\right)^T \Sigma^{-1}\left(\overrightarrow{x}-\overrightarrow{\mu}\right)\right) \tag{3.11}$$

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

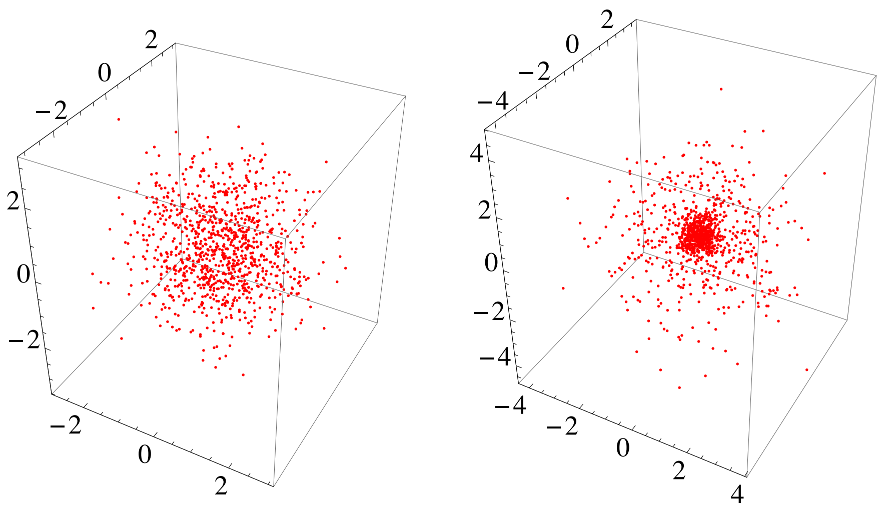


Figure 3.9: Multidimensional Fat Tails: For a 3 dimentional vector, thin tails (left) and fat tails (right) of the same variance. Instead of a bell curve with higher peak (the "tunnel") we see an increased density of points towards the center.

We can apply the same simplied variance preserving heuristic as in 3.4 to fatten the tails:

$$
\begin{aligned}
f_a\left(\vec{x}\right) &= \frac{1}{2}(2\pi)^{-m/2}|\Sigma_1|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T\Sigma_1{}^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \\
&\quad + \frac{1}{2}(2\pi)^{-m/2}|\Sigma_2|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T\Sigma_2{}^{-1}\left(\vec{x}-\vec{\mu}\right)\right)
\end{aligned}
\tag{3.12}
$$

Where a is a scalar that determines the intensity of stochastic volatility, $\Sigma_1 = \Sigma(1+a)$ and $\Sigma_2 = \Sigma(1-a)$.[5]

As we can see in Figure **??**, as with the one-dimensional case, we see concentration in the middle part of the distribution.

3.8 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.

5 We can simplify by assuming as we did in the single dimension case, without any loss of generality, that $\vec{\mu} = (0, \ldots, 0)$.

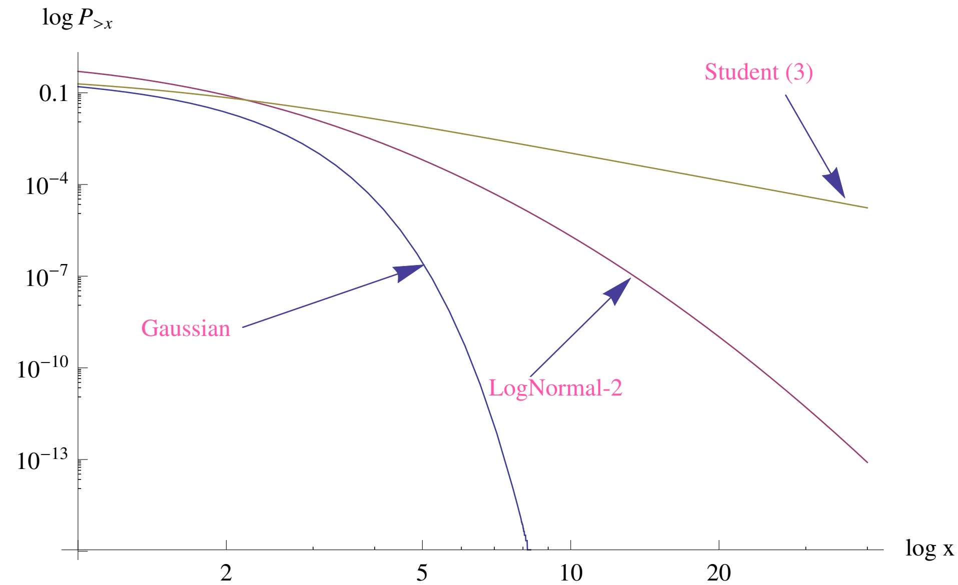


Figure 3.10: Three Types of Distributions. As we hit the tails, the Student remains scalable while the Standard Lognormal shows an intermediate position before eventually ending up getting an infinite slope on a log-log plot.

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

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

for all constants $t > 0$.

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

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

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.

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

Table 7: Scalability, comparing slowly varying functions to other distributions

We will return to the issue in Chapter 10.

3.9 SUBEXPONENTIAL AS A CLASS OF FAT TAILED 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 3 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 $X = (x_i)_{1\leq i\leq n}$ be realizations of i.i.d. random variables in $\mathbb{R}^+$, with cumulative distribution function F; then by the Teugels (1975)[113] and Pitman [89] (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. ÏĂ

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 (that is two-summed variables) scaled at a minimum value of 1 with tail exponent α, where the density of the non-convolved distribution

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

$x \geq 1$,
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, $\mathcal{N}(0,1)$ for short (0 mean and unitary standard deviation), the densities of the convolution will be Gaussian $\left(0, \sqrt{2}\right)$, the ratio of the exceedances

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

will rapidly explode.

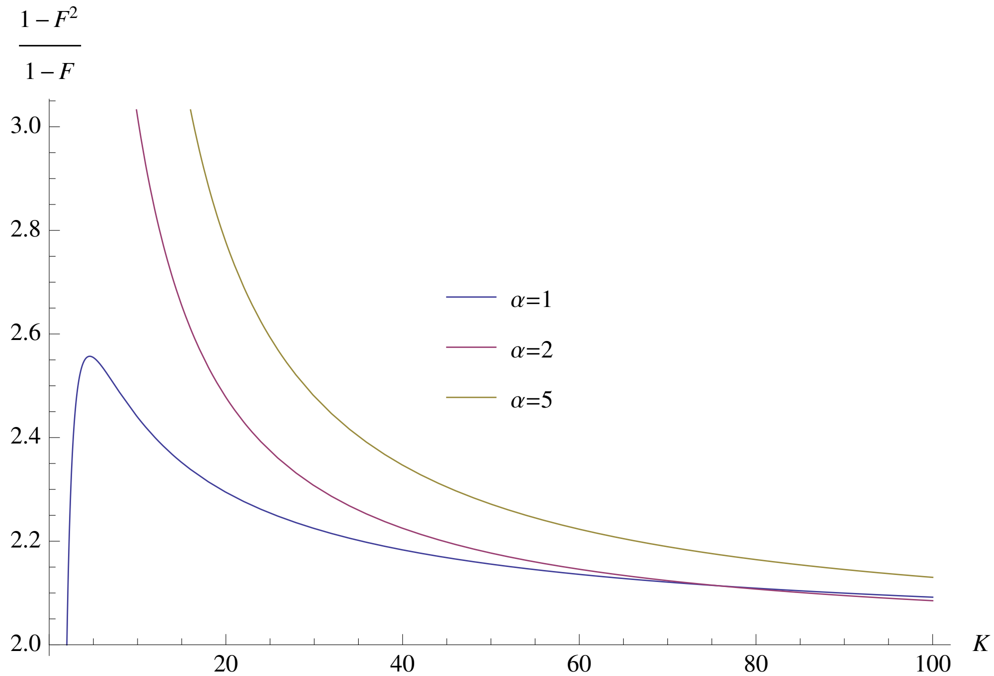


Figure 3.11: The ratio of the exceedance probabilities of a sum of two variables over a single one: power law

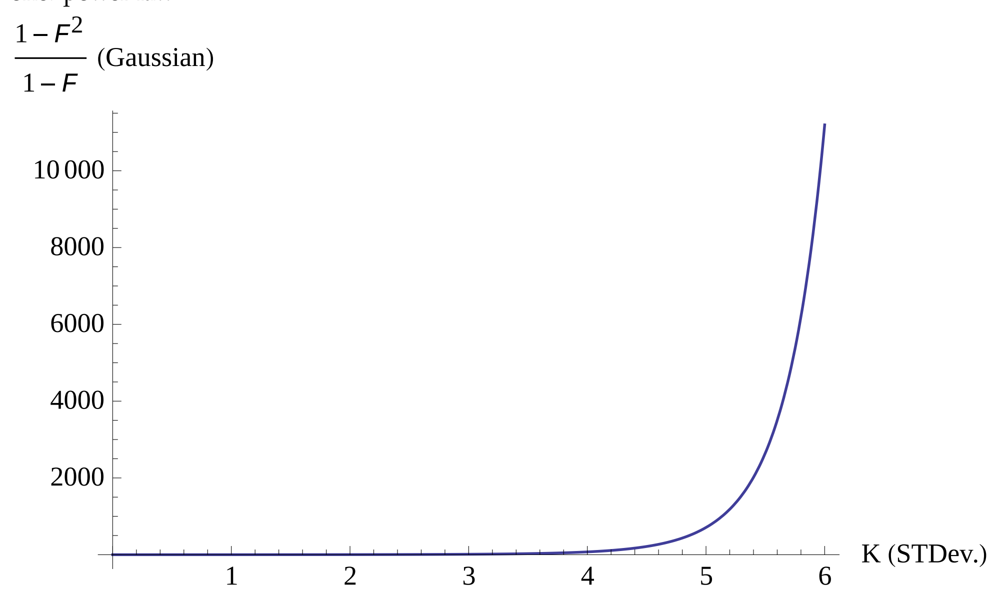


Figure 3.12: The ratio of the exceedance probabilities of a sum of two variables over a single one: Gaussian

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?

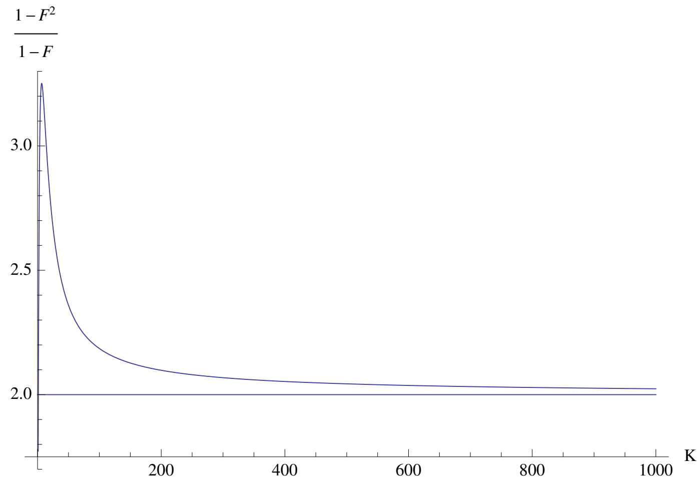


Figure 3.13: The ratio of the exceedance probabilities of a sum of two variables over a single one: Case of the Lognormal which in that respect behaves like a power law

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,

$$P[X_1 + X_2 = s] = f^2(s) = \int f(y)\ f(s-y)\, dy.$$

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

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

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

For a Gaussian, the product becomes

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

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 $\beta \geq \frac{s}{2} - x_{\min}$

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

The product of two densities decreases with β for the Gaussian[6], 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.

6 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) $\pm$ 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.

More General Approach to Subexponentiality

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

For a one-tailed random variable[7],

a) $\lim_{x\to\infty} \frac{P_{X>\Sigma x}}{P_{X>x}} = n$, (Christyakov, 1964, [19]), which is equivalent to

b) $\lim_{x\to\infty} \frac{P_{X>\Sigma x}}{P(X>\max(x))} = 1$, (Embrecht and Goldie, 1980,[33]).

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 3.14 and 3.15 illustrate 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.

7 for two-tailed variables, the result should be the same by splitting the observations in two groups around a center. BUT I NEED TO CHECK IF TRUE

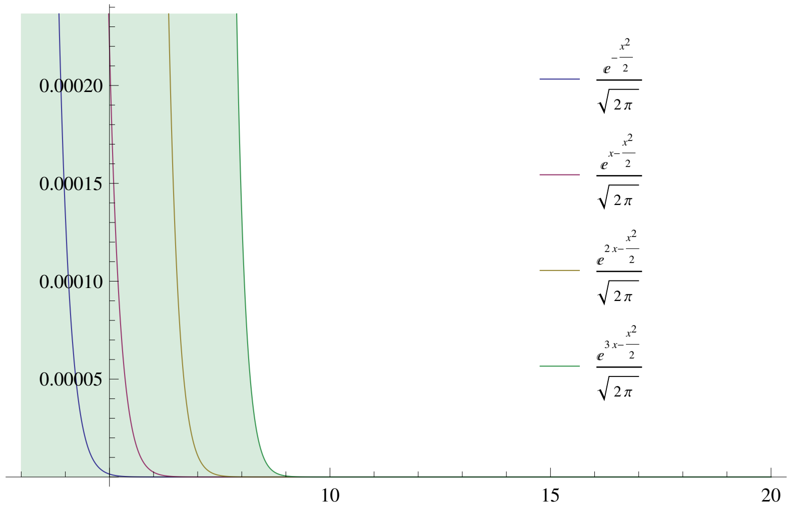


Figure 3.14: Multiplying the standard Gaussian density by e^{mx}, for $m = \{0, 1, 2, 3\}$.

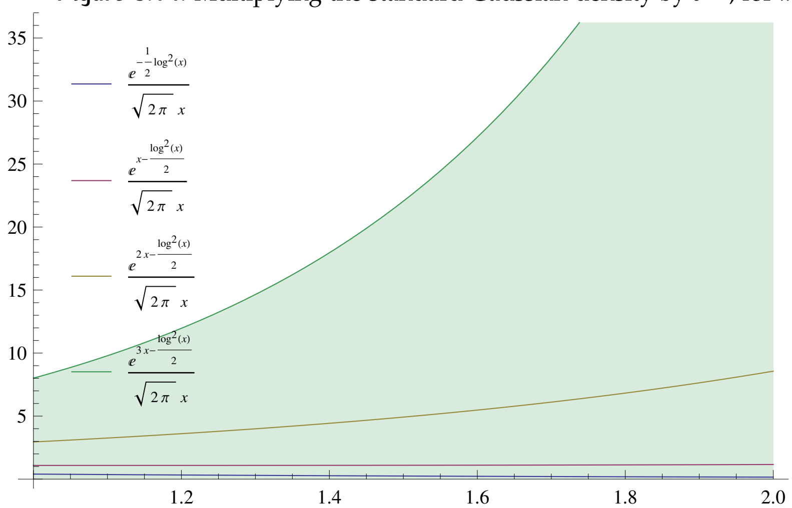


Figure 3.15: Multiplying the Lognormal (0,1) density by e^{mx}, for $m = \{0, 1, 2, 3\}$.

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 $\sim Ke^{-\beta(\log(x)-\mu)^{\gamma}}$ where γ=2)

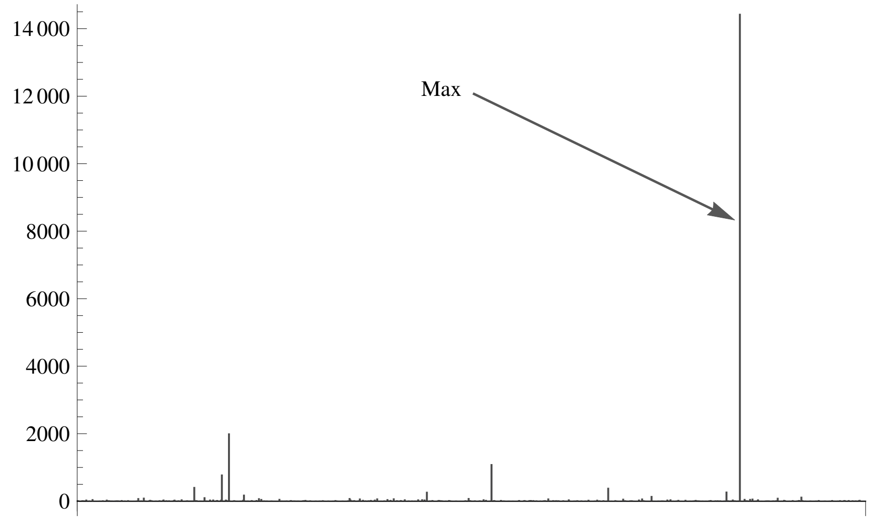


Figure 3.16: A time series of an extremely fat-tailed distribution (one-tailed). Given a long enough series, the contribution from the largest observation should represent the entire sum, dwarfing the rest.

3.10 JOINT FAT-TAILEDNESS AND ELLIPTICAL DISTRIBUTIONS

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

Definition of an Elliptical Distribution From the definition in [37], X, a p random vector is said to have an elliptical (or elliptical contoured) distribution with parameters μ, Σ and Ψ if its characteristic function is of the form $exp(it'\mu)\Psi(t\Sigma t')$.

The main property of the class of elliptical distribution is that it is closed under linear transformation. This leads to attractive properties in the building of portfolios, and in the results of portfolio theory (in fact one cannot have portfolio theory without ellitical distributions).

Note that (ironically) Levy-Stable distributions are elliptical.

Stochastic Parameters The problem of elliptical distributions is that they do not map the return of securities, owing to the absence of a single variance at any point in time, see Bouchaud and Chicheportiche (2010) [18]. When the scales of the distributions of the individuals move but not in tandem, the distribution ceases to be elliptical.

Figure 3.17 shows the effect of applying the equivalent of stochastic volatility methods: the more annoying stochastic correlation. Instead of perturbating the correlation matrix Σ as a unit as in section 3.7, we perturbate the correlations with surprising effect.

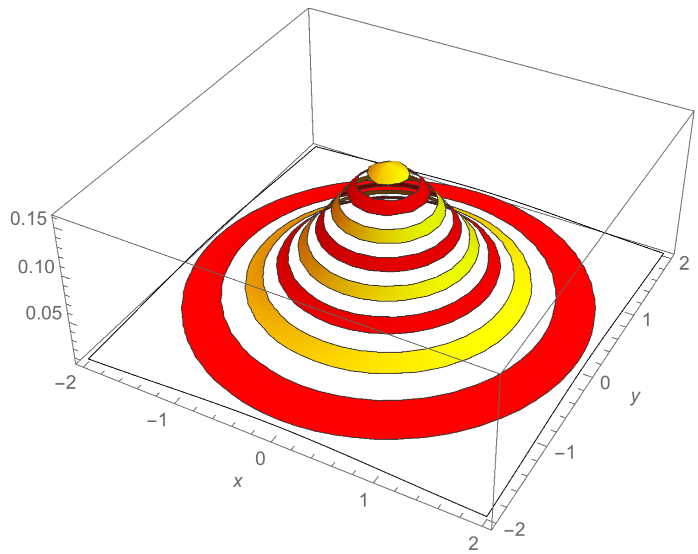


Figure 3.17: Elliptically Contoured Joint Returns of Powerlaw (Student T)

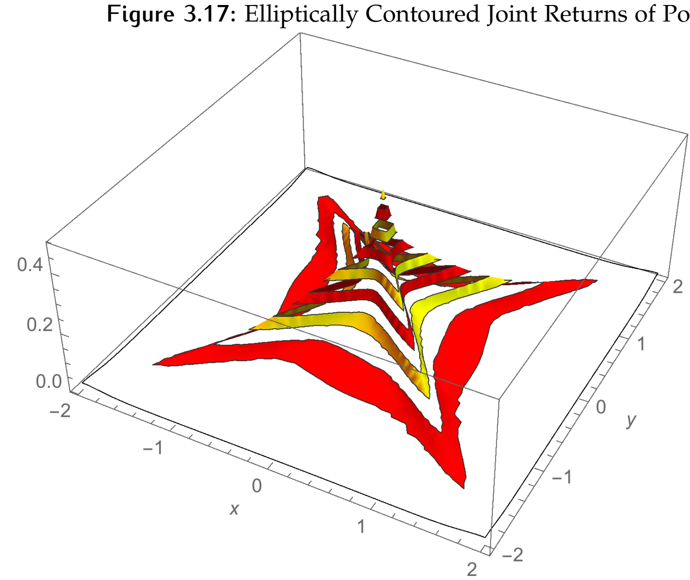


Figure 3.18: NonElliptical Joint Returns, from stochastic correlations

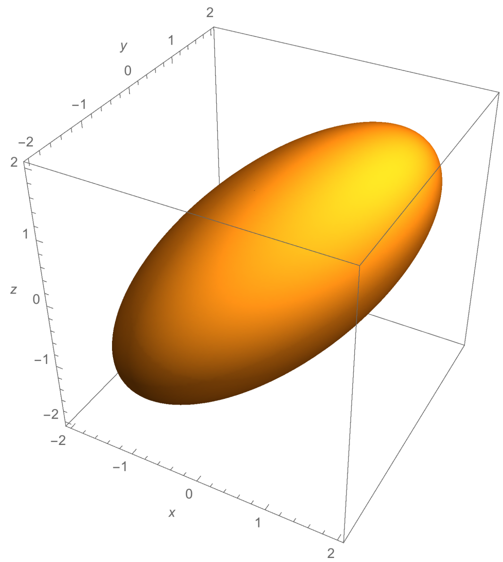


Figure 3.19: Elliptically Contoured Joint Returns for for a multivariate distribution (x, y, z) solving to the same density.

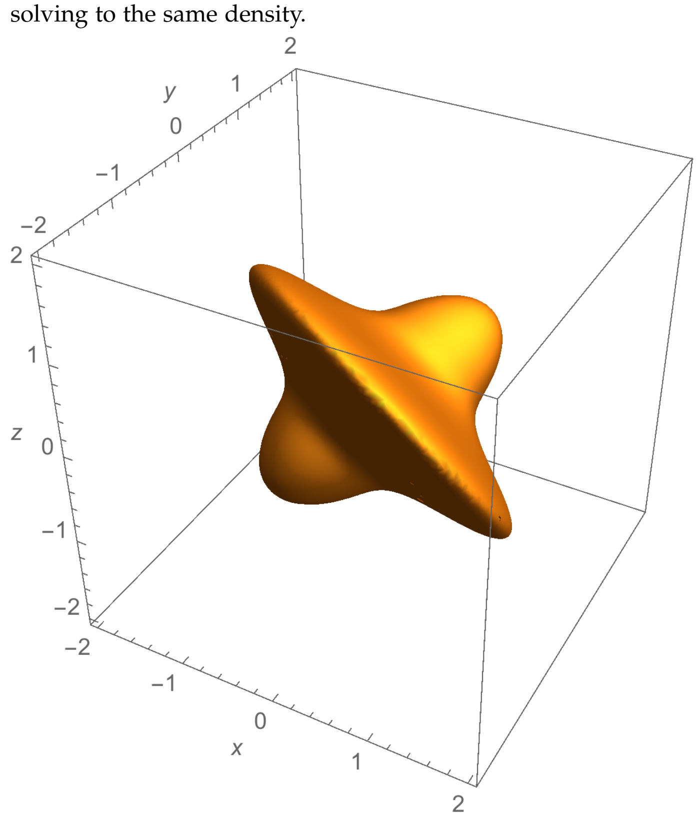


Figure 3.20: NonElliptical Joint Returns, from stochastic correlations, for a multivariate distribution (x, y, z) solving to the same density.

3.11 DIFFERENT APPROACHES FOR STATISTICAL 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 $M_T^X(A_z, f)$ will produce the following inequality:

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

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 $M_T^X(A_z, f)$ be the estimator of a function of x in the subset $A_z = (\delta_1, \delta_2)$ of the support of the variable. Let $\xi(M_T^X(A_z, f))$ be the mean absolute error in the estimation of the probability in the small subset $A_z = (\delta_1, \delta_2)$, i.e.,

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

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

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

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

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 3.1. *Under our conditions above, if for all $\lambda \in (0,1)$ and $f^{\{i,j\}}(x \pm \Delta) \in A_z$, $\frac{(1-\lambda)f^i(x-\Delta) + \lambda f^i(x+\Delta)}{f^i(x)} \geq \frac{(1-\lambda)f^j(x-\Delta) + \lambda f^j(x+\Delta)}{f^j(x)}$, ($f^i$ 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 3.2. *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 $\xi\left(M_{T,n_i}^X\left(A_{z_i}, f^i\right)\right) = \xi\left(M_{T,n_j}^X(A_{z_j}, f^j)\right)$, then*

$$n_i \geq n_j$$

This inequality becomes strict 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:

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

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

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

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 powerlaws, 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 $n_1 = 2\ n\ \sigma^2$. 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 support $v[x_{\min}^{\beta}, \infty)$,

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

Log characteristic functions allows us to deal with the difference in sums and extract the speed of convergence.

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

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

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 $M_T^X(A_{>z}, f)$ and basing it on calculations for $M_T^X(A_z, 1)$, 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 $\delta_1 \to$-∞, we may have an infinite error on $M_T^X(A_z, f)$, 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 3.83.9, then the problem becomes more acute.

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.

The Black Swan was understood by :
100% of Firemen
99.9% of skin-in-the-game risk-takers and businesspersons
85% of common readers
80% of hard scientists (except some complexity artists)
65% of psychologists (except Harvard psychologists)
60% of traders
25% of U.K. journalists
15% of money managers who manage money of others
1.5% of "Risk professionals"
1% of U.S. journalists
and
0% of economists (or perhaps, to be fair, .5%)
If is frequent that economists like Andrew Lo and Mueller [69] or Nicholas Barberis [7] play straw man by treating it as "popular" (to delegitimize is intellectual content) while both misunderstanding (and misrepresenting) its message and falling for the very errors it warns against, as in the confusion between binary and vanilla exposures.[a]

[a] Lo and Mueler: "... "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." Had they read *The Black Swan* they would have found the message is the exact opposite of "blissful ignorance".

3.12 ECONOMETRICS IMAGINES 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. Practitioners 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.

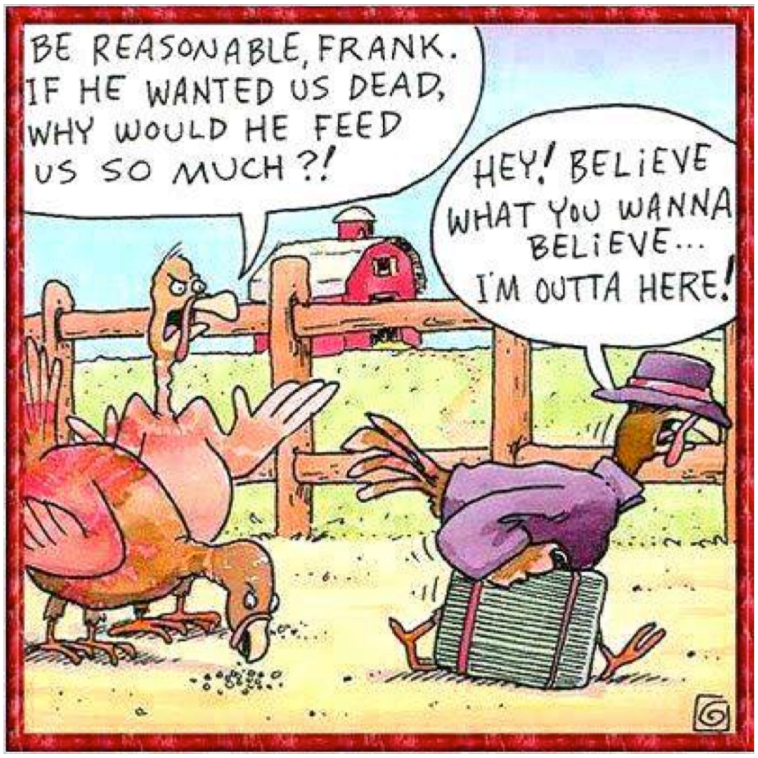


Figure 3.21: The Turkey Problem: This is the shortest explanation of the link between evidentiary and nonprecautionary risk management and the problem of induction. *Looking for the name of the author for credit/premission*.

3.13 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

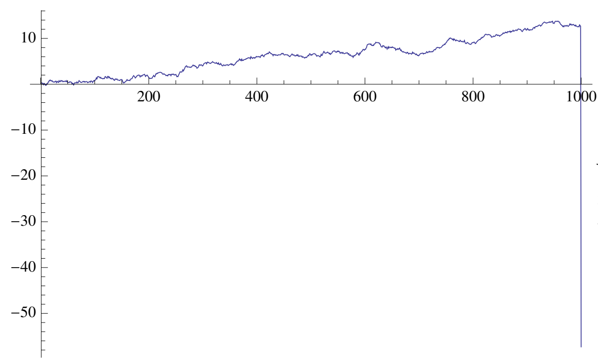


Figure 3.22: The Turkey Problem, where nothing in the past properties seems to indicate the possibility of the jump.

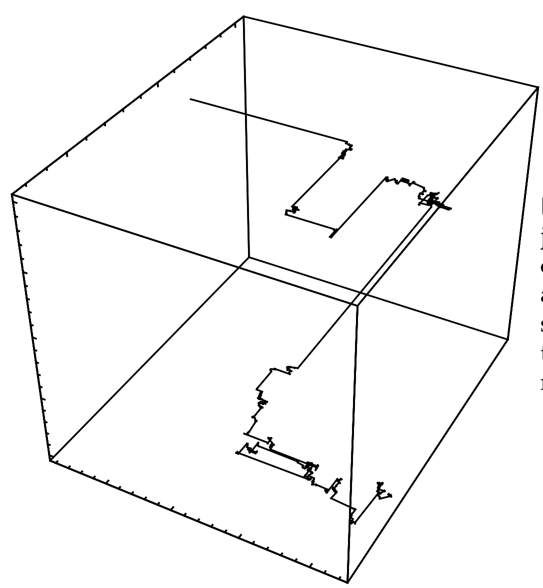


Figure 3.23: History moves by jumps: A fat tailed historical process, in which events are distributed according to a power law that corresponds to the "80/20", with $\alpha \simeq 1.2$, the equivalent of a 3-D Brownian motion.

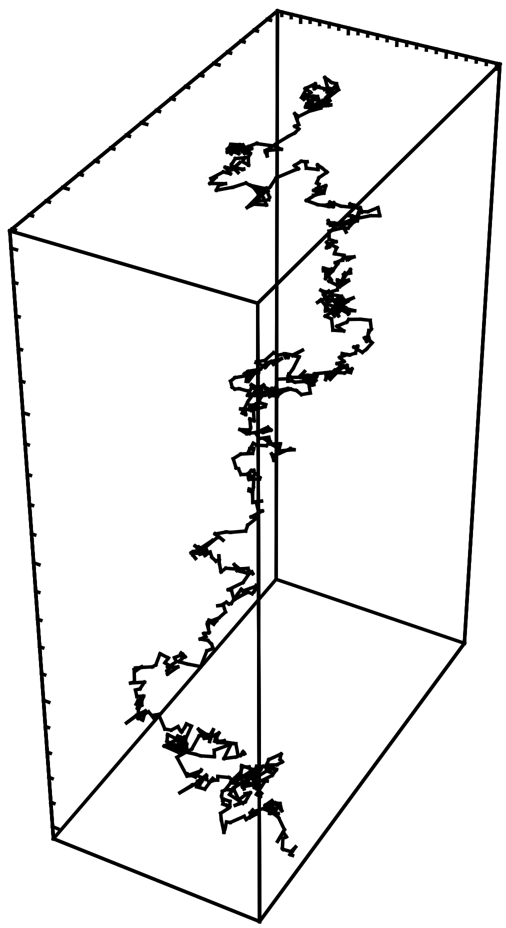


Figure 3.24: What the proponents of "great moderation" or "long peace" have in mind: history as a thin-tailed process.

Figure 3.25: High Water Mark in Palais de la Cité in Paris. The Latin poet Lucretius, who did not attend business school, wrote that we consider the biggest objeect of any kind that we have seen in our lives as the largest possible item: *et omnia de genere omni / Maxima quae vivit quisque, haec ingentia fingit.* The high water mark has been fooling humans for millennia: ancient Egyptians recorded the past maxima of the Nile, not thinking that the worst could be exceeded. The problem has recently affected the UK. floods with the "it never happened before" argument. Credit Tony Veitch

(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 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.

3.14 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). The first cumulant,

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

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

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

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

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

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.

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).

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

3.15 USING THE HILBERT TRANSFORM

In the cases where $|X|$ is hard to compute by integration, particularly with Lévy Stable distributions that do not allow no explicit densities, we can make use of the Hilbert Transform to extract the expected mean deviations.

$$\mathcal{H}(f) = \mathcal{F}^{-1}(-i\ \mathrm{sgn}(\cdot) \cdot \mathcal{F}(f)),$$

where

$$[\mathcal{H}(f)](x) \stackrel{\text{def}}{=} \text{p.v.} \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(t)}{x-t}\, dx.$$

p.v. means principal value in the Cauchy sense, so

$$\text{p.v.} \int_{-\infty}^{\infty} = \lim_{a\to\infty} \lim_{b\to 0} \int_{-a}^{-b} + \int_{b}^{a}$$

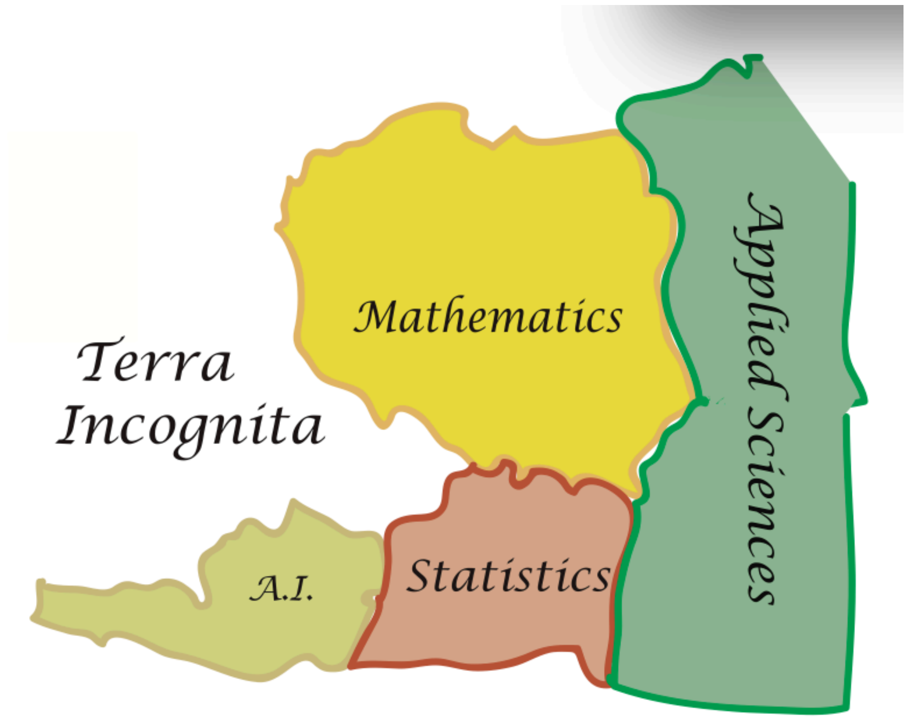


Figure 3.26: **Terra Incognita**: Brad Efron's positioning of the unknown that is certainly out of reach for any type of knowledge, which includes Bayesian inference.(Efron, via Susan Holmes)

3.16 A COMMENT ON BAYESIAN METHODS IN RISK MANAGEMENT

[This section will be developed further; how the statemennt "but this is my prior" can be nonsense with risk management if such a prior is not solid.]

Brad Efron (2013)[28]:

> Sorry. My own practice is to use Bayesian analysis in the presence of genuine prior information; to use empirical Bayes methods in the parallel cases situation; and otherwise to be cautious when invoking uninformative priors. In the last case, Bayesian calculations cannot be uncritically accepted and should be checked by other methods, which usually means frequentistically.

Diaconis and Friedman [24] show the difficulty for an agent to formulate a prior.

Further Reading

Pitman [89], Embrechts and Goldie (1982)[32]Embrechts (1979 Doctoral thesis?)[33], Chistyakov (1964) [19], Goldie (1978)[52], Pitman[89], Teugels [113], and, more general, [34].

B SPECIAL CASES OF FAT TAILS

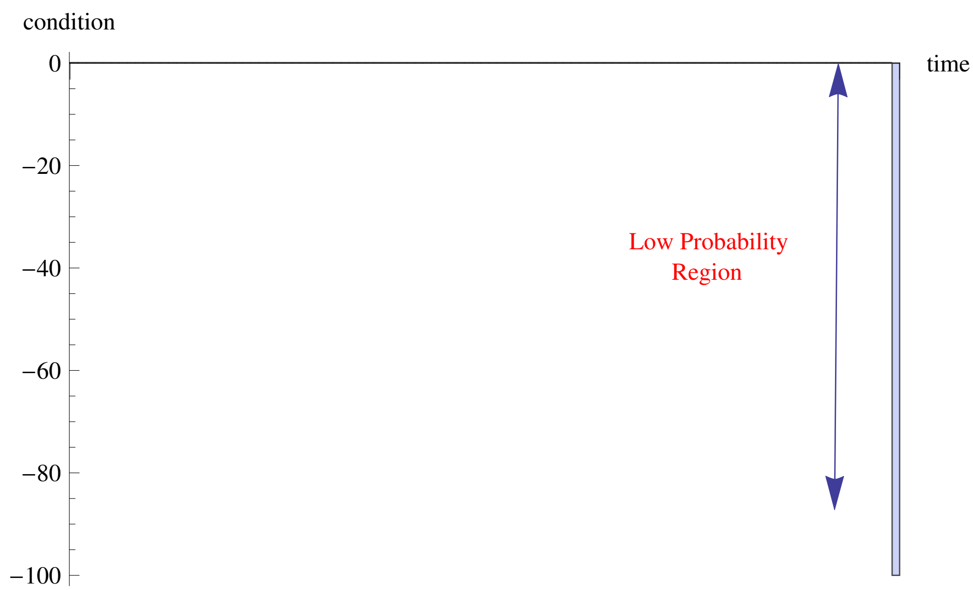


Figure B.1: The coffee cup is less likely to incur "small" than large harm; it is exposed to (almost) everything or nothing.

For monomodal distributions, fat tails are the norm: one can look at tens of thousands of time series of the socio-economic variables without encountering a single episode of "platykurtic" distributions. But for multimodal distributions, some surprises can occur.

B.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" (as measured by the fourth moment) would drop. And if we associate different variances with different means, we get a variety of "regimes", each with its set of probabilities.

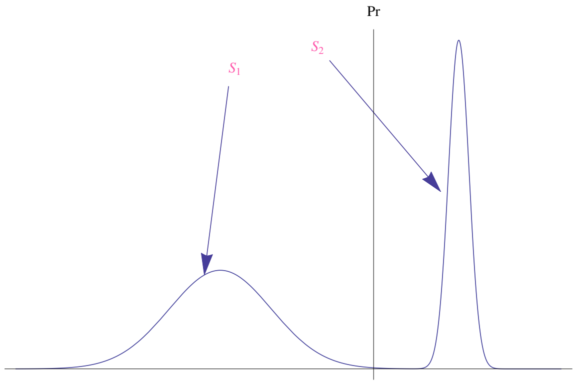


Figure B.2: The War and peace model. Kurtosis K=1.7, much lower than the Gaussian.

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.[6, 72].

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 becomes:

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

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

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

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

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

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

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

Variety 1: War and Peace. Calm period with positive mean and very low volatility, turmoil with negative mean and extremely low volatility.

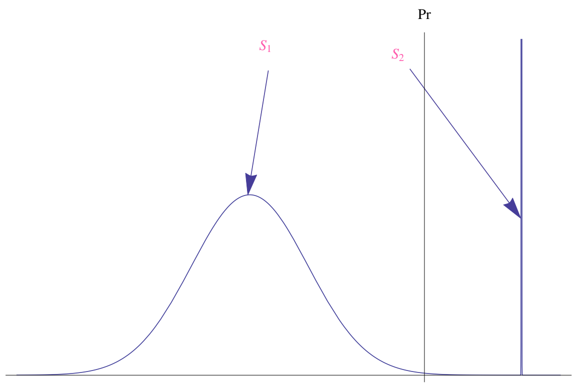


Figure B.3: The Bond payoff model. Absence of volatility, deterministic payoff in regime 2, mayhem in regime 1. Here the kurtosis K=2.5. Note that the coffee cup is a special case of both regimes 1 and 2 being degenerate.

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

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 see in figure B.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

B.2 TRANSITION PROBABILITES: WHAT CAN BREAK WILL BREAK

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

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

After n periods, that is, n steps,

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

Where

$$\begin{aligned}
a_n &= \frac{(p_{1,1}-1)\,(p_{1,1}+p_{2,2}-1)^n + p_{2,2}-1}{p_{1,1}+p_{2,2}-2} \\
b_n &= \frac{(1-p_{1,1})\,((p_{1,1}+p_{2,2}-1)^n - 1)}{p_{1,1}+p_{2,2}-2} \\
c_n &= \frac{(1-p_{2,2})\,((p_{1,1}+p_{2,2}-1)^n - 1)}{p_{1,1}+p_{2,2}-2} \\
d_n &= \frac{(p_{2,2}-1)\,(p_{1,1}+p_{2,2}-1)^n + p_{1,1}-1}{p_{1,1}+p_{2,2}-2}
\end{aligned}$$

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 $\pi = \begin{pmatrix} 1 & 1 \\ \frac{1-p_{1,1}}{1-p_{2,2}} & 1 \end{pmatrix}$.

C QUICK AND ROBUST MEASURE OF FAT TAILS

C.1 INTRODUCTION

We propose a new measure of fatness of tails. We also propose a quick heuristic to extract the tail exponent α and get distributions for a symmetric power law distributed variable. It is based on using whatever moments are believed to be reasonably finite, and replaces kurtosis which in financial data has proved to be unbearingly unstable ([109], [?]). The technique also remedies some of the instability of the Hill estimator, along with its natural tradoff between how much data one must discard in otder to retain in the tails that is relevant to draw the slope. Our estimators use the entire data available. This paper covers two situations:

1. Mild fat tails: a symmetric distribution with finite second moment, $\alpha > 2$, preferably in the neighborhood of 3. (Above 4 the measure of kurtosis becomes applicable again).
2. Extremely fat tails: a symmetric distribution with finite first moment, $1 < \alpha < 3$.

Let x be a r.v. on the real line. Let x be distributed according to a Student T distribution.

$$p(x) = \frac{\left(\frac{\alpha}{\alpha + \frac{(x-\mu)^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\,\sigma B\left(\frac{\alpha}{2}, \frac{1}{2}\right)} \tag{C.1}$$

We assume that $\mu = 0$ for data in high enough frequency as the mean will not have an effect on the estimation tail exponent.

C.2 FIRST METRIC, THE SIMPLE ESTIMATOR

Assume finite variance and the tail exponent $\alpha > 2$.

Define the ratio $\Xi(\alpha)$ as $\frac{\sqrt{\mathbb{E}(x^2)}}{\mathbb{E}(|x|)}$.

$$\Xi(\alpha) = \frac{\sqrt{\int_{-\infty}^{\infty} \frac{x^2 \left(\frac{\alpha}{\alpha+\frac{x^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha} B\left(\frac{\alpha}{2},\frac{1}{2}\right)} \, dx}}{\int_{-\infty}^{\infty} \frac{|x| \left(\frac{\alpha}{\alpha+\frac{x^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\ B\left(\frac{\alpha}{2},\frac{1}{2}\right)} \, dx} = \frac{\sqrt{\pi}\sqrt{\frac{\alpha}{\alpha-2}}\ \Gamma\left(\frac{\alpha}{2}\right)}{\sqrt{\alpha}\ \Gamma\left(\frac{\alpha-1}{2}\right)} \tag{C.2}$$

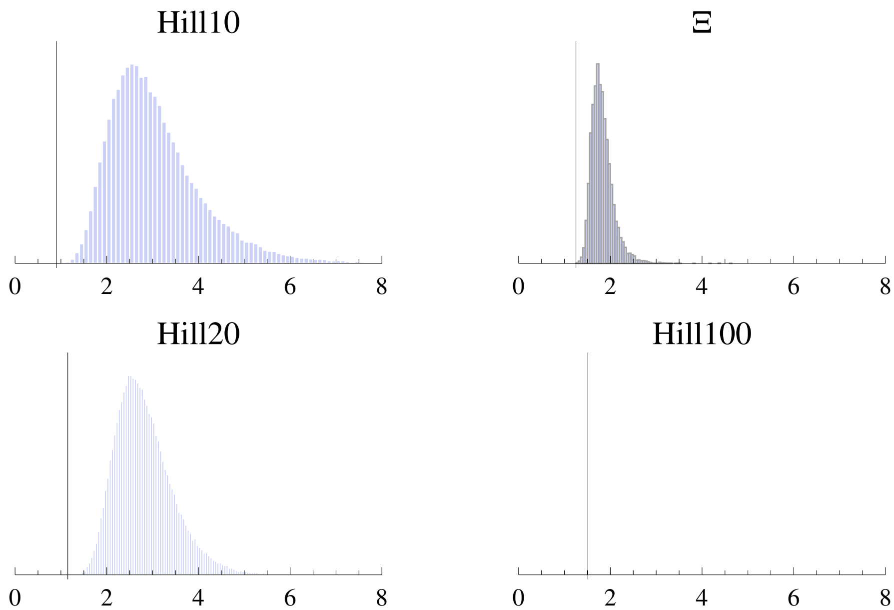


Figure C.1: Full Distribution of the estimators for $\alpha = 3$

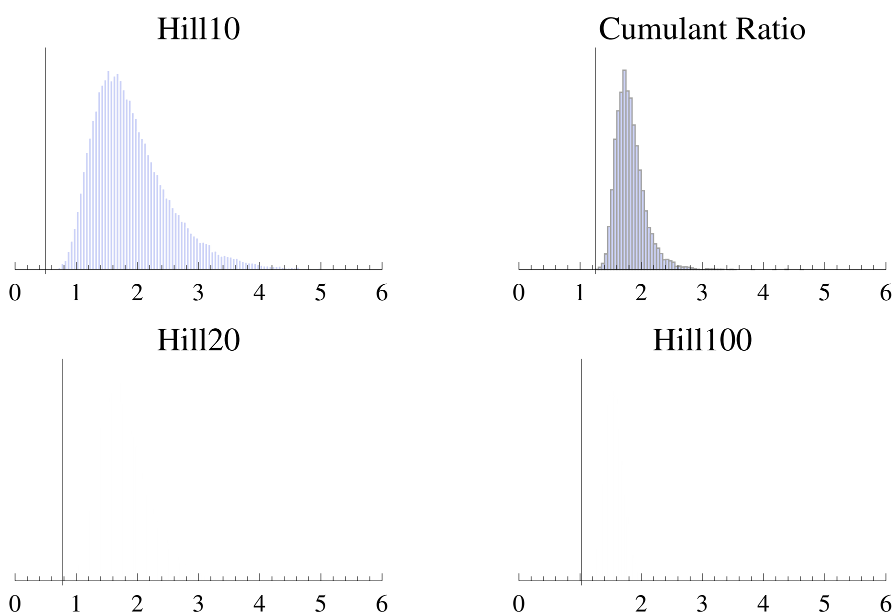


Figure C.2: Full Distribution of the estimators for $\alpha = 7/4$

The tail from the observations: Consider a random sample of size n, $(X_i)_{1 \leq i \leq n}$. Get a sample metric

Where STD and MAD are the sample standard and mean absolute deviations.

$$m = \frac{STD}{MAD}$$

for the sample (these measures do not necessarily need to be central). The estimation of m using maximum likelihood methods [FILL]

The recovered tail α_Ξ.

$$\alpha_\Xi = \Xi^{-1}(m) = \{\alpha : \Xi(\alpha) = m\}$$

which is computed numerically.

The H_m corresponds to the measure of the m largest deviation in the right tails= (a negative value for m means it is the left tail). We rank $X_{(1)} \geq X_{(2)} \geq ... \geq X_{(m)} \geq ... \geq X_{(n)}$. The Hill estimator

$$H_m = \left(\frac{\sum_{i=1}^{m} \log\left(\frac{X_i}{X_{m+1}}\right)}{m} \right)^{-1}$$

Table 8: Simulation for true $\alpha = 3$, $N = 1000$

Method	Estimate	STD Error
H_{10}	3.09681	1.06873
H_{20}	2.82439	0.639901
H_{50}	2.4879	0.334652
H_{100}	2.14297	0.196846
α^*_Ξ	3.26668	0.422277

C.3 SECOND METRIC, THE Ξ_2 ESTIMATOR

$$\Xi_2(\alpha) = \frac{\mathbb{E}(|x - E|x||)}{\mathbb{E}(|x|)}$$

$$\begin{aligned}\Xi_2(\alpha) = \left((\alpha-1)B\left(\frac{\alpha}{2},\frac{1}{2}\right)\right)^{\alpha-1}\Bigg(\left((\alpha-1)^2 B\left(\frac{\alpha}{2},\frac{1}{2}\right)^2+4\right)^{\frac{1-\alpha}{2}} - \\ \frac{2^{-\alpha}(\alpha-1)\,{}_2F_1\left(\frac{\alpha}{2},\frac{\alpha+1}{2};\frac{\alpha+2}{2};-\frac{1}{4}(\alpha-1)^2 B\left(\frac{\alpha}{2},\frac{1}{2}\right)^2\right)}{\alpha} \\ + \frac{2\,{}_2F_1\left(\frac{1}{2},\frac{\alpha+1}{2};\frac{3}{2};-\frac{4}{(\alpha-1)^2 B\left(\frac{\alpha}{2},\frac{1}{2}\right)^2}\right)}{(\alpha-1)B\left(\frac{\alpha}{2},\frac{1}{2}\right)^2}\Bigg) + \frac{1}{2}\end{aligned} \tag{C.3}$$

$$m' = \frac{1}{n}\frac{\sum_{i=1}^{n}|X_i - MAD|}{MAD}$$

Table 9: Simulation for true $\alpha = 7/4$, $N = 1000$

Method	Estimate	STD Error
H_{10}	1.92504	0.677026
H_{20}	1.80589	0.423783
H_{50}	1.68919	0.237579
H_{100}	1.56134	0.149595
$\alpha^*_{\Xi_2}$	1.8231	0.243436

4 HIERARCHY OF DISTRIBUTIONS FOR ASYMMETRIES

Chapter Summary **4: Using the asymptotic Radon-Nikodym derivatives of probability measures, we construct a formal methodology to avoid the "masquerade problem" namely that standard "empirical" tests are not empirical at all and can be fooled by fat tails, though not by thin tails, as a fat tailed distribution (which requires a lot more data) can masquerade as a low-risk one, but not the reverse. Remarkably this point is the statistical version of the logical asymmetry between *evidence of absence* and *absence of evidence*. We put some refinement around the notion of "failure to reject", as it may misapply in some situations. We show how such tests as Kolmogorov Smirnoff, Anderson-Darling, Jarque-Bera, Mardia Kurtosis, and others can be gamed and how our ranking rectifies the problem.**

4.1 PERMISSIBLE EMPIRICAL STATEMENTS

One can make statements of the type "This is not Gaussian", or "this is not Poisson"(many people don't realize that Poisson distributions are generally thin tailed owing to finite moments); but one cannot rule out a Cauchy tail or other similar power laws. So this chapter puts some mathematical structure around the idea of which "empirical" statements are permissible in acceptance and rejection and which ones are not. (One can violate these statements but not from data analysis, only basing oneself on *a priori* statement of what belongins to some probability distributions.)[1,2]

Let us get deeper into the masquerade problem, as it concerns the problem of induction and fat-tailed environments, 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 standard 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 for sure what is going

1 Classical statistical theory is based on rejection and failure to reject, which is inadequade as one can reject fat tails, for instance, which is not admissible here. Likewise this framework allows us to formally "accept" some statements.

2 This chapter was motivated by the findings in an article by Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." SIAM review 51.4 (2009): 661-703, deeming that wealth data "cannot plausibly be considered to follow a power law". The methodology they used is based on a class of "naive" power law fitting methods than ignore the properties of out-of-sample parts.

on. 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.

This chapter aims 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.

4.2 MASQUERADE EXAMPLE

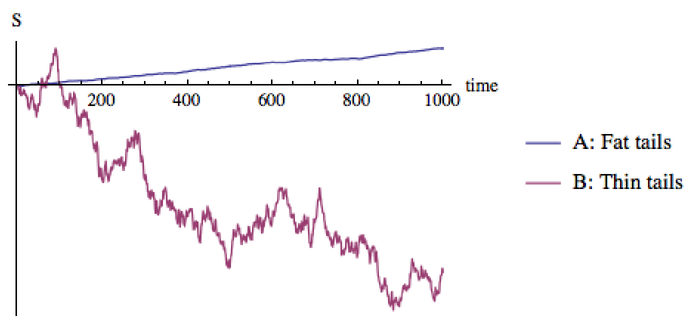


Figure 4.1: N=1000. Sample simulation. Both series have the exact same means and variances at the level of the generating process. Naive use of common metrics leads to the acceptance that the process A has thin tails.

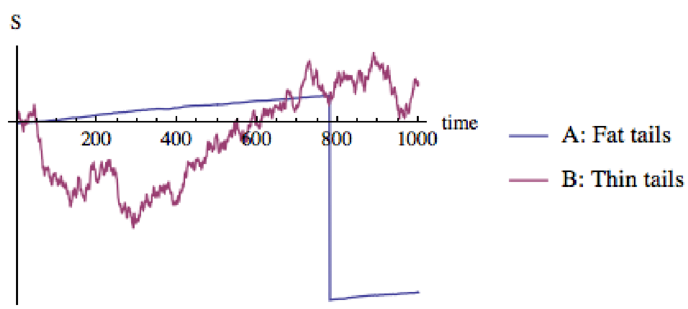


Figure 4.2: N=1000. **Rejection:** Another realization. there is 1/2 chance of seeing the real properties of A. We can now reject the hypothesis that the smoother process has thin tails.

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 $\kappa = \frac{3\left(\left(1-a^2\right)p-1\right)}{p-1}$ 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 $\kappa = 3$.

The two graphs in figures 4.1 and 4.2 show the realizations of the processes A (to repeat, produced with the switching process) and B, entirely Gaussian, both of the same variance.

4.3 THE PROBABILISTIC VERSION OF ABSENSE OF EVIDENCE

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 not, this pervades the entire literature.

Many 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 4.1 (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.[3]

4.4 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*[111], Glossary).

3 ab esse ad posse valet consequentia.

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).

4.5 HIERARCHY OF DISTRIBUTIONS IN TERM OF TAILS

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. $A_K^- = (-\infty, K]$ or $A_K^+ = [K, \infty)$, 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 4.1 (Acceptance and Rejection)**.**
We can define i) "right tail 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 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 $K > K_0$, and left tail acceptance if there exists a K_0 ("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 $K < K_0$.

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 quote Adrien Douady, it has compact support for all practical purposes.[4]

Let us focus on the right tail.

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 4.1.*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 d, the difference between the two densities,

$$d = \frac{\frac{e^{-\frac{\log^2(x)}{2k^2s^2}}}{ksx} - \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma}}{\sqrt{2\pi}}$$

by comparing the unscaled tail values of $\frac{e^{-\frac{\log^2(x)}{2k^2s^2}}}{ksx}$ and $\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma}$. Taking logarithms of the ratio, $\delta(x) = \frac{x^2}{2\sigma^2} - \frac{\log^2(x)}{2k^2s^2} - \log(ksx) + \log(\sigma)$, which is dominated by the first

4 Van Zwet,[cite]: Given two cumulative distribution functions $F(x)$ and $G(x)$, F has lighter tails than G (and G has heavier tails than F) if the function $G^{-1}(F(x))$ is convex for $x \geq 0$.

Table 10: Ranking distributions

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

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 4.2. *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 4.3. *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: [SEE CLASSIFICATION IN EMBRECHTS & ALL FOR COMPARISON]

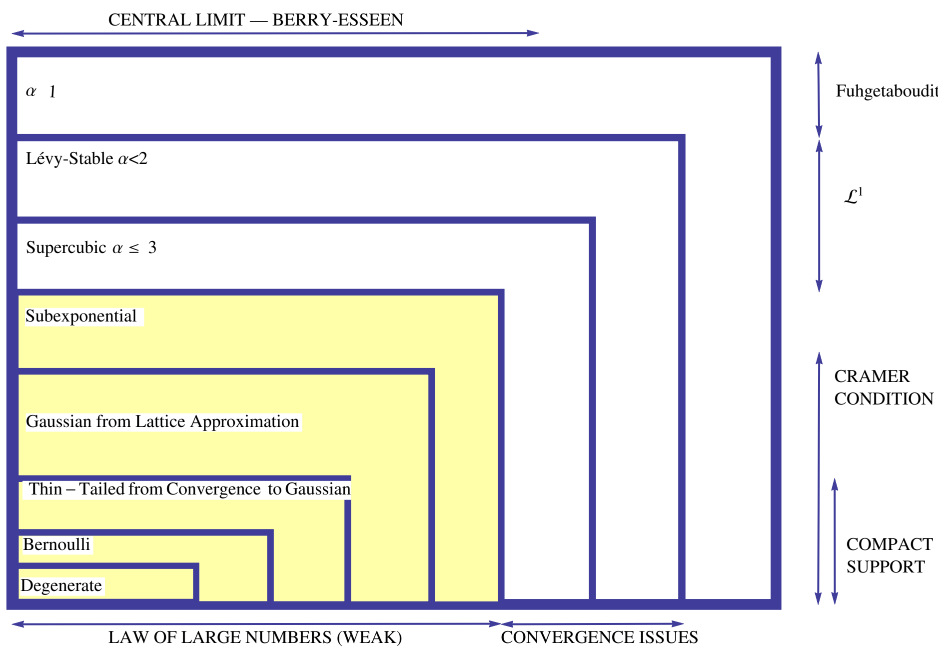


Figure 4.3: The tableau of Fat tails, along the various classifications for convergence purposes (i.e., convergence to the law of large numbers, etc.)A variation around Embrechts et al [31], but applied to the Radon-Nikodym derivatives.

A comment on 4.3

Gaussian From Convergence is Not Gaussian : We establish a demarcation between two levels of Gaussians. Adding Bernouilli variables or Binomials, according to the random walk idea (or similar mechanism that generate Gaussians) *always* leads to thinner tails to the true Gaussian.

Subgaussian domain for a review,[17], Kahane's "gaussian shift"[5]:

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 volatilities discussed in Chapter 6. Now a few rapidly stated rules.

Rule 4.4.*(General Decision Making Heuristic).* *For any information entailing nonbinary decision (see definition in Chapter x), 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*

5 J.P. Kahane, "Local properties of functions interms of random Fourier series," Stud. Math., 19, No. i, 1-25 (1960)

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.

Warning 1 : Always remember that one does not observe probability distributions, only realizations. Every probabilistic statement needs to be discounted by the probability of the parameter being away from the true one.

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

Rule 4.5.*(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 (from 3.5), there in increase in observations in the "tunnel"(a_2, a_3) in response to increase in fat-tailedness.

4.6 HOW TO ARBITRAGE KOLMOGOROV-SMIRNOV

Counterintuitively, when one raises the kurtosis, as in Figure 4.1.4.1 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,

Table 11: Comparing the Fake and genuine Gaussians (Figure 4.1.4.1) and subjecting them to a battery of tests. Note that some tests, such as the Jarque-Bera test, are more relevant to fat tails as they include the payoffs.

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).

		Statistic	P-Value
Fake Distr	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

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

Table 12: Table of the "fake" Gaussian when busted. 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.

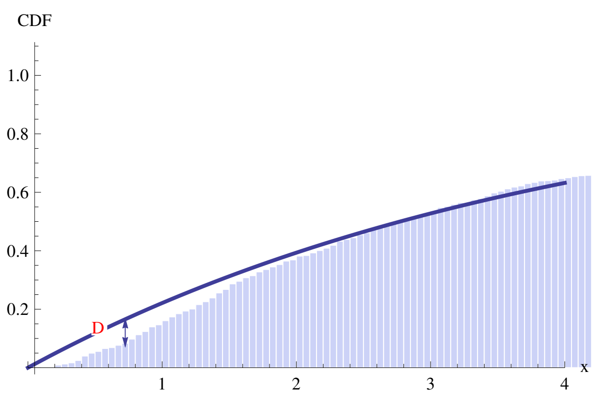


Figure 4.4: **The Kolmogorov-Smirnov Gap.** D is the measure of the largest absolute divergence between the candidate and the target distribution.

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:

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

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

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), $D_j = \left| \left(\frac{\sum_{i=1}^{J} X_{\Delta t i+t_0}}{j} - F\left(X_{\Delta t j+t_0}\right) \right) \right|$ 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.

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):

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

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.

4.7 MISTAKING EVIDENCE FOR ANECDOTES & THE REVERSE

Now some sad, very sad comments.

[MOVE TO CHAPTER ON SOCIAL SCIENCE] 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

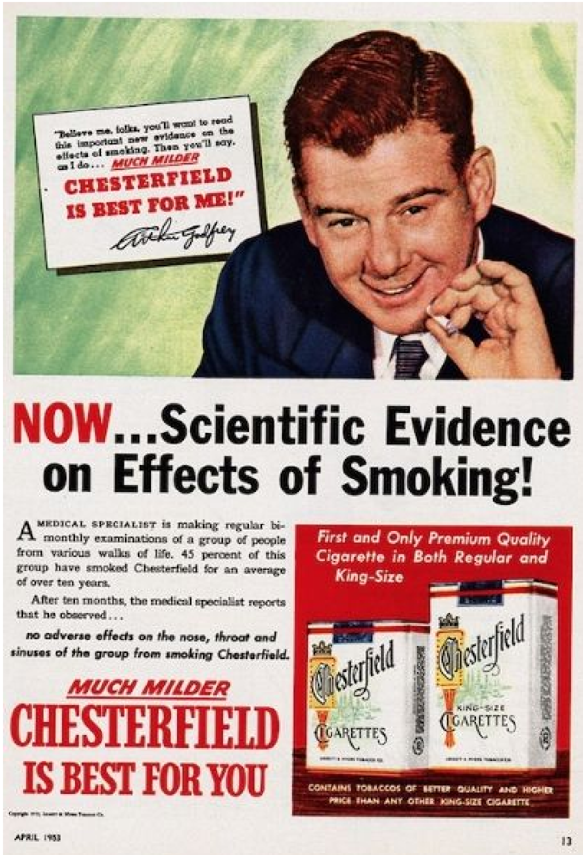


Figure 4.5: The good news is that we know exactly what not to call "evidence" in complex domains where one goes counter to the principle of "nature as a LLN statistician".

"scientists" 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". 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*.

The Good News

There are domains where "confirmatory evidence" works, or can be used for decisions. But for that one needs the LLN to operate rather quickly. The idea of "scientific evidence" in fat tailed domains leads to pathologies: it may work "for knowledge" and some limited applications, but not when it comes to risky decisions.

Further Reading

Doob (1949) [27].

5 EFFECTS OF HIGHER ORDERS OF UNCERTAINTY

Chapter Summary **5: The Spectrum Between Uncertainty and Risk. There has been a bit of discussions about the distinction between "uncertainty" and "risk". We believe in gradation of uncertainty at the level of the probability distribution itself (a "meta" or higher order of uncertainty.) One end of the spectrum, "Knightian risk", is not available for us mortals in the real world. We show how the effect on fat tails and on the calibration of tail exponents and reveal inconsistencies in models such as Markowitz or those used for intertemporal discounting (as many violations of "rationality" aren't violations .**

5.1 META-PROBABILITY DISTRIBUTION

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 "uncertainty in the Knightian sense" by decision theorisrs(Knight, 1923). "Risk" is when the probabilities are computable without an error rate. Such an animal does not exist in the real world. The entire distinction is a lunacy, since no parameter should be rationally computed witout an error rate. We find it preferable to talk about degrees of uncertainty about risk/uncertainty, using metadistribution, or metaprobability.

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 social science, economics (and elsewhere) take, as input, an average or expected parameter,

$$\bar{\alpha} = \int \alpha \; \phi(\alpha) \; d\alpha, \tag{5.1}$$

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

$$p(x) = \int \phi(\alpha) \, \mathrm{d}\alpha \tag{5.2}$$

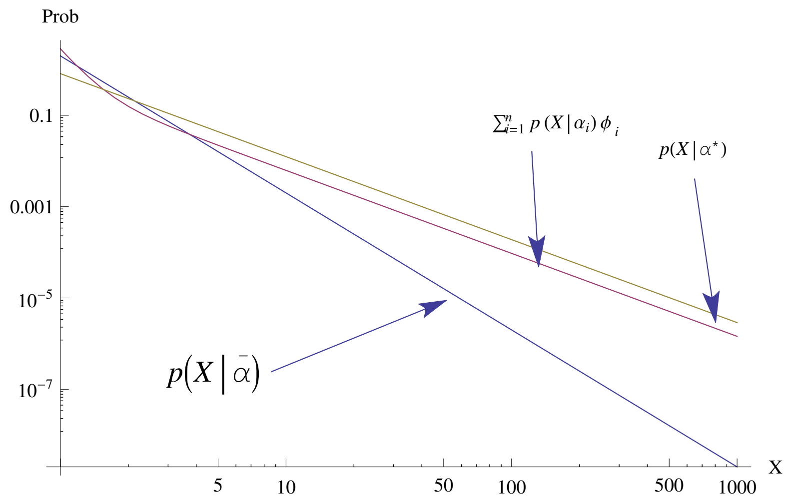


Figure 5.1: Log-log plot illustration of the asymptotic tail exponent with two states.

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

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

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

$$\omega_{A'} = \int \left(\int \phi(\alpha)\ f(x)\ p(x|\alpha)\ \ d\alpha \right) \mathrm{d}x - \int f(x)\ p\left(x|\int \alpha\ \phi(\alpha)\, \mathrm{d}\alpha\right) \mathrm{d}x \quad (5.3)$$

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 5.2 becomes

$$p(x) = \phi_i \left(\sum_{i=1}^{n} p\ (x\ |\alpha_i\ ) \right) \quad (5.4)$$

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

5.2 METADISTRIBUTION AND THE CALIBRATION OF POWER LAWS

Remark 5.1.
In the presence of a layer of metadistributions (from uncertainty about the parameters), the asymptotic tail exponent for a powerlaw corresponds to the lowest possible tail exponent regardless of its probability.

This 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 $\alpha^* = \underset{1 \le i \le n}{min}\, \alpha_i$. 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.

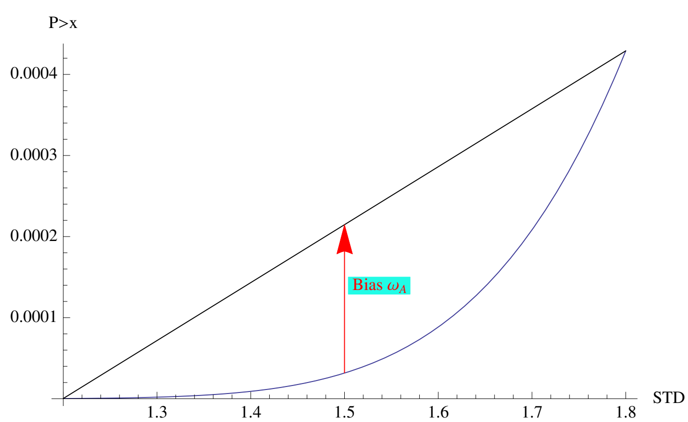


Figure 5.2: Illustration of the convexity bias for a Gaussian from raising small probabilities: The plot shows the STD effect on P>x, and compares P>6 with a STD of 1.5 compared to P> 6 assuming a linear combination of 1.2 and 1.8 (here a(1)=1/5).

Figure 5.1 shows the different situations: a) $p(x|\bar{\alpha})$, b) $\sum_{i=1}^{n} p\left(x\left|\alpha_i\right.\right)\phi_i$ and c) $p\left(x\left|\alpha^*\right.\right)$. We can see how the last two converge. The asymptotic Jensen Gap ω_A becomes $p(x|\alpha^*) - p(x|\bar{\alpha})$.

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")[121].

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.

5.3 THE EFFECT OF METAPROBABILITY ON FAT TAILS

Recall that the tail fattening methods in 3.4 and 3.6.These are based on randomizing the variance. Small probabilities rise precisely because they are convex to perturbations of the parameters (the scale) of the probability distribution.

5.4 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 Tohoku 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

5.5 THE MARKOWITZ INCONSISTENCY

Assume that someone tells you that the probability of an event is exactly zero. You ask him where he got this from. "Baal told me" is the answer. In such case, the person is coherent, but would be deemed unrealistic by non-Baalists. But if on the other hand, the person tells you "I estimated it to be zero," we have a problem. The person is both unrealistic and inconsistent. Something estimated needs to have an estimation error. So probability cannot be zero if it is estimated, its lower bound is linked to the estimation error; the higher the estimation error, the higher the probability, up to a point. As with Laplace's argument of total ignorance, an infinite estimation error pushes the probability toward $\frac{1}{2}$. We will return to the implication of the mistake; take for now that anything estimating a parameter and then putting it into an equation is different from estimating the equation across parameters. And Markowitz was inconsistent by starting his "seminal" paper with "Assume you know E and V" (that is, the expectation and the variance). At the end of the paper he accepts that they need to be estimated, and what is worse, with a combination of statistical techniques and the "judgment of practical men." Well, if these parameters need to be estimated, with an error, then the derivations need to be written differently and, of course, we would have no such model. Economic models are extremely fragilefragile to assumptions, in the sense that a slight alteration in these assumptions can lead to extremely consequential differences in the results. The perturbations can be seen as follows. Let $\overrightarrow{X} = (X_1, X_2, \ldots, X_m)$ be the vector of random variables representing returns. Consider the joint probability distribution $f(x_1, \ldots, x_m)$. We denote the m-variate multivariate Normal distribution by $N(\overrightarrow{\mu}, \Sigma)$, with mean vector $\overrightarrow{\mu}$, variance-covariance matrix Σ, and joint pdf,

$$f\left(\overrightarrow{x}\right) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\left(\overrightarrow{x}-\overrightarrow{\mu}\right)^T \Sigma^{-1}\left(\overrightarrow{x}-\overrightarrow{\mu}\right)\right) \tag{5.5}$$

where $\overrightarrow{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, and Σ is a symmetric, positive definite $(m \times m)$ matrix. The weights matrix $\overrightarrow{\Omega} = (\omega_1, \ldots, \omega_m)$,normalized, with $\sum_{i=1}^{N} \omega_i = 1$ (allowing exposures to be both positive and negative): The scalar of concern is; $r = \Omega^T.X$, which happens to be normally distributed, with variance

$$v = \vec{\omega}^T.\Sigma.\vec{\omega}$$

The Markowitz portfolio construction, through simple optimization, gets an optimal $\vec{\omega}^*$, obtained by, say, minimizing variance under constraints, getting the smallest $\vec{\omega}^T.\Sigma.\vec{\omega}$ under constraints of returns, a standard Lagrange multiplier. So done statically, the problem gives a certain result that misses the metadistribution. Now the problem is that the covariance matrix is a random object, and needs to be treated as so. So let us focus on what can happen under these conditions:

Route 1: The stochastic volatility route This route is insufficient but can reveal structural defects for the construction. We can apply the same simplied variance preserving heuristic as in 3.4 to fatten the tails. Where a is a scalar that determines the intensity of stochastic volatility, $\Sigma_1 = \Sigma(1 + a)$ and $\Sigma_2 = \Sigma(1 - a)$. Simply, given the conservation of the Gaussian distribution under weighted summation, maps to

$v(1+a)$ and $v(1-a)$ for a Gaussian and we could see the same effect as in 3.4. The corresponding increase in fragility is explained in Chapter 17.

Route 2: Full random parameters route Now one can have a fully random matrix —not just the overal level of the covariance matrix. The problem is working with matrices is cumbersome, particularly in higher dimensions, because one element of the covariance can vary unconstrained, but the degrees of freedom are now reduced for the matrix to remain positive definite. A possible technique is to extract the principal components, necessarily orthogonal, and randomize them without such restrictions.

5.6 PSYCHOLOGICAL PSEUDO-BIASES UNDER SECOND LAYER OF UNCERTAINTY.

Often psychologists and behavioral economists find "irrational behavior" (or call it under something more polite like "biased") as agents do not appear to follow a normative model and violate their models of rationality. But almost all these correspond to missing a second layer of uncertainty by a dinky-toy first-order model that doesn't get nonlinearities $-$ it is the researcher who is making a mistake, not the real-world agent. Recall that the expansion from "small world" to "larger world" can be simulated by perturbation of parameters, or "stochasticization", that is making something that appears deterministic a random variable itself. Benartzi and Thaler [8], for instance, find an explanation that agents are victims of a disease labelled "myopic loss aversion" in not investing enough in equities, not realizing that these agents may have a more complex, fat-tailed model. Under fat tails, no such puzzle exists, and if it does, it is certainly not from such myopia.

This approach invites "paternalism" in "nudging" the preferences of agents in a manner to fit professors-without-skin-in-the-game-using-wrong-models.

5.6.1 The pathologization fallacy

Today many mathematical or conceptual models that are claimed to be rigorous are based upon fixed parameters that miss a layer of uncertainty. Such models are deemed *rational* in the sense that they are logically derived from their assumptions, and, further, can be used to assess rationality by examining deviations from such models, as indicators of irrationality. Except that it is often the modeler who is using an incomplete representation of the reality, hence using an erroneous benchmark for rationality. Often the modelers are not familiar with the dynamics of complex systems or use antiquated statistical methods that do not take into account fat-tails and make inferences that would not be acceptable under different classes of probability distributions. Many biases, such as the ones used by Cass Sunstein, about the overestimation of the probabilities of rare events in fact correspond to the testers using a bad probability model that is thin-tailed.

It has became popular to claim irrationality for GMO and other skepticism on the part of the general public—not realizing that there is in fact an "expert problem" and such skepticism is healthy and even necessary for survival. For instance, in *The Rational Animal* **[?]**, the authors pathologize people for not accepting GMOs although "the World Health Organization has never found evidence of ill effects," a standard confusion of evidence of absence and absence of evidence. Such pathologizing is similar to behavioral researchers labeling hyperbolic discounting as "irrational" when in fact it is largely the researcher who has a very narrow model and richer models make the "irrationality" go away, as we will see further down.

These researchers fail to understand that humans may have precautionary principles against systemic risks, and can be skeptical of the untested consequences of policies for deeply rational reasons, even if they do not express such fears in academic format.

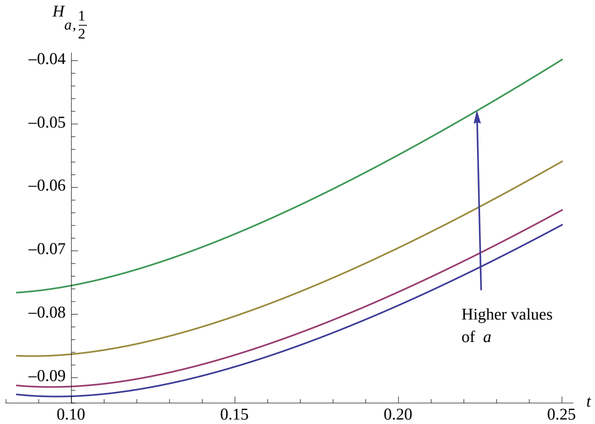


Figure 5.3: The effect of $H_{a,p}(t)$ "utility" or prospect theory of under second order effect on variance. Here $\sigma = 1, \mu = 1$ and t variable.

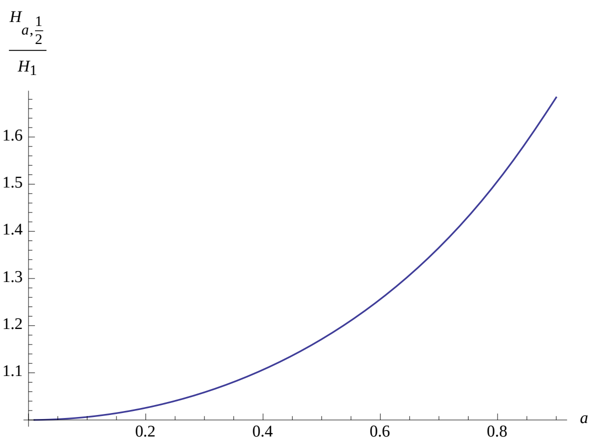


Figure 5.4: The ratio $\frac{H_{a,\frac{1}{2}}(t)}{H_0}$ or the degradation of "utility" under second order effects.

Hidden convexities Let us use our approach in detecting convexity to three specific problems: 1) the myopic loss aversion that we just discussed, 2) time preferences, 3) probability matching.

Myopic loss aversion

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

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

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

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

$$H_0(t) = \int_{-\infty}^{\infty} w_{\lambda,\alpha}(x)\phi_{\mu t,\sigma\sqrt{t}}(x)\, \mathrm{d}x \tag{5.6}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{\pi}} 2^{\frac{\alpha}{2}-2} \left(\frac{1}{\sigma^2 t}\right)^{-\frac{\alpha}{2}} \\
&\left(\Gamma\left(\frac{\alpha+1}{2}\right)\left(\sigma^\alpha t^{\alpha/2}\left(\frac{1}{\sigma^2 t}\right)^{\alpha/2} - \lambda\sigma\sqrt{t}\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(-\frac{\alpha}{2};\frac{1}{2};-\frac{t\mu^2}{2\sigma^2}\right)\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad + \frac{1}{\sqrt{2}\sigma}\mu\Gamma\left(\frac{\alpha}{2}+1\right) \\
&\left.\left(\sigma^{\alpha+1} t^{\frac{\alpha}{2}+1}\left(\frac{1}{\sigma^2 t}\right)^{\frac{\alpha+1}{2}} + \sigma^\alpha t^{\frac{\alpha+1}{2}}\left(\frac{1}{\sigma^2 t}\right)^{\alpha/2} + 2\lambda\sigma t\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(\frac{1-\alpha}{2};\frac{3}{2};-\frac{t\mu^2}{2\sigma^2}\right)\right)
\end{aligned} \tag{5.7}
$$

We can see from 5.7 that the more frequent sampling of the performance translates into worse utility. So what Benartzi and Thaler did was try to find the sampling period "myopia" that translates into the sampling frequency that causes the "premium" —the error being that they missed second order effects.

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

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

Rewriting 5.6

$$
H_{a,p}(t) = \int_{-\infty}^{\infty} w_{\lambda,\alpha}(x)\left(p\,\phi_{\mu\,t,\sqrt{1+a}\,\sigma\sqrt{t}}(x) + (1-p)\,\phi_{\mu\,t,\sqrt{1-a}\,\sigma\sqrt{t}}(x)\right)\mathrm{d}x \tag{5.8}
$$

Result Conclusively, as can be seen in figures 5.3 and 5.4, second order effects cancel the statements made from "myopic" loss aversion. This doesn't mean that myopia doesn't have effects, rather that it cannot explain the "equity premium", not from the outside (i.e. the distribution might have different returns", but from the inside, owing to the structure of the Kahneman-Tversky value function $v(x)$.

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

Rewriting 5.8 under that form,

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

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

Time preference under model error

This author once watched with a great deal of horror one Laibson [67] at a conference in Columbia University present the idea that having one massage today to two tomorrow, but reversing in a year from now is irrational and we need to remedy it with some policy. (For a review of time discounting and intertemporal preferences, see [43], as economists temps to impart what seems to be a varying "discount rate" in a simplified model).[1]

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

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

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

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

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

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

It is easy to prove the general case that under symmetric stochasticization of intensity $\Delta\lambda$ (that is, with probabilities $\frac{1}{2}$ around the center of the distribution) using the same technique we did in 3.4:

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

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

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

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

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

so

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

1 I discovered that [38] Farmer and Geanakoplos have applied a similar approach to Hyperbolic discounting

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

Further, fudging the distribution and normalizing it, when

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

we get the *normatively obtained* (not empirical pathology) so-called hyperbolic discounting:

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

6 LARGE NUMBERS AND CLT IN THE REAL WORLD

Chapter Summary **6: The Law of Large Numbers is the foundation of statistical knowledge –or, even (inductive) knowledge *tout court*. The behavior of the sum of random variables allows us to get to the asymptote and use handy asymptotic properties. However real life is more complicated. We cannot talk about LLN without figuring out the speed of convergence, which, when it is at $\sqrt{n}$, is only so asymptotically. Further, in some cases the LLN doesn't work at all. For very fat tailed, under the slightest parametric error, it will be more than 400 times slower than thought.**

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

Main Results In addition to explicit extractions of partial expectations for alpha stable distributions, one main result in this paper is the expression of how uncertainty about parameters (in terms of parameter volatility) translates into a larger (or smaller) required n. **Model Uncertainty** The practical import is that model uncertainty worsens inference, in a quantifiable way.

6.0.2 The "Pinker Problem"

It is also necessary to debunk a fallacy: we simply do not have enough data with commonly discussed fat-tailed processes to naively estimate a sum and make series of claims about stability of systems, pathology of people reacting to risks, etc. A surprising result: for the case with equivalent tails to the "Pareto 80/20 rule" (a tail exponent $\alpha = 1.16$) one needs 10^{11} more data than the Gaussian.

Take a certain sample size in the conventional Gaussian domain, say $n = 30$ or some other such heuristically used number. Assuming we are confortable with such a number of summands, how much larger (or smaller) n does one need for *the same error* under a different process? And how do we define errors in the absence of standard deviation which might not exist (power laws with exponents close to 2), or be too unreliable (power laws with exponents > 2, that is finite variance but infinite kurtosis).

It is strange that given the dominant role of fat tails nobody thought of calculating some practical equivalence table. How can people compare averages concerning street crime (very thin tailed) to casualties from war (very fat tailed) without some sample adjustment?[1]

1 **The Pinker Problem** A class of naive empiricism. It has been named so in reference to sloppy use of statistical techniques in social science and policy making, based on a theory promoted by the science

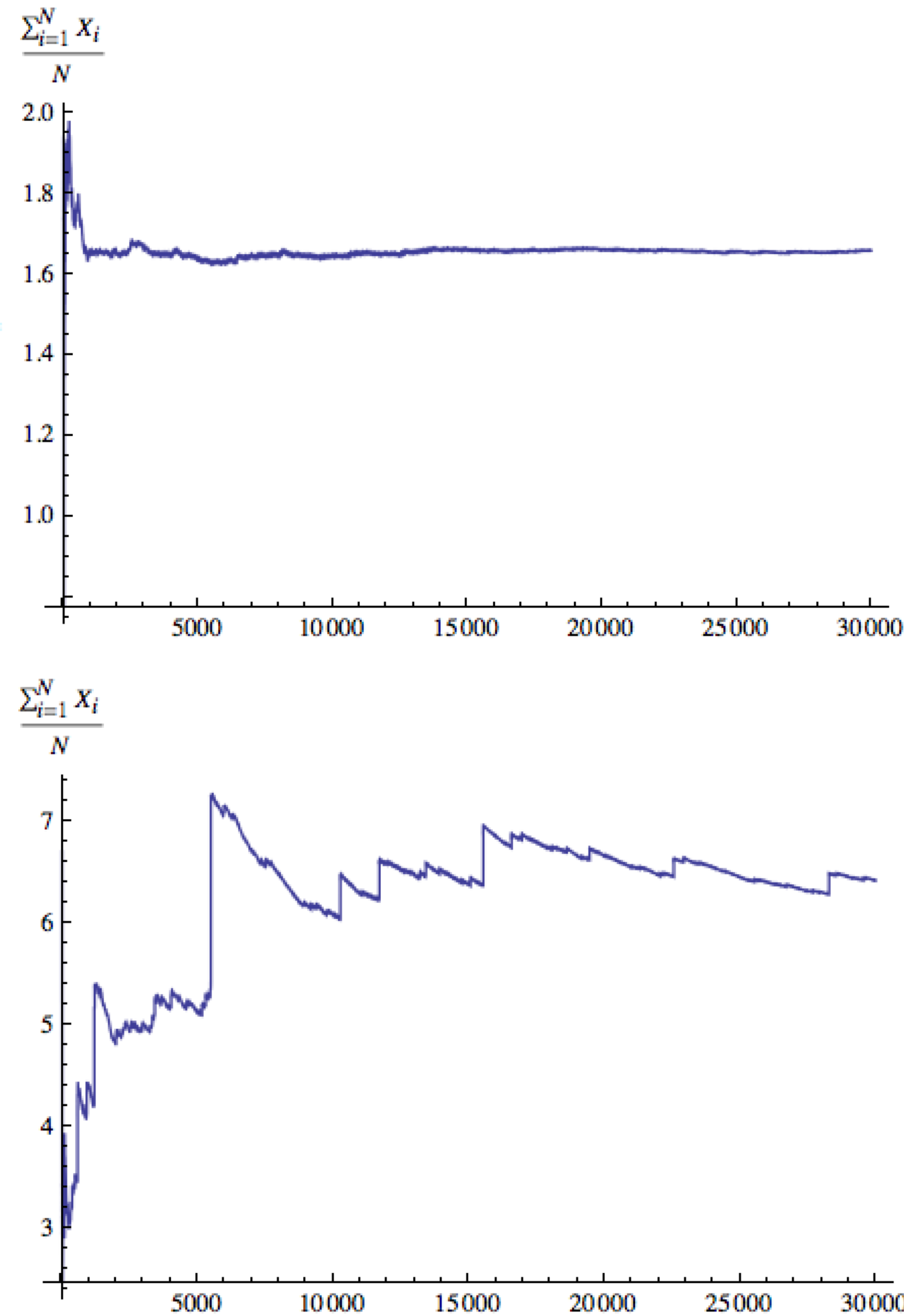


Figure 6.1: How thin tails (Gaussian) and fat tails (1< α ≤2) converge to the mean.

---

writer S. Pinker [88] about the drop of violence that is based on such statistical fallacies since wars –unlike domestic violence –are fat tailed. But this is a very general problem with the (irresponsible) mechanistic use of statistical methods in social science and biology.

Perhaps the problem lies at the core of the law of large numbers: the average is not as "visible" as other statistical dimentions; there is no sound statistical procedure to derive the properties of a powerlaw tailed data by estimating the mean – typically estimation is done by fitting the tail exponent (via, say, the Hill estimator or some other method), or dealing with extrema, yet it remains that many articles make comparisons about the mean since it is what descriptive statistics and, alas, decisions, are based on.

6.1 THE PROBLEM OF MATCHING ERRORS

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 $\to \infty$. And the idea is that we live with finite n.

We get most of the intuitions from closed-form and semi-closed form expressions working with:

- stable distributions (which allow for a broad span of fat tails by varying the α exponent, along with the asymmetry via the β coefficient
- stable distributions with mixed α exponent.
- other symmetric distributions with fat-tails (such as mixed Gaussians, Gamma-Variance Gaussians, or simple stochastic volatility)

More complicated situations entailing more numerical tinkering are also covered: Pareto classes, lognormal, etc.

Instability of Mean Deviation

Indexing with p the property of the variable X^p and g for X^g the Gaussian:

$$\left\{ n_p : \mathbb{E}\left(\left| \sum^{n_p} \frac{X_i^p - m_p}{n_p} \right| \right) = \mathbb{E}\left(\left| \sum^{n_g} \frac{X_i^g - m_g}{n_g} \right| \right) \right\} \tag{6.1}$$

And since we know that convergence for the Gaussian happens at speed $n^{\frac{1}{2}}$, we can compare to convergence of other classes.

We are expressing in Equation 6.1 the expected error (that is, a risk function) in $\mathbf{L}^1$ as mean absolute deviation from the observed average, to accommodate absence of variance –but assuming of course existence of first moment without which there is no point discussing averages.

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

Using mean deviations of the samples (when these exist) doesn't accommodate the fact that fat tailed data hide properties. The "volatility of volatility", or the dispersion around the mean deviation increases nonlinearly as the tails get fatter.

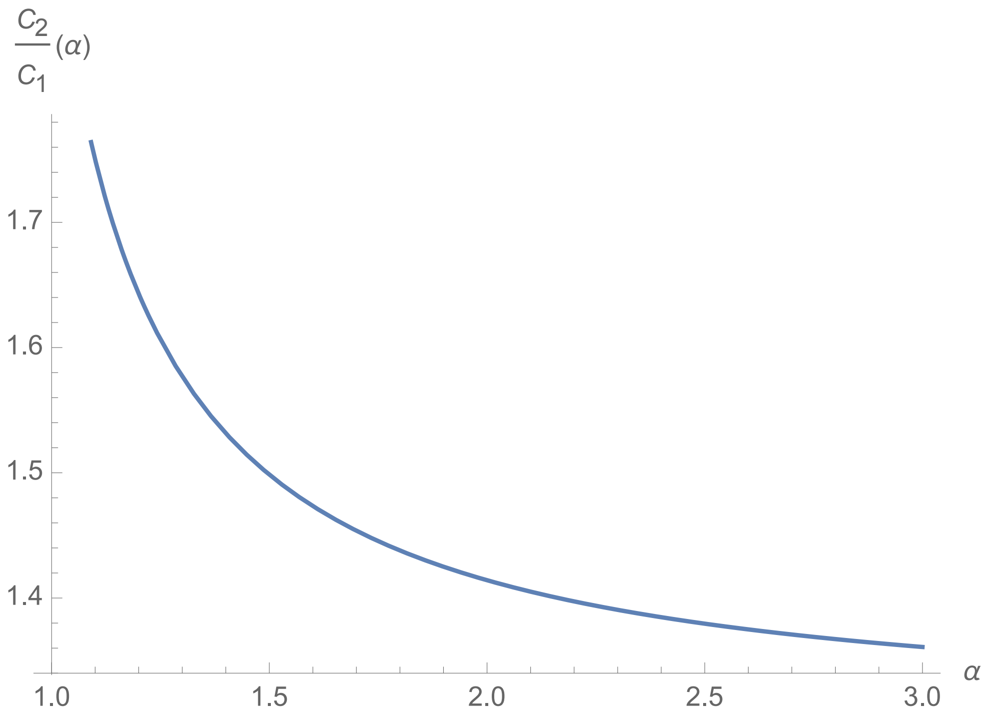


Figure 6.2: The ratio of cumulants for a symmetric powerlaw, as a function of the tail exponent.

For instance, a stable distribution with tail exponent at $\frac{3}{2}$ matched to exactly the same mean deviation as the Gaussian will deliver measurements of mean deviation 1.4 times as unstable as the Gaussian.

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

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

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

we have in the Gaussian case indexed by g:

$$C_2^g = \left(\operatorname{erf}(\frac{1}{\sqrt{\pi}} + e^{-1/\pi}\right) C_1^g$$

which is $\approx 1.30\, C_1^g$.

For a powerlaw distribution, cumulants are more unwieldy:

$$C_1^{\alpha=3/2} = \frac{2\sqrt{\frac{6}{\pi}}\Gamma\left(\frac{5}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}\sigma$$

Move to appendix

$$C_2^{\alpha=3/2} = \frac{1}{2\sqrt{6}\pi^{3/2}\Gamma_1^3\left(\pi\Gamma_1^2+\Gamma_3^2\right)^{5/4}}\sigma$$

$$\Big(384\pi^{5/4}\Gamma_2^3\Gamma_1^{5/2}+24\pi^{9/4}\Gamma_2\Gamma_1^{9/2}-2\pi^{9/4}\sqrt{\Gamma_2}\sqrt[4]{\pi\Gamma_1^2+\Gamma_3^2}\Gamma_1^{9/2}H_1$$
$$+1536\Gamma_2^5\sqrt[4]{\pi\Gamma_1^2+\Gamma_3^2}H_2+\pi^3\sqrt[4]{\pi\Gamma_1^2+\Gamma_3^2}\left(3\sqrt{2}\Gamma_1^3+3\Gamma_3\left(H_2+2\right)-2\sqrt[4]{2}\pi^{3/4}H_1\right)\Big)$$

where $\Gamma_1 = \Gamma\left(\frac{3}{4}\right)$, $\Gamma_2 = \Gamma\left(\frac{5}{4}\right)$, $\Gamma_3 = \Gamma\left(\frac{1}{4}\right)$, $H_1 = {}_2F_1\left(\frac{3}{4},\frac{5}{4};\frac{7}{4};-\frac{\pi\Gamma_1^2}{\Gamma_3^2}\right)$, and $H_2 = {}_2F_1\left(\frac{1}{2},\frac{5}{4};\frac{3}{2};-\frac{\Gamma_3^2}{\pi\Gamma_1^2}\right)$.

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

6.2 GENERALIZING MEAN DEVIATION AS PARTIAL EXPECTATION

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

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

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

And the special expectation becomes, by convoluting the Fourier transforms:

$$\mathbb{E}(X|_{X>K}) = -i\frac{\partial}{\partial t}\int_{-\infty}^{\infty}\Psi(t-u)\Psi^\theta(u)\mathrm{d}u|_{t=0} \tag{6.2}$$

Mean deviation becomes a special case of equation 6.2, $\mathbb{E}(|X|) = \mathbb{E}(X|_{X>\mu}) + \mathbb{E}(-X|_{X<\mu}) = 2\mathbb{E}(X|_{X>\mu})$.

6.3 CLASS OF STABLE DISTRIBUTIONS

Assume alpha-stable the class $\mathfrak{S}$ of probability distribution that is closed under convolution: $\mathbf{S}(\alpha,\beta,\mu,\sigma)$ represents the stable distribution with tail index $\alpha \in (0,2]$, symmetry parameter $\beta \in [0,1]$, location parameter $\mu \in \mathbb{R}$, and scale parameter

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

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

We note that the characteristic functions are real for all symmetric distributions. [We also note that the convergence is not clear across papers[119] but this doesn't apply to symmetric distributions.]

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

The characteristic function $\Psi(t)$ of a variable X^α with scale σ will be, using the expression for $\alpha > 1$, See Zolotarev[125], Samorodnitsky and Taqqu[98]:

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

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

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

6.3.1 Results

Let $X^\alpha \in \mathfrak{S}$, be the centered variable with a mean of zero, $X^\alpha = (Y^\alpha - \mu)$. We write. $\mathbb{E}(\alpha, \beta, \mu, \sigma, K) \equiv \mathbb{E}(X^\alpha|_{X^\alpha > K})$ under the stable distribution above. From Equation 6.2:

$$\mathbb{E}(X|_{X>K}) = \int_{-\infty}^{\infty} \frac{1}{u} i \left(\alpha u \sigma^\alpha |u|^{\alpha-2}\left(1 + i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(u)\right)\right.$$
$$\left. + i\mu\right)\exp\left(|u\sigma|^\alpha\left(-1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sgn}(u)\right) + iu(K-\mu)\right) du$$

with explicit solutions:

$$E(\alpha, \beta, 0, \sigma, 0) = -\sigma\frac{1}{\pi\alpha}\Gamma\left(-\frac{1}{\alpha}\right)\left(\left(1 + i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha} + \left(1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}\right). \tag{6.3}$$

Our formulation in Equation 6.3 generalizes and simplifies the commonly used one from Wolfe [122] from which Hardin [55] got the explicit form, promoted in Samorodnitsky and Taqqu [98] and Zolotarev[125]:

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

Which allows us to prove the following statements:

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

$$n_\alpha^\beta = 2^{\frac{\alpha}{1-\alpha}} \pi^{\frac{\alpha}{2-2\alpha}} \left(\Gamma\left(\frac{\alpha-1}{\alpha}\right) \sqrt{n_g} \left(\left(1 - i\beta \tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{1}{\alpha}} + \left(1 + i\beta \tan\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{1}{\alpha}} \right) \right)^{\frac{\alpha}{\alpha-1}} \tag{6.4}$$

with alternative expression:

$$n_\alpha^\beta = \pi^{\frac{\alpha}{2-2\alpha}} \left(\frac{\sec^2\left(\frac{\pi\alpha}{2}\right)^{-\frac{1}{2}/\alpha} \sec\left(\frac{\tan^{-1}\left(\tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}\right)}{\sqrt{n_g}\,\Gamma\left(\frac{\alpha-1}{\alpha}\right)} \right)^{\frac{\alpha}{1-\alpha}} \tag{6.5}$$

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

$$n_\alpha = \pi^{\frac{\alpha}{2(1-\alpha)}} \left(\frac{1}{\sqrt{n_g}\,\Gamma\left(\frac{\alpha-1}{\alpha}\right)} \right)^{\frac{\alpha}{1-\alpha}} \tag{6.6}$$

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

$$\mathbb{E}\left(\left| \sum^{kn_\alpha} \frac{X_i^\alpha - m_\alpha}{n_\alpha} \right| \right) / \mathbb{E}\left(\left| \sum^{n_\alpha} \frac{X_i^\alpha - m_\alpha}{n_\alpha} \right| \right) = k^{\frac{1}{\alpha}-1} \tag{6.7}$$

Table 13 shows the equivalence of summands between processes.

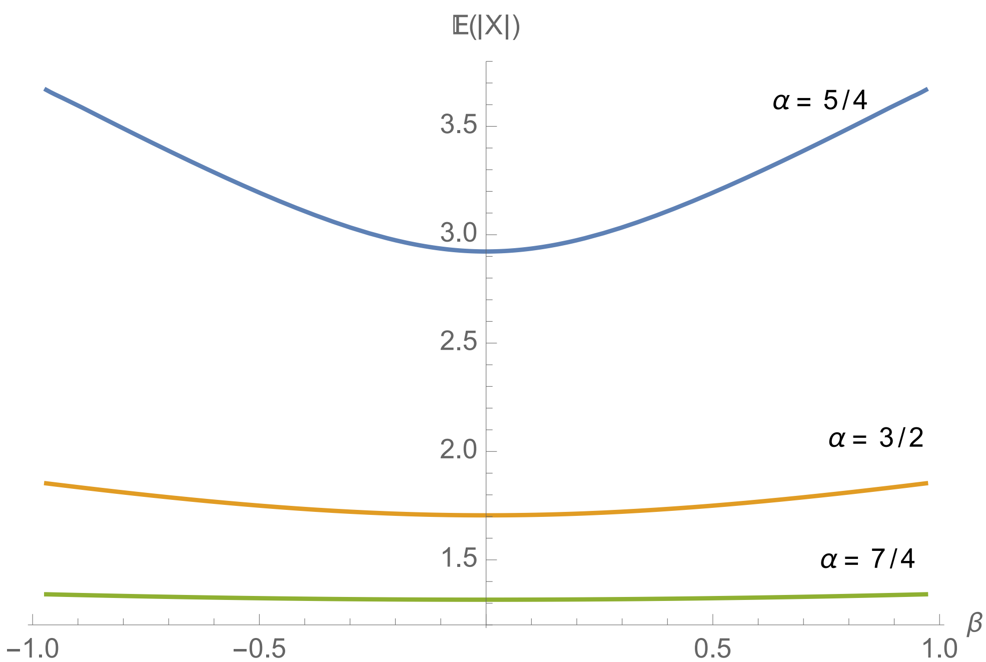


Figure 6.3: Asymmetries and Mean Deviation.

Table 13: Corresponding n_α, or how many for equivalent α-stable distribution. The Gaussian case is the $\alpha = 2$. For the case with equivalent tails to the 80/20 one needs 10^{11} more data than the Gaussian.

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

Remark 6.1.
The ratio mean deviation of distributions in $\mathfrak{S}$ is homogeneous of degree $k^{\frac{1}{\cdot}\alpha-1}$. This is not the case for other classes "nonstable".

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

6.3.2 Stochastic Alpha or Mixed Samples

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

Proposition 6.1.
For so and so

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

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

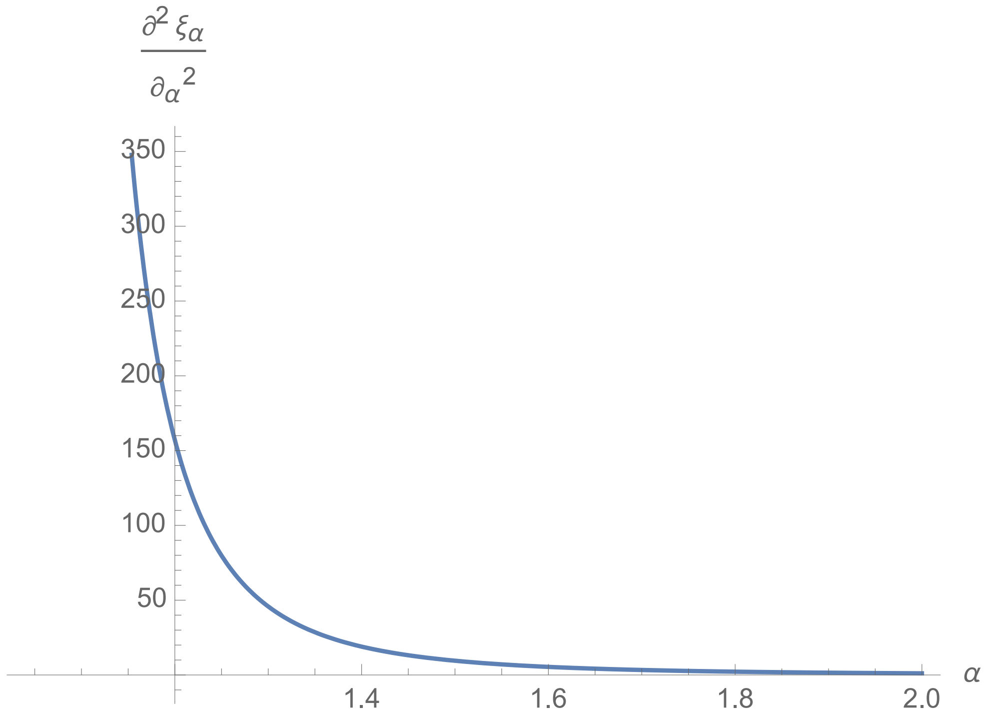


Figure 6.4: Mixing distributions: the effect is pronounced at lower values of α, as tail uncertainty creates more fat-tailedness.

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

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

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

Which is also negative with respect to $alpha$ as can be seen in Figure 6.4. The implication is that one's sample underestimates the required "n". (Commentary).

6.4 SYMMETRIC NONSTABLE DISTRIBUTIONS IN THE SUBEXPONENTIAL CLASS

6.4.1 Symmetric Mixed Gaussians, Stochastic Mean

While mixing Gaussians the kurtosis rises, which makes it convenient to simulate fattailedness. But mixing means has the opposite effect, as if it were more "stabilizing". We can observe a similar effect of "thin-tailedness" as far as the n required to match the standard benchmark. The situation is the result of multimodality, noting that stable distributions are unimodal (Ibragimov and Chernin) [60] and infinitely divisible Wolfe [123]. For X_i Gaussian with mean μ, $\mathbb{E} = \mu \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\mu^2}{2\sigma^2}}$, and keeping the average $\mu \pm \delta$ with probability $1/2$ each. With the perfectly symmetric case $\mu = 0$ and sampling with equal probability:

$$\frac{1}{2}(\mathbb{E}_{+\delta} + \mathbb{E}_{-\delta}) = \left(\frac{\sigma e^{-\frac{\delta^2}{2\sigma^2}}}{\sqrt{2\pi}} + \frac{1}{2}\delta \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right)\right) \operatorname{erf}\left(\frac{e^{-\frac{\delta^2}{2\sigma^2}}}{\sqrt{\pi}} + \frac{\delta \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right)}{\sqrt{2}\sigma}\right)$$
$$+ \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\left(\sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\delta^2}{2\sigma^2}} + \delta \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right)\right)^2}{2\sigma^2}\right)$$

6.4.2 Half cubic Student T (Lévy Stable Basin)

Relative convergence:

Theorem 6.1.
For all so and so, (details), etc.

$$c_1 \leq \frac{\mathbb{E}\left(\left|\sum^{kn} \frac{X_i^\alpha - m_\alpha}{n_\alpha}\right|\right)}{\mathbb{E}\left(\left|\sum^{n} \frac{X_i^\alpha - m_\alpha}{n_\alpha}\right|\right)} \leq c_2 \tag{6.8}$$

where:

$$c_1 = k^{\frac{1}{\alpha}-1}$$
$$c_2 = 2^{7/2}\pi^{1/2}\left(-\Gamma\left(-\frac{1}{4}\right)\right)^{-2}$$

Note that because the instability of distribution outside the basin, they end up converging to $S_{Min(\alpha,2)}$, so at $k = 2, n = 1$, equation 6.8 becomes an equality and $k \to \infty$ we satisfy the equalities in **??** and 6.7.

Proof. (Sketch)
The characteristic function for $\alpha = \frac{3}{2}$:

$$\Psi(t) = \frac{3^{3/8}\,|t|^{3/4}\,K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}\,|t|\right)}{\sqrt[8]{2}\,\Gamma\left(\frac{3}{4}\right)}$$

Leading to convoluted density p_2 for a sum $n = 2$:

$$p_2(x) = \frac{\Gamma\left(\frac{5}{4}\right)\,{}_2F_1\left(\frac{5}{4}, 2; \frac{7}{4}; -\frac{2x^2}{3}\right)}{\sqrt{3}\,\Gamma\left(\frac{3}{4}\right)^2\,\Gamma\left(\frac{7}{4}\right)}$$

□

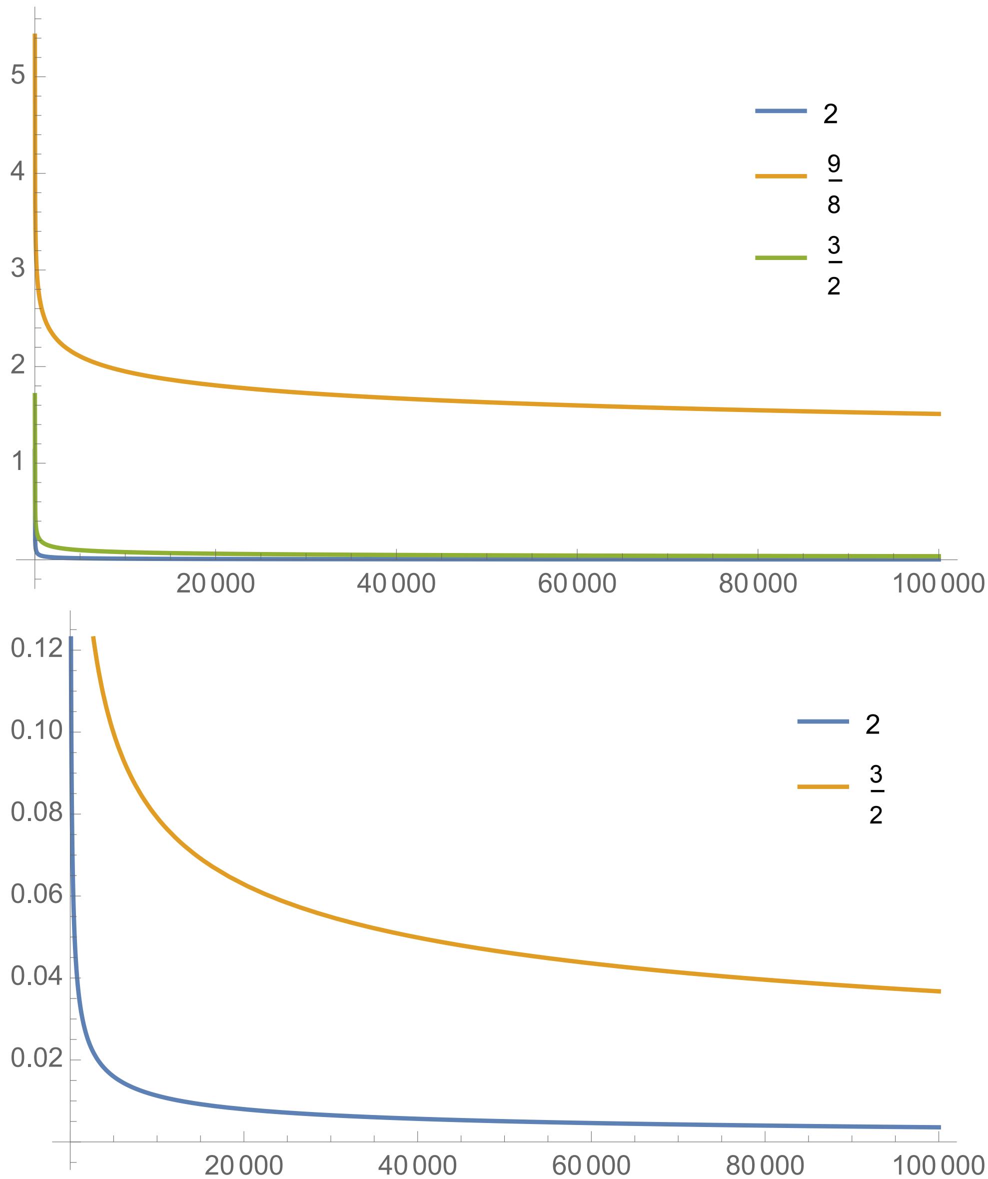


Figure 6.5: Different Speed: the fatter tailed processes are not just more uncertain; they also converge more slowly.

6.4.3 Cubic Student T (Gaussian Basin)

Student T with 3 degrees of freedom (higher exponent resembles Gaussian). We can get a semi-explicit density for the Cubic Student T.

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

we have:

$$\varphi(t) = \mathbb{E}[e^{itX}] = (1 + \sqrt{3}\,|t|)\, e^{-\sqrt{3}\,|t|}$$

hence the n-summed characteristic function is:

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

and the pdf of Y is given by:

$$p(x) = \frac{1}{\pi} \int_0^{+\infty} (1 + \sqrt{3}\,t)^n \, e^{-n\sqrt{3}\,t} \cos(tx)\, dt$$

using

$$\int_0^{\infty} t^k e^{-t} \cos(st)\, \mathrm{d}t = \frac{T_{1+k}(1/\sqrt{1+s^2})k!}{(1+s^2)^{(k+1)/2}}$$

where $T_a(x)$ is the T-Chebyshev polynomial,[2] the pdf p(x) can be writen:

$$p(x) = \frac{\left(n^2 + \frac{x^2}{3}\right)^{-n-1}}{\sqrt{3}\pi}$$

$$\sum_{k=0}^{n} \frac{\left(n! \left(n^2 + \frac{x^2}{3}\right)^{\frac{1-k}{2}+n}\right) T_{k+1}\left(\frac{1}{\sqrt{\frac{x^2}{3n^2}+1}}\right)}{(n-k)!}$$

which allows explicit solutions for specific values of n, not not for the general form:

$$\{\mathbb{E}_n\}_{1\,\leq n<\infty} = \left\{ \frac{2\sqrt{3}}{\pi}, \frac{3\sqrt{3}}{2\pi}, \frac{34}{9\sqrt{3}\pi}, \frac{71\sqrt{3}}{64\pi}, \frac{3138\sqrt{3}}{3125\pi}, \frac{899}{324\sqrt{3}\pi}, \frac{710162\sqrt{3}}{823543\pi}, \right.$$

$$\left. \frac{425331\sqrt{3}}{524288\pi}, \frac{33082034}{14348907\sqrt{3}\pi}, \frac{5719087\sqrt{3}}{7812500\pi} \right\}$$

2 With thanks to Abe Nassen and Jack D'Aurizio on Math Stack Exchange.

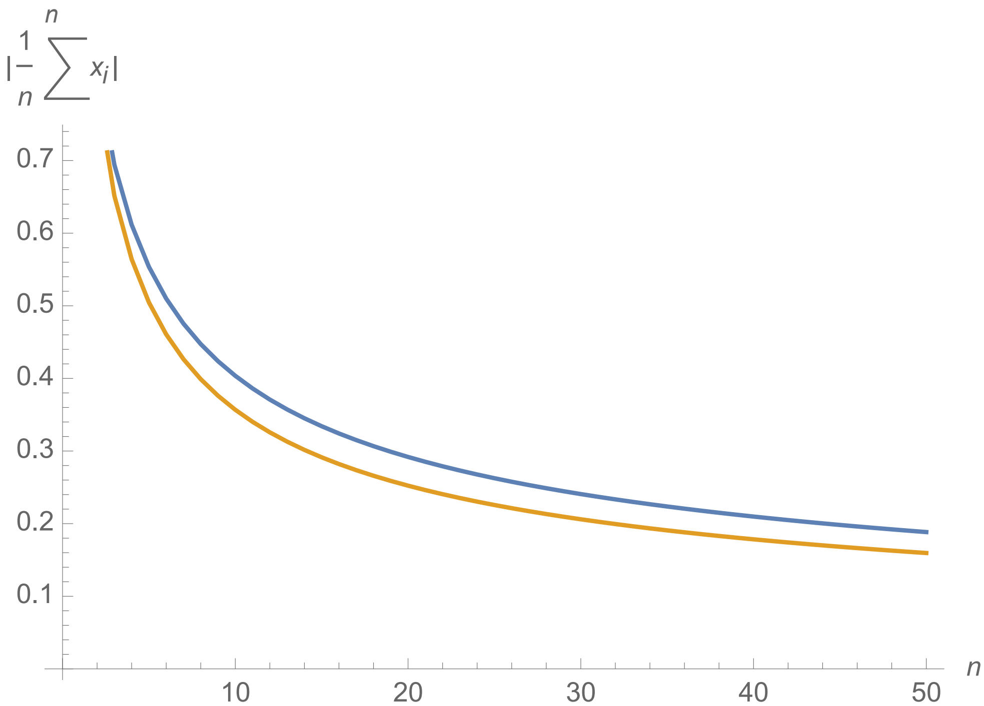


Figure 6.6: Student T with exponent =3. This applies to the general class of symmetric power law distributions.

6.5 ASYMMETRIC NONSTABLE DISTRIBUTIONS IN THE SUBEXPONETIAL CLASS

6.5.1 One-tailed Pareto Distributions

6.5.2 The Lognormal and Borderline Subexponential Class

6.6 ASYMMETRIC DISTRIBUTIONS IN THE SUPEREXPONENTIAL CLASS

6.6.1 Mixing Gaussian Distributions and Poisson Case

6.6.2 Skew Normal Distribution

This is the most untractable case mathematically, apparently though the most present when we discuss fat tails [124].

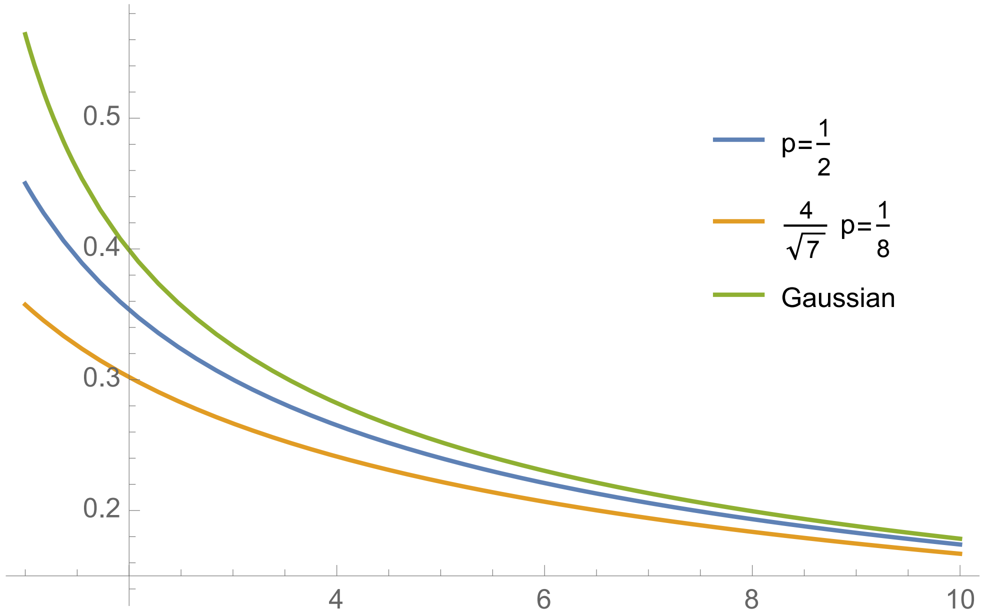


Figure 6.7: Sum of bets converge rapidly to Gaussian bassin but remain clearly subgaussian for small samples.

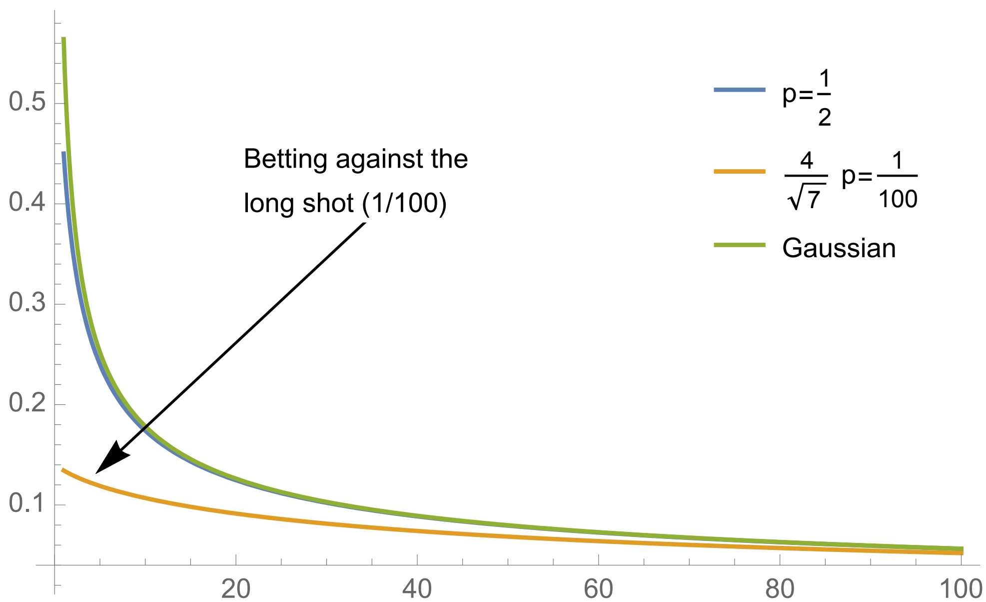


Figure 6.8: For asymmetric binary bets, at small values of p, convergence is slower.

6.6.3 Super-thin tailed distributions: Subgaussians

Consider a sum of Bernoulli variables X. The average $\sum_n \equiv \sum_{i \leq n} x_i$ follows a Binomial Distribution. Assuming $np \in \mathbb{N}^+$ to simplify:

$$\mathbb{E}\left(|\Sigma_n|\right) = -2 \sum_{i \leq 0 \leq np} (x - np)\, p^x \binom{n}{x} (1-p)^{n-x}$$

$$\mathbb{E}\left(|\Sigma_n|\right) = -2(1-p)^{n(-p)+n-2}p^{np+1}\Gamma(np+2)$$
$$\left((p-1)\binom{n}{np+1}\lambda_1 - p(np+2)\binom{n}{np+2}\lambda_2\right)$$

where:

$$\lambda_1 ={}_2\tilde{F}_1\left(1, n(p-1)+1; np+2; \frac{p}{p-1}\right)$$

and

$$\lambda_2 ={}_2\tilde{F}_1\left(2, n(p-1)+2; np+3; \frac{p}{p-1}\right)$$

6.7 ACKNOWLEDGEMENT

Colman Humphrey,...

APPENDIX: METHODOLOGY, PROOFS, ETC.

6.7.1 Derivations using explicit $\mathbb{E}(|X|)$

See Wolfe [122] from which Hardin got the explicit form[55].

6.7.2 Derivations using the Hilbert Transform and $\beta = 0$

Section obsolete since I found forms for asymmetric stable distributions. Some commentary on Hilbert transforms for symmetric stable distributions, given that for $Z = |X|$, $\mathrm{d}F_z(z) = \mathrm{d}F_X(x)(1-\mathrm{sgn}(x))$, that type of thing.

Hilbert Transform for a function f (see Hlusel, [57], Pinelis [87]):

$$H(f) = \frac{1}{\pi}\text{p.v.}\int_{-\infty}^{\infty}\frac{f(x)}{t-x}\mathrm{d}x$$

Here p.v. means principal value in the Cauchy sense, in other words

$$\text{p.v.}\int_{-\infty}^{\infty} = \lim_{a\to\infty}\lim_{b\to 0}\int_{-a}^{-b} + \int_{b}^{a}$$

$$\mathbb{E}(|X|) = \frac{\partial}{\partial t}H(\Psi(0)) = \frac{1}{\pi}\frac{\partial}{\partial t}\text{p.v.}\int_{-\infty}^{\infty}\frac{\Psi(z)}{t-z}\mathrm{d}z|_{t=0}$$

$$\mathbb{E}(|X|) = \frac{1}{\pi}\text{p.v.}\int_{-\infty}^{\infty}\frac{\Psi(z)}{z^2}\mathrm{d}z$$

In our case:

$$\mathbb{E}(|X|) = \frac{1}{\pi}\text{p.v.}\int_{-\infty}^{\infty} -\frac{e^{-|t\sigma|^{\alpha}}}{t^2}\,dt = \frac{2}{\pi}\Gamma\left(\frac{\alpha-1}{\alpha}\right)\sigma$$

D IN PROGRESS DERIVATIONS FOR LLN ACROSS FAT TAILS

D.1 COMMENTS ON LLN AND BOUNDS

Recall from Chapter 3 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 which deals with the limiting behavior of relative frequencies.

(Hint: we will need to look at the limit without the common route of Chebychev's inequality which requires $\mathbb{E}[X_i^2] < \infty$. 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.

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, $\underset{n\to\infty}{lim\,sup} \frac{\sum_{i=1}^n x_i}{\sqrt{2n\log\log(n)}} = 1$ a.s. (and by symmetry $\underset{n\to\infty}{lim\,inf} \frac{\sum_{i=1}^n x_i}{\sqrt{2n\log\log(n)}} = -1$), thus giving us an acceptably narrow cone limiting the fluctuation of the sum.

Chernoff Bounds For very, very thin tails, that is variations that are either fixed (binary such as in a Bernouilli) or hard bound to a maximum and a minimum, the tightest bound we can find is the Chernoff. See discussion section x.

D.1.1 Speed of Convergence for Simple Cases

Let us examine the speed of convergence of the average $\frac{1}{N}\sum_{1\le i\le N} X_i$. For a Gaussian distribution (m,σ), the characteristic function for the convolution is:

$$\varphi(t/N)^N = \left(e^{\frac{imt}{N}-\frac{s^2t^2}{2N^2}}\right)^N,$$

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 consists in expanding φ and letting N go to infinity

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

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 implies that convergence takes place almost everywhere except for a set of probability 0; for that the same result should be obtained for all values of 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 $\alpha > 1$.

Setting $L = 1$ to scale, the standard deviation $\sigma_\alpha(N)$ for the N-average becomes, for $\alpha > 2$

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

The trap After some tinkering, we get $\sigma_\alpha(N) = \frac{\sigma_\alpha(1)}{\sqrt{N}}$, the same as with the Gaussian, which is a 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, variance v, minimum value L and exponent α.

Infinite variance of variance The distribution of the variance, v can be obtained analytically: intuitively its asymptotic tail is $v^{-\frac{\alpha}{2}-1}$. Where g(.) is the probability density of the variance:

$$g(v) = \frac{\alpha L^\alpha \left(\frac{\sqrt{\frac{\alpha}{\alpha-2}}L}{\alpha-1} + \sqrt{v}\right)^{-\alpha-1}}{2\sqrt{v}}$$

with support:$[(L - \frac{\sqrt{\frac{\alpha}{\alpha-2}}L}{\alpha-1})^2, \infty)$.

Cleaner: Δ_α the expected mean deviation of the variance for a given α will be $\Delta_\alpha = \frac{1}{v}\int_L^\infty \left|(x-m)^2 - v\right| p(x) \mathrm{d}x$.

Absence of Useful Theory: As to situations, central situations, where $1 < \alpha < 2$, 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.

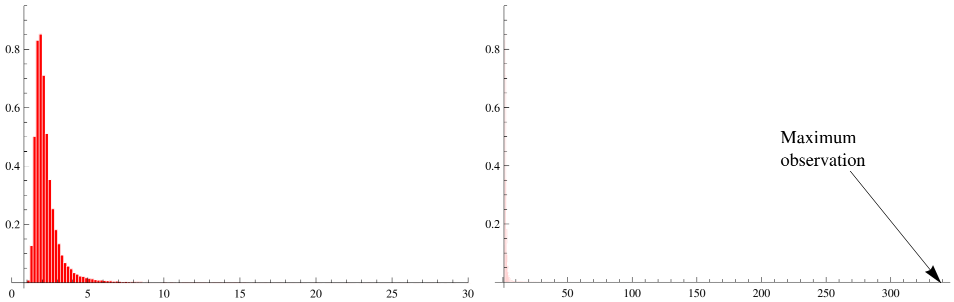


Figure D.1: The distribution (histogram) of the standard deviation of the sum of N=100 α=13/6. The second graph shows the entire span of realizations. If it appears to shows very little information in the middle, it is because the plot is stretched to accommodate the extreme observation on the far right.

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).

D.1.2 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 transform of ϕ,

$$p_{2,\alpha}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(t/2)^{2-i\;t\;x}\,\mathrm{d}t,$$

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:

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty}(a)_k(b)_k/(c)_k\;z^k\Big/k!$$

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

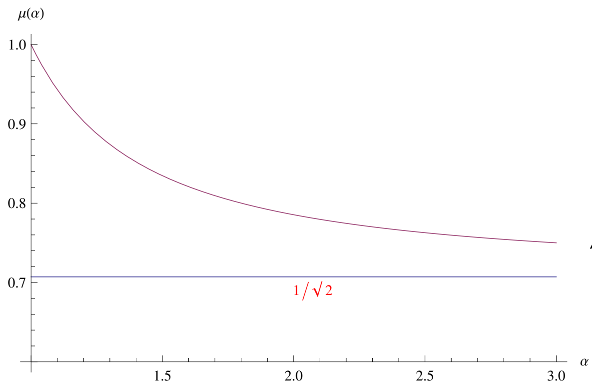


Figure D.2: Preasymptotics of the ratio of mean deviations for a symmetric power law (Student). But one should note that mean deviations themselves are extremely high in the neighborhood of ↓1. So we have a "sort of" double convergence to $\sqrt{n}$: convergence at higher n and convergence at higher 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.

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

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,

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

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

and

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

D.2 DIGRESSION INTO INVERSION OF CHARACTERISTIC FUNCTION OF NONSTABLE POWER LAW

The Characteristic function of the Student T with α degrees of freedom, $C(t) = \frac{2^{1-\frac{\alpha}{2}}\alpha^{\alpha/4}|t|^{\alpha/2}K_{\frac{\alpha}{2}}\left(\sqrt{\alpha}|t|\right)}{\Gamma\left(\frac{\alpha}{2}\right)}$ entails a modified Bessel function of the second kind $K_{\alpha/2}\left(\sqrt{\alpha}\,|t|\right)$. To invert the Fourier to get the probability density of the n-summed variable when α is not an integer poses problem as the equation below seems integrable otherwise. Of particular interest is the distribution for $\alpha = 3/2$ ("halfcubic"). With n an integer ($n > 2$):

$$f_n(x) = \left(\frac{3^{3/8}}{\sqrt[8]{2}\,\Gamma\left(\frac{3}{4}\right)}\right)^n \int_{-\infty}^{\infty} e^{-i\,t\,x}\,|t|^{\frac{3n}{4}}\,K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}\,|t|\right)^n\,dt$$

I tried all manner of expansions and reexpressions of the Bessel into other functions (Hypergeometric, Gamma) to no avail. One good news is that $n = 2$ works on Mathematica because the Wolfram library has the square of a Bessel function. It would be great to get the solution for at least $n = 3$.

Take the n-convoluted

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

α	C(t)
1	$\left(\frac{\pi}{2}\right)^{n/2}\left(\frac{e^{-t}}{\sqrt{t}}\right)^n$
$\frac{5}{4}$	$K_{\frac{5}{8}}\left(\frac{\sqrt{5}t}{2}\right)^n$
$\frac{3}{2}$	$K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}t\right)^n$
$\frac{7}{4}$	$K_{\frac{7}{8}}\left(\frac{\sqrt{7}t}{2}\right)^n$
2	$K_1\left(\sqrt{2}t\right)^n$
$\frac{9}{4}$	$K_{\frac{9}{8}}\left(\frac{3t}{2}\right)^n$
$\frac{5}{2}$	$K_{\frac{5}{4}}\left(\sqrt{\frac{5}{2}}t\right)^n$
$\frac{11}{4}$	$K_{\frac{11}{8}}\left(\frac{\sqrt{11}t}{2}\right)^n$
3	$3^{-5n/4}\left(\frac{\pi}{2}\right)^{n/2}\left(\frac{e^{-\sqrt{3}t}\left(3t+\sqrt{3}\right)}{t^{3/2}}\right)^n$

D.2.1 Integrable Characteristic Functions

α	Student T	Pareto
1	$e^{-t\,\text{sgn}(t)}$	$E_2(-iLt)$
$\frac{5}{4}$	$\frac{5^{5/16}\lvert t\rvert^{5/8}K_{\frac{5}{8}}\left(\frac{\sqrt{5}\lvert t\rvert}{2}\right)}{\sqrt[4]{2}\,\Gamma\left(\frac{5}{8}\right)}$	$\frac{5}{4}E_{\frac{9}{4}}(-iLt)$
$\frac{3}{2}$	$\frac{3^{3/8}\lvert t\rvert^{3/4}K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}\lvert t\rvert\right)}{\sqrt[8]{2}\,\Gamma\left(\frac{3}{4}\right)}$	$\frac{3}{2}E_{\frac{5}{2}}(-iLt)$
$\frac{7}{4}$	$\frac{7^{7/16}\lvert t\rvert^{7/8}K_{\frac{7}{8}}\left(\frac{\sqrt{7}\lvert t\rvert}{2}\right)}{2^{3/4}\,\Gamma\left(\frac{7}{8}\right)}$	$\frac{7}{4}E_{\frac{11}{4}}(-iLt)$
2	$\sqrt{2}\,\lvert t\rvert\,K_1\left(\sqrt{2}\,\lvert t\rvert\right)$	$2E_3(-iLt)$
$\frac{9}{4}$	$\frac{6\;2^{3/4}\sqrt[8]{3}\lvert t\rvert^{9/8}K_{\frac{9}{8}}\left(\frac{3\lvert t\rvert}{2}\right)}{\Gamma\left(\frac{1}{8}\right)}$	$\frac{9}{4}E_{\frac{13}{4}}(-iLt)$
$\frac{5}{2}$	$\frac{5^{5/8}\lvert t\rvert^{5/4}K_{\frac{5}{4}}\left(\sqrt{\frac{5}{2}}\lvert t\rvert\right)}{2^{7/8}\,\Gamma\left(\frac{5}{4}\right)}$	$\frac{5}{2}E_{\frac{7}{2}}(-iLt)$
$\frac{11}{4}$	$\frac{11^{11/16}\lvert t\rvert^{11/8}K_{\frac{11}{8}}\left(\frac{\sqrt{11}\lvert t\rvert}{2}\right)}{2\;2^{3/4}\,\Gamma\left(\frac{11}{8}\right)}$	$\frac{11}{4}E_{\frac{15}{4}}(-iLt)$
3	$e^{-\sqrt{3}\lvert t\rvert}\left(\sqrt{3}\,\lvert t\rvert+1\right)$	$3E_4(-iLt)$

D.3 PULLING THE PDF OF AN n-SUMMED STUDENT T

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

we have:

$$\varphi(t) = \mathbb{E}[e^{itX}] = (1+\sqrt{3}\,|t|)\,e^{-\sqrt{3}\,|t|}$$

hence the n-summed characteristic function is:

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

and the pdf of Y is given by:

$$f_Y(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}(1+\sqrt{3}\,|t|)^n\,e^{-n\sqrt{3}\,|t|}e^{-itx}\,dt = \frac{1}{\pi}\int_0^{+\infty}(1+\sqrt{3}\,t)^n\,e^{-n\sqrt{3}\,t}\cos(tx)\,dt$$

One can expand with the Newton formula

$$(1+\sqrt{3}t)^n = \sum_{k=0}^{n}\binom{n}{k}(\sqrt{3}t)^k$$

then the integrals can be performed using

$$\int_0^{\infty} t^k e^{-t}\cos(st)\,\mathrm{d}t = \frac{T_{1+k}(1/\sqrt{1+s^2})k!}{(1+s^2)^{(k+1)/2}}$$

where $T_a(x)$ is the T-Chebyshev polynomial. Then, the given sequence can be rewritten setting $s = x/\sqrt{3}$ as[1]

$$S_n = \frac{6\sqrt{3}\left(n^2+s^2\right)^{-n-1}}{\pi 18} \sum_{k=0}^{n} \frac{n!}{(n-k)!} \left(n^2+s^2\right)^{\frac{1-k}{2}+n} T_{k+1}\left(\frac{1}{\sqrt{\frac{s^2}{n^2}+1}}\right)$$

1 with thanks to Abe Nassen and Jack D'Aurizio on Math Stack Exchange.

7 PREASYMPTOTICS AND CENTRAL LIMIT IN THE REAL WORLD

Chapter Summary 7: **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). We examine how fat tailedness worsens the process.**

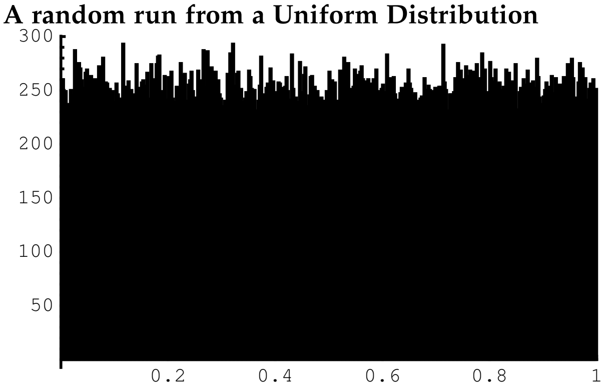
An intuition: how we converge mostly in the center of the distribution

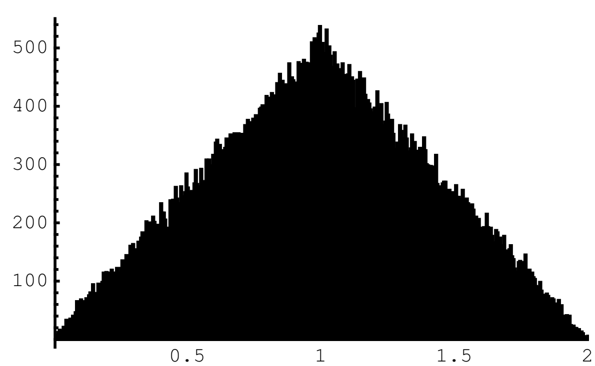
We start with the Uniform Distribution, patently the easiest of all.

$$f(x) = \begin{cases} \frac{1}{H-L} & L \leq x \leq H \\ 0 & \text{elsewhere} \end{cases}$$

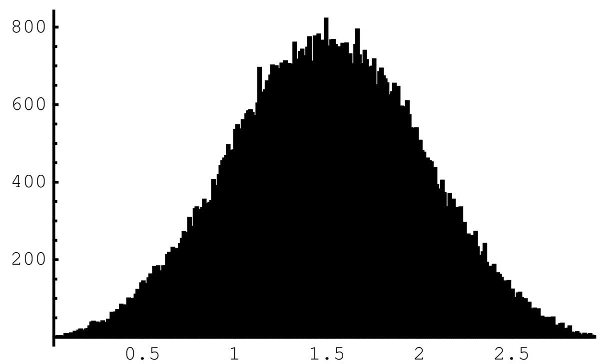
where L = 0 and H =1

A random run from a Uniform Distribution





The functioning of CLT is as follows: the convolution is a multiplication; it is the equivalent of weighting the probability distribution by a function that iteratively gives more weight to the body, and less weight to the tails, until it becomes round enough to dull the iterative effect. See how "multiplying" a flat distribution by something triangular as in Figure 7 produces more roundedness.



As we can see, we get more observations where the peak is higher. Now some math. By convoluting 2, 3, 4 times we can see the progress and the decrease of mass in the tails:

$$f_2(z_2) = \int_{-\infty}^{\infty} (f(z-x))(fx)\,\mathrm{d}x = \begin{cases} 2 - z_2 & 1 < z_2 < 2 \\ z_2 & 0 < z_2 \leq 1 \end{cases} \tag{7.1}$$

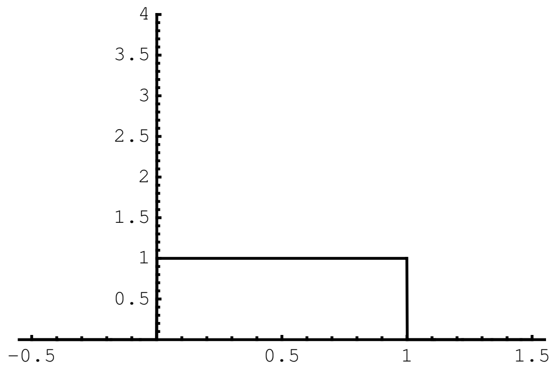
We have a triangle (piecewise linear).

$$f_3(z_3) = \int_0^3 (f_2(z_3-2))f(x_2)\,\mathrm{d}x_2 = \begin{cases} \frac{z_3^2}{2} & 0 < z_3 \leq 1 \\ -(z_3-3)z_3 - \frac{3}{2} & 1 < z_3 < 2 \\ -\frac{1}{2}(z_3-3)(z_3-1) & z_3 = 2 \\ \frac{1}{2}(z_3-3)^2 & 2 < z_3 < 3 \end{cases} \tag{7.2}$$

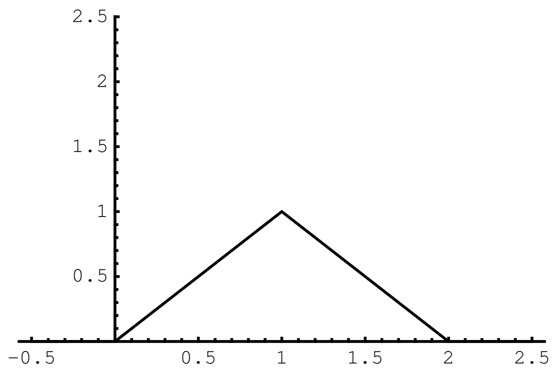
With $N = 3$ we square terms, and the familiar "bell" shape starts to emerge thanks to such squaring.

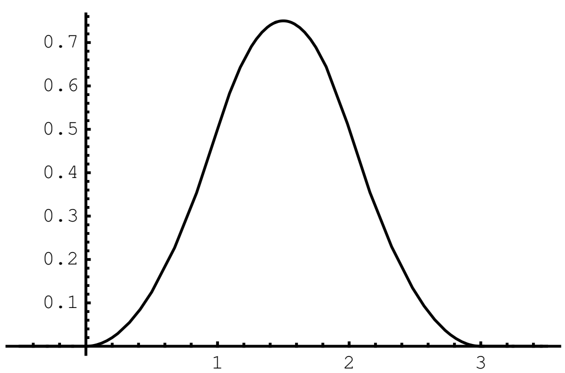
$$f_4 x = \int_0^4 (f_3(z_4 - x))(fx3)\,dx3 = \begin{cases} \frac{1}{4} & z_4 = 3 \\ \frac{1}{2} & z_4 = 2 \\ \frac{z_4^2}{4} & 0 < z_4 \leq 1 \\ \frac{1}{4}\left(-z_4^2 + 4z_4 - 2\right) & 1 < z_4 < 2 \vee 2 < z_4 < 3 \\ \frac{1}{4}(z_4-4)^2 & 3 < z_4 < 4 \end{cases} \tag{7.3}$$

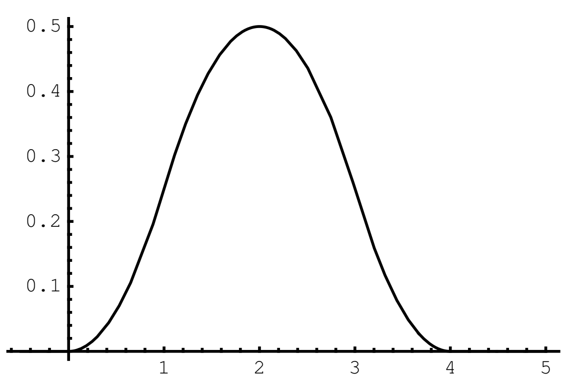
A simple Uniform Distribution



We can see how quickly, after one single addition, the net probabilistic "weight" is going to be skewed to the center of the distribution, and the vector will weight future densities..







Finite Variance: Necessary but Not Sufficient

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

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

Where $\xrightarrow{D}$ 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 [2] 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:

1 Feller 1971, Vol. II

2 See Loeve for a presentation of the method of truncation used by Kolmogorov in the early days before Lyapunov started using characteristic functions.

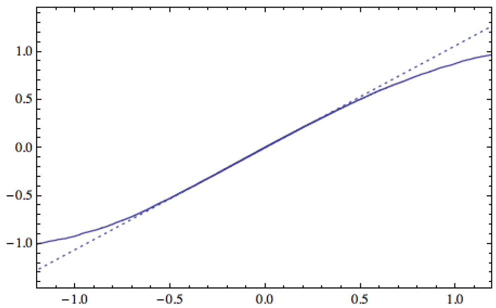


Figure 7.1: Q-Q Plot of N Sums of variables distributed according to the Student T with 3 degrees of freedom, N=50, compared to the Gaussian, rescaled into standard deviations. We see on both sides a higher incidence of tail events. 10^6simulations

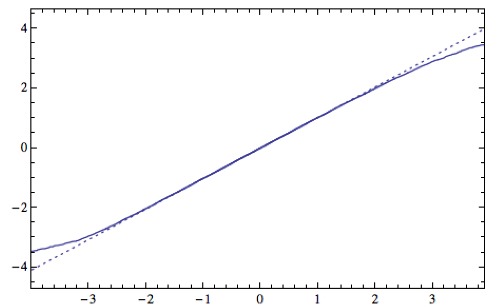


Figure 7.2: The Widening Center. Q-Q Plot of variables distributed according to the Student T with 3 degrees of freedom compared to the Gaussian, rescaled into standard deviation, N=500. We see on both sides a higher incidence of tail events. 10^7simulations.

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

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.

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.

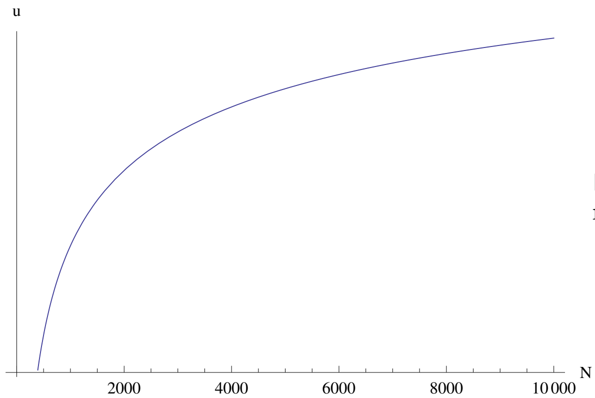


Figure 7.3: The behavior of the "tunnel" under summation

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 α= 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[3], particularly if instead of considering the sum u we rescaled by $1/a\sqrt{N}$.

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 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

3 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.

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 [NOT] $0 \leq x \leq \sqrt{\left(\frac{m}{2}-1\right) N \log(N)}$

Generalizing for all exponents > 2 More generally, using the reasoning for a broader set and getting the crossover for powelaws of all exponents:

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

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

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

7.1 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 $(3!\,\Lambda^2)/N^2$.

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:

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

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

Table 14: Table of Normalized Cumulants For Thin Tailed Distributions-Speed of Convergence (Dividing by Σ^n where n is the order of the cumulant).

Distr.	**Normal**(μ, σ)	**Poisson**(λ)	**Exponent'l**(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	$\frac{1}{N^8\lambda^8}$	$\frac{9!\lambda^8}{N^8}$	$\frac{9!}{a^8b^8N^8}$

Distr.	**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	$N\log\left(pe^{-\frac{z^2\sigma_1{}^2}{2}}+(1-p)e^{-\frac{z^2\sigma_2{}^2}{2}}\right)$	$N\left(\log\left(\sqrt{3}\lvert z\rvert+1\right)-\sqrt{3}\lvert z\rvert\right)$	$N\log\left(2\lvert z\rvert^2K_2(2\lvert z\rvert)\right)$
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"

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

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

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

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

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

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. $\mathbf{S}(\alpha, \beta, \mu, \sigma)$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 $1 - cx^{-\mu}$

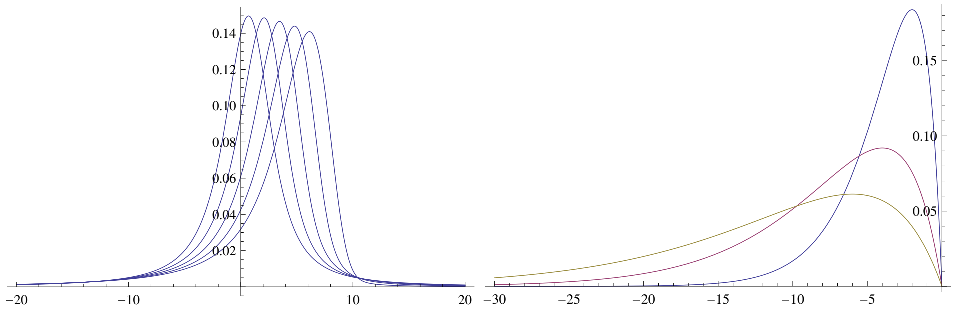


Figure 7.4: Disturbing the scale of the alpha stable and that of a more natural distribution, the gamma distribution. The alpha stable does not increase in risks! (risks for us in Chapter x is defined in thickening of the tails of the distribution). We will see later with "convexification" how it is rare to have an isolated perturbation of distribution without an increase in risks.

as $x\text{->}+\infty$ and $d(-x)^{-\mu}$ as $x\text{->}-\infty$ 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 $\longrightarrow$ CHEBYCHEV $\longrightarrow$ 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

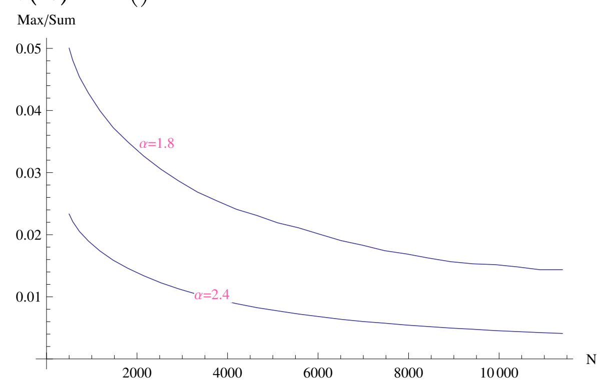
This table shows the inverse of the probability of exceeding a certain σ for the Gaussian and the lower on probability limit for any distribution with finite variance.

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

7.2 CONVERGENCE OF THE MAXIMUM OF A FINITE VARIANCE POWER LAW

An illustration of the following point. 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 $\alpha > 2$ or not. (See comments in Mandelbrot & Taleb, 2011)

$\tau(N) \equiv E\,()$



7.3 SOURCES AND FURTHER READINGS

Limits of Sums

Paul Lévy [68], Gnedenko and Kolmogorov [51], Prokhorov [92], [91], Hoeffding[58], Petrov[86], Blum[12].

For Large Deviations

Nagaev[81], [80], Mikosch and Nagaev[77], Nagaev and Pinelis [82]. In the absence of Cramér conditions, Nagaev [79], Brennan[16], Ramsay[93], Bennet[9].

Also, for dependent summands, Bernstein [10].

Discussions of Concentration functions Esseen [36], [**?**], Doeblin [26], [25], Darling [21], Kolmogorov [66], Rogozin [94], Kesten [63], Rogogin [95].

7.4 CONVERGENCE FOR NON-LÉVY STABLEPOWER LAWS TEMPORARILY HERE)

The Characteristic function of the Student T with α degrees of freedom, $C(t) = \frac{2^{1-\frac{\alpha}{2}}\alpha^{\alpha/4}|t|^{\alpha/2}K_{\frac{\alpha}{2}}\left(\sqrt{\alpha}|t|\right)}{\Gamma\left(\frac{\alpha}{2}\right)}$ entails a modified Bessel function of the second kind $K_{\alpha/2}\left(\sqrt{\alpha}\,|t|\right)$. To invert the Fourier to get the probability density of the n-summed variable when α is not an integer $\in \mathbb{Z}$ poses problem as $K_{\frac{\alpha}{2}}$ seems integrable otherwise. Of particular interest is the distribution for $\alpha = 3/2$. With n integer > 1:

$$f_n(x) = \left(\frac{3^{3/8}}{\sqrt[8]{2}\,\Gamma\left(\frac{3}{4}\right)}\right)^n \int_{-\infty}^{\infty} e^{-i\,tx}\,|t|^{3n/4}\,K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}\,|t|\right)^n dt$$

I tried all manner of expansions and reexpressions of the Bessel into other functions (Hypergeometric, Gamma) to no avail. One good news is that $n = 2$ works on Mathematica because the Wolfram library has the square of a Bessel function. It would be great to get the solution for at least $n = 3$.

α	$StudentT$	$Pareto(1,\alpha)$
1	$e^{-t\,\mathrm{sgn}(t)}$	$E_2(-iLt)$
$\frac{5}{4}$	$\frac{5^{5/16}\|t\|^{5/8}K_{\frac{5}{8}}\left(\frac{\sqrt{5}\|t\|}{2}\right)}{\sqrt[4]{2}\Gamma\left(\frac{5}{8}\right)}$	$\frac{5}{4}E_{\frac{9}{4}}(-iLt)$
$\frac{3}{2}$	$\frac{3^{3/8}\|t\|^{3/4}K_{\frac{3}{4}}\left(\sqrt{\frac{3}{2}}\|t\|\right)}{\sqrt[8]{2}\Gamma\left(\frac{3}{4}\right)}$	$\frac{3}{2}E_{\frac{5}{2}}(-iLt)$
$\frac{7}{4}$	$\frac{7^{7/16}\|t\|^{7/8}K_{\frac{7}{8}}\left(\frac{\sqrt{7}\|t\|}{2}\right)}{2^{3/4}\Gamma\left(\frac{7}{8}\right)}$	$\frac{7}{4}E_{\frac{11}{4}}(-iLt)$
2	$\sqrt{2}\,\|t\|\,K_1\left(\sqrt{2}\,\|t\|\right)$	$2E_3(-iLt)$
$\frac{9}{4}$	$\frac{6\;2^{3/4}\sqrt[8]{3}\|t\|^{9/8}K_{\frac{9}{8}}\left(\frac{3\|t\|}{2}\right)}{\Gamma\left(\frac{1}{8}\right)}$	$\frac{9}{4}E_{\frac{13}{4}}(-iLt)$
$\frac{5}{2}$	$\frac{5^{5/8}\|t\|^{5/4}K_{\frac{5}{4}}\left(\sqrt{\frac{5}{2}}\|t\|\right)}{2^{7/8}\Gamma\left(\frac{5}{4}\right)}$	$\frac{5}{2}E_{\frac{7}{2}}(-iLt)$
$\frac{11}{4}$	$\frac{11^{11/16}\|t\|^{11/8}K_{\frac{11}{8}}\left(\frac{\sqrt{11}\|t\|}{2}\right)}{2\;2^{3/4}\Gamma\left(\frac{11}{8}\right)}$	$\frac{11}{4}E_{\frac{15}{4}}(-iLt)$
3	$e^{-\sqrt{3}\|t\|}\left(\sqrt{3}\,\|t\|+1\right)$	$3E_4(-iLt)$

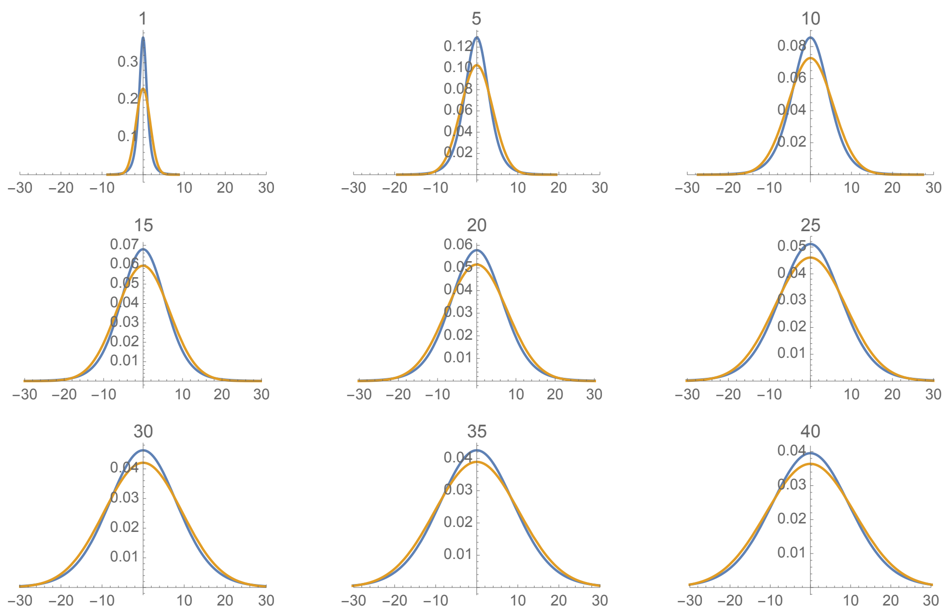


Figure 7.5: Convergence for summed Student T -3

E | WHERE STANDARD DIVERSIFICATION FAILS

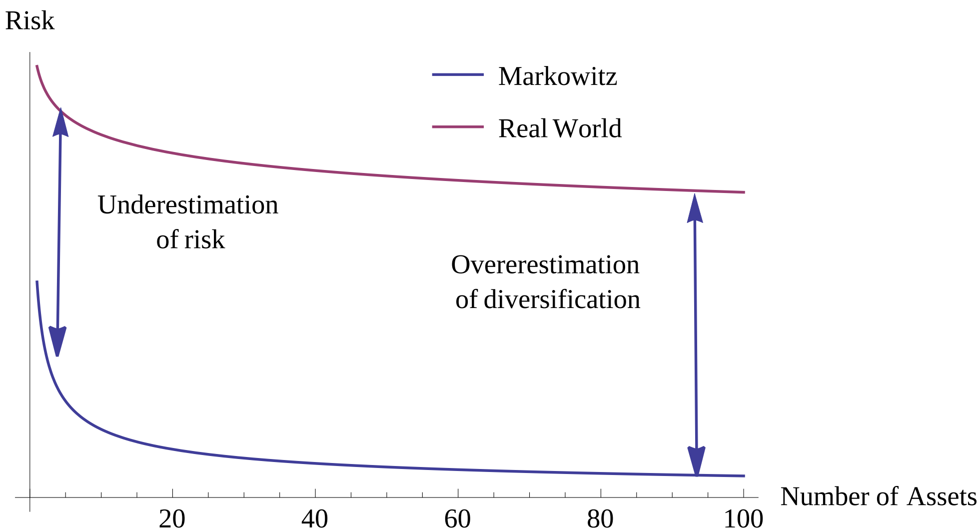


Figure E.1: The "diversification effect": difference between promised and delivered. Markowitz Mean Variance based portfolio construction will stand probably as one of the most empirically invalid theory ever used in modern times.

This is an analog of the problem with slowness of the law of large number: how a portfolio can track a general index (speed of convergence) and how high can *true* volatility be compared to the observed one (the base line).

Model Structure under Markowitz Mean Variance Historically, Modern Portfolio Theory (MPT), as a normative theory of risk bearing, has made the central assumption that, for a set return, the investor rationally minimizes the variance of his exposure, defined as the mean square variability (or, which was proven by MPT to be equivalent to maximizing the mean conditional on a set variance). The standard models, Markowitz (1952, 1959) [75] [76], and the extensions such as Treynor(1965) [116] and Sharpe (1966)[102], all base their criterion on minimum variance. [See Constantinides and Malliaris (1995)[20], Elton and Gruber (1997)[30] for a survey, Huang and Litzenberger (1988)[59] for an analytical discussion. **Distributional Condition**: The agent is supposed to have full grasp of the probability distribution with all the joint returns *known* and deemed to be Gaussian. Further, no error rate is allowed. **Utility Condition**: The agent is supposed to have a utility function that allows optimization via minimum variance, as the agent cares only about the first two moments of the distribution. So, conveniently, under quadratic utility of wealth, $U(W) \equiv aW - bW^2$, where W is wealth, a random variable, the expected wealth $E\left(U(W)\right) = aE(W) - bE(W^2)$ does not depent on higher orders of the random variable and allows therefore the maximization of $\frac{E(W)}{V(W)}$ without recomputing utilities for every state. (The same reasoning applies to the situation where in place of focusing on W the wealth, we focus on the return or relative changes in wealth, $\frac{\Delta W}{W}$).

Difficulty Knowing the Exact Structure of Returns The first difficulty arises in the presence of unknown structure to the future states of the world, as the MPT models are based on perfect, error-free knowledge of the probability distribution of future returns and its parameter, with constraints that the distribution should have specific properties. In the case of the exposure being a portfolio of assets, hence requiring the use of the standard classes of multivariate probability distributions, there are additional errors that grown nonlinearly with the number of assets (the nonlinearity of the covariance matrix): the investor would now need to estimate the correlation structure as well as all future returns. In order to implement a full Markowitz-style optimization, one needs to know the entire joint probability distribution of all assets for the entire future, plus the exact utility function for wealth at all future times -all that without errors. Estimation errors make the system highly unreliable, as small changes in parameters lead to extremely varying effects on the "optimal" allocation. The second difficulty lies in the specificity of the probability distribution, namely the reliance on the sufficiency of the first two moments in the formulation of preferences, and the neglect of higher moments of the payoffs, which, ironically necessitates models having all moments finite. "Exploding" higher moments lead to theoretical incompatibilities. Assuming finite variance, but infinite kurtosis (say a power law with tail exponent <4) results in the inability of the mean-variance equations to hold owing to the presence of incompressible higher order terms. It is not just that variance is not a good proxy for risk, it is that it is a bad one for variability –it has been shown that mean deviation, for instance, does a vastly better job out of sample. The first two difficulties we just exposed are empirical, not normative or logical (that is, in a world that is normally distributed with known probability distributions, assuming these exist, the problems would not arise); the next one is normative. So the third, and most severe difficulty is in the following incompatibility: the aim by MPT at lowering variance (for a given expected return) is inconsistent with the preferences of a rational investor, regardless of his risk aversion, since it also minimizes the variability in the profit domain. Minimum variance is indeed fundamentally incompatible with theoretically established risk preferences, see Arrow (1965, 1971)[2] [3] , Pratt (1964) [90], Machina and Rothchild (1987, 2008)[73] [74], except in the far-fetched case where the investor can only invest in symmetric probability distributions —and only under such assumption of symmetry. In other words, the idea of "risk" = variance necessitates symmetry between the profit and loss domain. If one assumes, realistically, that variance is a poor proxy for risk, constraining it for all states of the world becomes inconsistent with a rational portfolio strategy.

Mitigation via Assumption of Elliptical Distributions One may broaden the distribution to include elliptical distributions is that they do not map the return of stocks, owing to the absence of a single variance at any point in time, see Bouchaud and Chicheportiche (2010) [18]. See discussion in 3.17 of the "other fat tail" in the failure of ellipticity owing to unstable correlation structure.

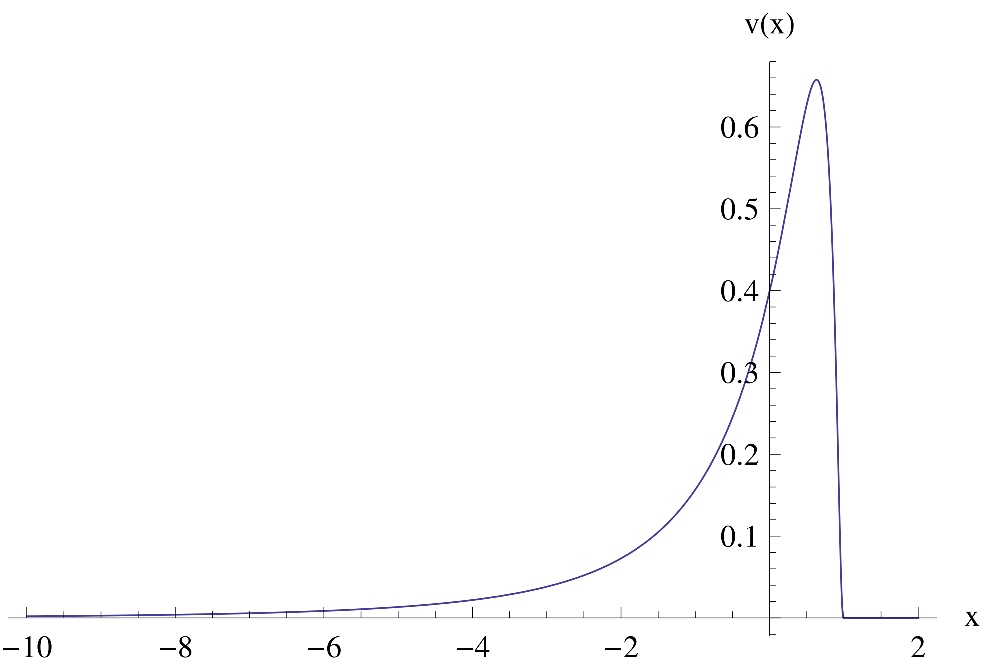
Difficulty With the Utility Structure There are problems with the utility structure and such risk aversion, as, once one fattens the left tail of the distribution, the concavity of the losses causes a severe degradation of expected utility. Quadratic utility is chosen to reverse engineer mean-variance and ignore higher moments even if these exist and are meaningful. Exponential utility can allow mean variance, but then under Gaussian distribution as the tails get heavily discounted. But absence of Gaussian outside of quadratic produces pathological expected utility returns in the presence of concavity (i.e., acceleration) in losses.

The distribution of the utility of losses can be captured by transforming the distribution. Take as an example the standard concave utility function $g(x) = 1 - e^{-ax}$. With a=1, 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 a fatter tailed distribution, such as the a standard powerlaw used in finance (Gabaix, 2008,[47]), where α is the tail exponent,

$$v(x) = \frac{x\left(\frac{\alpha}{\frac{(\log(1-x)-1)^2}{a^2}+\alpha}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}(a-ax)B\left(\frac{\alpha}{2},\frac{1}{2}\right)}$$



We can see With such a distribution of utility it would be absurd to do anything.

F FAT TAILS AND RANDOM MATRICES

[The equivalent of fat tails for matrices. This will be completed, but consider for now that the 4th moment reaching Gaussian levels (i.e. 3) in the chapter is equivalent to eigenvalues reaching Wigner's semicircle.]

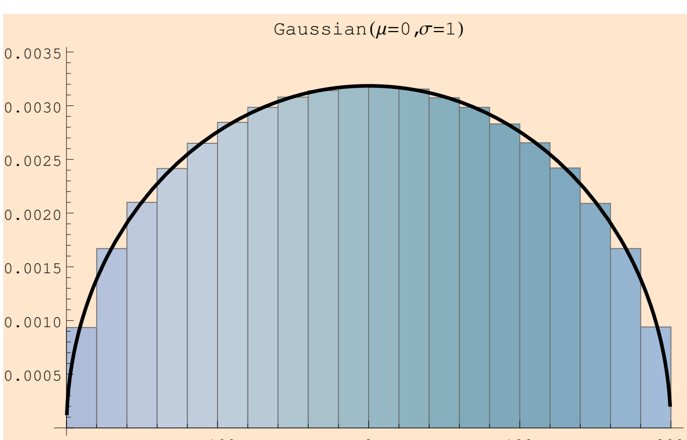


Figure F.1: Gaussian

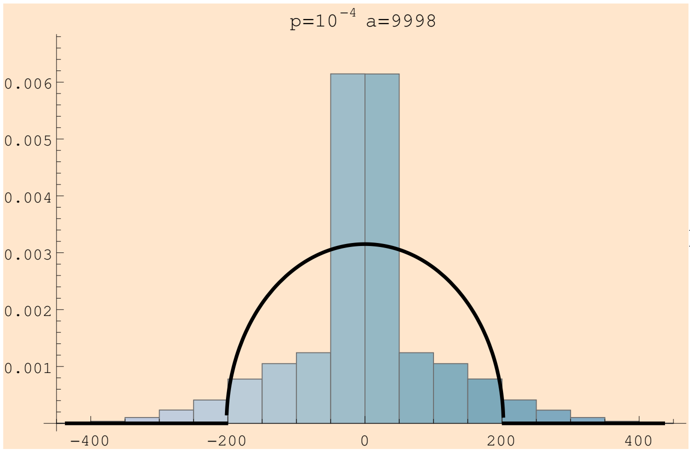


Figure F.2: Standard Tail Fattening

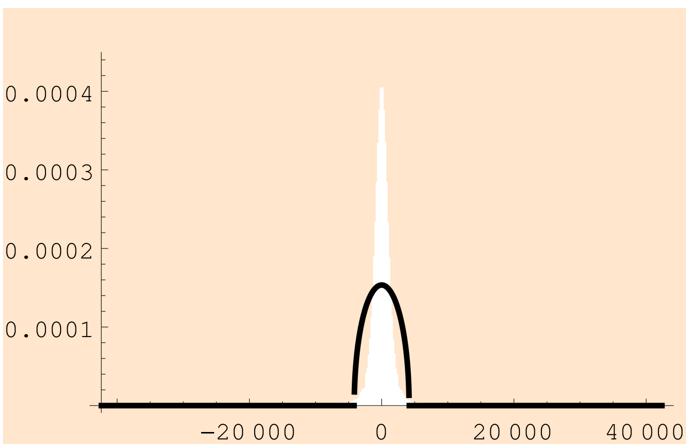


Figure F.3: Student T $\frac{3}{2}$

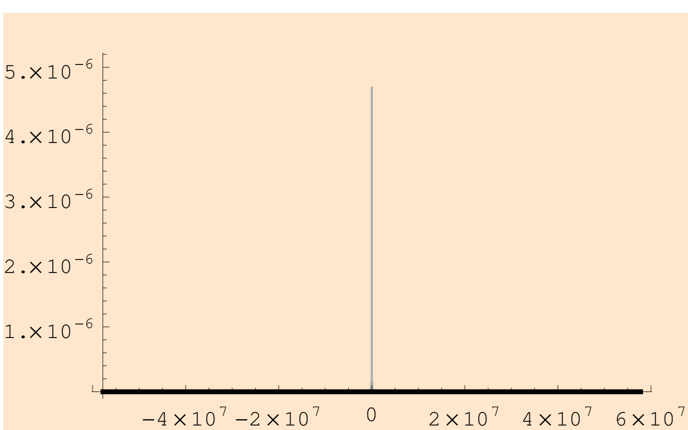


Figure F.4: Cauchy

8 SOME MISUSES OF STATISTICS IN SOCIAL SCIENCE

Chapter Summary **8: We apply the results of the previous chapter on the slowness of the LLN and list misapplication of statistics in social science, almost all of them linked to misinterpretation of the effects of fat-tailedness (and often from lack of awareness of fat tails), and how by attribute substitution researchers can substitute one measure for another. Why for example, because of chronic small-sample effects, the 80/20 is milder in-sample (less fat-tailed) than in reality and why regression rarely works.**

8.1 MECHANISTIC STATISTICAL STATEMENTS

Recall from the Introduction that the best way to figure out if someone is using an erroneous statistical technique is to use such technique on a dataset for which you have the answer. The best way to know the exact properties is to generate it by Monte Carlo. So the technique throughout the chapter is to generate fat-tailed data, the properties of which we know with precision, and check how such standard and mechanistic methods detect the *true* properties, then show the wedge between *observed* and *true* properties.

Also recall from Chapter 6 (D.1) that fat tails make it harder for someone to detect the true properties; for this we need a much, much larger dataset, more rigorous ranking techniques allowing inference in one direction not another (Chapter 4), etc. Hence this chapter is a direct application of the results and rules of Chapter 4.

8.2 ATTRIBUTE SUBSTITUTION

Attribute substitution 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 [118] , Hoggarth and Soyer, [104] and comment [107]) showing how statistical researchers overinterpret their own findings, as simplication leads to the *fooled by randomness* effect.

Dan Goldstein and this author (Goldstein and Taleb [53]) showed how professional researchers and practitioners substitute norms in the evaluation of higher order properties of time series, mistaking $\|x\|_1$ for $\|x\|_2$ (or $\frac{1}{n}\sum|x|$ for $\sqrt{\frac{\sum x^2}{n}}$). The common result is underestimating the randomness of the estimator M, in other words read too much into it (and, what is worse, underestimation of the tails, since, as we saw in 3.4, the ratio $\frac{\sqrt{\sum x^2}}{\sum|x|}$ increases with "fat-tailedness" to become infinite

The Small n Problem One often hears the statement "$n = 1$", or, worse, "the plural of anecdote is not data", a very, very representative (but elementary) violation of probability theory. It is very severe in effect for risk taking. For large deviations, $n = 1$ is plenty of data. To rule out large deviations, $n = 10^6$ can be small (as we saw with the law of large numbers under fat tails). Sample size should be a nonlinear proportion of the violation. The Chebychev distance, or norm $\mathcal{L}^\infty$ focuses on the largest measure (also see concentration functions, maximum of divergence (Lévy, Petrov), or even the standard and ubiquitous Kolmogorov-Smirnoff): looking at the extremum of a time series is not cherry picking since it is disconfirmatory evidence, the only true evidence one can get in statistics. Remarkably such people tend to also fall for the opposite mistake, the "n-large", in thinking that confirmatory observations provide "p-values". All these errors are magnified by fat tails.[a] [b]

[a] In addition to Paul Lévy and, of course, the Russians (see Petrov), there is an interesting literature on concentration functions, mostly in Italian (to wit, Gini): Finetti, Bruno (1953) : Sulla nozione di "dispersione" per distribuzioni a piu dimensioni, de Unione Roma. Gini, corrado (1914) : Sulla misura delia concentrazione delia variabilita dei caratteri. Atti del Reale Istituto Veneto di S. L. A., A. A. 1913-1914, 78, parte II, 1203-1248. Atti IV Edizioni- Congresso Cremonese,: La Matematica Italiana in (Taormina, 25-31 Ott. 1951), 587-596, astratto Giornale qualsiasi, (1955) deiristituto delle distribuzioni 18, 15-28. insieme translation in : de Finetti, Bruno struttura degli Attuari (1972).

[b] In ecology there is an interesting comedy of errors with the Séralini affair by which a collection of scientists (with some involvement from the firm Monsanto that has an evident track record of using lobbyists and a new breed of lobbyist-scientist) managed to get a safety-testing paper retracted from a journal (though subsequently republished in another one), allegedly because the claims made off small samples –although the samples were not particularly small compared to similar papers that were positive towards GMOs, and what is worse, the sample does not have to be particularly large for risk functions as the left tail grows with skepticism. The problem illustrates the failure to understand that disconfirmatory empiricism requires a different "n" than confirmatory ones.

under tail exponents $\alpha \geq 2$). 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.

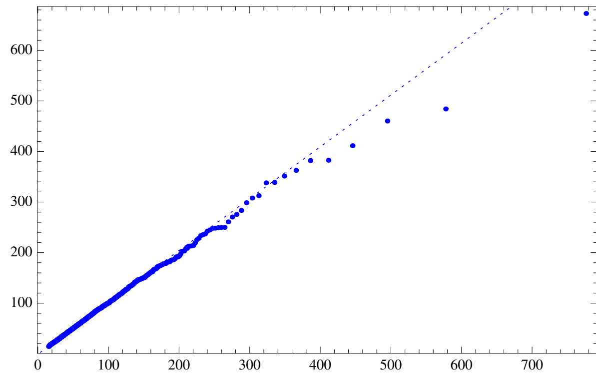


Figure 8.1: Q-Q plot" Fitting extreme value theory to data generated by its own process , the rest of course owing to sample insuficiency for extremely large values, a bias that typically causes the underestimation of tails, as the reader can see the points tending to fall to the right.

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 [88]), 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 $\mathbb{P}$.

8.3 THE TAILS SAMPLING PROPERTY

From the derivations in D.1, 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.

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)

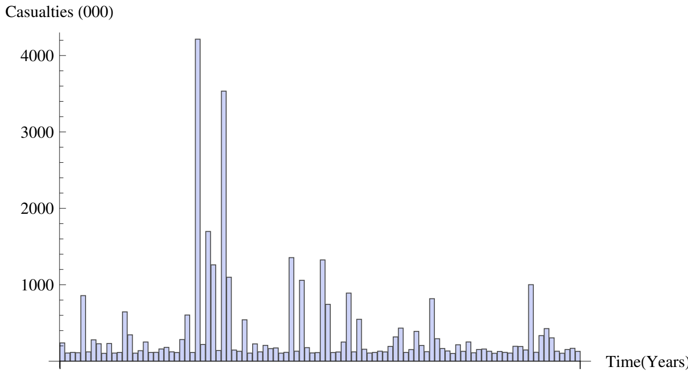


Figure 8.2: First 100 years (Sample Path): A Monte Carlo generated realization of a process for casualties from violent conflict of the "80/20 or 80/02 style", that is tail exponent α= 1.15

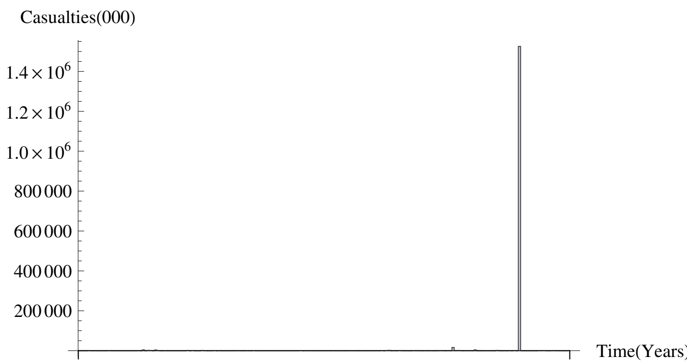


Figure 8.3: The Turkey Surprise: Now 200 years, the second 100 years dwarf the first; these are realizations of the exact same process, seen with a longer window and at a different scale.

Case Study: "Long Peace" Type 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 [88] claims that the generating process has a tail exponent $\sim$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.

Figures 8.2 and 8.3 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.

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.

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).

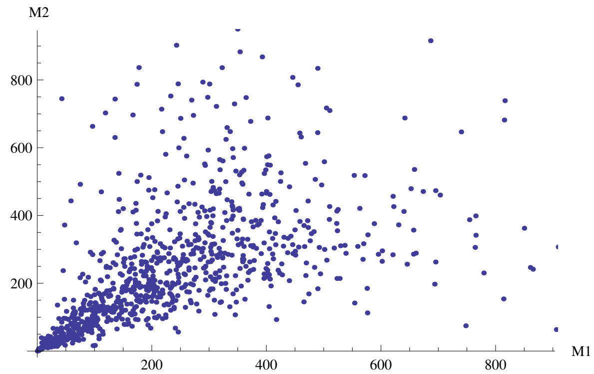


Figure 8.4: Does the past mean predict the future mean? Not so. M1 for 100 years,M2 for the next century. Seen at a narrow scale.

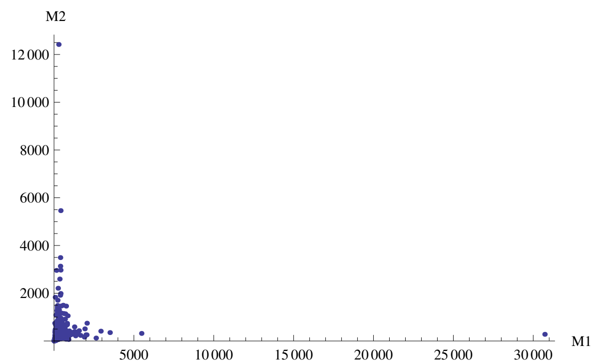


Figure 8.5: Does the past mean predict the future mean? Not so. M1 for 100 years,M2 for the next century. Seen at a wider scale.

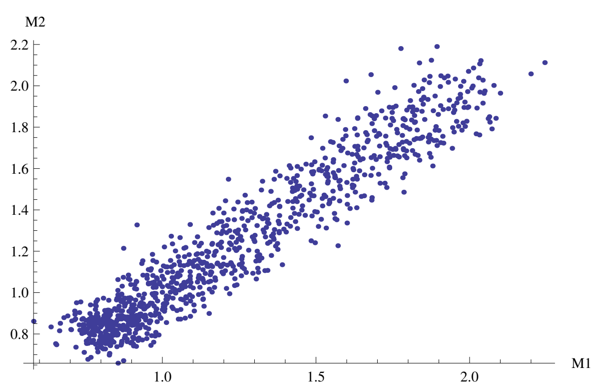


Figure 8.6: The same seen with a thin-tailed distribution.

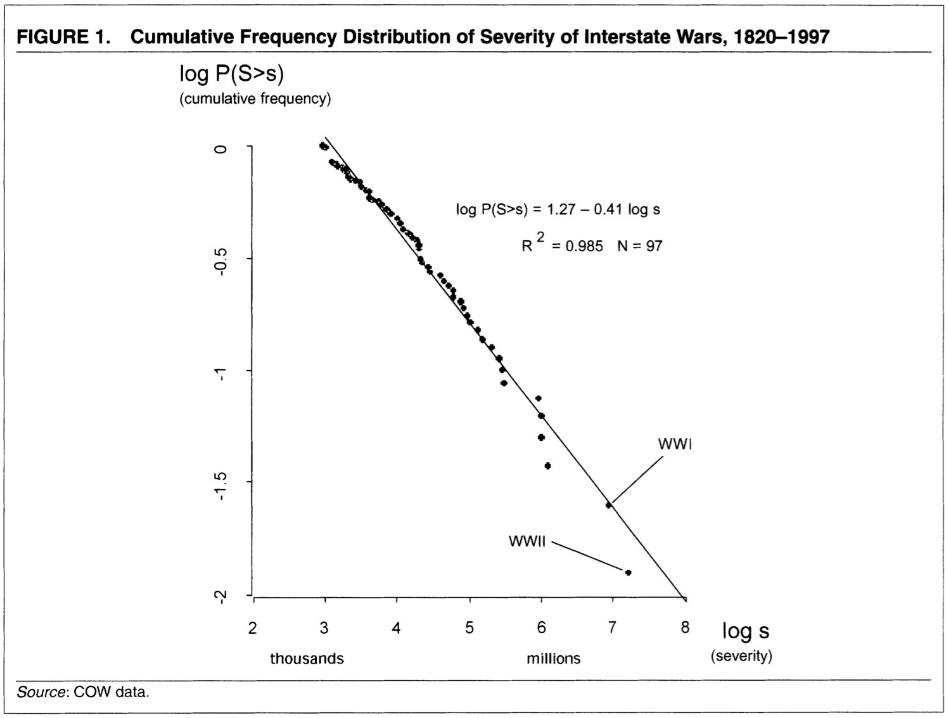


Figure 8.7: Cederman 2003, used by Pinker [88] . I wonder if I am dreaming or if the exponent α is really = .41. Chapters x and x show why such inference is centrally flawed, since *low exponents do not allow claims on mean of the variable*except to say that it is very, very high and not observable in 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.

Claims Made From Power Laws

The Cederman graph, Figure 8.7 shows exactly how *not* to make claims upon observing power laws.

8.4 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 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

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, α ≤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: Figure 8.8. 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).

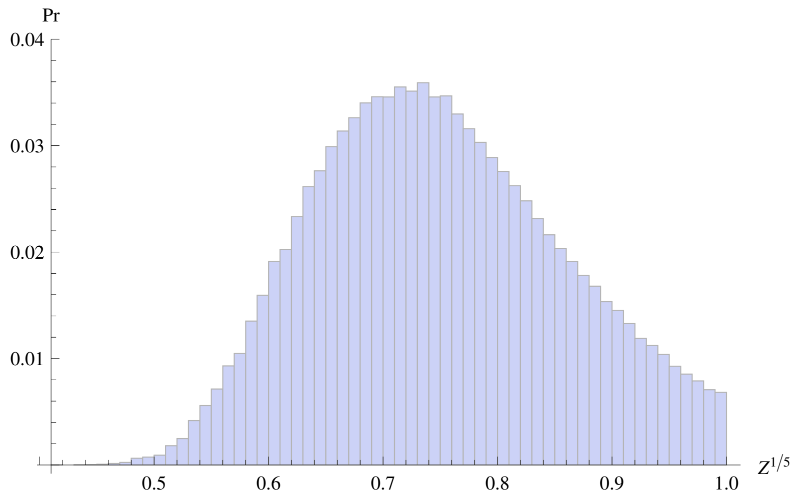


Figure 8.8: The difference betwen the generated (*ex ante*) and recovered (*ex post*) processes; $\nu = 20/100$, $N = 10^7$. Even when it should be 100/.0001, we tend to watch an average of 75/20

8.5 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 $S(t) = S(t - \Delta t) + Z_{\Delta t}$ where $Z_{\Delta t}$ is noise drawn from any distribution.

Geometric process: We can model it as $S(t) = S(t - \Delta t)e^{W_t}$ typically $S(t - \Delta t)e^{\mu \Delta t + s\sqrt{\Delta t} Z_t}$ but W_t can be noise drawn from any distribution. Typically, $\log\left(\frac{S(t)}{S(t-i\Delta t)}\right)$ 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)$e^{\mu \Delta t + s\sqrt{\Delta t} W_t}$, with $W_t \sim$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.

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

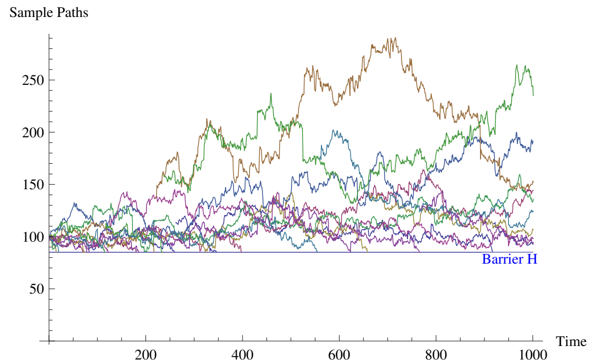


Figure 8.9: Counterfactual historical paths subjected to an absorbing barrier.

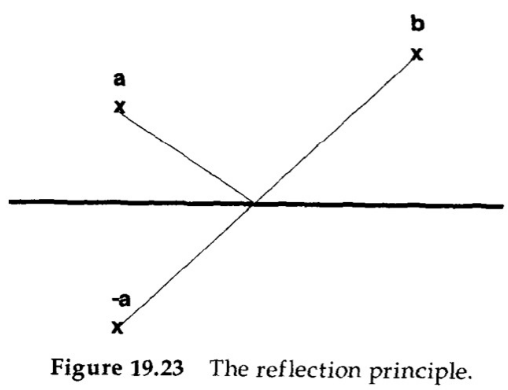


Figure 8.10: The reflection principle (graph from Taleb, 1997). The number of paths that go from point a to point b without hitting the barrier H is equivalent to the number of path from the point - a (equidistant to the barrier) to b.

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" $\geqslant$ "true mean"

Bias in the measurement of the volatility The "observed" variance (or mean deviation) $\leqslant$ "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.

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

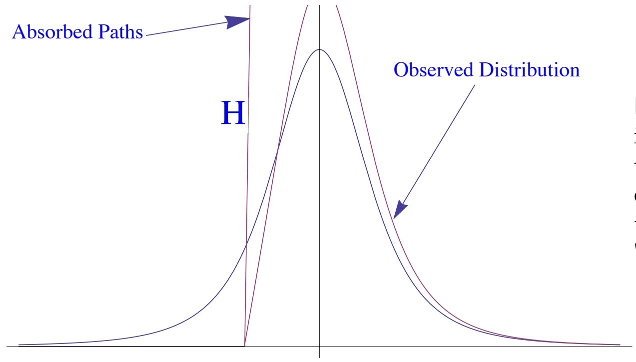


Figure 8.11: If you don't take into account the sample paths that hit the barrier, the observed distribution seems more positive, and more stable, than the "true" one.

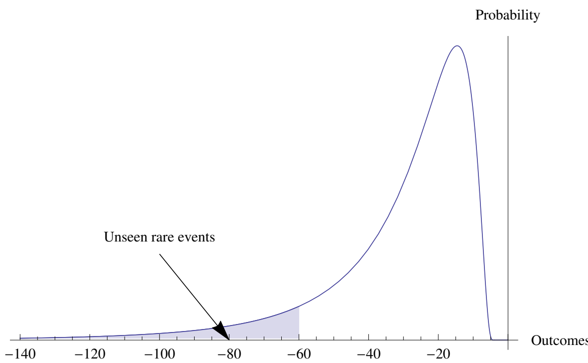


Figure 8.12: The left tail has fewer samples. The probability of an event falling below K in n samples is F(K), where F is the cumulative distribution.

$\int_{\infty}^{H} S\,(f(S) - f\,(S - 2\,(S_0 - H)))\,dS$ and perturbate $f(S)$ using the previously seen method to "fatten" the tail.

> 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.

8.6 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.

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 a sequence $(X_{1,j}, X_{2,j}, ...,X_{N,j})$ of random samples indexed by j as a designator of a certain history j. Measure

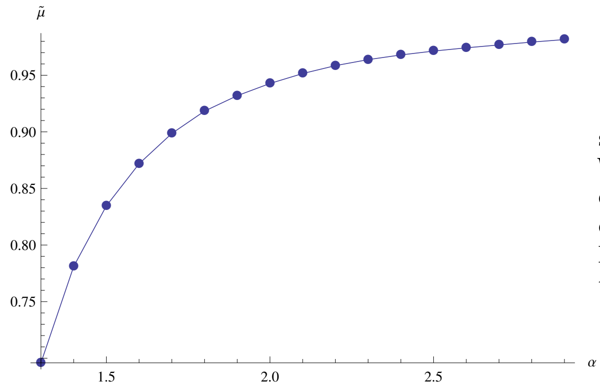


Figure 8.13: Median of $\sum_{j=1}^{T} \frac{\mu_j}{MT}$ in simulations (10^6 Monte Carlo runs). We can observe the underestimation of the mean of a skewed power law distribution as α exponent gets lower. Note that lower values of α imply fatter tails.

$\mu_j = \frac{\sum_{i=1}^{N} X_{i,j}}{N}$. We end up with the sequence of various sample means $\left(\mu_j\right)_{j=1}^{T}$, 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.

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.

8.7 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 [positive] skewness and fat tails, negative [positive] deviations from the mean are more informational than positive [negative] deviations.

8.8 THE INSTABILITY OF SQUARED VARIATIONS IN REGRESSIONS

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

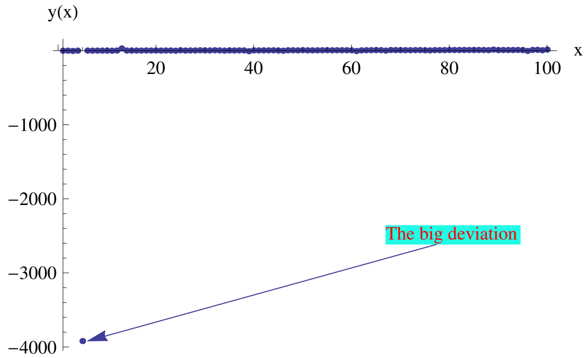


Figure 8.14: A sample regression path dominated by a large deviation. Most samples don't exhibit such deviation this, which is a problem. We know that with certainty (an application of the zero-one laws) that these deviations are certain as $n \to \infty$, so if one pick an arbitrarily large deviation, 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**.

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, $\mathbb{E}[z_i^n]$ = ∞ for all $n \geq 1$, 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.

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 sequence

$$\left(\left(\beta_0 + \beta_1 x_i + s z^{\alpha}_{i,j}\right)_{i=1}^{n}\right)_{j=1}^{T}.$$

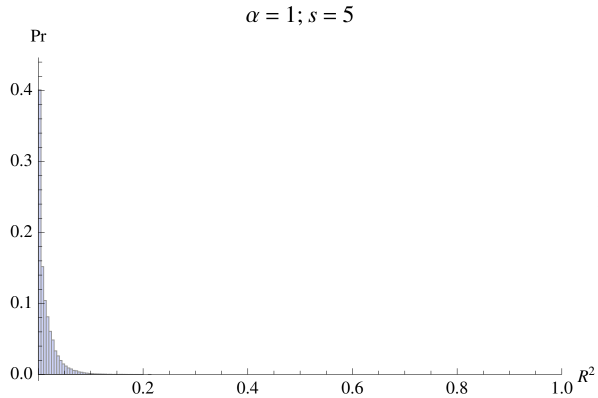


Figure 8.15: The histograms showing the distribution of R Squares; $T = 10^6$ simulations.The "true" R-Square should be 0. High scale of noise.

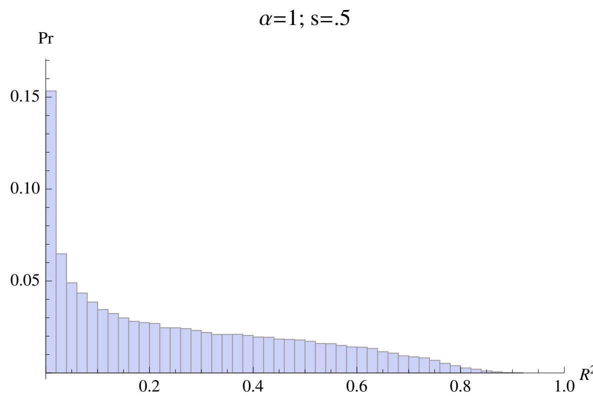


Figure 8.16: The histograms showing the distribution of R Squares; $T = 10^6$ simulations.The "true" R-Square should be 0. Low scale of noise.

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 $\text{E} \equiv (\epsilon_{i,j}) \equiv \left(\left(y_{i,j}^{\text{est}} - \beta_0^{\text{est}} - \beta_1^{\text{est}} x_{ij} \right)_{i=1}^{n} \right)_{j=1}^{T}$. We thus obtain T simulated values of $\rho \equiv (\rho_j)_{j=1}^{T}$, where $\rho_j \equiv 1 - \frac{\sum_{i=1}^{n} \epsilon_{i,j}{}^2}{\sum_{i=1}^{n} (y_{i,j} - \overline{y_j})^2}$, 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.

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	Std Error	T-Statistic	P-Value
1	4.99	0.417	11.976	7.8×10^{-33}
x	0.10	0.00007224	1384.68	9.3×10^{-11426}

Application to Economic Variables

We saw in G.G 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

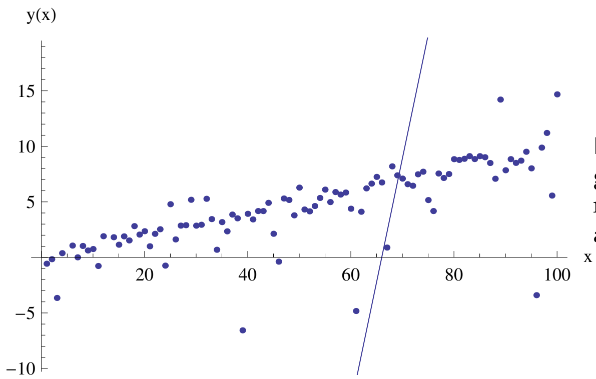


Figure 8.17: We can fit different regressions to the same story (which is no story). A regression that tries to accommodate the large deviation.

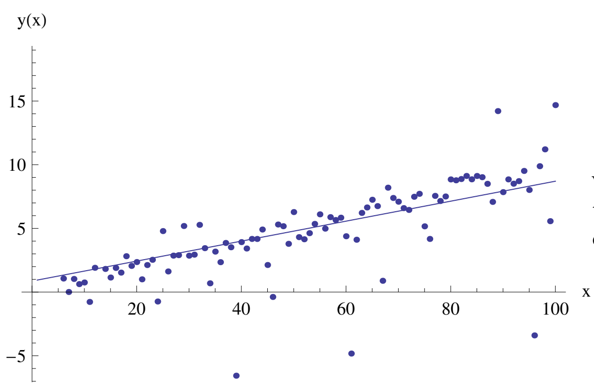


Figure 8.18: Missing the largest deviation (not necessarily voluntarily): the sample doesn't include the critical observation.

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.

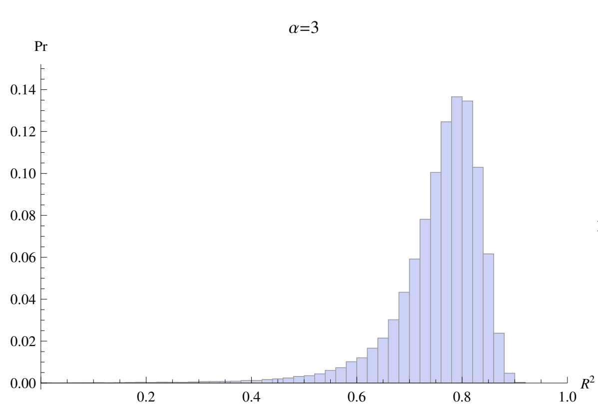


Figure 8.19: Finite variance but infinite kurtosis.

8.9 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 $\mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y)$ 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.

8.10 STUDYING THE STATISTICAL PROPERTIES OF BINARIES AND EXTENDING TO VANILLAS

See discussion in Chapter **??**. A lot of nonsense in discussions of rationality facing "dread risk" (such as terrorism or nuclear events) based on wrong probabilistic structures, such as comparisons of fatalities from falls from ladders to death from terrorism. The probability of falls from ladder doubling is 1 10^{20}. Terrorism is fat-tailed: similar claims cannot be made.

A lot of unrigorous claims like "long shot bias" is also discussed there.

8.11 WHY ECONOMICS TIME SERIES DON'T REPLICATE

(Debunking a Nasty Type of Misinference)

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.

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 [109]), using Log returns, $X_t \equiv \log\left(\frac{P(t)}{P(t-i\Delta t)}\right)$, 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 $\text{Max}(X_{t-i\Delta t}{}^4)_{i=0}^n$. $Q(n)$ is the contribution of the maximum quartic variations over n samples.

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

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 $\pm$ .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).

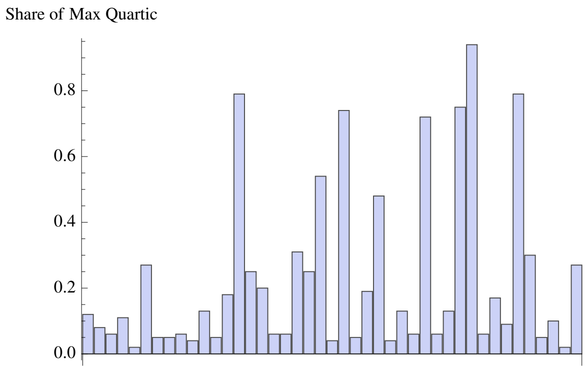


Figure 8.20: Max quartic across securities

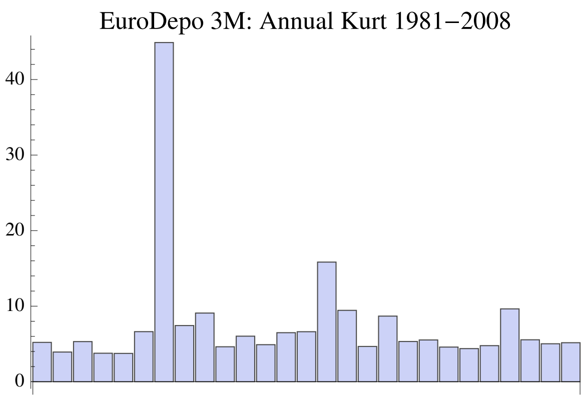


Figure 8.21: Kurtosis across nonoverlapping periods

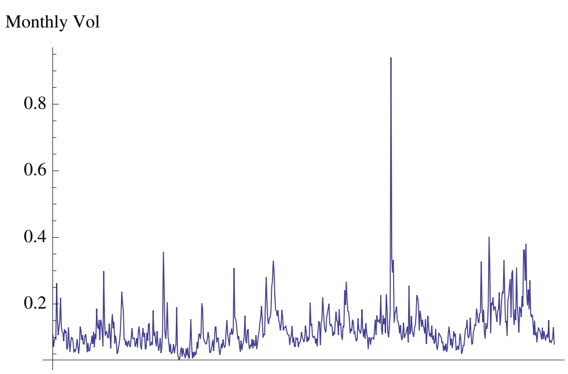


Figure 8.22: Monthly delivered volatility in the SP500 (as measured by standard deviations). The only structure it seems to have comes from the fact that it is bounded at 0. This is standard.

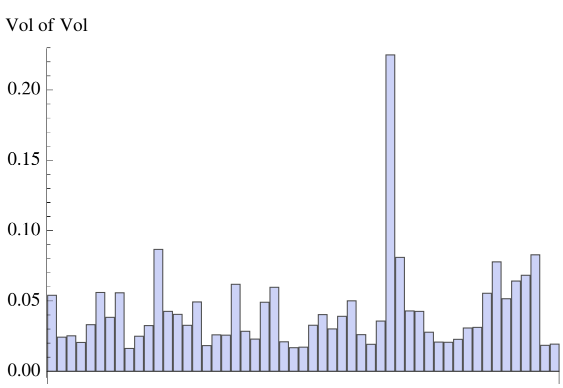


Figure 8.23: Montly volatility of volatility from the same dataset, predictably unstable.

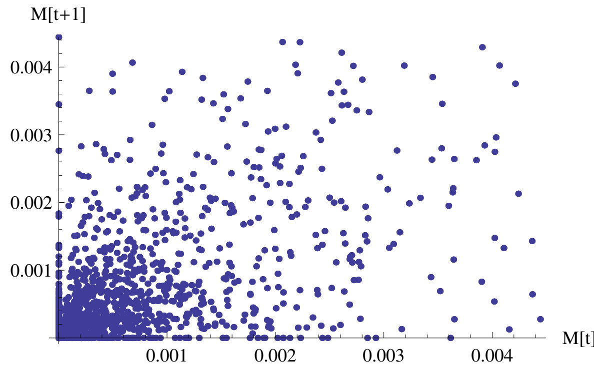


Figure 8.24: Comparing M[t-1, t] and M[t,t+1], where τ= 1year, 252 days, for macroeconomic data using extreme deviations, $A = (-\infty, -2$ STD (equivalent)], $f(x) = x$ (replication of data from *The Fourth Quadrant*, Taleb, 2009)

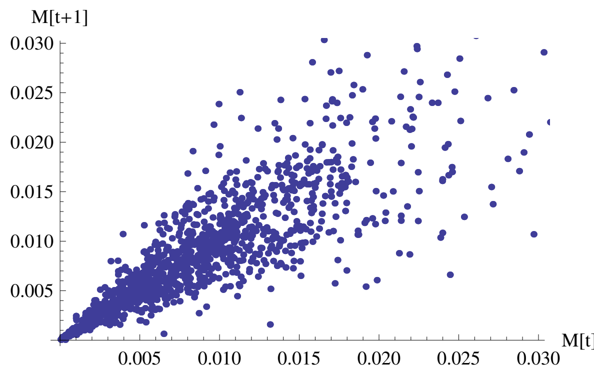


Figure 8.25: The "regular" is predictive of the regular, that is mean deviation. Comparing M[t] and M[t+1 year] for macroeconomic data using regular deviations, A= (-∞ ,∞), f(x)= |x|

Performance of Standard NonParametric 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.

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

When x_t are now in $\mathbb{R}^N$, the problems of sensitivity to changes in the covariance matrix makes the estimator M extremely unstable. Tail events for a vector are vastly more difficult to calibrate, and increase in dimensions.

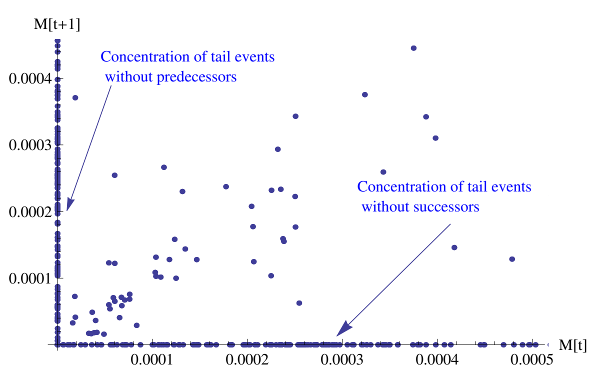


Figure 8.26: The figure shows how things get a lot worse for large deviations $A = (-\infty, -4)$ standard deviations (equivalent),$f(x) = x$

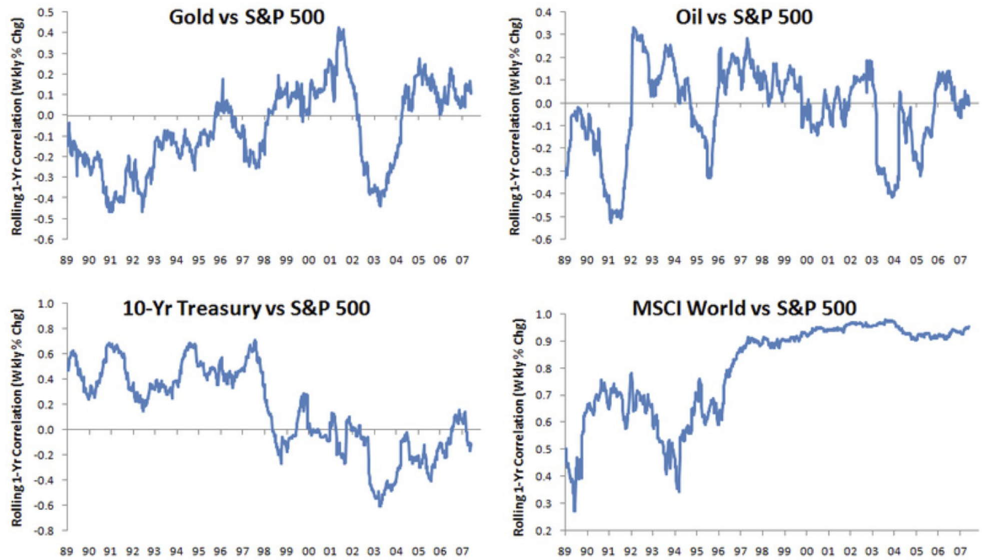


Figure 8.27: Correlations are also problematic, which flows from the instability of single variances and the effect of multiplication of the values of random variables.

The Responses so far by members of the economics/econometrics establishment : "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 [CITATION] 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 confusion about the Black Swan problem :

> "(...) "Black Swans" (Taleb, 2007). These cultural icons refer to disasters that occur so infrequently that they are virtually impossible to analyze using standard statistical 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 the common sense to look for robustness to these events instead of just focuing on probability).

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 acts of artistry. Basically these capture the ordinary and mask higher order effects. Since blowups are not frequent, these events do not show in data and the researcher looks smart most of the time while being fundamentally wrong. At the source, researchers, "quant" risk manager, and academic economist do not have skin in the game so they are not hurt by wrong risk measures: other people are hurt by them. And the artistry should continue perpetually so long as people are allowed to harm others with impunity. (More in Taleb and Sandis, 2013)

8.12 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.

8.13 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.

G ON THE INSTABILITY OF ECONOMETRIC DATA

Table 15: 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.

Table 15: (continued from previous page)

	K(1)	K(10)	K(66)	Max Quartic	Years
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.

9 FAT TAILS FROM RECURSIVE UNCERTAINTY

Second Version. An earlier version was presented at Benoit Mandelbrot's Scientific Memorial, New Haven, April 11, 2011, under the title: *The Future Will Be More Fat Tailed Than The Past*

Chapter Summary 9: **Error about Errors. Probabilistic representations require the inclusion of model (or representation) error (a probabilistic statement has to have an error rate), and, in the event of such treatment, one also needs to include second, third and higher order errors (about the methods used to compute the errors) and by a regress argument, to take the idea to its logical limit, one should be continuously reapplying the thinking all the way to its limit unless when one has a reason to stop, as a declared a priori that escapes quantitative and statistical method. We show how power laws emerge from nested errors on errors of the standard deviation for a Gaussian distribution. We also show under which regime regressed errors lead to non-power law fat-tailed distributions.**

9.1 LAYERING UNCERTAINTY

With the Central Limit Theorem: we start with a distribution and, under some conditions, end with a Gaussian. The opposite is more likely to be true. We start with a Gaussian and under error rates we end with a fat-tailed distribution.

Unlike with the Bayesian compounding the:

1. Numbers of recursions

   and

2. Structure of the error of the error (declining, flat, multiplicative or additive)

determine the final moments and the type of distribution.

Note that historically, 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 the projection of standard philosophical doubts into the future, in addition to regress arguments.

Missing the point

Savage, in his *Foundation of Statistics* [99]:

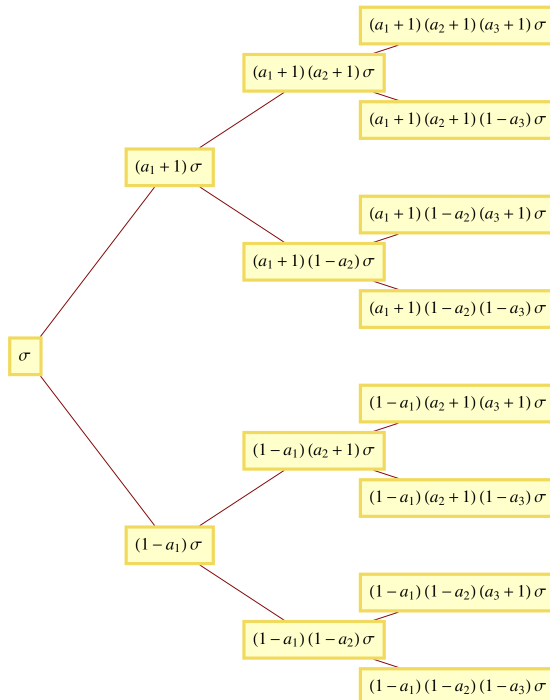


Figure 9.1: Three levels of multiplicative relative error rates for the standard deviation σ , with $(1 \pm a_n)$ the relative error on a_{n-1}

Estimate of the accuracy of estimates:
The doctrine is often expressed that a point estimate is of little, or no, value unless accompanied by an estimate of its own accuracy. This doctrine, which for the moment I will call the *doctrine of accuraty estimation*, may be a little old-fashioned, but 1 think some critical discussion of it here is in order for two reasons. In the first place, the doctrine is still widely considered to contain more than a grain of truth. For example, many readers will think it strange, and even remiss, that I have written a long chapter (Chapter 15) on estimation without even sugesting that an estimate should be accompanied by an estimate of its accuracy. In the second place, it seems to me that the concept of interval estimation, which is the subject of the next section, has largely evolved from the doctrine of accuracy estimation and that discussion of the doctrine will, for some, pave the way for discuasion of interval estimation. The doctrine of accuracy estimation is vague, even by the standards of the verbalistic tradition, for it does not say what should be taken as a measure of accuracy, that is, what an estimate of accuracy ahould estimate.

So we got diverted into the wrong direction for all these years as it did not hit Savage that we should perhaps see *what effect* would the estimation error have via structured perturbation.[1]

> Taking the doctrine literally, it evidently leads to endess regression for an estimate of the accuracy of an estimate should presumably be accompanied by an estimate of its own accuracy, and so on forever.

So as we will see in this chapter, we can actually see the effect with a great deal of clarity.

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 –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.

Main Results

Note that unless one stops the branching at an early stage, all the results raise small probabilities (in relation to their remoteness; the more remote the event, the worse the relative effect).

1. Under the first regime of proportional constant (or increasing) recursive layers of uncertainty about rates of uncertainty expressed as standard deviation, the distribution converges to a power law with infinite variance, even when one starts with a standard Gaussian.

2. Under the same first regime, expressing uncertainty about uncertainty in terms of variance, the distribution converges to a power law with finite variance but infinite (or undefined) higher moments.

3. Under the other regime, where the errors are decreasing (proportionally) for higher order errors, the ending distribution becomes fat-tailed but in a benign way as it retains its finite variance attribute (as well as all higher moments), allowing convergence to Gaussian under Central Limit.

We manage to set a boundary between these two regimes.

In both regimes the use of a thin-tailed distribution is not warranted unless higher order errors can be completely eliminated a priori.

1 I thank Dane Rook for the discussion.

Higher order integrals in the Standard Gaussian Case

We start with the case of a Gaussian and focus the uncertainty on the assumed standard deviation. Define $\phi(\mu,\sigma,x)$ as the Gaussian PDF for value x with mean μ and standard deviation σ.

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

$$f(x)_N = \int_0^{\infty} \dots \int_0^{\infty} \phi(\mu,\sigma,x) f\left(\bar{\sigma},\sigma_1,\sigma\right) \\ f\left(\overline{\sigma_1},\sigma_2,\sigma_1\right) \dots f\left(\overline{\sigma_{N-1}},\sigma_N,\sigma_{N-1}\right) \mathrm{d}\sigma\, \mathrm{d}\sigma_1\, \mathrm{d}\sigma_2 \dots \mathrm{d}\sigma_N \tag{9.1}$$

The problem is that this approach is parameter-heavy and requires the specifications of the subordinated distributions (in finance, the lognormal has been traditionally used for σ^2 (or Gaussian for the ratio $\text{Log}[\frac{\sigma_t^2}{\sigma^2}]$ since the direct use of a Gaussian allows for negative values). We would need to specify a measure f for each layer of error rate. Instead this can be approximated by using the mean deviation for σ, as we will see next[2].

Discretization using nested series of two-states for σ- a simple multiplicative process

There are quite effective simplifications to capture the convexity, the ratio of (or difference between) $\phi(\mu,\sigma,x)$ and $\int_0^{\infty} \phi(\mu,\sigma,x) f\left(\bar{\sigma},\sigma_1,\sigma\right) \mathrm{d}\sigma$ (the first order standard deviation) by using a weighted average of values of σ, say, for a simple case of one-order stochastic volatility:

$$\sigma(1 \pm a_1)$$

with $0 \leq a_1 < 1$, 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. Such a method does not aim at preserving the variance as in standard stochastic volatility modeling, rather the STD.

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

$$f(x)_1 = \frac{1}{2}\bigg(\phi(\mu,\sigma\,(1+a_1),x) + \phi(\mu,\sigma(1-a_1),x)\bigg) \tag{9.2}$$

2 A well developed technique for infinite (or non integrable) Gaussian cumulants, now, is the Wiener Chaos expansion [85].

Now assume uncertainty about the error rate a_1, expressed by a_2, in the same manner as before. Thus, as a first method, the multiplicative effect, in place of $1 \pm a_1$ we have $(1 \pm a_1)(1 \pm a_2)$. Later we will use the non-multiplicative (or, rather, weakly multiplicative) error expansion $\sigma(1 \pm (a_1(1 \pm (a_2(1 \pm a_3(\ldots)))$.

The second order stochastic standard deviation:

$$f(x)_2 = \frac{1}{4}\Bigg(\phi\Big(\mu, \sigma(1 + a_1)(1 + a_2), x\Big) +$$
$$\phi\Big(\mu, \sigma(1 - a_1)(1 + a_2), x) + \phi(\mu, \sigma(1 + a_1)(1 - a_2), x\Big) + \phi\Big(\mu, \sigma(1 - a_1)(1 - a_2), x\Big)\Bigg) \tag{9.3}$$

and the N^{th} order:

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

where M_i^N is the 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 $(1, 1, 1, \ldots)$ 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

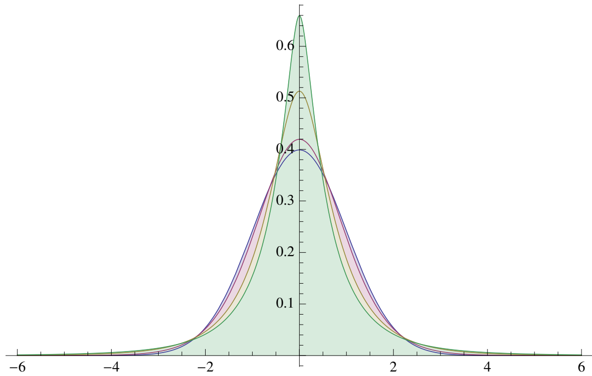


Figure 9.2: Thicker tails (higher peaks) for higher values of N; here $N = 0, 5, 10, 25, 50$, all values of a=$\frac{1}{10}$

$$M^3 = \begin{pmatrix} (1-a_1)\,(1-a_2)\,(1-a_3) \\ (1-a_1)\,(1-a_2)\,(a_3+1) \\ (1-a_1)\,(a_2+1)\,(1-a_3) \\ (1-a_1)\,(a_2+1)\,(a_3+1) \\ (a_1+1)\,(1-a_2)\,(1-a_3) \\ (a_1+1)\,(1-a_2)\,(a_3+1) \\ (a_1+1)\,(a_2+1)\,(1-a_3) \\ (a_1+1)\,(a_2+1)\,(a_3+1) \end{pmatrix}$$

So $M_1^3 = ((1-a_1)(1-a_2)(1-a_3))$, etc.

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.

9.2 REGIME 1 (EXPLOSIVE): CASE OF A CONSTANT ERROR 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.

$$f(x|\mu, \sigma, N) = 2^{-N} \sum_{j=0}^{N} \binom{N}{j} \phi\left(\mu, \sigma(a+1)^j (1-a)^{N-j}, x\right) \tag{9.4}$$

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*).

$M_1(N) = \mu$
$M_2(N) = \sigma^2 \left(a^2+1\right)^N + \mu^2$
$M_3(N) = 3\ \mu\sigma^2 \left(a^2+1\right)^N + \mu^3$
$M_4(N) = 6\ \mu^2\sigma^2 \left(a^2+1\right)^N + \mu^4 + 3\left(a^4+6a^2+1\right)^N \sigma^4$

For clarity, we simplify the table of moments, with μ=0

$M_1(N) = 0$
$M_2(N) = \left(a^2+1\right)^N \sigma^2$
$M_3(N) = 0$
$M_4(N) = 3\left(a^4+6a^2+1\right)^N \sigma^4$
$M_5(N) = 0$
$M_6(N) = 15\left(a^6+15a^4+15a^2+1\right)^N \sigma^6$
$M_7(N) = 0$
$M_8(N) = 105\left(a^8+28a^6+70a^4+28a^2+1\right)^N \sigma^8$

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.

- Even the smallest value of $a > 0$, since $\left(1+a^2\right)^N$ is unbounded, leads to the second moment going to infinity (though not the first) when N$\to \infty$. So

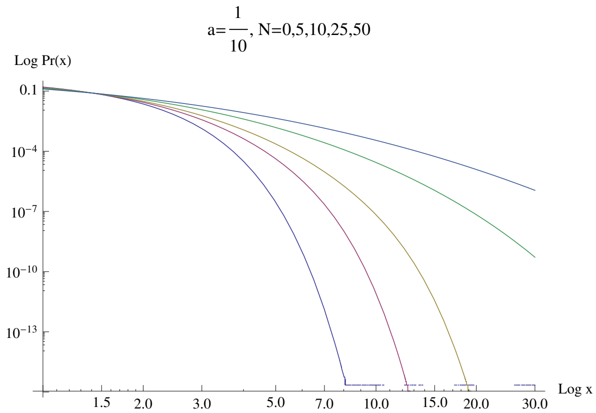


Figure 9.3: LogLog Plot of the probability of exceeding x showing power law-style flattening as N rises. Here all values of a= 1/10

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.

- 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.

9.3 CONVERGENCE TO POWER LAWS

Convergence to power law would require the following from the limit distribution. Where $P_{>x}$ is short for $P(X > x)$, $P_{>x} = L(x)\ x^{-\alpha^*}$ and $L(x)$ is a slowly varying function.

$$\alpha^* = \lim_{x\to\infty} \lim_{N\to\infty} \alpha(x, N)$$

We know from the behavior of moments that, if convergence is satisfied, $\alpha^* \in (1, 2)$.

We can have a visual idea with the Log-Log plot (Figure 9.3) how, at higher orders of stochastic volatility, with equally proportional stochastic coefficient, (where $a_1 = a_2 = \ldots = a_n = \frac{1}{10}$) the density approaches that of a power law, as shown in flatter density on the LogLog plot. The probabilities keep rising in the tails as we add layers of uncertainty until they seem to reach the boundary of the power law, while ironically the first moment remains invariant.

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

$$\alpha(x, N) = -1 - \frac{\frac{\partial \log f(x|\mu,\sigma,N)}{\partial x}}{\frac{\partial \log(x)}{\partial x^1}}$$

Simplifying and normalizing, with $\mu = 0$, $\sigma = 1$,

$$\alpha(x, N) = -1 - \frac{x\ \kappa_1(N)}{\kappa_2(N)} \tag{9.5}$$

where

$$\kappa_1(N) = \sum_{j=0}^{K} x(a+1)^{-3j} \left(-(1-a)^{3j-3K}\right) \binom{K}{j} \exp\left(-\frac{1}{2}x^2(a+1)^{-2j}(1-a)^{2j-2K}\right)$$

$$\kappa_2(N) = \sum_{j=0}^{K} (a+1)^{-j}(1-a)^{j-K} \binom{K}{j} \exp\left(-\frac{1}{2}x^2(a+1)^{-2j}(1-a)^{2j-2K}\right)$$

Making the variable continuous (binomial as ratio of gamma functions) makes it equivalent, at large N, to:

$$\alpha(x, N) = 1 - \frac{x(1-a)^N \kappa_1(N)}{\sqrt{2}\ \kappa_2(N)} \tag{9.6}$$

where

$$\kappa_1^*(N) = \int_0^N -\frac{x(a+1)^{-3y}\Gamma(N+1)(1-a)^{3(y-N)}}{\Gamma(y+1)\Gamma(N-y+1)} \exp\left(-\frac{1}{2}x^2(a+1)^{-2y}(1-a)^{2y-2N}\right) \mathrm{d}y$$

$$\kappa_2^*(N) = \int_0^N \frac{\left(\frac{2}{a+1}-1\right)^y \Gamma(N+1)}{\sqrt{2}\,\Gamma(y+1)\Gamma(N-y+1)} \exp\left(-\frac{1}{2}x^2(a+1)^{-2y}(1-a)^{2y-2N}\right) \mathrm{d}y$$

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:

$$P > K|N = \sum_{j=0}^{N} 2^{-N-1} \binom{N}{j} \operatorname{erfc}\left(\frac{K}{\sqrt{2}\sigma(a+1)^j(1-a)^{N-j}}\right) \tag{9.7}$$

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

Convexity effect The next two tables shows the ratio of exceedant probability under different values of N divided by the probability in the case of a standard Gaussian.

Table 16: 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 17: 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}

9.4 REGIME 1B: PRESERVATION OF VARIANCE

$$\begin{aligned} M_1(N) &= \mu \\ M_2(N) &= \mu^2 + \sigma^2 \\ M_3(N) &= \mu^3 + 3\sigma^2\mu \\ M_4(N) &= 3\sigma^4 \left(a^2+1\right)^N + \mu^4 + 6\mu^2\sigma^2 \end{aligned}$$

Hence $\alpha \in (3,4)$

9.5 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).

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 $a_n = \lambda\, a_{N-1}$, hence $a_N = \lambda^N a_1$, with a_1 the conventional intensity of stochastic standard deviation. Assume $\mu = 0$.

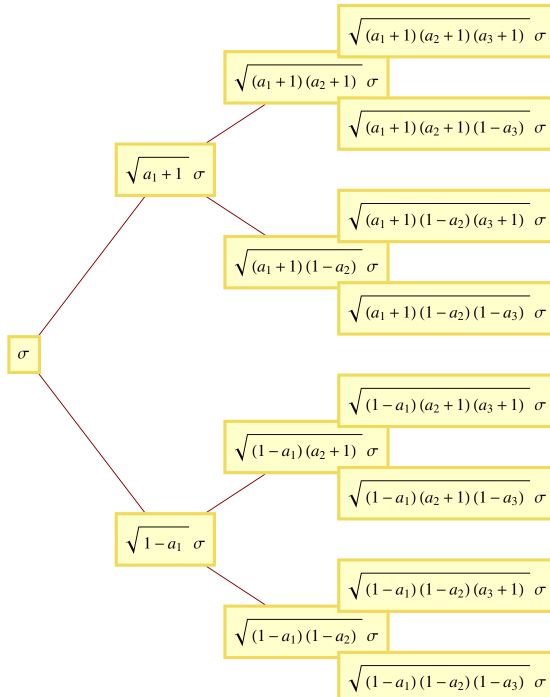


Figure 9.4: Preserving the variance

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:

$$M_3(N) = \left(a_1^2+1\right)\sigma^2\prod_{i=1}^{N-1}\left(a_1^2\lambda^{2i}+1\right) \tag{9.8}$$

We can reexpress (9.8) 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:

$$M_4(N) = 3\sigma^4 \prod_{i=0}^{N-1} \left(6a_1^2\lambda^{2i} + a_1^4\lambda^{4i} + 1\right)$$

$$M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2}-3\right) a_1^2; \lambda^2\right)_N \left(-\left(3+2\sqrt{2}\right) a_1^2; \lambda^2\right)_N \quad (9.9)$$

$$\lim_{N\to\infty} M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2}-3\right) a_1^2; \lambda^2\right)_\infty \left(-\left(3+2\sqrt{2}\right) a_1^2; \lambda^2\right)_\infty \quad (9.10)$$

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.

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

In place of $(1 \pm a_1)(1 \pm a_2)$, we use, for N recursions,

$$\sigma(1 \pm (a_1(1 \pm (a_2(1 \pm a_3(\dots)))$$

Assume $a_1 = a_2 = \dots = a_N$

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

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

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

So for instance, for $N = 3$, $\mathbf{T} = \left(1, a, a^2, a^3\right)$

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

The moments are as follows:

$$M_1(N) = \mu$$

$$M_2(N) = \mu^2 + 2\sigma$$

$$M_4(N) = \mu^4 + 12\mu^2\sigma + 12\sigma^2 \sum_{i=0}^{N} a^{2i}$$

At the limit:

$$\lim_{N\to\infty} M_4(N) = \frac{12\sigma^2}{1-a^2} + \mu^4 + 12\mu^2\sigma$$

which is very mild.

9.6 CONCLUSION AND SUGGESTED APPLICATION

Counterfactuals, Estimation of the Future v/s Sampling Problem

Note that it is hard to escape higher order uncertainties, even outside of the use of counterfactual: even when sampling from a conventional population, an error rate can come from the production of information (such as: is the information about the sample size correct? is the information correct and reliable?), etc. These higher order errors exist and could be severe in the event of convexity to parameters, but they are qualitatively different with forecasts concerning events that have not taken place yet.

This discussion is about an epistemic situation that is markedly different from a sampling problem as treated conventionally by the statistical community, particularly the Bayesian one. In the classical case of sampling by Gosset ("Student", 1908) from a normal distribution with an unknown variance (Fisher, 1925), the Student T Distribution (itself a power law) arises for the estimated mean since the square of the variations (deemed Gaussian) will be Chi-square distributed. The initial situation is one of relatively unknown variance, but that is progressively discovered through sampling; and the degrees of freedom (from an increase in sample size) rapidly shrink the tails involved in the underlying distribution.

The case here is the exact opposite, as we have an a priori approach with no data: we start with a known priorly estimated or "guessed" standard deviation, but with an unknown error on it expressed as a spread of branching outcomes, and, given the a priori aspect of the exercise, we have no sample increase helping us to add to the information and shrink the tails. We just deal with nested "counterfactuals" (actually the equivalent of couterfactuals projected into the future).

Note that given that, unlike the Gosset's situation, we have a finite mean (since we don't hold it to be stochastic and know it a priori) hence we necessarily end in a situation of finite first moment (hence escape the Cauchy distribution), but, as we will see, a more complicated second moment. [3] [4]

The Future is Fatter Tailed Than The Past

A simple application of these derivations: It shows why any uncertainty about the link between the past and the future leads to underestimation of fat tails.

3 See the discussion of the Gosset and Fisher approach in Chapter 3 of Mosteller and Tukey [78].
4 I thank Andrew Gelman and Aaron Brown for the discussion.

10 PARAMETRIZATION AND TAILS

Chapter Summary **10: We present 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 one believed the r.v. was power law distributed, one still would not be able to make a precise statement on tail risks. Shows weaknesses of calibration of Extreme Value Theory.**

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.

10.1 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.

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" $S_+ = \int_K^\infty x\phi(x)\mathrm{d}x$ 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,

$$\frac{\partial S_+}{\partial \alpha} = -\frac{K^{1-\alpha}((\alpha-1)\alpha\log(K)+1)}{(\alpha-1)^2}.$$

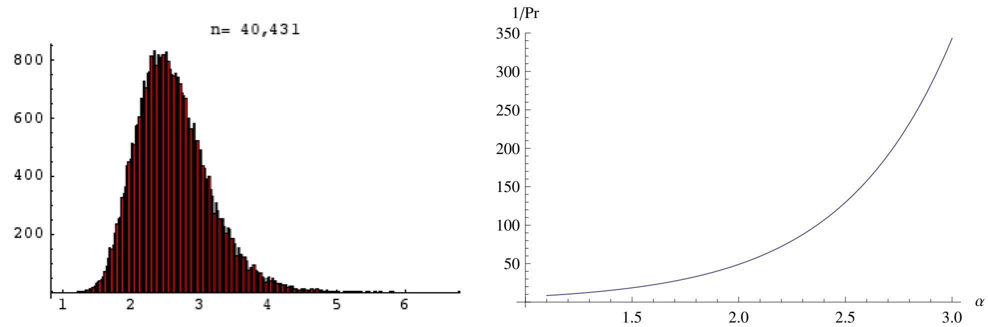


Figure 10.1: The effect of small changes in tail exponent on a probability of exceeding a certain point. To the left, a histogram of possible tail exponents across >4 10^3 variables. To the right the probability, probability of exceeding 7 times the scale of a power law ranges from 1 in 10 to 1 in 350. For further in the tails the effect is more severe.

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.

10.2 EXTREME VALUE THEORY: NOT A PANACEA

We saw earlier how difficult it is to compute risks using power laws, owing to excessive model sensitivity. Let us apply this to the Extreme Value Theory, EVT. (The idea is that is useable by the back door as test for nonlinearities exposures not to get precise probabilities).

On its own it can mislead. 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.

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

$$M_j = \max\left(X_{i,j}\right)_{i=1}^n$$

Assume we do so p times, to get p samples of maxima for the sequence M, $M = \max\left(\left(X_{i,j}\right)_{i=1}^n\right)_{j=1}^p$.

Figure 10.2 and 10.2 plot a histogram of the result of both the simulation and the fitting of a distribution.

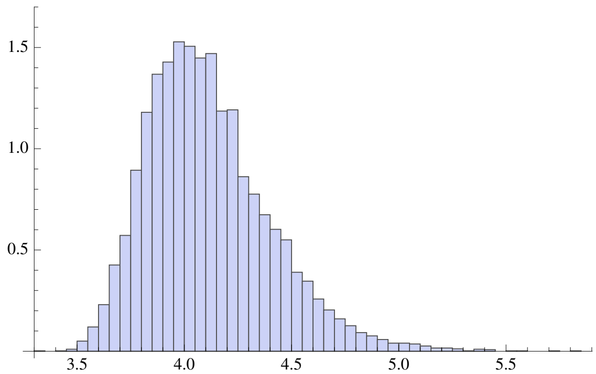


Figure 10.2: Taking p samples of Gaussian maxima; here $N = 30K$, $M = 10K$. We get the Mean of the maxima = 4.11159, Standard Deviation= 0.286938; Median = 4.07344

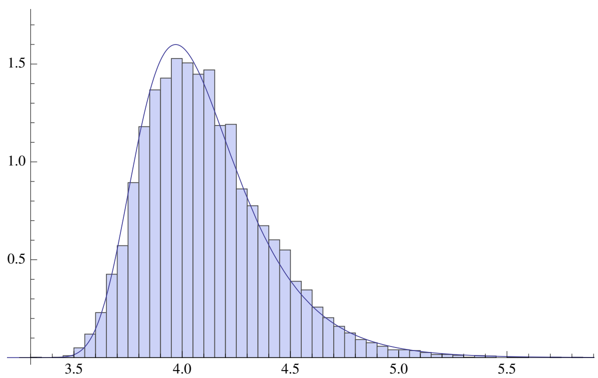


Figure 10.3: Fitting an extreme value distribution (Gumbel for the maxima) α= 3.97904, β= 0.235239

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: $g(x;\alpha,\beta) = \frac{e^{\frac{\alpha-x}{\beta}-e^{\frac{\alpha-x}{\beta}}}}{\beta}$.

Some Intuition: 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.

$$P\left(X_1 < x, X_2 < x, \ldots, X_n < x\right) = \bigcap_{i=1}^{n} P(X_i) = F(x)^n,$$

where F is the cumulative d.f of the Gaussian. Taking the first derivative of the cumulative distribution to get the density of the distribution of the maximum,

$$p_n(x) \equiv \partial_x \left(F(x)^n\right) = -\frac{2^{\frac{1}{2}-n} n e^{-\frac{x^2}{2}} \left(\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)+1\right)^{n-1}}{\sqrt{\pi}}$$

Now we have norming constants $a_n$and b_n such that

$$G(x) \equiv P\left(\frac{M(n)-a_n}{b_n} > x\right).$$

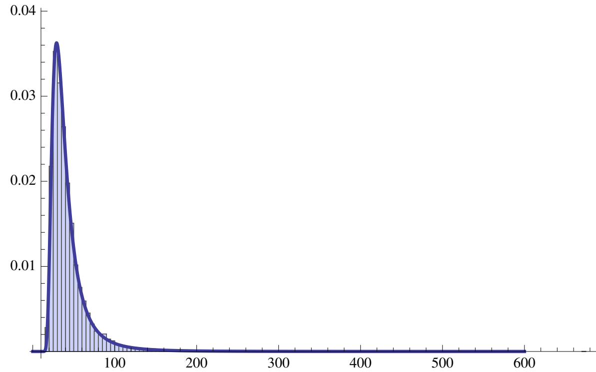


Figure 10.4: Fitting a Fréchet distribution to the Student T generated with α=3 degrees of freedom. The Frechet distribution α=3, β=32 fits up to higher values of E.But next two graphs shows the fit more closely.

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 $\text{erfc}^{-1}$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 + $o\left((\log n)^{-\frac{1}{2}}\right)$

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

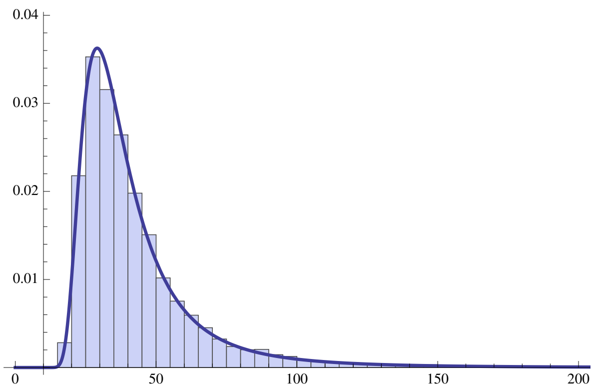


Figure 10.5: Seen more closely.

α	$\frac{1}{P_{>3\beta}}$	$\frac{1}{P_{>10\beta}}$	$\frac{1}{P_{>20\beta}}$	$\frac{1}{P_{>40\beta}}$	$\frac{1}{P_{>80\beta}}$
1.	4.	11.	21.	41.	81.
1.25	4.	18.	43.	101.	240.
1.5	6.	32.	90.	253.	716.
1.75	7.	57.	190.	637.	2140.
2	10.	101.	401.	1601.	**6400**
2.25	12.	178.	846.	4024.	19141.
2.5	16.	317.	1789.	10120.	57244.
2.75	21.	563.	3783.	25449.	171198.
3.	28.	1001.	8001.	64001.	512001.
3.25	36.	1779.	16918.	160952.	$\mathbf{1.5 \times 10^6}$
3.5	47.	3163.	35778.	404772.	4.5×10^6
3.75	62.	5624.	75660.	1.01×10^6	1.3×10^7
4.	82.	10001.	160001.	2.56×10^6	4.0×10^7
4.25	107.	17783.	338359.	6.43×10^6	1.2×10^8
4.5	141.	31623.	715542.	1.61×10^7	3.6×10^8
4.75	185.	56235.	1.5×10^6	4.07×10^7	1.1×10^9
5.	244.	100001.	3.2×10^6	1.02×10^8	3.27×10^9

Table 18: EVT for different tail parameters α. We can see how a perturbation of α moves the probability of a tail event from 6,000 to 1.5×10^6 . [ADDING A TABLE FOR HIGHER DIMENSION WHERE THINGS ARE A LOT WORSE]

A Severe Inverse Problem for EVT

In the previous case we started with the distribution, with the assumed parameters, then obtained the corresponding values, just 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.

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.

10.3 USING POWER LAWS WITHOUT BEING HARMED BY MISTAKES

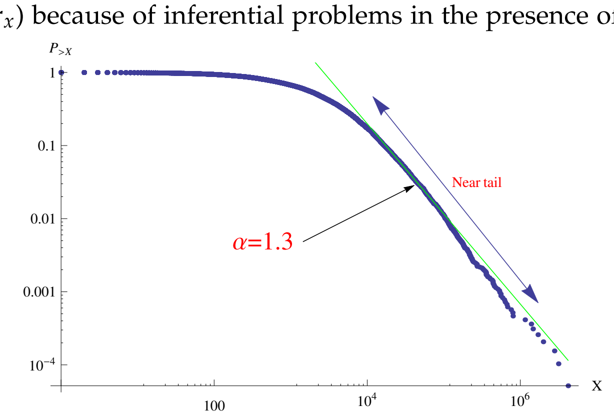
We can use power laws in the "near tails" for information, not risk management. That is, not pushing outside the tails, staying within a part of the distribution for which errors are not compounded.

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.



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".

H | POISSON VS. POWER LAW TAILS

H.1 BEWARE THE POISSON

By the **masquerade problem**, any power law can be seen backward as a Gaussian plus a series of simple (that is, noncompound) Poisson jumps, the so-called jump-diffusion process. So the use of Poisson is often just a backfitting problem, where the researcher fits a Poisson, happy with the "evidence".

The next exercise aims to supply convincing evidence of scalability and NonPoissonness of the data (the Poisson here is assuming a standard Poisson). Thanks to the need for the probabililities add up to 1, scalability in the tails is the sole possible model for such data. We may not be able to write the model for the full distribution –but we know how it looks like in the tails, where it matters.

The Behavior of Conditional Averages With a scalable (or "scale-free") distribution, when K is "in the tails" (say you reach the point when $1 - F(X > x) = Cx^{-\alpha}$ where C is a constant and α the power law exponent), the relative conditional expectation of X (knowing that X $>K$) divided by K, that is, $\frac{E[X|X>K]}{K}$ is a constant, and does not depend on K. More precisely, the constant is $\frac{\alpha}{\alpha-1}$.

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

This provides for a handy way to ascertain scalability by raising K and looking at the averages in the data.

Note further that, for a standard Poisson, (too obvious for a Gaussian): not only the conditional expectation depends on K, but it "wanes", i.e.

$$\lim_{K\to\infty} \left(\frac{\int_K^\infty \frac{m^x}{\Gamma(x)}\, dx}{\int_K^\infty \frac{m^x}{x!}\, dx} \Big/ K \right) = 1$$

Calibrating Tail Exponents In addition, we can calibrate power laws. Using K as the cross-over point, we get the α exponent above it –the same as if we used the Hill estimator or ran a regression above some point.

We heuristically defined fat tails as the contribution of the low frequency events to the total properties. But fat tails can come from different classes of distributions. This chapter will present the difference between two broad classes of distributions.

This brief test using 12 million pieces of exhaustive returns shows how equity prices (as well as short term interest rates) do not have a characteristic scale. No other possible method than a Paretan tail, albeit of unprecise calibration, can charaterize them.

H.2 LEAVE IT TO THE DATA

This exercise was done using about every piece of data in sight: single stocks, macro data, futures, etc.

Equity Dataset We collected the most recent 10 years (as of 2008) of daily prices for U.S. stocks (no survivorship bias effect as we included companies that have been delisted up to the last trading day), n= 11,674,825 , deviations expressed in logarithmic returns.
We scaled the data using various methods.

The expression in "numbers of sigma" or standard deviations is there to conform to industry language (it does depend somewhat on the stability of sigma). In the "MAD" space test we used the mean deviation.

$$\text{MAD}(i) = \frac{\log \frac{S^i_t}{S^i_{t-1}}}{\frac{1}{N}\sum_{t\leq n}\left|\log \frac{S^i_{t-j}}{S^i_{-j+t-1}}\right|}$$

We focused on negative deviations. We kept moving K up until to 100 MAD (indeed) –and we still had observations.

$$\text{Implied}\alpha|_K = \frac{E\left[X|_{X<K}\right]}{E\left[X|_{X<K}\right] - K}$$

MAD	$E\left[X\vert_{X<K}\right]$	$n(\text{for}X < K)$	$\frac{E[X\vert_{X<K}]}{K}$	Impliedα
$-1.$	-1.75	1.32×10^6	1.75	2.32
$-2.$	-3.02	300806.	1.51	2.95
$-5.$	-7.96	19285.	1.59	2.68
$-10.$	-15.32	3198.	1.53	2.87
$-15.$	-22.32	1042.	1.48	3.04
$-20.$	-30.24	418.	1.51	2.95
$-25.$	-40.87	181.	1.63	2.57
$-50.$	-101.75	24.	2.03	1.96
$-70.$	-156.70	11.	2.23	1.80
$-75.$	-175.42	9.	2.33	1.74
$-100.$	-203.99	7.	2.03	1.96

Sigma-Space In the "sigma space" test we used a rolling 22 day window scaled by the noncentral standard deviations. We did not add a mean for reasons explained elsewhere.

Short term Interest Rates Literally, you do not even have a large number K for which scalability drops from a small sample effect.

STD	$E[X\|_{X<K}]$	$n(\text{for} X < K)$	$\frac{E[X\|_{X<K}]}{K}$	Impliedα
−2.	−3.01	343952.	1.50	2.97
−5.	−8.02	21156.	1.60	2.65
−10.	−15.60	3528.	1.56	2.78
−20.	−30.41	503.	1.52	2.91
−50.	−113.324	20.	2.26	1.78
−70.	−170.105	10.	2.43	1.69
−80.	−180.84	9.	2.26	1.79
−90.	−192.543	8.	2.13	1.87
−100.	−251.691	5.	2.51	1.65

EuroDollars Front Month 1986-2006
n=4947

MAD	$E[X\|_{X<K}]$	$n(\text{for} X < K)$	$\frac{E[X\|_{X<K}]}{K}$	Impliedα
−0.5	−1.8034	1520	3.6068	1.38361
−1.	−2.41323	969	2.41323	1.7076
−5.	−7.96752	69	1.5935	2.68491
−6.	−9.2521	46	1.54202	2.84496
−7.	−10.2338	34	1.46197	3.16464
−8.	−11.4367	24	1.42959	3.32782

Global Macroeconomic data

UK Rates 1990-2007
n=4143

MAD	$E[X\|_{X<K}]$	$n(\text{for} X < K)$	$\frac{E[X\|_{X<K}]}{K}$	Impliedα
0.5	1.68802	1270	3.37605	1.42087
1.	2.23822	806	2.23822	1.80761
3.	4.97319	140	1.65773	2.52038
5.	8.43269	36	1.68654	2.45658
6.	9.56132	26	1.59355	2.68477
7.	11.4763	16	1.63947	2.56381

Conditional expectation for moves > K, 43 economic variables.

K, Mean deviations	Mean move (in MAD) in excess of K	n
1	2.01443	65,958
2	3.0814	23,450
3	4.19842	8,355
4	5.33587	3,202
5	6.52524	1,360
6	7.74405	660
7	9.10917	340
8	10.3649	192
9	11.6737	120
10	13.8726	84
11	15.3832	65
12	19.3987	47
13	21.0189	36
14	21.7426	29
15	24.1414	21
16	25.1188	18
17	27.8408	13
18	31.2309	11
19	35.6161	7
20	35.9036	6

11 BROWNIAN MOTION IN THE REAL WORLD

Chapter Summary **11: Much of the work concerning martingales and Brownian motion has been idealized; we look for holes and pockets of mismatch to reality, with consequences. Infinite (or undefined) higher moments are not compatible with Ito calculus −outside the asymptote. Path dependence as a measure of fragility.**

11.1 PATH DEPENDENCE AND HISTORY AS REVELATION OF ANTIFRAGILITY

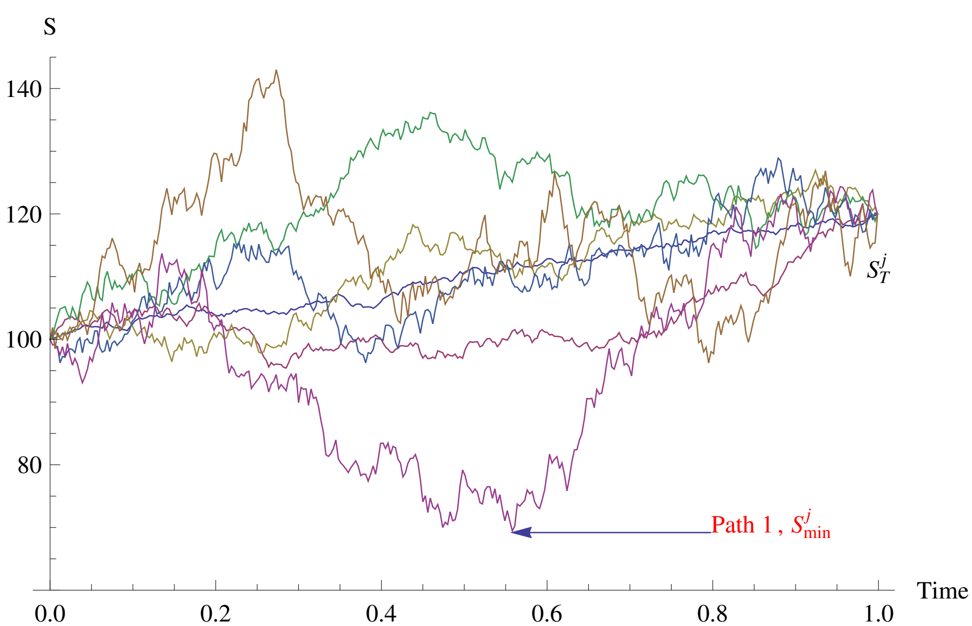


Figure 11.1: Brownian Bridge Pinned at 100 and 120, with multiple realizations $\{S_0^j, S_1^j, .., S_T^j\}$, each indexed by j ; the idea is to find the path j that satisfies the maximum distance $D_j = \left| S_T - S^j_{\min} \right|$

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 $X_{N|\{X_1,X_2,...X_{N-1}\}} = X_{N|\{X_{N-1}\}}$, that is the last realization is the only one that matters. Now if we take fat tailed

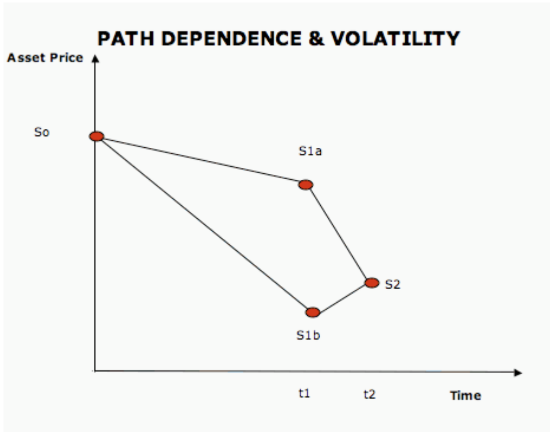


Figure 11.2: The recovery theorem requires the pricing kernel to be transition independent. So the forward kernel at S2 depends on the path. Implied vol at S2 via S1b is much lower than implied vol at S2 via S1a.

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 ,

$$S_i^j = \left(\left(S^j_{\mathrm{i}\Delta\mathrm{t}+t_0} \right)_{i=0}^{N} \right)_{j=1}^{M}$$

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.

11.2 SP AND PATH DEPENDENCE (INCOMPLETE)

For time series sampled at $(t_0, t_{0+\Delta t}, ..., t \equiv t_{0+n\Delta t})$, the minimum distance δ:

$$S^* \left(t_0, t, \Delta \mathrm{t} \right) \equiv \min \left(S_{\mathrm{i}\Delta\mathrm{t}+t_0} - \min \left(S_{\mathrm{j}\Delta\mathrm{t}+t_0} \right)_{j=i+1}^{N} \right)_{i=0}^{N}$$

We have the stopping time $\{\tau : S_\tau = S^* \left(t_0, t, \Delta \mathrm{t} \right)\}$ and the distance from the worst becomes $\delta \left(t_0, t, \Delta \mathrm{t} \right) \equiv S_t - S_\tau$

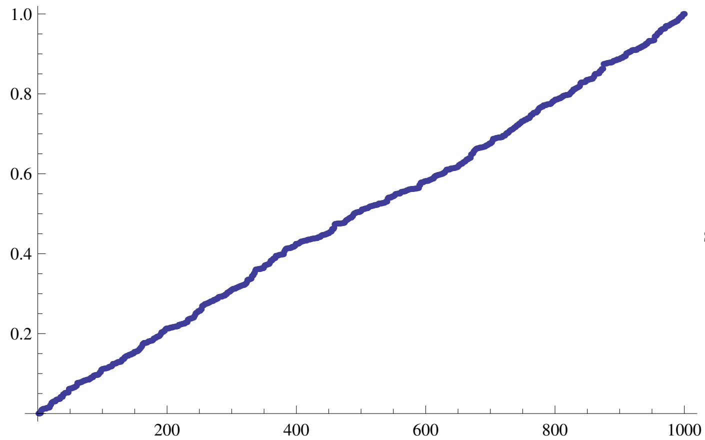


Figure 11.3: $C(n)$, Gaussian Case

11.3 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),

$$C(n) = \frac{\sum_{i=1}^{n} (W[i\Delta t] - W[(i-1)\Delta t])^2}{\sum_{i=1}^{T/\Delta t} (W[i\Delta t] - W[(i-1)\Delta t])^2}.$$

Let us play with a finite variance situations.

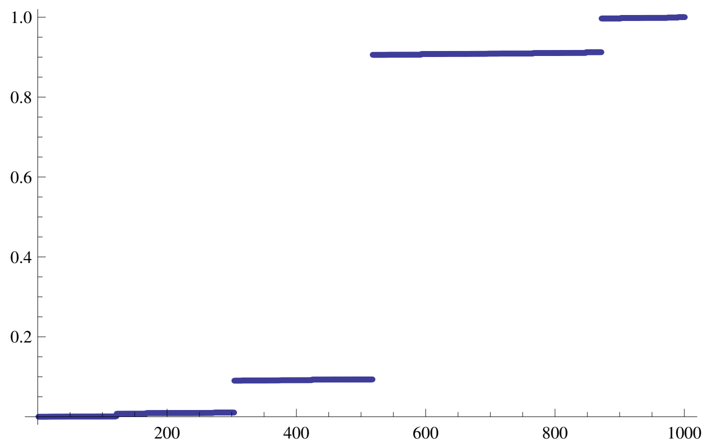


Figure 11.4: $\alpha = 1.16$

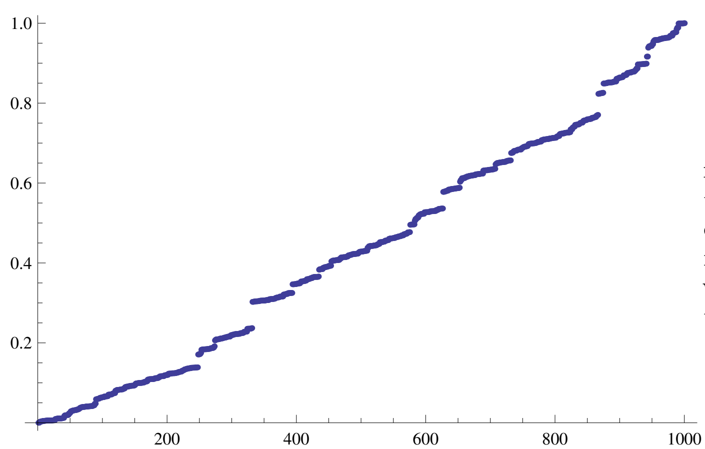


Figure 11.5: $\alpha = 3$: Even finite variance does not lead to the smoothing of discontinuities except in the infinitesimal limit, another way to see failed asymptotes.

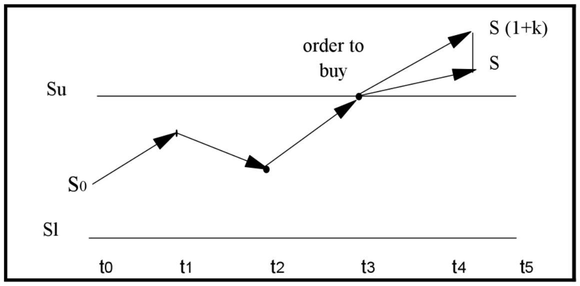


Figure 11.6: Asymmetry between a convex and a concave strategy

11.4 STOCHASTIC 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.

11.5 FINITE VARIANCE NOT NECESSARY FOR ANYTHING ECOLOGICAL (INCL. QUANT FINANCE)

[Summary of article in Complexity (2008)

12 THE FOURTH QUADRANT "SOLUTION"

Chapter Summary **12: A less technical demarcation between Black Swan Domains and others.**

Let us return to $M[A, f(x)]$ of Chapter 3. 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 $M[A, x^m]$ 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*, Part II), where the payoff is bounded, $A = [L, H]$, or the integral $\int_L^H f(x)p(x)dx$ will necessarily converge.

12.1 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.

Table 19: 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

M1$^+$Complex, depend on the 1st 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 20: Tableau of Decisions

*M*0 "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")

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 either on the left or on the right.

13 RISK AND PRECAUTION

Chapter Summary 13: We present the difference between ruin problems, particularly when systemic, and other risk management of a more computable nature.

A more technical exposition of the fourth quadrant, in which we replace the 4th Q with the precautionary principle. Simply one need to "show" that a given policy does not belong to IV in Table 21, or, alternatively exert more vigilance there.

Table 21: The Four Quadrants and Exposure Classes $\mathcal{H}^A$ and $\mathcal{H}^B$

	$Z = \sum_i \omega_i f_i(X_i) \in \mathcal{H}^A$	$Z = \sum_i \omega_i f_i(X_i) \in \mathcal{H}^B$
$X \in \mathcal{A}$	I	II
$X \in \mathcal{B}$	III	IV: Domain of PP

I: First Quadrant, safe
II: Second Quadrant, safe but calculated risks
III: Quadrant III, safe but rigorous risk management
IV: Quadrant Where PP should be exercized

Let $\mathbf{X} = (X_i)_{1 \leq i \leq n}$ be a sequence of random variables with support in ($\mathbb{R}^+$), with cumulative distribution function F. Let $S_n = \Sigma_{i=1}^n x_i$ and $M_n = \max_{1 \leq i \leq n} x_i$. Without making any statement as to the probability distribution nor independence:

Definition 13.1 (Membership of a r.v. in Class of Fat Tailed, convolution criterion)**.**
$\mathcal{B} = \{X \in \mathcal{B} : \lim_{x \to +\infty} \frac{1-F^{*2}(x)}{1-F(x)} = 2\}$, where $F^{*2} = F' * F$ is the cumulative distribution of $X_1 + X_2$, the sum of two copies of X.
Or, equivalently, For a given $n \geq 2$, a) $\lim_{x \to \infty} \frac{P(S_n > x)}{P(X > x)} = n$, b) $\lim_{x \to \infty} \frac{P(S_n > x)}{P(M_n > x)} = 1$.

Definition 13.2 (Membership in Class of Thin Tailed)**.**
$\mathcal{A} = \{X \in \mathcal{A} : X \notin \mathcal{B}\}$

Let $H^X \in (0, \infty)$ be a predefined "ruin" barrier associated with exposure to variable X and Z be an n-summed of mixing of functions of variables X in quantities ω_i with $f_i : (0, \infty) \to [0, H_i)$ as the loss(risk) function:

Definition 13.3 (NonSystemic Risk, Strong Condition)**.**

$$\mathcal{H}^A = \{Z \in \mathcal{H}^A : \sum \omega_i max_{(X_i \in \mathbf{R}^+, i \leq n)} (f_i(X_i)) < H^X\}. \tag{13.1}$$

13.0.1 Stopping time equivalence

We used a static representation instead of stopping time for clarity. Let $\tau = \{\inf t : f(X_{i,t}) > L_i\}$ be the stopping time where L_i is the local aborbing barrier. In that case in Equation 13.1, replace $\max_{(X_i \in \mathbf{R}^+, i \leq n)} (f_i(X_i))$ with $(f_i(X_{i,\tau}))$.

13.0.2 Sign of exposure

If we choose to indicate deviations as negative values of the variable x, the same result holds by symmetry for extreme negative values, replacing $x \to +\infty$ with $x \to -\infty$ (and using the complementary of the exceedance probability). For two-tailed variables, we can separately consider x^+ and x^- positive and negative domains.

13.0.3 layering

H^X is a layer attached specifically to variable X, for which "ruin" is specifically defined. In analyzing systems, we may have telescope-style multi-layering. This is for an isolated level (say ruin for a given continent, industry), etc.

13.1 WHAT IS THE PRECAUTIONARY PRINCIPLE

The precautionary principle (PP) states that if an action or policy has a suspected risk of causing severe harm to the public domain (affecting general health or the environment globally), the action should not be taken in the absence of scientific near-certainty about its safety. Under these conditions, the burden of proof about absence of harm falls on those proposing an action, not those opposing it. PP is intended to deal with uncertainty and risk in cases where the absence of evidence and the incompleteness of scientific knowledge carries profound implications and in the presence of risks of "black swans", unforeseen and unforeseable events of extreme consequence.

This non-naive version of the PP allows us to avoid paranoia and paralysis by confining precaution to specific domains and problems. Here we formalize PP, placing it within the statistical and probabilistic structure of “ruin” problems, in which a system is at risk of total failure, and in place of risk we use a formal"fragility" based approach. In these problems, what appear to be small and reasonable risks accumulate inevitably to certain irreversible harm. Traditional cost-benefit analyses, which seek to quantitatively weigh outcomes to determine the best policy option, do not apply, as outcomes may have infinite costs. Even high-benefit, high-probability outcomes do not outweigh the existence of low probability, infinite cost options—i.e. ruin. Uncertainties result in sensitivity analyses that are not mathematically well behaved. The PP is increasingly relevant due to man-made dependencies that propagate impacts of policies across the globe. In contrast, absent humanity the biosphere engages in natural experiments due to random variations with only local impacts.

The aim of the precautionary principle (PP) is to prevent decision makers from putting society as a whole—or a significant segment of it—at risk from the unexpected side effects of a certain type of decision. The PP states that if an action or policy has a suspected risk of causing severe harm to the public domain (such as general health or the environment), and in the absence of scientific near-certainty about the safety of the action, the burden of proof about absence of harm falls on those proposing the action. It is meant to deal with effects of absence of evidence and the incompleteness of scientific knowledge in some risky domains.[1]

We believe that the PP should be evoked only in extreme situations: when the potential harm is systemic (rather than localized) and the consequences can involve total irreversible ruin, such as the extinction of human beings or all life on the planet.

Standard Risk Management	Precautionary Approach
localized harm	systemic ruin
nuanced cost-benefit	avoid at all costs
statistical	fragility based
statistical	probabilistic non-statistical
variations	ruin
convergent probabibilities	divergent probabilities
recoverable	irreversible
independent factors	interconnected factors
evidence based	precautionary
thin tails	fat tails
bottom-up, tinkering	top-down engineered
evolved	human-made

Table 22: Two different types of risk and their respective characteristics compared

13.2 WHY RUIN IS SERIOUS BUSINESS

13.3 SKEPTICISM AND PRECAUTION

We show in Figures 13.2 and 13.3 that an increase in uncertainty leads to an increase in the probability of ruin, hence "skepticism" is that its impact on decisions should lead to increased, not decreased conservatism in the presence of ruin. More skepticism about models implies more uncertainty about the tails, which necessitates more precaution about newly implemented techniques, or larger size of exposures. As we said, Nature might not be smart, but its longer track record means smaller uncertainty in following its logic.

Mathematically, more uncertainty about the future –or about a model –increases the scale of the distribution, hence thickens the "left tail" (as well as the "right one")

1 The Rio Declaration on Environment and Development presents it as follows: "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."

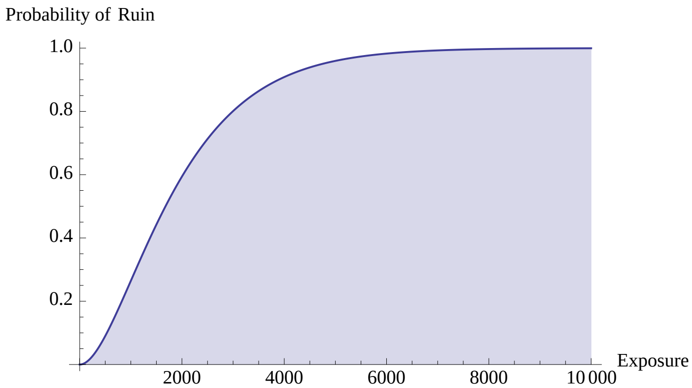


Figure 13.1: Why Ruin is not a Renewable Resource. No matter how small the probability, in time, something bound to hit the ruin barrier is about guaranteed to hit it.

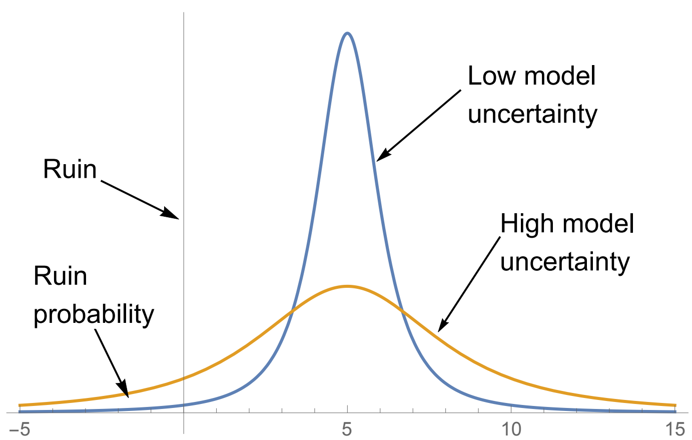


Figure 13.2: The more uncertain or skeptical one is of "scientific" models and projections, the higher the risk of ruin, which flies in the face of the argument of the style "skeptical of climate models". No matter how increased the probability of benefits, ruin as an absorbing barrier, i.e. causing extinction without further recovery, can more than cancels them out. This graph assumes changes in uncertainty without changes in benefits (a mean-preserving sensitivity) –the next one isolates the changes in benefits.

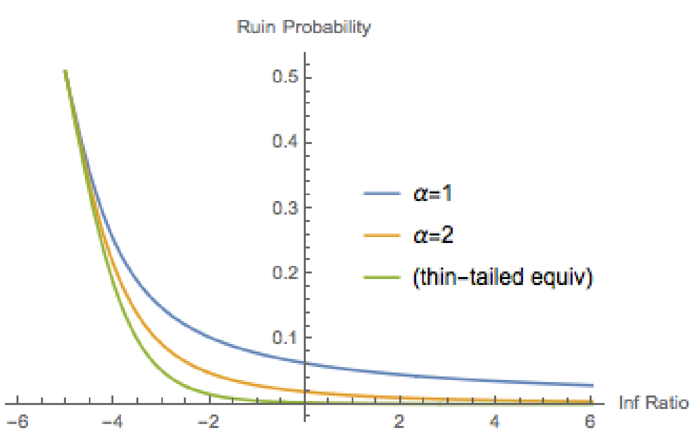


Figure 13.3: The graph shows the asymmetry between benefits and harm and the effect on the ruin probabilities. Shows the effect on ruin probability of changes the Information Ratio, that is, $\frac{\text{expected benefit}}{\text{uncertainty}}$ (or signal divided by noise). Benefits are small compared to negative effects. Three cases are considered, two from Extremistan: extremely fat-tailed ($\alpha = 1$), and less fat-tailed ($\alpha = 2$), and one from Mediocristan.

which raises the potential ruin. The survival probability is reduced no matter what takes place in the right tail. Hence skepticim about climate models should lead to more precautionary policies.

In addition, such increase uncertainty matters far more in Extremistan –and has benign effects in Mediocristan. Figure 13.3 shows th asymmetries between costs and benefits as far as ruin probabilities, and why these matter more for fat-tailed domains than thin-tailed ones. In thin-tailed domains, an increase in uncertainty changes the probability of ruin by several orders of magnitude, but the effect remains small: from say 10^{-40} to 10^{-30} is not quite worrisome. In fat-tailed domains, the effect is sizeable as we start with a substantially higher probability of ruin (which is typically underestimated, see **[?**]).

13.4 FALLACIOUS ARGUMENTS IN RISK MANAGEMENT

13.4.1 Crossing the road (the paralysis fallacy)

Many have countered against risk measures with "nothing is ever totally safe." "I take risks crossing the road every day, so according to you I should stay home in a state of paralysis." The answer is that we don't cross the street blindfolded, we use sensory information to mitigate risks and reduce exposure to extreme shocks.[2]

Even more importantly, the probability distribution of death from road accidents at the population level is thin-tailed; I do not incur the risk of generalized human extinction by crossing the street—a human life is bounded in duration and its unavoidable termination is an inherent part of the bio-social system. The error of my crossing the street at the wrong time and meeting an untimely demise in general does not cause others to do the same; the error does not spread. If anything, one might expect the opposite effect, that others in the system benefit from my mistake by adapting their behavior to avoid exposing themselves to similar risks. Equating risks a person takes with his or her own life with risking the existence of civilization is an inappropriate ego trip. In fact, the very idea of the PP is to avoid such a frivolous focus.

The paralysis argument is often used to present that precaution as incompatible with progress. This is untrue: tinkering, bottom-up progress where mistakes are bounded is *how* progress has taken place in history. The non-naive PP simply asserts that the risks we take as we innovate must not extend to the entire system; local failure serves as information for improvement. Global failure does not.

This fallacy illustrates the misunderstanding between systemic and idiosyncratic risk in the literature. Individuals are fragile and mortal. The idea of sustainability is to stike to make systems as close to immortal as possible.

13.4.2 The Psychology of Risk and Thick Tailed Distributions

One concern about the utility of the PP is that its evocation may become commonplace because of risk aversion. Is it true that people overreact to small probabilities and the PP would feed into human biases? While we have carefully identified

2 Actually the actuarial risk of death for pedestrians is one in 47,000 years, I thank Hanoz Kalwachwala.

the scope of the domain of applicability of the PP, it is also helpful to review the evidence of risk aversion, which we find not to be based upon sound studies.

Certain empirical studies appear to support the existence of a bias toward risk aversion, claiming evidence that people choose to avoid risks that are beneficial, inconsistent with cost-benefit analyses. The relevant experiments ask people questions about single probability events, showing that people overreact to small probabilities. However, those researchers failed to include the consequences of the associated events which humans underestimate. Thus, this empirical strategy as a way of identifying effectiveness of response to risk is fundamentally flawed [**?**].

The proper consideration of risk involves both probability and consequence, which should be multiplied together. Consequences in many domains have thick tails, i.e. much larger consequences can arise than are considered in traditional statistical approaches. Overreacting to small probabilities is not irrational when the effect is large, as the product of probability and harm is larger than expected from the traditional treatment of probability distributions.

13.4.3 The Loch Ness fallacy

Many counter that we have no evidence that the Loch Ness monster doesn't exist, and, to take the argument of *evidence of absence* being different from *absence of evidence*, we should act as if the Loch Ness monster existed. The argument is a corruption of the absence of evidence problem.

The relevant question is whether the existence of the Loch Ness monster has implications for decisions about actions that are being taken. We are not considering a decision to swim in the Loch Ness. If the Loch Ness monster did exist, there would still be no reason to invoke the PP, as the harm he might cause is limited in scope to Loch Ness itself, and does not present the risk of ruin.

13.4.4 The fallacy of misusing the naturalistic fallacy

Some people invoke "the naturalistic fallacy," a philosophical concept that is limited to the moral domain. According to this critique, we should not claim that natural things are necessarily good; human innovation can be equally valid. We do not claim to use nature to derive a notion of how things "ought" to be organized. Rather, as scientists, we respect nature for the extent of its experimentation. The high level of statistical significance given by a very large sample cannot be ignored. Nature may not have arrived at the best solution to a problem we consider important, but there is reason to believe that it is smarter than our technology based only on statistical significance.

The question about what kinds of systems *work* (as demonstrated by nature) is different than the question about what working systems ought to do. We can take a lesson from nature—and time—about what kinds of organizations are robust against, or even benefit from, shocks, and in that sense systems should be structured in ways that allow them to function. Conversely, we cannot derive the structure of a functioning system from what we believe the outcomes *ought* to be.

To take one example, Cass Sunstein—who has written an article critical of the PP [**?**]—claims that there is a "false belief that nature is benign." However, his conceptual discussion fails to distinguish between thin and fat tails, local harm and global ruin. The method of analysis misses both the statistical significance of nature and the fact that it is not necessary to believe in the perfection of nature, or in its "benign" attributes, but rather in its track record, its sheer statistical power as a risk evaluator and as a risk manager in avoiding ruin.

13.4.5 The "Butterfly in China" fallacy

The statement “if I move my finger to scratch my nose, by the butterfly-in-China effect, owing to non-linearities, I may terminate life on earth," is known to be flawed. The explanation is not widely understood. The fundamental reason arises because of the existence of a wide range in levels of predictability and the presence of a large number of fine scale degrees of freedom for every large scale one [**?**]. Thus, the traditional deterministic chaos, for which the butterfly effect was named, applies specifically to low dimensional systems with a few variables in a particular regime. High dimensional systems, like the earth, have large numbers of fine scale variables for every large scale one. Thus, it is apparent that not all butterfly wing flaps can cause hurricanes. It is not clear that any one of them can, and, if small perturbations can influence large scale events, it happens only under specific conditions where amplification occurs.

Empirically, our thesis rebuts the butterfly fallacy with the argument that, in the aggregate, nature has experienced trillions of small variations and yet it survives. Therefore, we know that the effects of scratching one’s nose fall into the thin tailed domain and thus do not warrant the precautionary principle.

As described previously, barriers in natural systems lead to subsystems having a high-degree of independence. Understanding how modern systems with a high-degree of connectivity have cascading effects is essential for understanding when it is and isn’t appropriate to use the PP.

13.4.6 The potato fallacy

Many species were abruptly introduced into the Old World starting in the 16th Century that did not cause environmental disasters (perhaps aside from diseases affecting Native Americans). Some use this observation in defense of GMOs. However, the argument is fallacious at two levels:

First, by the fragility argument, potatoes, tomatoes and similar "New World" goods were developed locally through progressive, bottom-up tinkering in a complex system in the context of its interactions with its environment. Had they had an impact on the environment, it would have caused adverse consequences that would have prevented their continual spread.

Second, a counterexample is not evidence in the risk domain, particularly when the evidence is that taking a similar action previously did not lead to ruin. Lack of ruin due to several or even many trials does not indicate safety from ruin in the next one. This is also the Russian roulette fallacy, detailed below.

13.4.7 The Russian roulette fallacy (the counterexamples in the risk domain)

The potato example, assuming potatoes had not been generated top-down by some engineers, would still not be sufficient. Nobody says "look, the other day there was no war, so we don't need an army," as we know better in real-life domains. Nobody argues that a giant Russian roulette with many barrels is "safe" and a great money making opportunity because it didn't blow up someone's brains last time.

There are many reasons a previous action may not have led to ruin while still having the *potential* to do so. If you attempt to cross the street with a blindfold and earmuffs on, you may make it across, but this is not evidence that such an action carries no risk.

More generally, one needs a large sample for claims of *absence of risk* in the presence of a small probability of ruin, while a single "$n = 1$" example would be sufficient to counter the claims of safety—this is the Black Swan argument [**?**]. Simply put, systemic modifications require a very long history in order for the evidence of lack of harm to carry any weight.

13.4.8 The Carpenter Fallacy

Risk managers skeptical of the understanding of risk of biological processes, such as GMOs, by the experts are sometimes asked "are you a biologist?" But nobody asks a probabilist dealing with roulette sequences if he is a carpenter. To understand the gambler's ruin problem by roulette betting, we know to ask a probabilist, not a carpenter. No amount of expertise in carpentry can replace rigor in understanding the properties of long sequences of small probability bets. Likewise, no amount of expertise in the details of biological processes can be a substitute for probabilistic rigor.

The context for evaluating risk is the extent of knowledge or lack of knowledge. Thus, when considering GMO risks, a key question is what is the extent to which we know the impacts of genetic changes in organisms. Claims that geneticists know these consequences as a basis for GMOs do not recognize either that their knowledge is not complete in its own domain nor is genetics complete as a body of knowledge. Geneticists do not know the developmental, physiological, medical, cognitive and environmental consequences of genetic changes in organisms. Indeed, most of these are not part of their training or competency. Neither are they trained in recognizing the impact of the limitations of knowledge on risk.

Consistent with these points, the track record of the experts in understanding biological and medical risks has been extremely poor. We need policies to be robust to such miscalculations. The "expert problem" in medicine by which experts mischaracterize the completeness of their own knowledge is manifest in a very poor historical record of risks taken with innovations in biological products. These range from biofuels to transfat to nicotine, etc. Consider the recent major drug recalls such as Thalidomide, Fen-Phen, Tylenol and Vioxx—all of these show blindness on the part of the specialist to large scale risks associated with absence of knowlege, i.e., Black Swan events. Yet most of these risks were local and not systemic (with the exception of biofuel impacts on global hunger and social unrest). Since systemic risks would result in a recall happening too late, we need the strong version of the PP.

13.4.9 The technological salvation fallacy

Iatrogenics is harm done by a healer despite positive intentions, see Appendix A for a list of innovations in care that have extensive documentation of adverse consequences. Each of these underwent best practices testing that did not reveal the iatrogenic consequences prior to widespread application. The controlled tests that are used to evaluate innovations for potential harm cannot replicate the large number of conditions in which interventions are applied in the real world. Adverse consequences are exposed only by extensive experience with the combinatorial number of real world conditions. Natural, i.e. evolutionary, selection implements as a strategy the use of selection of lack of harm under such conditions in a way that bounds the consequences because the number of replicates is increased only gradually during the process in which success is determined. In contrast, traditional engineering of technological solutions does not. Thus, the more technological a solution to a current problem—the more it departs from solutions that have undergone evolutionary selection—the more exposed one becomes to iatrogenics owing to combinatorial branching of conditions with adverse consequences.

Our concern here isn't mild iatrogenics, but the systemic case.

13.4.10 The pathologization fallacy

Today many mathematical or conceptual models that are claimed to be rigorous are based upon unvalidated and incorrect assumptions and are not robust to changes in these assumptions. Such models are deemed rational in the sense that they are logically derived from their assumptions, and, further, can be used to assess rationality by examining deviations from such models, as indicators of irrationality. Except that it is often the modeler who is using an incomplete representation of the reality, hence using an erroneous benchmark for rationality. Often the modelers are not familiar with the dynamics of complex systems or use antiquated statistical methods that do not take into account fat-tails and make inferences that would not be acceptable under different classes of probability distributions. Many biases, such as the ones used by Cass Sunstein (mentioned above), about the overestimation of the probabilities of rare events in fact correspond to the testers using a bad probability model that is thin-tailed. See Ref. [**?**] for a deeper discussion.

It has became popular to claim irrationality for GMO and other skepticism on the part of the general public—not realizing that there is in fact an "expert problem" and such skepticism is healthy and even necessary for survival. For instance, in *The Rational Animal* [**?**], the authors pathologize people for not accepting GMOs although "the World Health Organization has never found evidence of ill effects," a standard confusion of evidence of absence and absence of evidence. Such pathologizing is similar to behavioral researchers labeling hyperbolic discounting as "irrational" when in fact it is largely the researcher who has a very narrow model and richer models make the "irrationality" go away.

These researchers fail to understand that humans may have precautionary principles against systemic risks, and can be skeptical of the untested consequences of policies for deeply rational reasons, even if they do not express such fears in academic format.

14 SKIN IN THE GAME AND RISK TAKING

Chapter Summary **14: Standard economic theory makes an allowance for the agency problem, but not the compounding of moral hazard in the presence of informational opacity, particularly in what concerns high-impact events in fat tailed domains (under slow convergence for the law of large numbers). Nor did it look at exposure as a filter that removes nefarious risk takers from the system so they stop harming others. (In the language of probability, skin in the game creates an absorbing state for the agent, not just the principal). But the ancients did; so did many aspects of moral philosophy. We propose a global and morally mandatory heuristic that anyone involved in an action which can possibly generate harm for others, even probabilistically, should be required to be exposed to some damage, regardless of context. While perhaps not sufficient, the heuristic is certainly necessary hence mandatory. It is supposed to counter voluntary and involuntary risk hiding − and risk transfer − in the tails.**

The literature in risk, insurance, and contracts has amply dealt with the notion of information asymmetry (see Ross, 1973, Grossman and Hart, 1983, 1984, Tirole 1988, Stiglitz 1988), but not with the consequences of deeper information opacity (in spite of getting close, as in HÃűlmstrom, 1979), by which tail events are impossible to figure out from watching time series and external signs: in short, in the "real world" (Taleb, 2013), the law of large numbers works very slowly, or does not work at all in the time horizon for operators, hence statistical properties involving tail events are completely opaque to the observer. And the central problem that is missing behind the abundant research on moral hazard and information asymmetry is that these rare, unobservable events represent the bulk of the properties in some domains. We define a fat tailed domain as follows: a large share of the statistical properties come from the extremum; for a time series involving n observations, as n becomes large, the maximum or minimum observation will be of the same order as the sum. Excursions from the center of the distributions happen brutally and violently; the rare event dominates. And economic variables are extremely fat tailed (Mandelbrot, 1997). Further, standard economic theory makes an allowance for the agency problem, but not for the combination of agency problem, informational opacity, and fat-tailedness. It has not yet caught up that tails events are not predictable, not measurable statistically unless one is *causing* them, or involved in increasing their probability by engaging in a certain class of actions with small upside and large downside. (Both parties may not be able to gauge probabilities in the tails of the distribution, but the agent knows which tail events do not affect him.) Sadly, the economics literature's treatment of tail risks , or "peso problems" has been to see them as outliers to mention *en passant* but hide under the rug, or remove from analysis, rather than a core center of the modeling and decision-making, or to think in terms of robustness and sensitivity to unpredictable events. Indeed, this pushing under the rug the determining statistical properties explains the failures of economics in mapping the real world, as witnessed by the inability of the economics establishment to see the accumulation of tail risks leading up to the financial crisis of 2008 (Taleb, 2009). The parts of the risk and insurance literature that have focused on tail events and extreme value theory, such as Embrechts (1997), accepts the large role of the tails, but then the users of these theories (in the applications) fall for the logical insonsistency of assuming that they can be figured out somehow: naively, since they are rare what do we know about them? The law of large numbers cannot be of help. Nor do theories have the required robustness. Alarmingly, very little has been done to make the leap that small calibration errors in models can change the probabilities (such as those involving the risks taken in Fukushima's nuclear project) from 1 in 10^6 to 1 in 50.

Add to the fat-tailedness the asymmetry (or skewness) of the distribution, by which a random variable can take very large values on one side, but not the other. An operator who wants to hide risk from others can exploit skewness by creating a situation in which he has a small or bounded harm to him, and exposing others to large harm; thus exposing others to the bad side of the distributions by fooling them with the tail properties.

Finally, the economic literature focuses on incentives as encouragement or deterrent, but not on disincentives as potent filters that remove incompetent and nefarious risk takers from the system. Consider that the symmetry of risks incurred on the road causes the bad driver to eventually exit the system and stop killing others. An unskilled forecaster with skin-in-the-game would eventually go bankrupt or out of business. Shielded from potentially (financially) harmful exposure, he would continue contributing to the buildup of risks in the system. [1]

Hence there is no possible risk management method that can replace skin in the game in cases where informational opacity is compounded by informational asymmetry viz. the principal-agent problem that arises when those who gain the upside resulting from actions performed under some degree of uncertainty are not the same as those who incur the downside of those same acts[2]. For example, bankers and corporate managers get bonuses for positive "performance", but do not have to pay out reverse bonuses for negative performance. This gives them an incentive to bury risks in the tails of the distribution, particularly the left tail, thereby delaying blowups.

The ancients were fully aware of this incentive to hide tail risks, and implemented very simple but potent heuristics (for the effectiveness and applicability of fast and frugal heuristics both in general and in the moral domain, see Gigerenzer, 2010). But we find the genesis of both moral philosophy and risk management concentrated within the same rule [3] . About 3,800 years ago, Hammurabi's code

1 The core of the problem is as follows. There are two effects: "crooks of randomness" and "fooled of randomness" (Nicolas Tabardel, private communication). Skin in the game eliminates the first effect in the short term (standard agency problem), the second one in the long term by forcing a certain class of harmful risk takers to exit from the game.

2 Note that Pigovian mechanisms fail when, owing to opacity, the person causing the harm is not easy to identify

3 Economics seems to be born out of moral philosophy (mutating into the philosophy of action via decision theory) to which was added naive and improper 19th C. statistics (Taleb, 2007, 2013). We are trying to go back to its moral philosophy roots, to which we add more sophisticated probability theory and risk management.

specified that if a builder builds a house and the house collapses and causes the death of the owner of the house, that builder shall be put to death. This is the best risk-management rule ever.

What the ancients understood very well was that the builder will always know more about the risks than the client, and can hide sources of fragility and improve his profitability by cutting corners. The foundation is the best place to hide such things. The builder can also fool the inspector, for the person hiding risk has a large informational advantage over the one who has to find it. The same absence of personal risk is what motivates people to only appear to be doing good, rather than to actually do it.

Note that Hammurabi's law is not necessarily literal: damages can be "converted" into monetary compensation. Hammurabi's law is at the origin of the *lex talonis* ("eye for eye", discussed further down) which, contrary to what appears at first glance, it is not literal. *Tractate Bava Kama* in the Babylonian Talmud [4], builds a consensus that "eye for eye" has to be figurative: what if the perpetrator of an eye injury were blind? Would he have to be released of all obligations on grounds that the injury has already been inflicted? Wouldn't this lead him to inflict damage to other people's eyesight with total impunity? Likewise, the Quran's interpretation, equally, gives the option of the injured party to pardon or alter the punishment[5]. This nonliteral aspect of the law solves many problems of asymmetry under specialization of labor, as the deliverer of a service is not required to have the same exposure in kind, but incur risks that are costly enough to be a disincentive.

The problems and remedies are as follows:

First, consider policy makers and politicians. In a decentralized system, say municipalities, these people are typically kept in check by feelings of shame upon harming others with their mistakes. In a large centralized system, the sources of error are not so visible. Spreadsheets do not make people feel shame. The penalty of shame is a factor that counts in favour of governments (and businesses) that are small, local, personal, and decentralized versus ones that are large, national or multi-national, anonymous, and centralised. When the latter fail, everybody except the culprit ends up paying the cost, leading to national and international measures of endebtment against future generations or "austerity "[6].These points against "big government " models should not be confused with the standard libertarian argument against states securing the welfare of their citizens, but only against doing so in a centralized fashion that enables people to hide behind bureaucratic anonymity. Much better to have a communitarian municipal approach:in situations in which we cannot enforce skin-in-the game we should change the system to lower the consequences of errors.

Second, we misunderstand the incentive structure of corporate managers. Counter to public perception, corporate managers are not entrepreneurs. They are not what one could call agents of capitalism. Between 2000 and 2010, in the United States, the stock market lost (depending how one measures it) up to two trillion dollars for investors, compared to leaving their funds in cash or treasury bills. It is tempting to think that since managers are paid on incentive, they would be incurring losses.

4 *Tractate Bava Kama*, 84a, Jerusalem: Koren Publishers, 2013.
5 Quran, *Surat Al-Ma'idat*, 45: "Then, whoever proves charitable and gives up on his right for reciprocation, it will be an atonement for him." (our translation).
6 See McQuillan (2013) and Orr (2013); cf. the "many hands " problem discussed by Thompson (1987)

Not at all: there is an irrational and unethical asymmetry. Because of the embedded option in their profession, managers received more than four hundred billion dollars in compensation. The manager who loses money does not return his bonus or incur a negative one[7].The built-in optionality in the compensation of corporate managers can only be removed by forcing them to eat some of the losses[8].

Third, there is a problem with applied and academic economists, quantitative modellers, and policy wonks. The reason economic models do not fit reality (fat-tailed reality) is that economists have no disincentive and are never penalized for their errors. So long as they please the journal editors, or produce cosmetically sound "scientific" papers, their work is fine. So we end up using models such as portfolio theory and similar methods without any remote empirical or mathematical reason. The solution is to prevent economists from teaching practitioners, simply because they have no mechanism to exit the system in the event of causing risks that harm others. Again this brings us to decentralization by a system where policy is decided at a local level by smaller units and hence in no need for economists.

Fourth, the predictors. Predictions in socioeconomic domains don't work. Predictors are rarely harmed by their predictions. Yet we know that people take more risks after they see a numerical prediction. The solution is to ask —and only take into account— what the predictor has done (what he has in his portfolio), or is committed to doing in the future. It is unethical to drag people into exposures without incurring losses. Further, predictors work with binary variables (Taleb and Tetlock, 2013), that is, "true" or "false" and play with the general public misunderstanding of tail events. They have the incentives to be right more often than wrong, whereas people who have skin in the game do not mind being wrong more often than they are right, provided the wins are large enough. In other words, predictors have an incentive to play the skewness game (more on the problem in section 2). The simple solution is as follows: predictors should be exposed to the variables they are predicting and should be subjected to the dictum "do not tell people what you think, tell them what you have in your portfolio" (Taleb, 2012, p.386) . Clearly predictions are harmful to people as, by the psychological mechanism of anchoring, they increases risk taking.

Fifth, to deal with warmongers, Ralph Nader has rightly proposed that those who vote in favor of war should subject themselves (or their own kin) to the draft.

We believe *Skin in the game* is a heuristic for a safe and just society. It is even more necessary under fat tailed environments. Opposed to this is the unethical practice of taking all the praise and benefits of good fortune whilst disassociating oneself from the results of bad luck or miscalculation. We situate our view within the framework of ethical debates relating to the moral significance of actions whose effects result from ignorance and luck. We shall demonstrate how the idea of skin in the game can effectively resolve debates about (a) moral luck and (b) egoism

7 There can be situations of overconfidence by which the CEOs of companies bear a disproportionately large amount of risk, by investing in their companies, as shown by Malmendier and Tate(2008, 2009), and end up taking more risk because they have skin in the game. But it remains that CEOs have optionality, as shown by the numbers above. Further, the heuristic we propose is necessary, but may not be sufficient to reduce risk, although CEOs with a poor understanding of risk have an increased probability of personal ruin.

8 We define "optionality" as an option-like situation by which an agent has a convex payoff, that is, has more to gain than to lose from a random variable, and thus has a positive sensitivity to the scale of the distribution, that is, can benefit from volatility and dispersion of outcomes.

vs. altruism, while successfully bypassing (c) debates between subjectivist and objectivist norms of action under uncertainty, by showing how their concerns are of no pragmatic concern.

Reputational Costs in Opaque Systems: Note that our analysis includes costs of reputation as skin in the game, with future earnings lowered as the result of a mistake, as with surgeons and people subjected to visible malpractice and have to live with the consequences. So our concern is situations in which cost hiding is effective over and above potential costs of reputation, either because the gains are too large with respect to these costs, or because these reputation costs can be "arbitraged ", by shifting blame or escaping it altogether, because harm is not directly visible. The latter category includes bureaucrats in non-repeat environments where the delayed harm is not directly attributable to them. Note that in many domains the payoff can be large enough to offset reputational costs, or, as in finance and government, reputations do not seem to be aligned with effective track record. (To use an evolutionary argument, we need to avoid a system in which those who make mistakes stay in the gene pool, but throw others out of it.)

Application of The Heuristic: The heuristic implies that one should be the first consumer of one's product, a cook should test his own food, helicopter repairpersons should be ready to take random flights on the rotorcraft that they maintain, hedge fund managers should be maximally invested in their funds. But it does not naively imply that one should always be using one's product: a barber cannot cut his own hair, the maker of a cancer drug should not be a user of his product unless he is ill. So one should use one's products *conditionally* on being called to use them. However the rule is far more rigid in matters entailing sytemic risks: simply some decisions should never be taken by a certain class of people.

Heuristic vs Regulation: A heuristic, unlike a regulation, does not require state intervention for implementation. It is simple contract between willing individuals: "I buy your goods if you use them", or "I will listen to your forecast if you are exposed to losses if you are wrong" and would not require the legal system any more than simple commercial transaction. It is bottom-up. (The ancients and more-or-less ancients effectively understood the contingency and probabilistic aspect in contract law, and asymmetry under opacity, as reflected in the works of Pierre de Jean Olivi. Also note that the foundation of maritime law has resided in skin-the-game unconditional sharing of losses, even as far in the past as 800 B.C. with the *Lex Rhodia*, which stipulates that all parties involved in a transaction have skin in the game and share losses in the event of damage. The rule dates back to the Phoenician commerce and caravan trades among Semitic people. The idea is still present in Islamic finance commercial law, see WardÃľ, 2010 .)

The rest of this chapter is organized as follows. First we present the epistemological dimension of the hidden payoff, expressed using the mathematics of probability, showing the gravity of the problem of hidden consequences. We conclude with the notion of heuristic as simple "convex" rule, simple in its application.

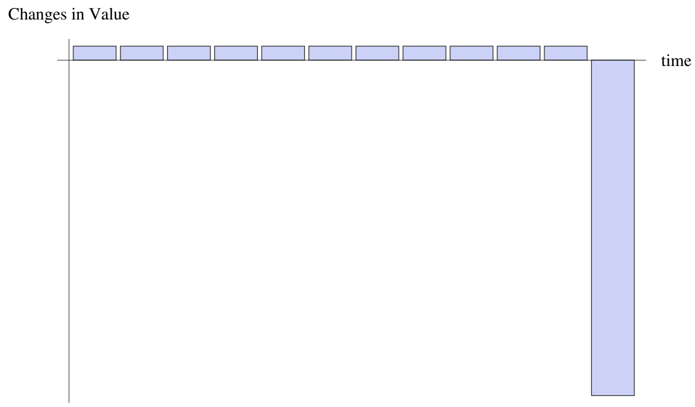


Figure 14.1: The most effective way to maximize the expected payoff to the agent at the expense of the principal.

14.1 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.

Transfer of Harm: *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

$$P(K, M) \equiv \gamma \sum_{i=1}^{M} q_{t+(i-1)\Delta\mathrm{t}} \left(x^{j}_{t+i\Delta t} - K \right)^{+} \mathbf{1}_{\Delta \mathrm{t}(i-1)+t<\tau} \tag{14.1}$$

with $X^j = (x^j_{t+i\Delta t})_{i=1}^{M} \in \mathbb{R}$, 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", τ= $\inf\left\{s : \left(\sum_{z\leq s} x_z\right) < x_{\min}\right\}$ 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 ∈(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 $q_{t+(i-1)\Delta\mathrm{t}} \in [1,\infty)$ 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 $f_j(x)\,\mathrm{d}x$, so $F_j^{+}=\int_K^{\infty} f_j(x)\ \mathrm{d}x$ and $F_j^{-}=\int_{-\infty}^{K} f_j(x)\ \mathrm{d}x$, the "upper" and "lower" distributions, each corresponding to certain conditional expectation $\mathbb{E}_j^+ \equiv \frac{\int_K^\infty x f_j(x)\,\mathrm{d}x}{\int_K^\infty f_j(x)\,\mathrm{d}x}$ and $\mathbb{E}_j^- \equiv \frac{\int_{-\infty}^K x\, f_j(x)\,\mathrm{d}x}{\int_{-\infty}^K f_j(x)\,\mathrm{d}x}$.

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

$$m^+ + m^- = m.$$

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 larger than $-K$ in the previous periods, $\tau = \inf\{(t + i\Delta\mathrm{t})) \colon x_{\Delta\mathrm{t}(i-1)+t} < K\}$, which leads to

$$\mathbb{E}(P(K, M)) = \gamma\ \mathbb{E}_j^+ \times \mathbb{E}\left(\sum_{i=1}^{M} \mathbf{1}_{t+i\Delta t<\tau}\right) \tag{14.2}$$

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 $\gamma\ \mathbb{E}_j{}^+$. And $\mathbb{E}\left(\sum_{i=1}^M \mathbf{1}_{\Delta\mathrm{t}(i-1)+t<\tau}\right) = \mathbb{E}\left(\left(\sum_{i=1}^M \mathbf{1}_{\Delta\mathrm{t}(i-1)+t<\tau}\right) \wedge M\right)$.

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

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

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:

$$\{(1 - F_j^+),\ F_j^+\left(1 - F_j^+\right),\ F_j^{+2}\left(1 - F_j^+\right),\ ...,,\ F_j^{+\,M-1}\left(1 - F_j^+\right)\},$$

$$\sum_{i=1}^{M} F_j^{+(i-1)}\left(1 - F_j^+\right) = 1 - F_j^{+\,M} \simeq 1 \tag{14.3}$$

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

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

Finally, the expected payoff for the agent:

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

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.

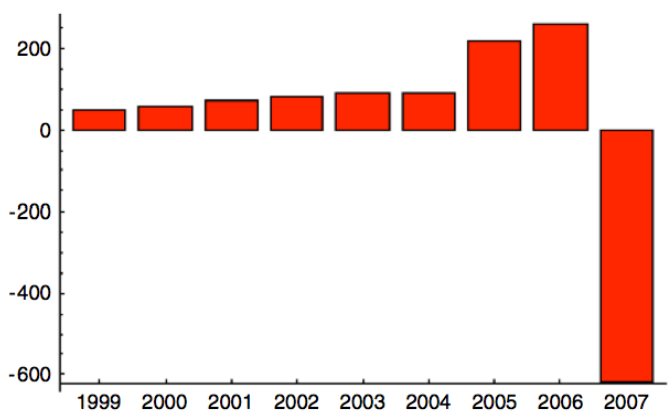


Figure 14.2: Indy Mac, a failed firm during the subprime crisis (from Taleb 2009). It is a representative of risks that keep increasing in the absence of losses, until the explosive blowup.

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 becomes 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[9]. 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:

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

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

Equation 14.1 becomes:

$$P(K,M) \equiv \gamma \sum_{i=1}^{M} q\ \omega(i) \left(x^j_{t+i\Delta t} - K\right)^+ \mathbf{1}_{t+(i-1)\Delta t<\tau}, \tag{14.4}$$

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

$$\mathbb{E}(P(K,M)) = \gamma\ \mathbb{E}^+_j\ q\ \mathbb{E}\left(\sum_{i=1}^{M} \omega(i)\ \mathbf{1}_{\Delta t(i-1)+t<\tau}\right). \tag{14.5}$$

Assuming the rate of conditional growth is a constant $r \in [0,\infty)$, and making the replacement ω(i)$\equiv e^{ri}$, we can call the last term in equation 14.5 the multiplier of the expected return to the agent:

$$\mathbb{E}\left(\sum_{i=1}^{M} e^{ir} \mathbf{1}_{\Delta t(i-1)+t<\tau}\right) = \sum_{i=1}^{M} (i-1)\ \ F_j^{\ +} e^{ir} \mathbb{E}(\mathbf{1}_{x_{\Delta t(i-1)+t}>K}) \tag{14.6}$$

$$= \frac{(F^+-1)\left((F^+)^M\left(Me^{(M+1)r}-F^+(M-1)e^{(M+2)r}\right)-F^+e^{2r}\right)}{(F^+e^r-1)^2} \tag{14.7}$$

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

9 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 [109]).

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 $(-\infty, x_{\text{min}}]$, with exceedance probability $P(X > x) = \; - x^{-\alpha} x_{\text{min}}^{\alpha}$, and mean $-\frac{\alpha x_{\text{min}}}{\alpha-1}$, with $\alpha > 1$, have a proportion of $1 - \frac{\alpha-1}{\alpha}$ of its realizations rosier than the true mean. Note that fat-tailedness increases at lower values of α. The popular "eighty-twenty", with tail exponent $\alpha = 1.15$, has $>$ 90 percent of observations above the true mean[10] –if anything, it should be called a "payoff" not a distribution. Likewise, to consider a thinner tailed skewed distribution, for a Lognormal distribution with support $(-\infty, 0)$, with mean $m = -e^{\mu + \frac{\sigma^2}{2}}$, the probability of exceeding the mean is $P(X > m = \frac{1}{2}\text{erfc}\left(-\frac{\sigma}{2\sqrt{2}}\right)$, which for $\sigma = 1$ is at 69%, and for $\sigma = 2$ is at 84%.

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” $\int f_j(x)\,\mathrm{d}x$ and full distribution, called “vanilla payoffs”, $\int x f_j(x)\,\mathrm{d}x$, see Taleb and Tetlock (2013)[11].

Opacity and Risk Hiding: NonMathematical Summary We will next proceed to summarize the mathematical argument in verbal form.

A) If an agent has the upside of the payoff of the random variable, with no downside [OR A DISPROPORTIONATE SHARE OF UPSIDE WITH RESPECT TO THE 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 oneâĂŹs actions.

B) Further, even if it is not intentional, i.e., the agent does not aim at probabilistic rent at the expense of the principal (at variance with the way agents are treated in the economics literature); by a survival argument, those agents without skin in the game who tend to engage in strategies that hide risk in the tail tend to fare better and longer and populate the agent population. So the argument is not one of incentive driving the agents, but one of survival.

We can sketch a demonstration of these statements with the following reasoning. Assume that an agent has a payoff as a proportional cut of his performance or the benefits to the principal, and can get a percentage at year end, his

10 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.

11 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) .

compensation being tied to the visible income. The timing of the compensation is periodic, with no total claw back (subsequent obligation to completely return past compensation). The expected value to the agent is that of a stream, a sum of payoffs over time, extending indefinitely (or bounded by the life of the agent). Assume that a loss will reduce his future risk-taking, or even terminate it, in terms of shrinking of such contracts, owing to change in reputation. A loss would hurt the track record, revealing it so to speak, making such a stream of payoffs stop. In addition, the payoff of the agent is compounded over time as the contracts get larger in response to the track record.

Critically, the principal does not observe statistical properties, only realizations of the random variable. However the agent has an edge over the principal, namely that he can select negatively skewed payoffs. All he needs to do is to figure out the shape of the probability distribution, not its expected returns, nothing else. More technically, the expectation for the agent does not depend on the size of the loss: a small loss or a large loss are the same to him. So the agent can benefit by minimizing the probability of the loss, not the expectation. Minimizing one not the other results in the most possibly negatively skewed distribution.

This result can be extended to include any situation in which the compensation or reward (in any form) to the agent depends on the probability, rather than the true expectation.

In an evolutionary setting, downside harm via skin-in-the-game would create an absorbing state, with the system failing to be ergodic, hence would clean up this class of risk takers.

Part III

(ANTI)FRAGILITY AND NONLINEAR RESPONSES TO RANDOM VARIABLES

15 EXPOSURES AS TRANSFORMED RANDOM VARIABLES

Chapter Summary **15: Deeper into the conflation between a random variable and exposure to it.**

15.1 THE CONFLATION PROBLEM REDUX: EXPOSURES TO X CONFUSED WITH KNOWLEDGE ABOUT X

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.

15.1.1 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.

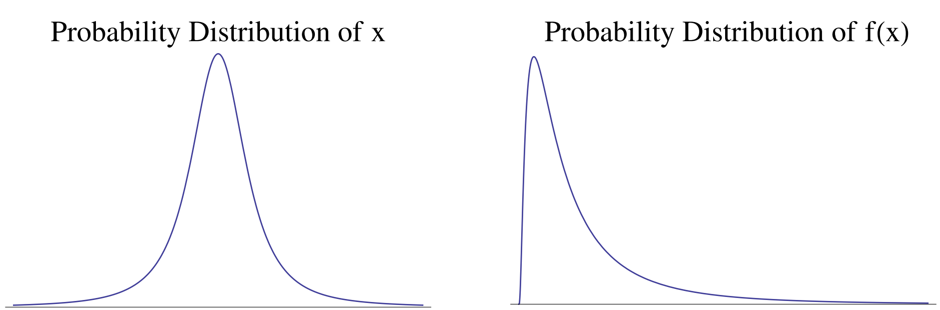


Figure 15.1: The Conflation

The danger with the treatment of the Black Swan problem is as follows: people focus on x ("predicting x"). My point is that, although we do not understand x, we can deal with it by working on F which we can understand, while others work on predicting x which we can't because small probabilities are incomputable, particularly in "fat tailed" domains. f(x) is how the end result affects you.

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)[120].

15.1.2 Bad news

F is almost always nonlinear, often "S curved", that is convex-concave (for an increasing function).

15.1.3 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.

15.1.4 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.

15.2 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

$$g(z)dz = p(x)dx$$

[ASSUMING f is Borel measurable, i.e. has an inverse that is a Borel Set...]

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:

$$g(z) = \frac{p\left(f^{(-1)}(z)\right)}{f'(u)|u = \left(f^{(-1)}(z)\right)} \tag{15.1}$$

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), $\frac{1}{2}(f(x-\Delta \mathrm{x}) + f(\Delta \mathrm{x} + x)) \geq f(x)$, concave if $\frac{1}{2}(f(x-\Delta \mathrm{x}) + f(\Delta \mathrm{x} + x)) \leq f(x)$. 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. [WILL DISCUSS OTHER CASES WHERE WE NEED TO SPLIT THE R.V. IN TWO DOMAINS BECAUSE INVERSE NOT UNIQUE]

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.

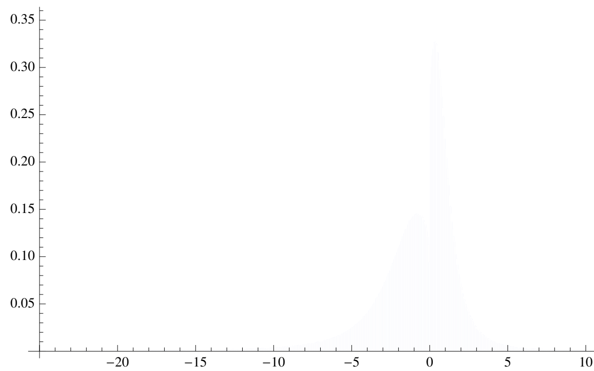


Figure 15.2: Simulation, first. The distribution of the utility of changes of wealth, when the changes in wealth follow a power law with tail exponent =2 (5 million Monte Carlo simulations).

15.3 APPLICATION 1: HAPPINESS ($f(x)$) IS DIFFERENT FROM WEALTH (x)

There is a conflation of fat-tailedness of Wealth and Utility: Happiness ($f(x)$)does not have the same statistical properties as wealth (x)

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 and λ calibrated $a = 0.88$ and $\lambda = 2.25$

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)

$$z(x|\alpha,a,\lambda) = \begin{cases} \dfrac{x^{\frac{1-a}{a}}\left(\frac{\alpha}{\alpha+x^{2/a}}\right)^{\frac{\alpha+1}{2}}}{a\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)} & x \geq 0 \\ \dfrac{\left(-\frac{x}{\lambda}\right)^{\frac{1-a}{a}}\left(\frac{\alpha}{\alpha+\left(-\frac{x}{\lambda}\right)^{2/a}}\right)^{\frac{\alpha+1}{2}}}{a\lambda\sqrt{\alpha}B\left(\frac{\alpha}{2},\frac{1}{2}\right)} & x < 0 \end{cases}$$

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.

v(x) has thinner tails than x $\Leftrightarrow$ more robust.

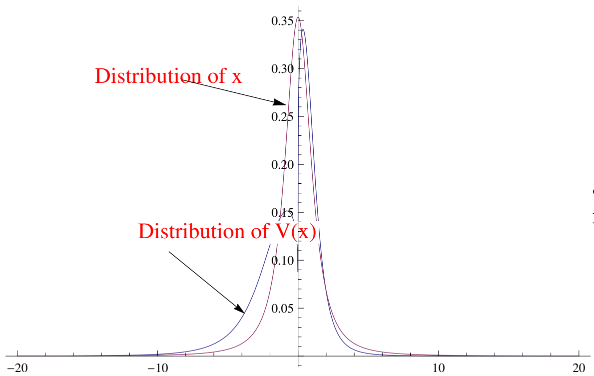


Figure 15.3: The same result derived analytically, *after* the Monte Carlo runs.

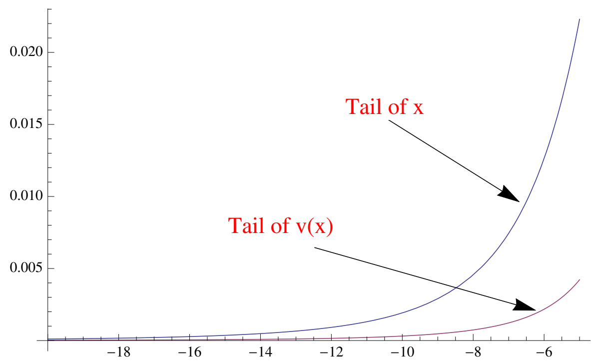


Figure 15.4: Left tail and fragility

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

$$z(x) \sim Kx^{-\frac{\alpha}{a}-1}$$

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.

> The Kahneman-Tversky Prospect function is part of the class of the generalized "S curves", bounded on the left or the right, or with waning derivatives that effectively gives the same result of boundedness (or more exactly, soft-boundedness). Utility theory is unbounded on the left, with preserving derivatives.

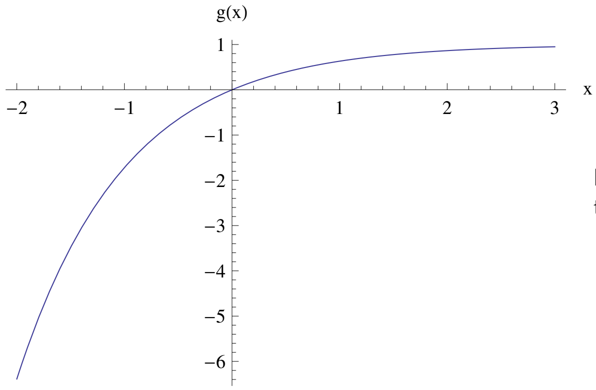


Figure 15.5: Plot of $1 - e^{-ax}$, illustrative of standard utility theory.

Researchers tend to consider that K-T is "empirical" while general utility is "normative". This distinction is absurd: K-T corresponds to normal left-boundedness, as with all dose-responses present in nature. For an organism, death is the *bounded* worst outcome. You cannot keep having detriment unless you have infinite resources or capabilities.

The next section illustrates the more general statement that anything with concave-left exposure has a mean that degrades enormously under increases in uncertainty. Utility theory, under transformation, gives absurd results *mathematically*.

Take the standard concave utility function g(x)= 1- e^{-ax}. With a=1
The distribution of $v(x)$ will be

$$v(x) = -\frac{e^{-\frac{(\mu+\log(1-x))^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma(x-1)}$$

Which can be tolerated owing to the rapid drop in probabilities in the Gaussian tail. But with a fatter tailed distribution, such as the standard powerlaw (a Student T Distribution) (Gabaix, 2008,[47]), where α is the tail exponent,

$$v(x) = \frac{x\left(\frac{\alpha}{\frac{(\log(1-x)-1)^2}{a^2}+\alpha}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}(a-ax)B\left(\frac{\alpha}{2},\frac{1}{2}\right)}$$

With such a distribution of utility it would be absurd to do anything.

15.4 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

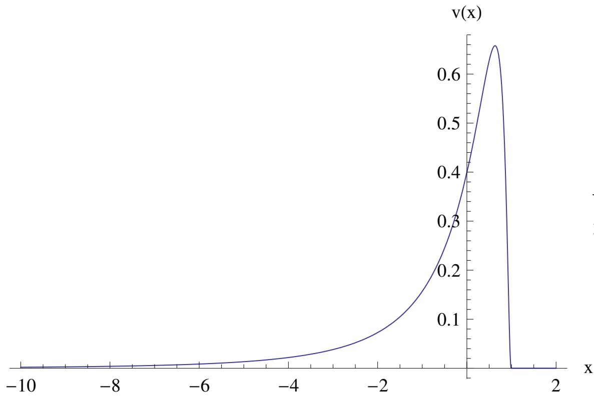
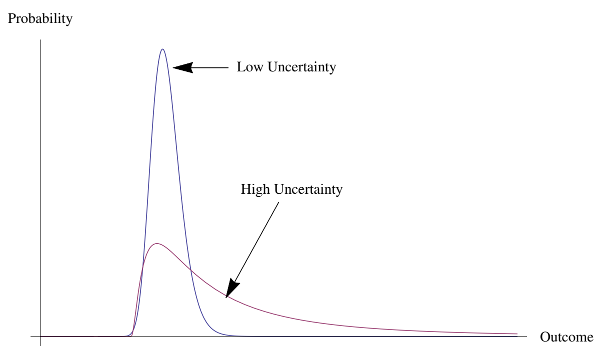


Figure 15.6: Distribution of utility of wealth under probabilistic transformation



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 $\sqrt{\alpha - 2}\sqrt{\alpha} \times$ Standard Deviation, $\frac{\sqrt{\frac{\alpha}{\alpha-2}}L}{\alpha-1}$

15.5 ESTIMATION METHODS WHEN THE PAYOFF IS CONVEX

A simple way to see the point that convex payoffs have larger estimation errors: 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. [1]

1 The same flaw, namely missing convexity, is present in Bodarenko **??**.

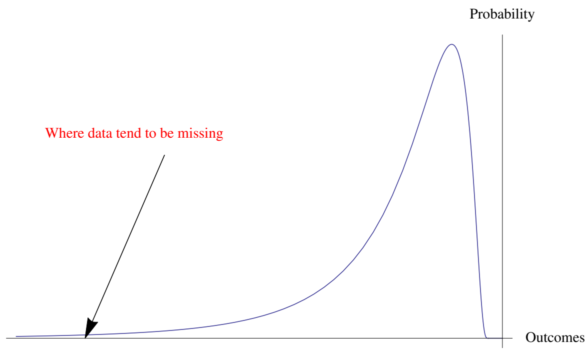


Figure 1: The Small Sample Effect and Naive Empiricism: When one looks at historical returns that are skewed to the left, most missing observations are in the left tails, causing an overestimation of the mean. The more skewed the payoff, and the thicker the left tail, the worst the gap between observed 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 $N\sigma_K$.

$$\begin{aligned}\Pi(N,K) \equiv \frac{1}{\theta_{S_0,t_0},\delta}\Big(&O(S_0 e^{N\sigma_K\sqrt{\delta}},K,T-t_0,\sigma_K)\\ &-O\left(S_0,K,T-t_0-\delta,\sigma_K\right)-\Delta_{S_0,t_0}\left(1-S_0\right)e^{N\sigma_K\sqrt{\delta}}\Big),\end{aligned} \tag{15.2}$$

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 δ = $1/252$, 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 15.2 did is re-express the Fokker-Plank-Kolmogorov differential equation (Black Scholes), in discrete terms, away from the limit of $\delta \rightarrow 0$. In the standard Black-Scholes World, the expectation of $\Pi(N,K)$ should be zero, as N follows a Gaussian distribution with mean 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%, *e*

15.5.1 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 $K = S_0 e^{\Lambda \sigma_K \sqrt{\delta}}$, 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.

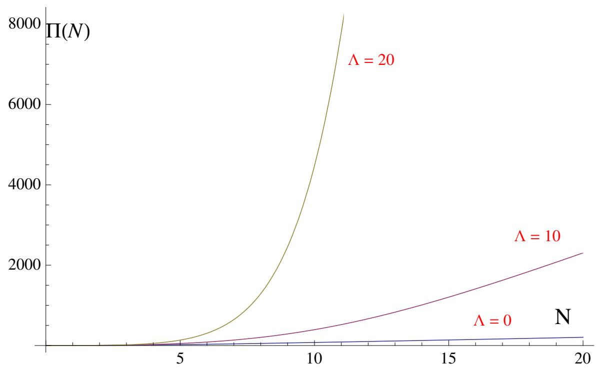
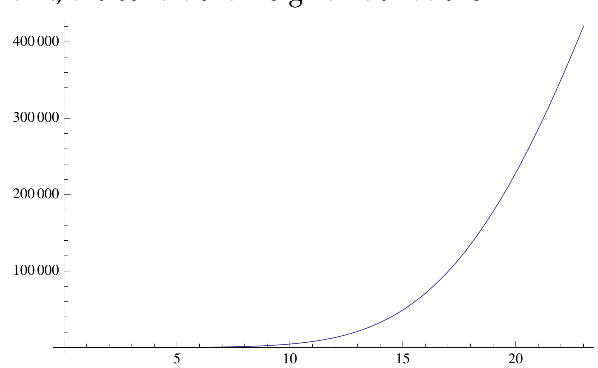


Figure 2: Returns for package $\Pi(N,K= S_0\text{Exp}[\Lambda\ \sigma_K]$) at values of Λ= 0,10,20 and N, the conditional "sigma" deviations.



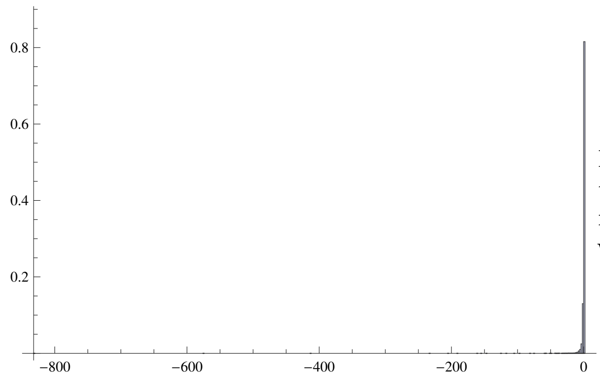


Figure 15.7: In probability space. Histogram of the distribution of the returns Λ=10 using powerlaw returns for underlying distribution with α tail exponent =3.

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 skewed it is.
[2]

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)[105] , Taleb Canetti et al (2012)[100] 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.

2 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.

Table 23: 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

	$\times 2$	$\times 3$	$\times 4$
ATM	2	3	4
$\Lambda = 5$	5	10	16
$\Lambda = 10$	27	79	143
$\Lambda = 20$	7686	72741	208429

15.5.2 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 or payoffs, 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.

16 | AN UNCERTAINTY APPROACH TO FRAGILITY

Chapter Summary 16: **We provide a mathematical approach to fragility as negative sensitivity to a semi-measure of dispersion and volatility (a variant of negative or positive "vega") and examine the link to nonlinear effects. We link to the litterature on model "robustness" and show how we add nonlinearity to the conventional approaches.**

16.1 A REVIEW OF THE GENERAL NOTION OF "ROBUSTNESS"

This section is incomplete; it will present a general review the literature and the variety of definitions of "robustness" in:

- Loss (risk) functions in statistical fitting
- Loss (risk) functions in risk and insurance
- Decision theory (minimax, etc.)
- Statistical robustness -
- Control theory-
  - Stochastic control
- Dynamical systems-
- Economic modelling (see Hansen and Sargent) -

We explain why what we do is add a nonlinear dimension to the loss models and solve some of the issues with loss models since our funtion (nonlinearity of exposure) solves the problem of mini-max with unbounded losses, and how our payoff function ("exposure") maps to the loss function.

16.2 INTRODUCTION

In short, *fragility* as we define it is related to how a system suffers from the variability of its environment beyond a certain preset threshold (when threshold is K, it is called K-fragility), while *antifragility* refers to when it benefits from this variability —in a similar way to "vega" of an option or a nonlinear payoff, that is, its sensitivity to volatility or some similar measure of scale of a distribution.

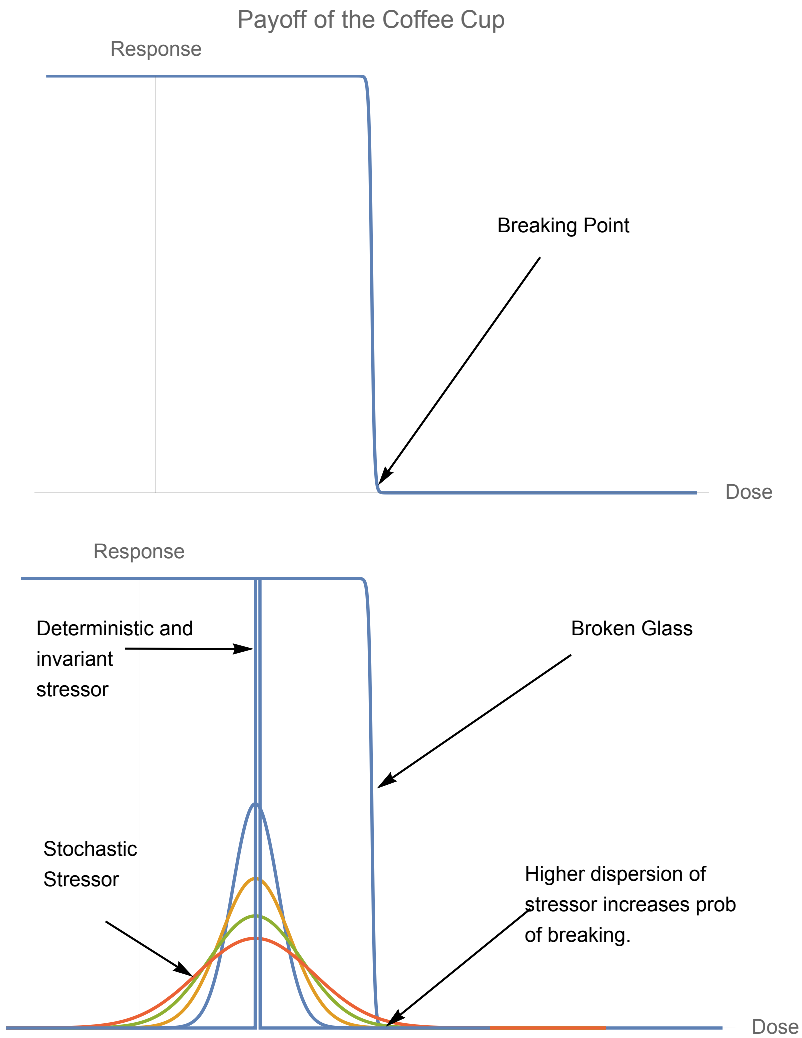


Figure 16.1: Illustrates why the coffee cup is fragile *because* it doesn't like variability. Imagine a stressor of intensity k. A deterministic stressor at k will always be more favorable than a stochastic stressor with average $\bar{k}$. The cup breaks at $k+\delta$, even if there are compensating effects at $k-\delta$ that lower the average. The more dispersion around $\bar{k}$, given the same average, the higher the probability of breakage. This illustrates the dependence of fragility on dispersion ("vega") and the theorems showing how fragility lies in the second order effect. [INCLUDE STOCHASTIC RESONANCE AS AN OPPOSITE EFFECT]

We are not saying that are no other definitions and representations of fragility (although we could not find any that clashes with, or does not fit within our variability approach). Our point is that such a definition allows us to perform analyses based on nonlinearity. Our method is, in a way, inverse "real option" theory ([117],[1]), by with studies of contingent claims are generalized to all decision-making under uncertainty that entail asymmetry of payoff.

Simply, a coffee cup on a table suffers more from large deviations than from the cumulative effect of some shocks—conditional on being unbroken, it has to suffer

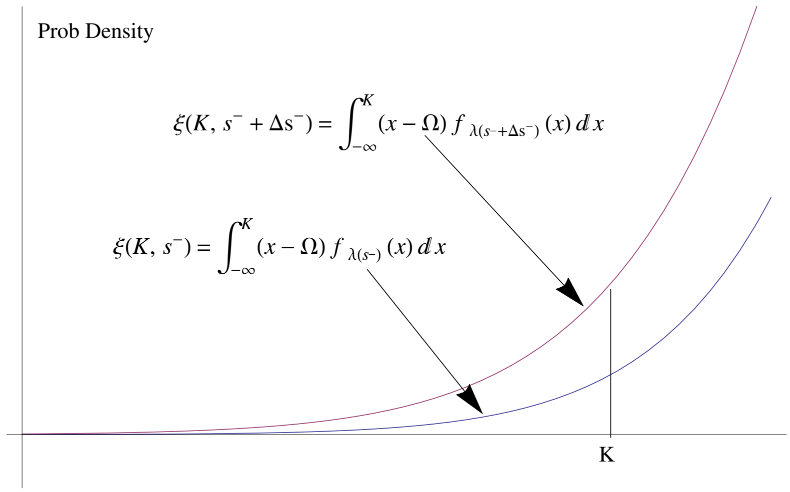


Figure 16.2: A definition of fragility as left tail-vega sensitivity; the figure shows the effect of the perturbation of the lower semi-deviation s^- on the tail integral ξ of $(x - \Omega)$ below K, Ω being a centering constant. Our detection of fragility does not require the specification of f the probability distribution.

more from "tail" events than regular ones around the center of the distribution, the "at the money" category. This is the case of elements of nature that have survived: conditional on being in existence, then the class of events around the mean should matter considerably less than tail events, particularly when the probabilities decline faster than the inverse of the harm, which is the case of all used monomodal probability distributions. Further, what has exposure to tail events suffers from uncertainty; typically, when systems – a building, a bridge, a nuclear plant, an airplane, or a bank balance sheet– are made robust to a certain level of variability and stress but may fail or collapse if this level is exceeded, then they are particularly *fragile* to uncertainty about the distribution of the stressor, hence to model error, as this uncertainty increases the probability of dipping below the robustness level, bringing a higher probability of collapse. In the opposite case, the natural selection of an evolutionary process is particularly *antifragile*, indeed, a more volatile environment increases the survival rate of robust species and eliminates those whose superiority over other species is highly dependent on environmental parameters.

Figure 16.2 show the "tail vega" sensitivity of an object calculated discretely at two different lower absolute mean deviations. We use for the purpose of fragility and antifragility, in place of measures in L^2 such as standard deviations, which restrict the choice of probability distributions, the broader measure of absolute deviation, cut into two parts: lower and upper semi-deviation above the distribution center Ω.

This article aims at providing a proper mathematical definition of fragility, robustness, and antifragility and examining how these apply to different cases where this notion is applicable. [1]

1 Hansen and Sargent, in [54]:"A long tradition dating back to Friedman (...) advocates framing macroeconomic policy rules and interpreting econometric findings in light of doubts about model specification, though how those doubts have been formalized in practice has varied". In fact what we are adding to

16.2.1 Intrinsic and Inherited Fragility:

Our definition of fragility is two-fold. First, of concern is the intrinsic fragility, the shape of the probability distribution of a variable and its sensitivity to s^-, a parameter controlling the left side of its own distribution. But we do not often directly observe the statistical distribution of objects, and, if we did, it would be difficult to measure their tail-vega sensitivity. Nor do we need to specify such distribution: we can gauge the response of a given object to the volatility of an external stressor that affects it. For instance, an option is usually analyzed with respect to the scale of the distribution of the "underlying" security, not its own; the fragility of a coffee cup is determined as a response to a given source of randomness or stress; that of a house with respect of, among other sources, the distribution of earthquakes. This fragility coming from the effect of the underlying is called inherited fragility. The transfer function, which we present next, allows us to assess the effect, increase or decrease in fragility, coming from changes in the underlying source of stress.

Transfer Function: A nonlinear exposure to a certain source of randomness maps into tail-vega sensitivity (hence fragility). We prove that

Inherited Fragility $\Leftrightarrow$ Concavity in exposure on the left side of the distribution

and build H, a transfer function giving an exact mapping of tail vega sensitivity to the second derivative of a function. The transfer function will allow us to probe parts of the distribution and generate a fragility-detection heuristic covering both physical fragility and model error.

16.2.2 Fragility As Separate Risk From Psychological Preferences

16.2.3 Avoidance of the Psychological

We start from the definition of fragility as tail vega sensitivity, and end up with nonlinearity as a necessary attribute of the source of such fragility in the inherited case —a cause of the disease rather than the disease itself. However, there is a long literature by economists and decision scientists embedding risk into psychological preferences —historically, risk has been described as derived from risk aversion as a result of the structure of choices under uncertainty with a concavity of the muddled concept of "utility" of payoff, see Pratt (1964)[90], Arrow (1965) [2], Rothchild and Stiglitz(1970,1971) [96],[97]. But this "utility" business never led anywhere except the circularity, expressed by Machina and Rothschild (1987,2008)[73],[74], "risk is what risk-averters hate." Indeed limiting risk to aversion to concavity of choices is a quite unhappy result —the utility curve cannot be possibly monotone concave, but rather, like everything in nature necessarily bounded on both sides, the left and the right, convex-concave and, as Kahneman and Tversky (1979)[62] have debunked, both path dependent and mixed in its nonlinearity. (See Van Zwet 1964 for the properties of mixed convexities [120].)

---

the story as far as economics here is local and global convexity of the variation and the asymmetry in one side or the other.

16.2.4 Beyond Jensen's Inequality

The economics and decision-theory literature reposes on the effect of Jensen's inequality, an analysis which requires monotone convex or concave transformations [?]—in fact limited to the expectation operator. The world is unfortunately more complicated in its nonlinearities. Thanks to the transfer function, which focuses on the tails, we can accommodate situations where the source is not merely convex, but convex-concave and any other form of mixed nonlinearities common in exposures, which includes nonlinear dose-response in biology. For instance, the application of the transfer function to the Kahneman-Tversky value function, convex in the negative domain and concave in the positive one, shows that its decreases fragility in the left tail (hence more robustness) and reduces the effect of the right tail as well (also more robustness), which allows to assert that we are psychologically "more robust" to changes in wealth than implied from the distribution of such wealth, which happens to be extremely fat-tailed.

Accordingly, our approach relies on nonlinearity of exposure as detection of the vega-sensitivity, not as a definition of fragility. And nonlinearity in a source of stress is necessarily associated with fragility. Clearly, a coffee cup, a house or a bridge don't have psychological preferences, subjective utility, etc. Yet they are concave in their reaction to harm: simply, taking z as a stress level and $\Pi(z)$ the harm function, it suffices to see that, with $n > 1$,

$$\Pi(nz) < n\,\Pi(z) \text{ for all } 0 < n\,z < Z^*$$

where Z^* is the level (not necessarily specified) at which the item is broken. Such inequality leads to $\Pi(z)$ having a negative second derivative at the initial value z.

So if a coffee cup is less harmed by n times a stressor of intensity Z than once a stressor of nZ, then harm (as a negative function) needs to be concave to stressors up to the point of breaking; such stricture is imposed by the structure of survival probabilities and the distribution of harmful events, and has nothing to do with subjective utility or some other figments. Just as with a large stone hurting more than the equivalent weight in pebbles, if, for a human, jumping one millimeter caused an exact linear fraction of the damage of, say, jumping to the ground from thirty feet, then the person would be already dead from cumulative harm. Actually a simple computation shows that he would have expired within hours from touching objects or pacing in his living room, given the multitude of such stressors and their total effect. The fragility that comes from linearity is immediately visible, so we rule it out because the object would be already broken and the person already dead. The relative frequency of ordinary events compared to extreme events is the determinant. In the financial markets, there are at least ten thousand times more events of 0.1% deviations than events of 10%. There are close to 8,000 micro-earthquakes daily on planet earth, that is, those below 2 on the Richter scale —about 3 million a year. These are totally harmless, and, with 3 million per year, you would need them to be so. But shocks of intensity 6 and higher on the scale make the newspapers. Accordingly, we are necessarily immune to the *cumulative* effect of small deviations, or shocks of very small magnitude, which implies that these affect us disproportionally less (that is, nonlinearly less) than larger ones.

Model error is not necessarily mean preserving. s^-, the lower absolute semi-deviation does not just express changes in overall dispersion in the distribution, such as for instance the "scaling" case, but also changes in the mean, i.e. when the upper semi-deviation from Ω to infinity is invariant, or even decline in a compensatory manner to make the overall mean absolute deviation unchanged. This would be the case when we shift the distribution instead of rescaling it. Thus the same vega-sensitivity can also express sensitivity to a stressor (dose increase) in medicine or other fields in its effect on either tail. Thus $s^{-}(l)$ will allow us to express the negative sensitivity to the "disorder cluster" (see *Antifragile*): 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.

Detection Heuristic Finally, thanks to the transfer function, this paper proposes a risk heuristic that "works " in detecting fragility even if we use the wrong model/pricing method/probability distribution. The main idea is that *a wrong ruler will not measure the height of a child; but it can certainly tell us if he is growing***.** Since risks in the tails map to nonlinearities (concavity of exposure), second order effects reveal fragility, particularly in the tails where they map to large tail exposures, as revealed through perturbation analysis. More generally every nonlinear function will produce some kind of positive or negative exposures to volatility for some parts of the distribution.

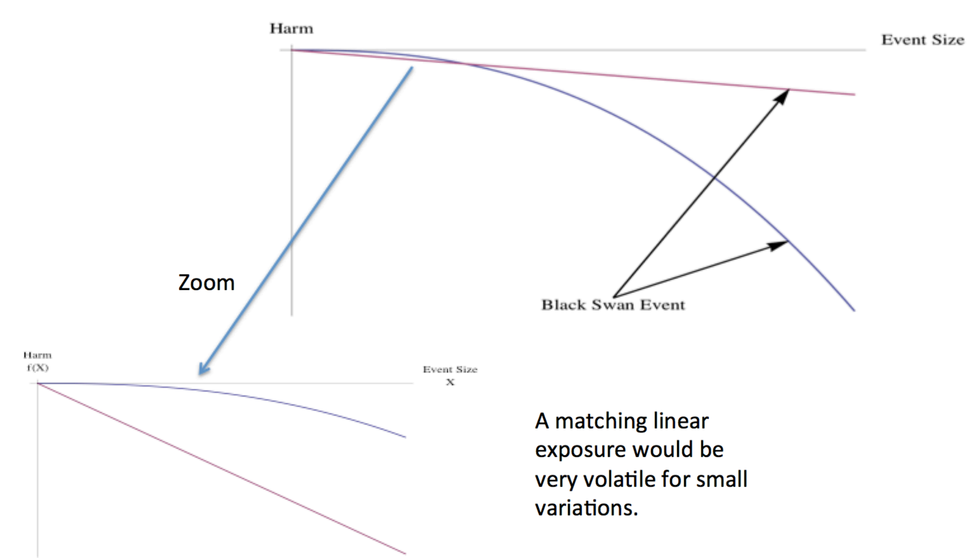


Figure 16.3: Disproportionate effect of tail events on nonlinear exposures, illustrating the necessary character of the nonlinearity of the harm function and showing how we can extrapolate outside the model to probe unseen fragility.

Fragility and Model Error

As we saw this definition of fragility extends to model error, as some models produce negative sensitivity to uncertainty, in addition to effects and biases under variability. So, beyond physical fragility, the same approach measures model fragility, based on the difference between a *point estimate* and stochastic value (i.e., full distribution). Increasing the variability (say, variance) of the estimated value (but not the mean), may lead to one-sided effect on the model —just as an increase of volatility causes porcelain cups to break. Hence sensitivity to the volatility of such value, the "vega" of the model with respect to such value is no different from the vega of other payoffs. For instance, the misuse of thin-tailed distributions (say Gaussian) appears immediately through perturbation of the standard deviation, no longer used as point estimate, but as a distribution with its own variance. For instance, it can be shown how fat-tailed (e.g. power-law tailed) probability distributions can be expressed by simple nested perturbation and mixing of Gaussian ones. Such a representation pinpoints the fragility of a wrong probability model and its consequences in terms of underestimation of risks, stress tests and similar matters.

Antifragility

It is not quite the mirror image of fragility, as it implies positive vega above some threshold in the positive tail of the distribution and absence of fragility in the left tail, which leads to a distribution that is skewed right. Table 24 introduces the Exhaustive Taxonomy of all Possible Payoffs $y = f(x)$

Table 24: Payoffs and Mixed Nonlinearities

Condition	**Left Tail (Loss Domain)**	**Right Tail (Gain Domain)**	**Nonlinear Payoff Function $y = f(x)$ "derivative" where x is a random variable**	**Derivatives Equivalent**	**Effect of fatailedness of $f(x)$ compared to primitive x.**
Fragile (type 1)	Fat (regular or absorbing barrier)	Fat	Mixed concave left, convex right (fence)	Long up-vega, short down-vega	More fragility if absorbing barrier, neutral otherwise
Fragile (type 2)	Thin	Thin	concave	Short vega	More fragility
Robust	Thin	Thin	Mixed convex left, concave right (digital, sigmoid)	Short up - vega, long down - vega	No effect
Antifragile	Thin	Fat (thicker than left)	Convex	Long vega	More antifragility

The central Table, Table 1 introduces the exhaustive map of possible outcomes, with 4 mutually exclusive categories of payoffs. Our steps in the rest of the paper are as follows: a. We provide a mathematical definition of fragility, robustness and antifragility. b. We present the problem of measuring tail risks and show the presence of severe biases attending the estimation of small probability and its nonlinearity (convexity) to parametric (and other) perturbations. c. We express the concept of model fragility in terms of left tail exposure, and show correspondence to the concavity of the payoff from a random variable. d. Finally, we present our simple heuristic to detect the possibility of both fragility and model error across a broad range of probabilistic estimations.

Conceptually, *fragility* resides in the fact that a small – or at least reasonable – uncertainty on the macro-parameter of a distribution may have dramatic consequences on the result of a given stress test, or on some measure that depends on the left tail of the distribution, such as an out-of-the-money option. This hypersensitivity of what we like to call an "out of the money put price" to the macro-parameter, which is *some* measure of the volatility of the distribution of the underlying source of randomness.

Formally, fragility is defined as the sensitivity of the left-tail shortfall (non-conditioned by probability) below a certain threshold K to the overall left semi-deviation of the distribution.

Examples

i- A porcelain coffee cup subjected to random daily stressors from use.

ii- Tail distribution in the function of the arrival time of an aircraft.

iii- Hidden risks of famine to a population subjected to monoculture —or, more generally, fragilizing errors in the application of Ricardo's comparative advantage without taking into account second order effects.

iv- Hidden tail exposures to budget deficits' nonlinearities to unemployment.

v- Hidden tail exposure from dependence on a source of energy, etc. ("squeezability argument").

It also shows why this is necessarily linked to accelerated response, how "size matters". The derivations explain in addition:

- How spreading risks are dangerous compared to limited one we need to weave into the derivations the notion of risk spreading as a non-concave response to make links clearer.
- Why error is a problem in the presence of nonlinearity.
- Why polluting "a little" is qualitatively different from pollution "a lot".
- Eventually, why fat tails arise from accelerating response.

17 | THE FRAGILITY THEOREMS

Chapter Summary **17: Presents the fragility theorems and the transfer function between nonlinear response and the benefits and harm from increased uncertainty.**

The following offers a formal definition of fragility as "vega", negative expected response from uncertainty.

17.1 TAIL SENSITIVITY TO UNCERTAINTY

We construct a measure of "vega", that is, the sensitivity to uncertainty, in the left tails of the distribution that depends on the variations of s the semi-deviation below a certain level W, chosen in the L^1 norm in order to ensure its existence under "fat tailed" distributions with finite first semi-moment. In fact s would exist as a measure even in the case of undefined moments to the right side of W.

Let X be a random variable, the distribution of which is one among a one-parameter family of pdf, $f_\lambda, \lambda \in I \subset \mathbb{R}$. We consider a fixed reference value Ω and, from this reference, the "raw" left-semi-absolute deviation:[1]

$$s^-(\lambda) = \int_{-\infty}^{\Omega} (\Omega - x) f_\lambda(x) \mathrm{d}x \tag{17.1}$$

We assume that $\lambda \to s^-(\lambda)$ is continuous, strictly increasing and spans the whole range $\mathbb{R}_+ = [0, +\infty)$, so that we may use the left-semi-absolute deviation s^- as a parameter by considering the inverse function $\lambda(s) : \mathbb{R}_+ \to I$, defined by $s^-(\lambda(s)) = s$ for $s \in \mathbb{R}_+$.

This condition is for instance satisfied if, for any given $x < \Omega$, the probability is a continuous and increasing function of λ. Indeed, denoting

$$F_\lambda(x) = P_{f_\lambda}(X < x) = \int_{-\infty}^{x} f_\lambda(t)\, \mathrm{d}t, \tag{17.2}$$

an integration by parts yields:

$$s^-(\lambda) = \int_{-\infty}^{\Omega} F_\lambda(x)\, \mathrm{d}x$$

1 We use a measure related to the left-semi absolute deviation, or roughly the half the mean absolute deviation (the part in the negative domain) because 1) distributions are not symmetric and might have changes on the right of Ω that are not of interest to us, 2) standard deviation requires finite second moment.
Further, we do not adjust s^- by its probability –with no loss of generality. Simply, probability in the negative domain is close to $\frac{1}{2}$ and would not change significantly in response to changes in parameters. Probabilities in the tails are nonlinear to changes, not those in the body of the distribution.

This is the case when λ is a scaling parameter, i.e., $X \sim \Omega + \lambda(X_1 - \Omega)$ indeed one has in this case

$$F_\lambda(x) = F_1\left(\Omega + \frac{x-\Omega}{\lambda}\right),$$

$$\frac{\partial F_\lambda}{\partial \lambda}(x) = \frac{\Omega - x}{\lambda^2} f_\lambda(x) \text{ and } s^-(\lambda) = \lambda\, s^-(1).$$

It is also the case when λ is a shifting parameter, i.e. $X \sim X_0 - \lambda$, indeed, in this case $F_\lambda(x) = F_0(x+\lambda)$ and $\frac{\partial s^-}{\partial \lambda}(x) = F_\lambda(\Omega)$.

For $K < \Omega$ and $s \in \mathbb{R}^+$, let:

$$\xi(K, s^-) = \int_{-\infty}^{K} (\Omega - x) f_{\lambda(s^-)}(x) \mathrm{d}x \tag{17.3}$$

In particular, $\xi(\Omega, s^-) = s^-$. We assume, in a first step, that the function $\xi(K,s^-)$ is differentiable on $(-\infty, \Omega] \times \mathbb{R}_+$. The *K-left-tail-vega sensitivity* of X at stress level $K < \Omega$ and deviation level $s^- > 0$ for the pdf f_λ is:

$$V(X, f_\lambda, K, s^-) = \frac{\partial \xi}{\partial s^-}(K, s^-) = \left(\int_{-\infty}^{\Omega} (\Omega - x) \frac{\partial f_\lambda)}{\partial \lambda} \mathrm{d}x\right) \left(\frac{\mathrm{d}s^-}{\mathrm{d}\lambda}\right)^{-1} \tag{17.4}$$

As in the many practical instances where threshold effects are involved, it may occur that ξ does not depend smoothly on s^-. We therefore also define a *finite difference* version of the *vega-sensitivity* as follows:

$$\begin{aligned} V(X, f_\lambda, K, s^-) &= \frac{1}{2\delta s}\left(\xi(K, s^- + \Delta s) - \xi(K, s^- - \Delta s)\right) \\ &= \int_{-\infty}^{K} (\Omega - x) \frac{f_\lambda(s^- + \Delta s)(x) - f_\lambda(s^- - \Delta s)(x)}{2\,\Delta\, s}\, \mathrm{d}x \end{aligned} \tag{17.5}$$

Hence omitting the input Δs implicitly assumes that $\Delta s \to 0$.

Note that $\xi(K, s^-) = -\mathbb{E}(X|X < K)\, \mathbb{P}_{f_\lambda}(X < K)$. It can be decomposed into two parts:

$$\xi\left(K, s^-(\lambda)\right) = (\Omega - K) F_\lambda(K) + P_\lambda(K) \tag{17.6}$$

$$P_\lambda(K) = \int_{-\infty}^{K} (K - x) f_\lambda(x)\, \mathrm{d}x \tag{17.7}$$

Where the first part $(\Omega - K)F_\lambda(K)$ is proportional to the probability of the variable being below the stress level K and the second part $P_\lambda(K)$ is the expectation of the amount by which X is below K (counting 0 when it is not). Making a parallel with financial options, while $s^-(\lambda)$ is a "put at-the-money", $\xi(K,s^-)$ is the sum of a put struck at K and a digital put also struck at K with amount $\Omega - K$; it can equivalently be seen as a put struck at Ω with a down-and-in European barrier at K.

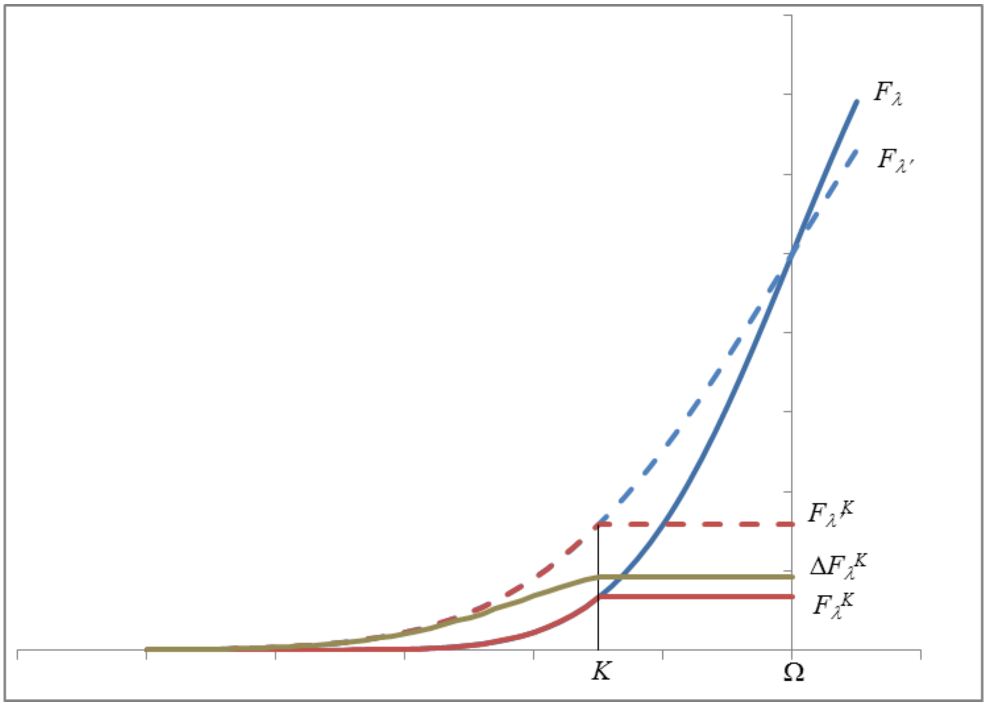


Figure 17.1: The different curves of $F_\lambda(K)$ and $F_{\lambda'}(K)$ showing the difference in sensitivity to changes at different levels of K.

Letting $\lambda = \lambda(s^-)$ and integrating by part yields

$$\xi\left(K, s^-(\lambda)\right) = (\Omega - K)F_\lambda(K) + \int_{-\infty}^{K} F_\lambda(x)\mathrm{d}x = \int_{-\infty}^{\Omega} F_\lambda^K(x)\,\mathrm{d}x \quad (17.8)$$

Where $F_\lambda^K(x) = F_\lambda\left(\min(x, K)\right) = \min\left(F_\lambda(x), F_\lambda(K)\right)$, so that

$$V(X, f_\lambda, K, s^-) = \frac{\partial \xi}{\partial s}(K, s^-) = \frac{\int_{-\infty}^{\Omega} \frac{\partial F_\lambda^K}{\partial \lambda}(x)\,\mathrm{d}x}{\int_{-\infty}^{\Omega} \frac{\partial F_\lambda}{\partial \lambda}(x)\,\mathrm{d}x} \quad (17.9)$$

For finite differences

$$V(X, f_\lambda, K, s^-, \Delta s) = \frac{1}{2\Delta\, s} \int_{-\infty}^{\Omega} \Delta F_{\lambda,\Delta s}^K(x)\mathrm{d}x \quad (17.10)$$

where λ_s^+ and λ_s^- are such that $s(\lambda_{s^-}^+) = s^- + \Delta s$, $s(\lambda_{s^-}^-) = s^- - \Delta s$ and $\Delta F_{\lambda,\Delta s}^K(x) = F_{\lambda_{s^+}}^K(x) - F_{\lambda_{s^-}}^K(x)$.

17.1.1 Precise Expression of Fragility

In essence, fragility is the sensitivity of a given risk measure to an error in the estimation of the (possibly one-sided) deviation parameter of a distribution, especially due to the fact that the risk measure involves parts of the distribution – tails – that are away from the portion used for estimation. The risk measure then assumes certain extrapolation rules that have first order consequences. These consequences are even more amplified when the risk measure applies to a variable that is derived from that used for estimation, when the relation between the two variables is strongly nonlinear, as is often the case.

Definition of Fragility: The *Intrinsic* Case *The local fragility of a random variable X_λ depending on parameter λ, at stress level K and semi-deviation level $s^-(\lambda)$ with pdf f_λ is its K-left-tailed semi-vega sensitivity $V(X, f_\lambda, K, s^-)$.*

The finite-difference fragility of X_λ at stress level K and semi-deviation level $s^-(\lambda) \pm \Delta s$ with pdf f_λ is its K-left-tailed finite-difference semi-vega sensitivity $V(X, f_\lambda, K, s^-, \Delta s)$.

In this definition, the *fragility* relies in the unsaid assumptions made when extrapolating the distribution of X_λ from areas used to estimate the semi-absolute deviation $s^-(\lambda)$, around Ω, to areas around K on which the risk measure ξ depends.

Definition of Fragility: The *Inherited* Case Next we consider the particular case where a random variable $Y = \varphi(X)$ depends on another source of risk X, itself subject to a parameter λ. Let us keep the above notations for X, while we denote by g_λ the pdf of Y,$\Omega_Y = \varphi(\Omega)$ and $u^-(\lambda)$ the left-semi-deviation of Y. Given a "strike" level

$L = \varphi(K)$, let us define, as in the case of X :

$$\zeta\left(L, u^-(\lambda)\right) = \int_{-\infty}^{K} (\Omega_Y - y) g_\lambda(y)\, \mathrm{d}y \tag{17.11}$$

The inherited fragility of Y with respect to X at stress level $L = \varphi(K)$ and left-semi-deviation level $s^-(\lambda)$ of X is the partial derivative:

$$V_X\left(Y, g_\lambda, L, s^-(\lambda)\right) = \frac{\partial \zeta}{\partial s}\left(L, u^-(\lambda)\right) = \left(\int_{-\infty}^{K} (\Omega_Y - Y) \frac{\partial g_\lambda}{\partial \lambda}(y) \mathrm{d}y\right) \left(\frac{\mathrm{d}s^-}{\mathrm{d}\lambda}\right)^{-1} \tag{17.12}$$

Note that the stress level and the pdf are defined for the variable Y, but the parameter which is used for differentiation is the left-semi-absolute deviation of X, $s^-(\lambda)$. Indeed, in this process, one first measures the distribution of X and its left-semi-absolute deviation, then the function φ is applied, using some mathematical model of Y with respect to X and the risk measure ζ is estimated. If an error is made when measuring $s^-(\lambda)$, its impact on the risk measure of Y is amplified by the ratio given by the "inherited fragility".

Once again, one may use finite differences and define the *finite-difference inherited fragility* of Y with respect to X, by replacing, in the above equation, differentiation by finite differences between values λ^+ and λ^-, where $s^-(\lambda^+) = s^- + \Delta s$ and $s^-(\lambda^-) = s^- - \Delta s$.

17.2 EFFECT OF NONLINEARITY ON INTRINSIC FRAGILITY

Let us study the case of a random variable $Y = \varphi(X)$; the pdf g_λ of which also depends on parameter λ, related to a variable X by the nonlinear function φ. We are now interested in comparing their *intrinsic fragilities*. We shall say, for instance, that Y is *more fragilefragile* at the stress level L and left-semi-deviation level $u^-(\lambda)$ than the random variable X, at stress level K and left-semi-deviation level $s^-(\lambda)$ if the L-left-tailed semi-vega sensitivity of Y_λ is higher than the K-left-tailed semi-vega sensitivity of X_λ:

$$V(Y, g_\lambda, L, \mu^-) > V(X, f_\lambda, K, s^-) \tag{17.13}$$

One may use finite differences to compare the fragility of two random variables:$V(Y, g_\lambda, L, \Delta\mu) > V(X, f_\lambda, K, \Delta s)$. In this case, finite variations must be comparable in size, namely $\Delta u/u^- = \Delta s/s^-$.

Let us assume, to start, that φ is differentiable, strictly increasing and scaled so that $\Omega_Y = \varphi(\Omega) = \Omega$. We also assume that, for any given $x < \Omega$, $\frac{\partial F_\lambda}{\partial \lambda}(x) > 0$.

In this case, as observed above, $\lambda \to s^-(\lambda)$ is also increasing.

Let us denote $G_y(y) = \mathbb{P}_{g_\lambda}(Y < y)$. We have:

$$G_\lambda\left(\phi(x)\right) = \mathbb{P}_{g_\lambda}\left(Y < \phi(y)\right) = \mathbb{P}_{f_\lambda}(X < x) = F_\lambda(x). \tag{17.14}$$

Hence, if $\zeta(L, u^-)$ denotes the equivalent of $\xi(K), s^-$ with variable (Y, g_λ) instead of (X, f_λ), we have:

$$\zeta\left(L, u^-(\lambda)\right) = \int_{-\infty}^{\Omega} F_\lambda^K(x)\frac{d\phi}{dx}(x)\mathrm{d}x \tag{17.15}$$

Because φ is increasing and min($\varphi(x)$,$\varphi(K)$) = φ(min(x,K)). In particular

$$\mu^-(\lambda) = \zeta\left(\Omega, \mu^-(\lambda)\right) = \int_{-\infty}^{\Omega} F_\lambda^K(x)\frac{d\phi}{\mathrm{d}x}(x)\,\mathrm{d}x \tag{17.16}$$

The L-left-tail-vega sensitivity of Y is therefore:

$$V\left(Y, g_\lambda, L, u^-(\lambda)\right) = \frac{\int_{-\infty}^{\Omega} \frac{\partial F_\lambda^K}{\partial\lambda}(x)\frac{d\phi}{dx}(x)\,\mathrm{d}x}{\int_{-\infty}^{\Omega} \frac{\partial F_\lambda}{\partial\lambda}(x)\frac{d\phi}{dx}(x)\,\mathrm{d}x} \tag{17.17}$$

For finite variations:

$$V(Y, g_\lambda, L, u^-(\lambda), \Delta u) = \frac{1}{2\Delta u}\int_{-\infty}^{\Omega} \Delta F_{\lambda,\Delta u}^K(x)\frac{d\phi}{dx}(x)\mathrm{d}x \tag{17.18}$$

Where $\lambda_{u^-}^+$ and $\lambda_{u^-}^-$ are such that $u(\lambda_{u^-}^+) = u^- + \Delta u$, $u(\lambda_{u^-}^+) = u^- - \Delta u$ and $F_{\lambda,\Delta u}^K(x) = F_{\lambda_u^+}^K(x) - F_{\lambda_u^-}^K(x)$.

Next, Theorem 1 proves how a concave transformation $\varphi(x)$ of a random variable x produces fragility.

Fragility Transfer Theorem

Theorem 17.1.
Let, with the above notations, $\varphi : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function such that $\varphi(\Omega) = \Omega$ and for any $x < \Omega$, $\frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) > 0$. The random variable $Y = \varphi(X)$ is more fragile at level $L = \varphi(K)$ and pdf g_λ than X at level K and pdf f_λ if, and only if, one has:

$$\int_{-\infty}^{\Omega} H_\lambda^K(x) \frac{d^2\varphi}{dx^2}(x)\mathrm{d}x < 0$$

Where

$$H_\lambda^K(x) = \frac{\partial P_\lambda^K}{\partial \lambda}(x) \Big/ \frac{\partial P_\lambda^K}{\partial \lambda}(\Omega) - \frac{\partial P_\lambda}{\partial \lambda}(x) \Big/ \frac{\partial P_\lambda}{\partial \lambda}(\Omega) \tag{17.19}$$

and where

$$P_\lambda(x) = \int_{-\infty}^{x} F_\lambda(t)dt \tag{17.20}$$

is the price of the "put option" on X_λ with "strike" x and

$$P_\lambda^K(x) = \int_{-\infty}^{x} F_\lambda^K(t)dt$$

is that of a "put option" with "strike" x and "European down-and-in barrier" at K.

H can be seen as a *transfer function*, expressed as the difference between two ratios. For a given level x of the random variable on the left hand side of Ω, the second one is the ratio of the vega of a put struck at x normalized by that of a put "at the money" (i.e. struck at Ω), while the first one is the same ratio, but where puts struck at x and Ω are "European down-and-in options" with triggering barrier at the level K.

The proof is detailed in [105] and [**?**].
Fragility Exacerbation Theorem

Theorem 17.2.
With the above notations, there exists a threshold $\Theta_\lambda < \Omega$ such that, if $K \leq \Theta_\lambda$ then $H_\lambda^K(x) > 0$ for $x \in (\infty, \kappa_\lambda]$ with $K < \kappa_lambda < \Omega$.As a consequence, if the change of variable φ is concave on $(-\infty, \kappa_\lambda]$ and linear on $[\kappa_\lambda, \Omega]$, then Y is more fragile at $L = \varphi(K)$than X at K.

One can prove that, for a monomodal distribution, $\Theta_\lambda < \kappa_\lambda < \Omega$ (see discussion below), so whatever the stress level K below the threshold Θ_λ, it suffices that the change of variable φ be concave on the interval $(-\infty, \Theta_\lambda]$ and linear on $[\Theta_lambda, \Omega]$ for Y to become more fragile at L than X at K. In practice, as long as the change of variable is concave around the stress level K and has limited convexity/concavity away from K, the fragility of Y is greater than that of X.

Figure 17.2 shows the shape of $H_\lambda^K(x)$ in the case of a Gaussian distribution where λ is a simple scaling parameter (λ is the standard deviation σ) and $\Omega = 0$. We represented $K = -2\lambda$ while in this Gaussian case, $\Theta_\lambda = -1.585\lambda$.

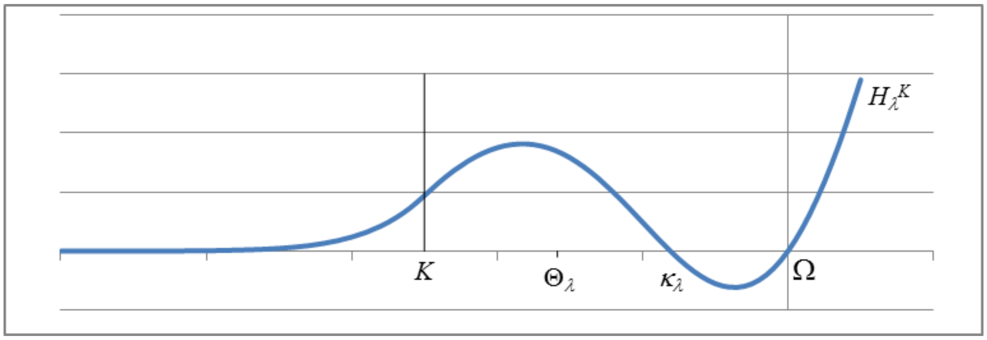


Figure 17.2: The Transfer function H for different portions of the distribution: its sign flips in the region slightly below Ω

Discussion

Monomodal case

We say that the family of distributions (f_λ) is *left-monomodal* if there exists $K_\lambda < \Omega$ such that $\frac{\partial f_\lambda}{\partial \lambda} \geqslant 0$ on $(-\infty, \kappa_\lambda]$ and $\frac{\partial f_\lambda}{\partial \lambda} \leqslant 0$ on $[\mu_\lambda, \Omega]$. In this case $\frac{\partial P_\lambda}{\partial \lambda}$ is a convex function on the left half-line $(-\infty, \mu_\lambda]$, then concave after the inflexion point μ_λ. For $K \leq \mu_\lambda$, the function $\frac{\partial P_\lambda^K}{\partial \lambda}$ coincides with $\frac{\partial P_\lambda}{\partial \lambda}$ on $(-\infty, K]$, then is a linear extension, following the tangent to the graph of $\frac{\partial P_\lambda}{\partial \lambda}$ in K (see graph below). The value of $\frac{\partial P_\lambda^K}{\partial \lambda}(\Omega)$ corresponds to the intersection point of this tangent with the vertical axis. It increases with K, from 0 when $K \to -\infty$ to a value above $\frac{\partial P_\lambda}{\partial \lambda}(\Omega)$ when $K = \mu_\lambda$. The threshold Θ_λ corresponds to the unique value of K such that $\frac{\partial P_\lambda^K}{\partial \lambda}(\Omega) = \frac{\partial P_\lambda}{\partial \lambda}(\Omega)$. When $K < \Theta_\lambda$ then $G_\lambda(x) = \frac{\partial P_\lambda}{\partial \lambda}(x) \Big/ \frac{\partial P_\lambda}{\partial \lambda}(\Omega)$ and $G_\lambda^K(x) = \frac{\partial P_\lambda^K}{\partial \lambda}(x) \Big/ \frac{\partial P_\lambda^K}{\partial \lambda}(\Omega)$ are functions such that $G_\lambda(\Omega) = G_\lambda^K(\Omega) = 1$ and which are proportional for $x \leq K$, the latter being linear on $[K, \Omega]$. On the other hand, if $K < \Theta_\lambda$ then $\frac{\partial P_\lambda^K}{\partial \lambda}(\Omega) < \frac{\partial P_\lambda}{\partial \lambda}(\Omega)$ and $G_\lambda(K) < G_\lambda^K(K)$, which implies that $G_\lambda(x) < G_\lambda^K(x)$ for $x \leq K$. An elementary convexity analysis shows that, in this case, the equation $G_\lambda(x) = G_\lambda^K(x)$ has a unique solution κ_λ with $\mu_l ambda < \kappa_\lambda < \Omega$. The "transfer" function $H_\lambda^K(x)$ is positive for $x < \kappa_\lambda$, in particular when $x \leq \mu_\lambda$ and negative for $\kappa_\lambda < x < \Omega$.

Scaling Parameter

We assume here that λ is a scaling parameter, i.e. $X_\lambda = \Omega + \lambda(X_1 - \Omega)$. In this case, as we saw above, we have

$$f_\lambda(x) = \frac{1}{\lambda} f_1 \left(\Omega + \frac{x - \Omega}{\lambda} \right), F_\lambda(x) = F_1 \left(\Omega + \frac{x - \Omega}{\lambda} \right)$$

$$P_\lambda(x) = \lambda P_1 \left(\Omega + \frac{x - \Omega}{\lambda} \right) \text{ and } s^-(\lambda) = \lambda s^-(1).$$

Hence

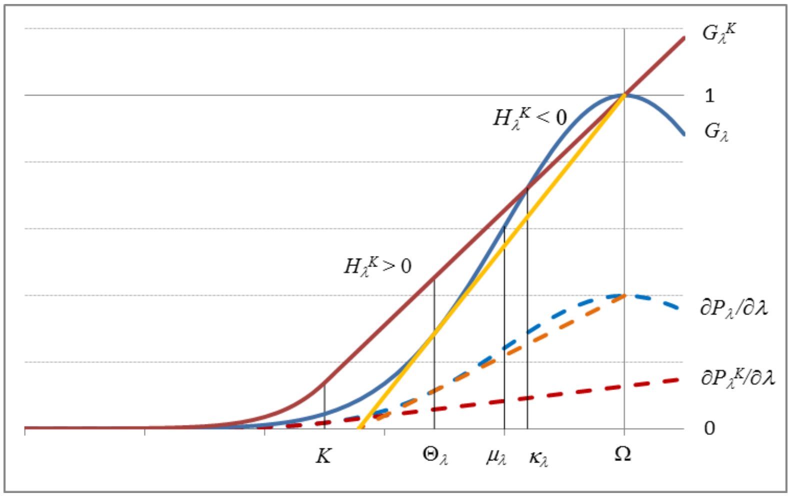


Figure 17.3: The distribution of G_λ and the various derivatives of the unconditional shortfalls

$$\xi(K, s^-(\lambda)) = (\Omega - K)F_1\left(\Omega + \frac{K-\Omega}{\lambda}\right) + \lambda P_1\left(\Omega + \frac{K-\Omega}{\lambda}\right) \quad (17.21)$$

$$\frac{\partial \xi}{\partial s^-}(K, s^-) = \frac{1}{s^-(1)}\frac{\partial \xi}{\partial \lambda}(K, \lambda) = \frac{1}{s^-(\lambda)}\left(P_\lambda(K) + (\Omega - K)F_\lambda(K) + (\Omega - K)^2 f_\lambda(K)\right) \quad (17.22)$$

When we apply a nonlinear transformation φ, the action of the parameter λ is no longer a scaling: when small negative values of X are multiplied by a scalar λ, so are large negative values of X. The scaling λ applies to small negative values of the transformed variable Y with a coefficient $\frac{d\varphi}{dx}(0)$, but large negative values are subject to a different coefficient $\frac{d\varphi}{dx}(K)$, which can potentially be very different.

17.3 FRAGILITY DRIFT

To summarize, textitFragility is defined at as the sensitivity – i.e. the first partial derivative – of the tail estimate ξ with respect to the left semi-deviation s^-. Let us now define the *fragility drift*:

$$V_K'(X, f_\lambda, K, s^-) = \frac{\partial^2 \xi}{\partial K \partial s^-}(K, s^-) \quad (17.23)$$

In practice, fragility always occurs as the result of *fragility*, indeed, by definition, we know that $\xi(\Omega, s^-) = s^-$, hence $V(X, f_\lambda, \Omega, s^-) = 1$. The *fragility drift* measures the speed at which fragility departs from its original value 1 when K departs from the center Ω.

17.3.1 Second-order Fragility

The *second-order fragility* is the second order derivative of the tail estimate ξ with respect to the semi-absolute deviation s^-:

$$V'_{s^-}(X, f_\lambda, K, s^-) = \frac{\partial^2 \xi}{(\partial s^-)^2}(K, s^-)$$

As we shall see later, the *second-order fragility* drives the bias in the estimation of stress tests when the value of s^- is subject to uncertainty, through Jensen's inequality.

17.4 EXPRESSIONS OF ROBUSTNESS AND ANTIFRAGILITY

Antifragility is not the simple opposite of fragility, as we saw in Table 1. Measuring antifragility, on the one hand, consists of the flipside of fragility on the right-hand side, but on the other hand requires a control on the *robustness* of the probability distribution on the left-hand side. From that aspect, unlike fragility, antifragility cannot be summarized in one single figure but necessitates at least two of them.

When a random variable depends on another source of randomness: $Y_\lambda = \varphi(X_\lambda)$, we shall study the antifragility of Y_λ with respect to that of X_λ and to the properties of the function φ.

17.4.1 Definition of Robustness

Let (X_λ) be a one-parameter family of random variables with pdf f_λ. Robustness is an upper control on the *fragility* of X, which resides on the left hand side of the distribution.

We say that f_λ is b-robust beyond stress level $K < \Omega$ if $V(X_\lambda, f_\lambda, K', s(\lambda)) \leq b$ for any $K' \leq K$. In other words, the robustness of f_λ on the half-line $(-\infty, K]$ is

$$R_{(-\infty,K]}(X_\lambda, f_\lambda, K, s^-(\lambda)) = \max_{K' \leqslant K} V(X_\lambda, f_\lambda, K', s^-(\lambda)), \tag{17.24}$$

so that b-robustness simply means

$$R_{(-\infty,K]}(X_\lambda, f_\lambda, K, s^-(\lambda)) \leqslant b$$

We also define *b-robustness over a given interval* $[K_1, K_2]$ by the same inequality being valid for any $K' \in [K_1, K_2]$. In this case we use

$$R_{[K_1,K_2]}(X_\lambda, f_\lambda, K, s^-(\lambda)) = \max_{K_1 \leqslant K' \leqslant K_2} V(X_\lambda, f_\lambda, K', s^-(\lambda)). \quad (17.25)$$

Note that the *lower R*, the tighter the control and the *more robust* the distribution f_λ.

Once again, the definition of *b-robustness* can be transposed, using finite differences $V(X_\lambda, f_\lambda, K', s^-(\lambda), \Delta s)$.

In practical situations, setting a material upper bound b to the fragility is particularly important: one need to be able to come with actual estimates of the impact of the error on the estimate of the left-semi-deviation. However, when dealing with certain class of models, such as Gaussian, exponential of stable distributions, we may be lead to consider asymptotic definitions of robustness, related to certain classes.

For instance, for a given decay exponent $a > 0$, assuming that $f_\lambda(x) = O(e^{ax})$ when $x \to -\infty$, the a-exponential asymptotic robustness of X_λ below the level K is:

$$R_{\exp}(X_\lambda, f_\lambda, K, s^-(\lambda), a) = \max_{K' \leqslant K} \left(e^{a(\Omega - K')} V(X_\lambda, f_\lambda, K', s^-(\lambda)) \right) \quad (17.26)$$

If one of the two quantities $e^{a(\Omega-K')} f_\lambda(K')$ or $e^{a(\Omega-K')} V(X_\lambda, f_\lambda, K', s^-(\lambda))$ is not bounded from above when $K \to -\infty$, then $R_{\exp} = +\infty$ and X_λ is considered as not a-exponentially robust.

Similarly, for a given power $\alpha > 0$, and assuming that $f_\lambda(x) = O(x^{-\alpha})$ when $x \to -\infty$, the α-power asymptotic robustness of X_λ below the level K is:

$$R_{\text{pow}}(X_\lambda, f_\lambda, K, s^-(\lambda), a) = \max_{K' \leqslant K} \left((\Omega - K')^{\alpha-2} V(X_\lambda, f_\lambda, K', s^-(\lambda)) \right)$$

If one of the two quantities

$$(\Omega - K')^{\alpha} f_\lambda(K')$$

$$(\Omega - K')^{\alpha-2} V(X_\lambda, f_\lambda, K', s^-(\lambda))$$

is not bounded from above when $K' \to -\infty$, then $R_{pow} = +\infty$ and X_λ is considered as not α-power robust. Note the exponent $\alpha - 2$ used with the fragility, for homogeneity reasons, e.g. in the case of stable distributions, when a random variable $Y_\lambda = \varphi(X_\lambda)$ depends on another source of risk X_λ.

Definition 17.1.
Left-Robustness (monomodal distribution). A payoff $y = \varphi(x)$ is said (a, b)-robust below

$L = \varphi(K)$ for a source of randomness X with pdf f_λ assumed monomodal if, letting g_λ be the pdf of $Y = \varphi(X)$, one has,for any $K' \leq K$ and $L = \varphi(K)$:

$$V_X\left(Y, g_\lambda, L', s^-(\lambda)\right) \leqslant aV\left(X, f_\lambda, K', s^-(\lambda)\right) + b \tag{17.27}$$

The quantity b is of order deemed of "negligible utility" (subjectively), that is, does not exceed some tolerance level in relation with the context, while a is a scaling parameter between variables X and Y.

Note that robustness is in effect impervious to changes of probability distributions. Also note that this measure of robustness ignores first order variations since owing to their higher frequency, these are detected (and remedied) very early on.

Example of Robustness (Barbells):

a. trial and error with bounded error and open payoff

b. for a "barbell portfolio " with allocation to numeraire securities up to 80% of portfolio, no perturbation below K set at 0.8 of valuation will represent any difference in result, i.e. $q = 0$. The same for an insured house (assuming the risk of the insurance company is not a source of variation), no perturbation for the value below K, equal to minus the insurance deductible, will result in significant changes.

c. a bet of amount B (limited liability) is robust, as it does not have any sensitivity to perturbations below 0.

17.4.2 Antifragility

The second condition of *antifragility* regards the *right hand side* of the distribution. Let us define the *right-semi-deviation* of X :

$$s^+(\lambda) = \int_\Omega^{+\infty} (x - \Omega) f_\lambda(x) \mathrm{d}x$$

And, for $H > L > \Omega$:

$$\xi^+(L, H, s^+(\lambda)) = \int_L^H (x - \Omega) f_\lambda(x) \mathrm{d}x$$

$$\begin{aligned} W(X, f_\lambda, L, H, s^+) &= \frac{\partial \xi^+(L, H, s^+)}{\partial s^+} \\ &= \left(\int_L^H (x - \Omega) \frac{\partial f_\lambda}{\partial \lambda}(x) \mathrm{d}x\right) \left(\int_\Omega^{+\infty} (x - \Omega) \frac{\partial f_\lambda}{\partial \lambda}(x) \mathrm{d}x\right)^{-1} \end{aligned}$$

When $Y = \varphi$ is a variable depending on a source of noise X,we define:

$$\begin{aligned} &W_X(Y, g_\lambda, \varphi(L), \varphi(H), s^+) = \\ &\quad \left(\int_{\varphi(L)}^{\varphi(H)} (y - \varphi(\Omega)) \frac{\partial g_\lambda}{\partial \lambda}(y) dy\right) \left(\int_\Omega^{+\infty} (x - \Omega) \frac{\partial f_\lambda}{\partial \lambda}(x) \mathrm{d}x\right)^{-1} \end{aligned} \tag{17.28}$$

Definition 2b, Antifragility (monomodal distribution). *A payoff $y = \varphi(x)$ is locally antifragile over the range $[L, H]$ if*

1. *It is b-robust below* Ω *for some* $b > 0$

2. $W_X\left(Y, g_\lambda, \varphi(L), \varphi(H), s^+(\lambda)\right) \geqslant aW\left(X, f_\lambda, L, H, s^+(\lambda)\right)$ *where* $a = \frac{u^+(\lambda)}{s^+(\lambda)}$

The scaling constant a provides homogeneity in the case where the relation between X and y is linear. In particular, nonlinearity in the relation between X and Y impacts robustness.

The second condition can be replaced with finite differences Δu and Δs, as long as $\Delta u/u = \Delta s/s$.

17.4.3 Remarks

Fragility is K-specific We are only concerned with adverse events below a certain pre-specified level, the breaking point. Exposures A can be more fragile than exposure B for $K = 0$, and much less fragile if K is, say, 4 mean deviations below 0. We may need to use finite Δs to avoid situations as we will see of vega-neutrality coupled with short left tail.

Effect of using the wrong distribution or model Comparing $V(X, f, K, s^-, \Delta s)$ and the alternative distribution $V(X, f^*, K, s^*, \Delta s)$, where $f*$ is the "true" distribution, the measure of fragility provides an acceptable indication of the sensitivity of a given outcome – such as a risk measure – to model error, provided no "paradoxical effects" perturbate the situation. Such "paradoxical effects" are, for instance, a change in the direction in which certain distribution percentiles react to model parameters, like s^-. It is indeed possible that nonlinearity appears between the core part of the distribution and the tails such that when s^- increases, the left tail starts fattening – giving a large measured fragility – then steps back – implying that the real fragility is lower than the measured one. The opposite may also happen, implying a dangerous under-estimate of the fragility. These nonlinear effects can stay under control provided one makes some regularity assumptions on the actual distribution, as well as on the measured one. For instance, paradoxical effects are typically avoided under at least one of the following three hypotheses:

1. The class of distributions in which both f and f^* are picked are all monomodal, with monotonous dependence of percentiles with respect to one another.

2. The difference between percentiles of f and f^* has constant sign (i.e. f^* is either *always* wider or *always* narrower than f at any given percentile)

3. For any strike level K (in the range that matters), the fragility measure V monotonously depends on s^- on the whole range where the true value s^* can be expected. This is in particular the case when partial derivatives $\partial^k V/\partial s^k$ all have the same sign at measured s^- up to some order n, at which the partial derivative has that same constant sign over the whole range on which the true value s^* can be expected. This condition can be replaced by an assumption on finite differences approximating the higher order partial derivatives, where n is large enough so that the interval $[s^-, n\Delta s]$covers the range of possible values of s^*. Indeed, in this case, f difference estimate of fragility uses evaluations of ξ at points spanning this interval. [REWRITE LAST SENTENCE]

-¶

17.4.4 Unconditionality of the shortfall measure ξ

Many, when presenting shortfall,deal with the conditional shortfall $\int_{-\infty}^{K} x\, f(x)\, \mathrm{d}x \Big/ \int_{-\infty}^{K} f(x)\, \mathrm{d}x$; while such measure might be useful in some circumstances, its sensitivity is not indicative of fragility in the sense used in this discussion. The unconditional tail expectation $\xi = \int_{-\infty}^{K} x f(x)\, \mathrm{d}x$ is more indicative of exposure to fragility. It is also preferred to the raw probability of falling below K, which is $\int_{-\infty}^{K} f(x)\, \mathrm{d}x$, as the latter does not include the consequences. For instance, two such measures $\int_{-\infty}^{K} f(x)\, \mathrm{d}x$ and $\int_{-\infty}^{K} g(x)\, \mathrm{d}x$ may be equal over broad values of K; but the expectation $\int_{-\infty}^{K} x f(x)\, \mathrm{d}x$ can be much more consequential than $\int_{-\infty}^{K} x\, g(x)\, \mathrm{d}x$ as the cost of the break can be more severe and we are interested in its "vega" equivalent.

18 APPLICATIONS TO MODEL ERROR

In the cases where Y depends on X, among other variables, often x is treated as non-stochastic, and the underestimation of the volatility of x maps immediately into the underestimation of the left tail of Y under two conditions:

1. X is stochastic and its stochastic character is ignored (as if it had zero variance or mean deviation)

2. Y is concave with respect to X in the negative part of the distribution, below Ω

"Convexity Bias " or Jensen's Inequality Effect: Further, missing the stochasticity under the two conditions a) and b) , in the event of the concavity applying above Ω leads to the negative convexity bias from the lowering effect on the expectation of the dependent variable Y.

18.0.5 Example:Application to Budget Deficits

Example: A government estimates unemployment for the next three years as averaging 9%; it uses its econometric models to issue a forecast balance B of 200 billion deficit in the local currency. But it misses (like almost everything in economics) that unemployment is a stochastic variable. Employment over 3 years periods has fluctuated by 1% on average. We can calculate the effect of the error with the following: âĂć Unemployment at 8% , Balance B(8%) = -75 bn (improvement of 125bn) âĂć Unemployment at 9%, Balance B(9%)= -200 bn âĂć Unemployment at 10%, Balance B(10%)= –550 bn (worsening of 350bn)

The convexity bias from underestimation of the deficit is by -112.5bn, since

$$\frac{B(8\%) + B(10\%)}{2} = -312.5$$

Further look at the probability distribution caused by the missed variable (assuming to simplify deficit is Gaussian with a Mean Deviation of 1%)

Adding Model Error and Metadistributions: Model error should be integrated in the distribution as a stochasticization of parameters. f and g should subsume the distribution of all possible factors affecting the final outcome (including the metadistribution of each). The so-called "perturbation " is not necessarily a change in the parameter so much as it is a means to verify whether f and g capture the full shape of the final probability distribution.

Any situation with a bounded payoff function that organically truncates the left tail at K will be impervious to all perturbations affecting the probability distribution below K.

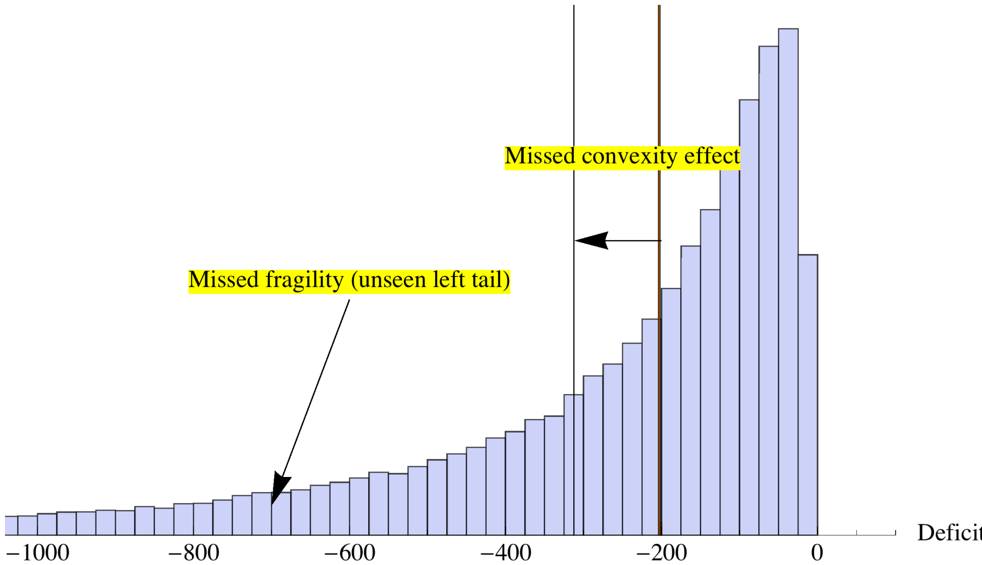


Figure 18.1: Histogram from simulation of government deficit as a left-tailed random variable as a result of randomizing unemployment of which it is a convex function. The method of point estimate would assume a Dirac stick at -200, thus underestimating both the **expected** deficit (-312) and the skewness (i.e., fragility) of it.

For $K = 0$, the measure equates to mean negative semi-deviation (more potent than negative semi-variance or negative semi-standard deviation often used in financial analyses).

18.0.6 Model Error and Semi-Bias as Nonlinearity from Missed Stochasticity of Variables

Model error often comes from missing the existence of a random variable that is significant in determining the outcome (say option pricing without credit risk). We cannot detect it using the heuristic presented in this paper but as mentioned earlier the error goes in the opposite direction as model tend to be richer, not poorer, from overfitting.

But we can detect the model error from missing the stochasticity of a variable or underestimating its stochastic character (say option pricing with non-stochastic interest rates or ignoring that the "volatility" σ can vary).

Missing Effects: The study of model error is not to question whether a model is precise or not, whether or not it tracks reality; it is to ascertain the first and second order effect from missing the variable, insuring that the errors from the model don't have missing higher order terms that cause severe unexpected (and unseen) biases in one direction because of convexity or concavity, in other words, whether or not the model error causes a change in z.

18.1 MODEL BIAS, SECOND ORDER EFFECTS, AND FRAGILITY

Having the right model (which is a very generous assumption), but being uncertain about the parameters will invariably lead to an increase in model error in the presence of convexity and nonlinearities.

As a generalization of the deficit/employment example used in the previous section, say we are using a simple function:

$$f\left(x \mid \overline{\alpha}\right)$$

Where $\overline{\alpha}$ is supposed to be the average expected rate, where we take φ as the distribution of α over its domain $\wp_\alpha$

$$\overline{\alpha} = \int_{\wp_\alpha} \alpha\ \varphi(\alpha)\ d\alpha$$

The mere fact that α is uncertain (since it is estimated) might lead to a bias if we perturb from the outside (of the integral), i.e. stochasticize the parameter deemed fixed. Accordingly, the convexity bias is easily measured as the difference between a) f integrated across values of potential α and b) f estimated for a single value of α deemed to be its average. The convexity bias ω_A becomes:

$$\omega_A \equiv \int_{\wp_x} \int_{\wp_\alpha} f\left(x \mid \alpha\right) \varphi\left(\alpha\right)\ d\alpha\, \mathrm{d}x - \int_{\wp_x} f(x \left| \left(\int_{\wp_\alpha} \alpha\ \varphi\left(\alpha\right)\ d\alpha \right) \right) \mathrm{d}x \tag{18.1}$$

And ω_B the missed fragility is assessed by comparing the two integrals below K, in order to capture the effect on the left tail:

$$\omega_B(K) \equiv \int_{-\infty}^{K} \int_{\wp_\alpha} f\left(x \mid \alpha\right) \varphi\left(\alpha\right)\ d\alpha\, \mathrm{d}x - \int_{-\infty}^{K} f(x \left| \left(\int_{\wp_\alpha} \alpha\ \varphi\left(\alpha\right)\ d\alpha \right) \right) \mathrm{d}x \tag{18.2}$$

Which can be approximated by an interpolated estimate obtained with two values of α separated from a mid point by $\Delta\alpha$ a mean deviation of α and estimating

$$\omega_B(K) \equiv \int_{-\infty}^{K} \frac{1}{2} \left(f\left(x \mid \bar{\alpha} + \Delta\alpha\right) + f\left(x \mid \bar{\alpha} - \Delta\alpha\right) \right) \mathrm{d}x - \int_{-\infty}^{K} f(x \mid \bar{\alpha})\, \mathrm{d}x \tag{18.3}$$

We can probe ω_B by point estimates of f at a level of $X \leq K$

$$\omega_B'(X) = \frac{1}{2} \left(f\left(X \mid \bar{\alpha} + \Delta\alpha\right) + f\left(X \mid \bar{\alpha} - \Delta\alpha\right) \right) - f(X \mid \bar{\alpha}) \tag{18.4}$$

So that

$$\omega_B(K) = \int_{-\infty}^{K} \omega'_B(x) \mathrm{d}x \tag{18.5}$$

which leads us to the fragility heuristic. In particular, if we assume that $\omega_B(X)'$ has a constant sign for $X \leq K$, then $\omega_B(K)$ has the same sign.

The fragility heuristic is presented in the next Chapter.

19 THE FRAGILITY MEASUREMENT HEURISTICS

Chapter Summary **18: Presents the IMF fragility heuristics, particularly in the improvement of stress testing.**

19.0.1 The Fragility/Model Error Detection Heuristic (detecting ω_A and ω_B when cogent)

19.1 EXAMPLE 1 (DETECTING RISK NOT SHOWN BY STRESS TEST)

Or detecting ω_a and ω_b when cogent. The famous firm Dexia went into financial distress a few days after passing a stress test "with flying colors".

If a bank issues a so-called "stress test" (something that has not proven very satisfactory), off a parameter (say stock market) at -15%. We ask them to recompute at -10% and -20%. Should the exposure show negative asymmetry (worse at -20% than it improves at -10%), we deem that their risk increases in the tails. There are certainly hidden tail exposures and a definite higher probability of blowup in addition to exposure to model error.

Note that it is somewhat more effective to use our measure of shortfall in Definition, but the method here is effective enough to show hidden risks, particularly at wider increases (try 25% and 30% and see if exposure shows increase). Most effective would be to use power-law distributions and perturb the tail exponent to see symmetry.

Example 2 (Detecting Tail Risk in Overoptimized System, ω_B). Raise airport traffic 10%, lower 10%, take average expected traveling time from each, and check the asymmetry for nonlinearity. If asymmetry is significant, then declare the system as overoptimized. (Both ω_A and ω_B as thus shown.

The same procedure uncovers both fragility and consequence of model error (potential harm from having wrong probability distribution, a thin- tailed rather than a fat-tailed one). For traders (and see Gigerenzer's discussions, in Gigerenzer and Brighton (2009)[49], Gigerenzer and Goldstein(1996)[50]) simple heuristics tools detecting the magnitude of second order effects can be more effective than more complicated and harder to calibrate methods, particularly under multi-dimensionality. See also the intuition of fast and frugal in Derman and Wilmott (2009)[23], Haug and Taleb (2011)[56].

The Fragility Heuristic Applied to Model Error

1- First Step (first order). Take a valuation. Measure the sensitivity to all parameters p determining V over finite ranges Δp. If materially significant, check if stochasticity of parameter is taken into account by risk assessment. If not, then stop and declare the risk as grossly mismeasured (no need for further risk assessment). 2- Second Step (second order). For all parameters p compute the ratio of first to second order effects at the initial range Δp = estimated mean deviation. $H(\Delta p) \equiv \frac{\mu'}{\mu}$, where

$$\mu'(\Delta p) \equiv \frac{1}{2}\left(f\left(p+\frac{1}{2}\Delta p\right)+f\left(p-\frac{1}{2}\Delta p\right)\right)$$

2-Third Step. Note parameters for which H is significantly > or < 1. 3- Fourth Step: Keep widening Δp to verify the stability of the second order effects.

19.2 THE HEURISTIC APPLIED TO A STRESS TESTING

[INSERT FROM IMF PAPER TALEB CANETTI ET AL]

In place of the standard, one-point estimate stress test S1, we issue a "triple", S1, S2, S3, where S2 and S3 are S1 $\pm$ Δp. Acceleration of losses is indicative of fragility.

Remarks a. Simple heuristics have a robustness (in spite of a possible bias) compared to optimized and calibrated measures. Ironically, it is from the multiplication of convexity biases and the potential errors from missing them that calibrated models that work in-sample underperform heuristics out of sample (Gigerenzer and Brighton, 2009). b. Heuristics allow to detection of the effect of the use of the wrong probability distribution without changing probability distribution (just from the dependence on parameters). c. The heuristic improves and detects flaws in all other commonly used measures of risk, such as CVaR, "expected shortfall", stress-testing, and similar methods have been proven to be completely ineffective (Taleb, 2009). d. The heuristic does not require parameterization beyond varying δp.

19.2.1 Further Applications Investigated in Next Chapters

[TO EXPAND]

In parallel works, applying the *"**simple heuristic**"* allows us to detect the following “hidden short options” problems by merely perturbating a certain parameter p:

i- Size and pseudo-economies of scale.

ii- Size and squeezability (nonlinearities of squeezes in costs per unit).

iii- Specialization (Ricardo) and variants of globalization.

iv- Missing stochasticity of variables (price of wine).

v- Portfolio optimization (Markowitz).

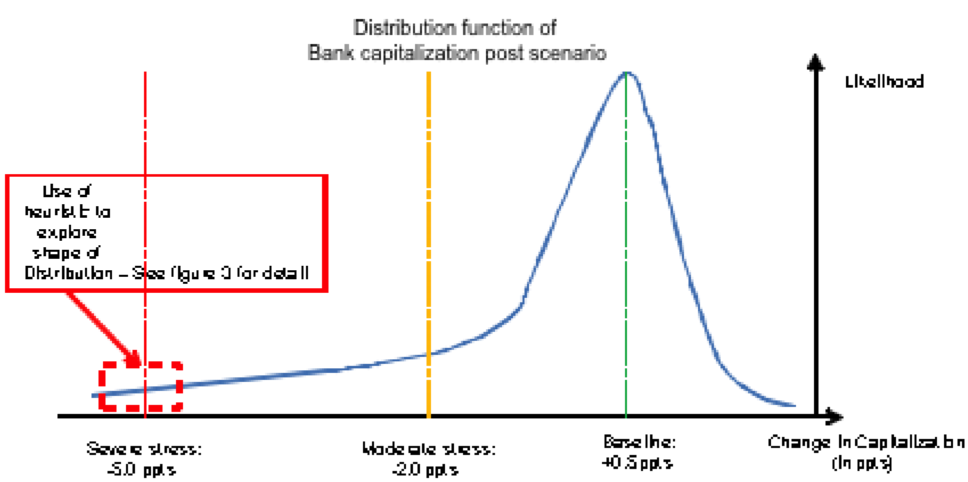
vi- Debt and tail exposure.

vii- Budget Deficits: convexity effects explain why uncertainty lengthens, doesn’t shorten expected deficits.

viii- Iatrogenics (medical) or how some treatments are concave to benefits, convex to errors.

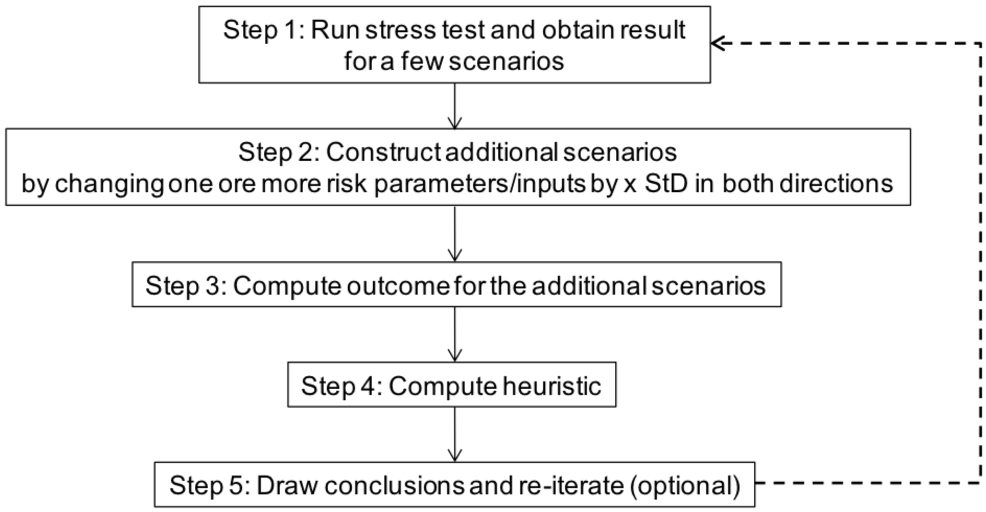
ix- Disturbing natural systems.[1]

19.3 STRESS TESTS



1 Acknowledgments: Bruno Dupire, Emanuel Derman, Jean-Philippe Bouchaud, Elie Canetti. Presented at JP Morgan, New York, June 16, 2011; CFM, Paris, June 17, 2011; GAIM Conference, Monaco, June 21, 2011; Max Planck Institute, BERLIN, Summer Institute on Bounded Rationality 2011 - *Foundations of an Interdisciplinary Decision Theory*- June 23, 2011; Eighth International Conference on Complex Systems - BOSTON, July 1, 2011, Columbia University September 24 2011.

19.4 GENERAL METHODOLOGY



20 FRAGILITY AND ECONOMIC MODELS

20.1 THE MARKOWITZ INCONSISTENCY

Assume that someone tells you that the probability of an event is exactly zero. You ask him where he got this from."Baal told me" is the answer. In such case, the person is coherent, but would be deemed unrealistic by non-Baalists. But if on the other hand, the person tells you"I estimated it to be zero," we have a problem. The person is both unrealistic and inconsistent. Some- thing estimated needs to have an estimation error. So probability cannot be zero if it is estimated, its lower bound is linked to the estimation error; the higher the estima- tion error, the higher the probability, up to a point. As with Laplace's argument of total ignorance, an infinite estimation error pushes the probability toward $\frac{1}{2}$.

We will return to the implication of the mistake; take for now that anything estimating a parameter and then putting it into an equation is different from estimating the equation across parameters (same story as the health of the grandmother, the average temperature, here "estimated" is irrelevant, what we need is average health across temperatures). And Markowitz showed his incoherence by starting his "seminal" paper with "Assume you know E and V" (that is, the expectation and the variance). At the end of the paper he accepts that they need to be estimated, and what is worse, with a combination of statistical techniques and the "judgment of practical men." Well, if these parameters need to be estimated, with an error, then the deriva- tions need to be written differently and, of course, we would have no paper–and no Markowitz paper, no blowups, no modern finance, no fragilistas teaching junk to students. . . . Economic models are extremely fragile to assumptions, in the sense that a slight alteration in these assumptions can, as we will see, lead to extremely consequential differences in the results. And, to make matters worse, many of these models are "back-fit" to assumptions, in the sense that the hypotheses are selected to make the math work, which makes them ultrafragile and ultrafragilizing.

20.2 APPLICATION: RICARDIAN MODEL AND LEFT TAIL EXPOSURE

For almost two hundred years, we've been talking about an idea by the economist David Ricardo called "comparative advantage." In short, it says that a country should have a certain policy based on its comparative advantage in wine or clothes. Say a country is good at both wine and clothes, better than its neighbors with whom it can trade freely. Then the visible optimal strategy would be to specialize in either wine or clothes, whichever fits the best and minimizes opportunity costs. Everyone would then be happy. The analogy by the economist Paul Samuelson is that if someone happens to be the best doctor in town and, at the same time, the best secretary, then it would be preferable to be the higher –earning doctor –as it would minimize opportunity losses–and let someone else be the secretary and buy secretarial ser- vices from him.

We agree that there are benefits in some form of specialization, but not from the models used to prove it. The flaw with such reasoning is as follows. True, it would be inconceivable for a doctor to become a part-time secretary just because he is good at it. But, at the same time, we can safely assume that being a doctor insures some professional stability: People will not cease to get sick and there is a higher social status associated with the profession than that of secretary, making the profession more desirable. But assume now that in a two-country world, a country specialized in wine, hoping to sell its specialty in the market to the other country, and that suddenly the price of wine drops precipitously. Some change in taste caused the price to change. Ricardo's analysis assumes that both the market price of wine and the costs of production remain constant, and there is no "second order" part of the story.

RICARDO'S ORIGINAL EXAMPLE (COSTS OF PRODUCTION PER UNIT)

	Cloth	Wine
Britain	100	110
Portugal	90	80

The logic The table above shows the cost of production, normalized to a selling price of one unit each, that is, assuming that these trade at equal price (1 unit of cloth for 1 unit of wine). What looks like the paradox is as follows: that Portugal produces cloth cheaper than Britain, but should buy cloth from there instead, using the gains from the sales of wine. In the absence of transaction and transportation costs, it is efficient for Britain to produce just cloth, and Portugal to only produce wine.

The idea has always attracted economists because of its paradoxical and counter-intuitive aspect. Clearly one cannot talk about returns and gains without discounting these benefits by the offsetting risks. Many discussions fall into the critical and dangerous mistake of confusing function of average and average of function. Now consider the price of wine and clothes variable–which Ricardo did not assume–with the numbers above the unbiased average long-term value. Further assume that they follow a fat-tailed distribution. Or consider that their costs of production vary according to a fat-tailed distribution.

If the price of wine in the international markets rises by, say, 40 %, then there are clear benefits. But should the price drop by an equal percentage, âĹŠ40 %, then massive harm would ensue, in magnitude larger than the benefits should there be an equal rise. There are concavities to the exposure–severe concavities.

And clearly, should the price drop by 90 percent, the effect would be disastrous. Just imagine what would happen to your household should you get an instant and unpredicted 40 percent pay cut. Indeed, we have had problems in history with coun- tries specializing in some goods, commodities, and crops that happen to be not just volatile, but extremely volatile. And disaster does not necessarily come from varia- tion in price, but problems in production: suddenly, you can't produce the crop be- cause of a germ, bad weather, or some other hindrance.

A bad crop, such as the one that caused the Irish potato famine in the decade around 1850, caused the death of a million and the emigration of a million more (Ireland's entire population at the time of this writing is only about six million, if one includes the northern part). It is very hard to reconvert resources–unlike the case in the doctor-typist story, countries don't have the ability to change. Indeed, monocul- ture (focus on a single crop) has turned out to be lethal in history–one bad crop leads to devastating famines.

The other part missed in the doctor-secretary analogy is that countries don't have family and friends. A doctor has a support community, a circle of friends, a collective that takes care of him, a father-in-law to borrow from in the event that he needs to reconvert into some other profession, a state above him to help. Countries don't. Further, a doctor has savings; countries tend to be borrowers.

So here again we have fragility to second-order effects.

Probability Matching The idea of comparative advantage has an analog in probability: if you sample from an urn (with replacement) and get a black ball 60 percent of the time, and a white one the remaining 40 percent, the optimal strategy, according to textbooks, is to bet 100 percent of the time on black. The strategy of betting 60 percent of the time on black and 40 percent on white is called "probability matching" and considered to be an error in the decision-science literature (which I remind the reader is what was used by Triffat in Chapter 10). People's instinct to engage in probability matching appears to be sound, not a mistake. In nature, probabilities are unstable (or unknown), and probability matching is similar to redundancy, as a buf- fer. So if the probabilities change, in other words if there is another layer of random- ness, then the optimal strategy is probability matching.

How specialization works: The reader should not interpret what I am saying to mean that specialization is not a good thing–only that one should establish such specialization after addressing fragility and second-order effects. Now I do believe that Ricardo is ultimately right, but not from the models shown. Organically, systems without top-down controls would specialize progressively, slowly, and over a long time, through trial and error, get the right amount of specialization–not through some bureaucrat using a model. To repeat, systems make small errors, design makes large ones.

So the imposition of Ricardo's insight-turned-model by some social planner would lead to a blowup; letting tinkering work slowly would lead to efficiency–true efficiency. The role of policy makers should be to, via negativa style, allow the emergence of specialization by preventing what hinders the process.

Portfolio fallacies Note one fallacy promoted by Markowitz users: portfolio theory entices people to diversify, hence it is better than nothing. Wrong, you finance fools: it pushes them to optimize, hence overallocate. It does not drive people to take less risk based on diversification, but causes them to take more open positions owing to perception of offsetting statistical properties–making them vulnerable to model error, and especially vulnerable to the underestimation of tail events. To see how, consider two investors facing a choice of allocation across three items: cash, and se- curities A and B. The investor who does not know the statistical properties

MODEL	SOURCE OF FRAGILITY	REMEDY
Portfolio theory, mean-variance, etc.	Assuming knowledge of the parameters, not integrating models across parameters, relying on (very unstable) correlations. Assumes ω_A (bias) and ω_B (fragility) = 0	1/n (spread as large a number of exposures as manageable), barbells, progressive and organic construction, etc.
Ricardian comparative advantage	Missing layer of randomness in the price of wine may imply total reversal of allocation. Assumes ω_A (bias) and ω_B (fragility) = 0	Natural systems find their own allocation through tinkering
Samuelson optimization	Concentration of sources of randomness under concavity of loss function. Assumes ω_A (bias) and ω_B (fragility) = 0	Distributed randomness
Arrow-Debreu lattice state-space	Ludic fallacy: assumes exhaustive knowledge of outcomes and knowledge of probabilities. Assumes ω_A (bias), ω_B (fragility), and ω_C (antifragility) = 0	Use of metaprobabilities changes entire model implications
Dividend cash flow models	Missing stochasticity causing convexity effects. Mostly considers ÏL'C (antifragility) =0	Heuristics

of A and B and knows he doesn't know will allocate, say, the portion he does not want to lose to cash, the rest into A and B–according to whatever heuristic has been in traditional use. The investor who thinks he knows the statistical properties, with parameters σ_a, σ_B, $\rho_{A,B}$, will allocate ω_A , ω_B in a way to put the total risk at some target level (let us ignore the expected return for this). The lower his perception of the correlation $\rho_{A,B}$, the worse his exposure to model error. Assuming he thinks that the correlation $\rho_{A,B}$, is 0, he will be overallocated by $\frac{1}{3}$ for extreme events. But if the poor investor has the illusion that the correlation is 1, he will be maximally overallocated to his investments A and B. If the investor uses leverage, we end up with the story of Long-Term Capital Management, which turned out to be fooled by the parameters. (In real life, unlike in economic papers, things tend to change; for Baal's sake, they change!) We can repeat the idea for each parameter $\sigma$and see how lower perception of this σleads to overallocation.

I noticed as a trader–and obsessed over the idea–that correlations were never the same in different measurements. Unstable would be a mild word for them: 0.8 over a long period becomes 0.2 over another long period. A pure sucker game. At times of stress, correlations experience even more abrupt changes–without any reliable regularity, in spite of attempts to model "stress correlations." Taleb (1997) deals with the effects of stochastic correlations: One is only safe shorting a correlation at 1, and buying it at âĹŠ1–which seems to correspond to what the 1/n heuristic does. Kelly Criterion vs. Markowitz: In order to implement a full Markowitz-style optimization, one needs to know the entire joint probability distribution of all assets for the entire future, plus the exact utility function for wealth at all future times. And with- out errors! (We saw that estimation errors make the system explode.) Kelly's method, developed around the same period, requires no joint distribution or utility function. In practice one needs the ratio of expected profit to worst-case return–dynamically adjusted to avoid ruin. In the case of barbell transformations, the worst case is guar- anteed. And model error is much, much milder under Kelly criterion. Thorp (1971, 1998), Haigh (2000).

The formidable Aaron Brown holds that Kelly's ideas were rejected by economists–in spite of the practical appeal–because of their love of general theories for all asset prices.

Note that bounded trial and error is compatible with the Kelly criterion when one has an idea of the potential return–even when one is ignorant of the returns, if losses are bounded, the payoff will be robust and the method should outperform that of Fragilista Markowitz.

Corporate Finance: In short, corporate finance seems to be based on point projections, not distributional projections; thus if one perturbates cash flow projections, say, in the Gordon valuation model, replacing the fixed–and known–growth (and other parameters) by continuously varying jumps (particularly under fat-tailed distributions), companies deemed âĂIJexpensive,âĂİ or those with high growth, but low earnings, could markedly increase in expected value, something the market prices heuristically but without explicit reason.

Conclusion and summary: Something the economics establishment has been missing is that having the right model (which is a very generous assumption), but being un- certain about the parameters will invariably lead to an increase in fragility in the presence of convexity and nonlinearities.

20.2.1 Error and Probabilities

21 THE ORIGIN OF THIN-TAILS

Chapter Summary **19: 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 and 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 $S^N(x)$: $\mathbb{R} \to [k_L, k_R]$, $S^N \in C^\infty$, 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:

$$S^N(x) \equiv \sum_{i=1}^{N} \frac{a_k}{1 + \exp\left(-b_k x + c_k\right)}$$

(21.1)

where a_k, b_k, c_k are scaling constants $\in \mathbb{R}$, satisfying:
i) S^N(-∞) =k_L
ii) S^N(∞) =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 $x \in (-\infty, k_1)$, $\frac{\partial^2 S^N}{\partial x^2} < 0$ for $x \in (k_2, k_{>2})$, and $\frac{\partial^2 S^N}{\partial x^2} \geq 0$ for $x \in (k_{>2}, \infty)$, 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.

21.1 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 $f_{\sigma,\alpha}(x)$:

$$x f_{\sigma,\alpha} \equiv \frac{\left(\frac{\alpha}{\alpha + \frac{x^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\sigma B\left(\frac{\alpha}{2}, \frac{1}{2}\right)} \tag{21.2}$$

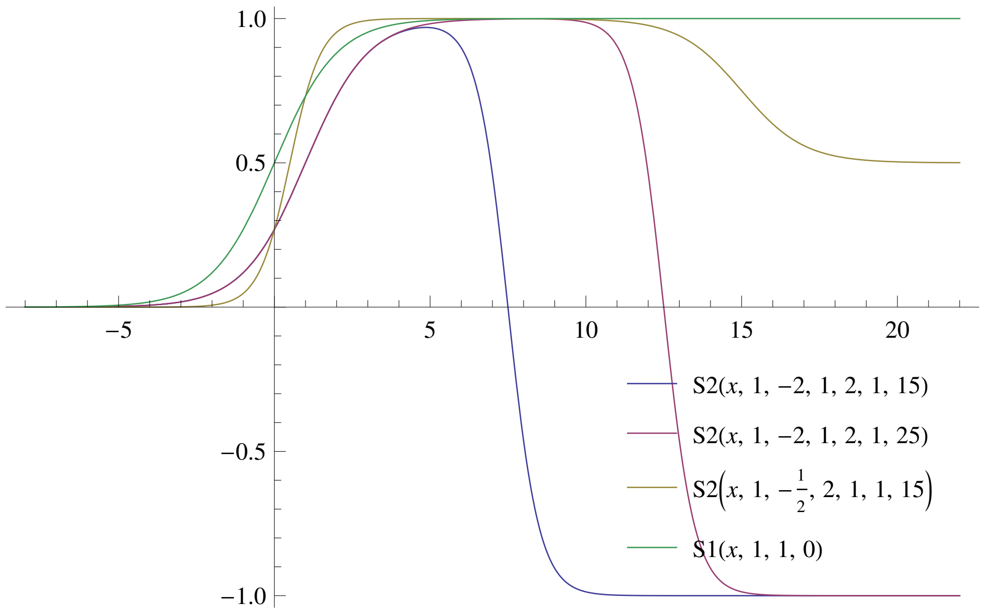


Figure 21.1: The Generalized Response Curve, $S^2\left(x, a_1, a_2, b_1, b_2, c_1, c_2\right)$, $S^1\left(x, a_1, b_1, c_1\right)$ The convex part with positive first derivative has been designated as "antifragile"

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.

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., $N = 1$

$$S(x) \equiv \frac{a_1}{1 + \exp(-b_1 x + c_1)}$$

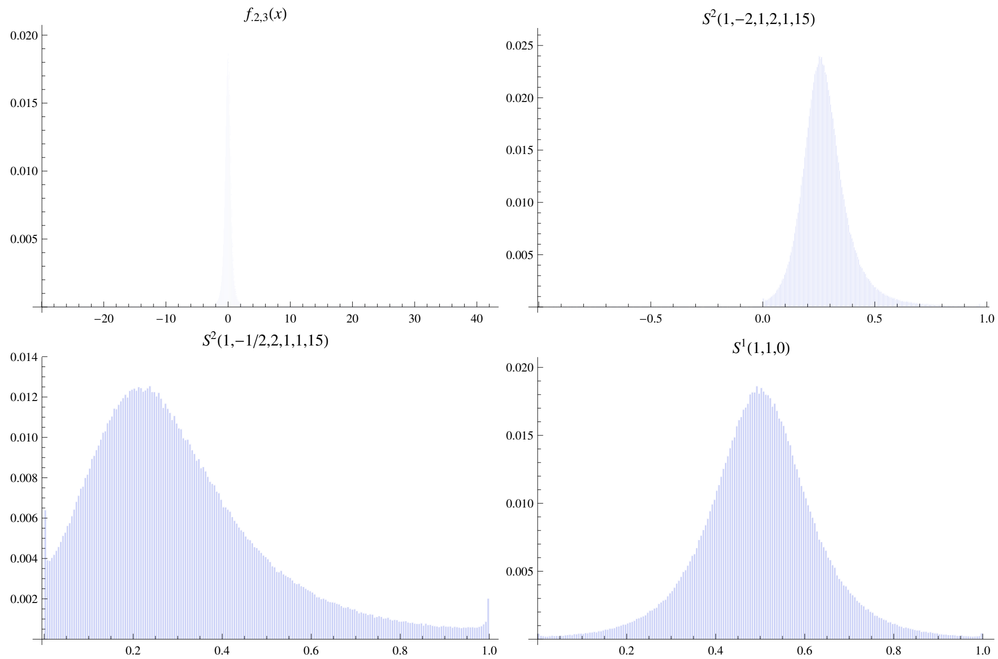


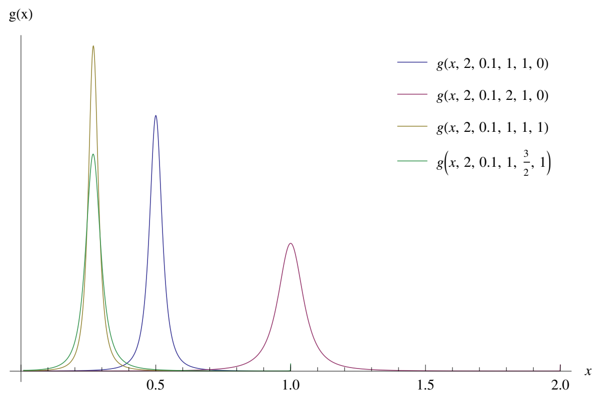
Figure 21.2: Histograms for the different inherited probability distributions (simulations,$N = 10^6$)

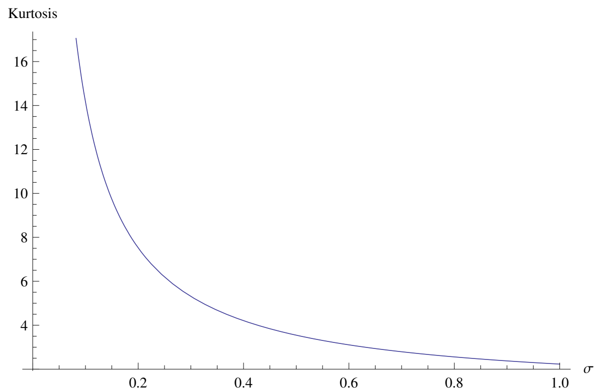
(21.3)

g(x) is the inherited distribution, which can be shown to have a scaled domain $\mathcal{D}^g$= (k_L, k_R). It becomes

$$g(x) = \frac{a1\left(\frac{\alpha}{\alpha+\frac{\left(\log\left(\frac{x}{a1-x}\right)+c1\right)^2}{b1^2\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}b1\sigma x B\left(\frac{\alpha}{2},\frac{1}{2}\right)(a1-x)}$$

(21.4)





Remark 1: The inherited distribution from S(x) will have a compact support regardless of the probability distribution of x.

21.2 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.

22 SMALL IS BEAUTIFUL: RISK, SCALE AND CONCENTRATION

Chapter Summary 20: **We extract the effect of size on the degradation of the expectation of a random variable, from nonlinear response. The method is general and allows to show the "small is beautiful" or "decentralized is effective" or "a diverse ecology is safer" effect from a response to a stochastic stressor and prove stochastic diseconomies of scale and concentration (with as example the Irish potato famine and GMOs). We apply the methodology to environmental harm using standard sigmoid dose-response to show the need to split sources of pollution across independent**
(nonsynergetic) pollutants.

22.1 INTRODUCTION: THE TOWER OF BABEL

Diseconomies and Harm of scale Where is small beautiful and how can we detect, even extract its effect from nonlinear response? [1] Does getting larger makes an entity more vulnerable to errors? Does polluting or subjecting the environment with a large quantity cause disproportional "unseen" stochastic effects? We will consider different types of dose-response or harm-response under different classes of probability distributions.

The situations convered include:

1. Size of items falling on your head (a large stone vs small pebbles).
2. Losses under strain.
3. Size of animals (The concavity stemming from size can be directly derived from the difference between allometic and isometric growth, as animals scale in a specific manner as they grow, an idea initially detected by Haldane,[46] (on the "cube law"(TK)).
4. Quantity in a short squeeze
5. The effect of crop diversity
6. Large vs small structures (say the National Health Service vs local entities)
7. Centralized government vs municipalities
8. Large projects such as the concentration of health care in the U.K.
9. Stochastic environmental harm: when, say, polluting with K units is more than twice as harmful than polluting with $K/2$ units.

1 The slogan "small is beautiful" originates with the works of Leonard Kohr [65] and his student Schumacher who thus titled his influential book.

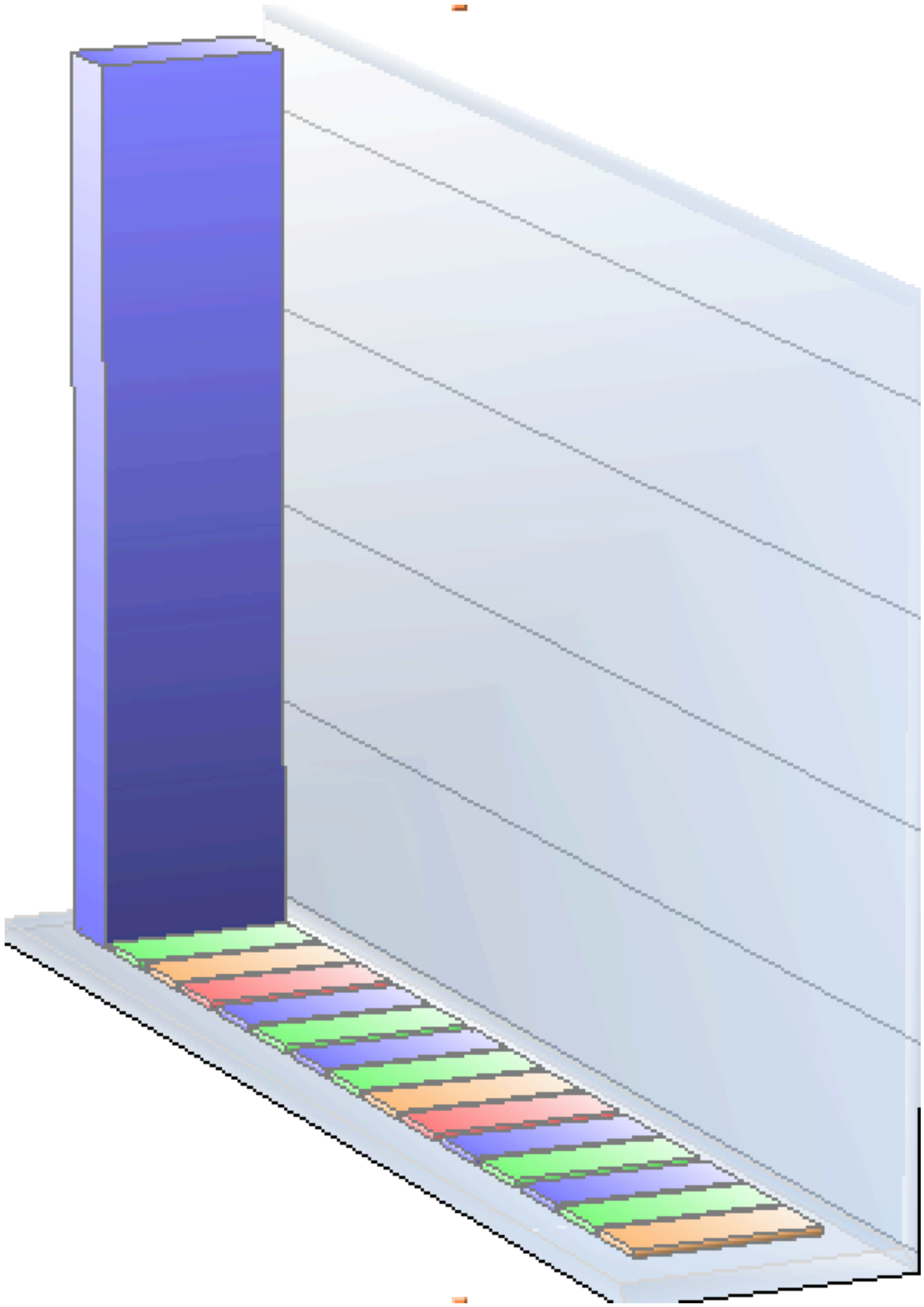


Figure 22.1: The Tower of Babel Effect: Nonlinear response to height, as taller towers are disproportionately more vulnerable to, say, earthquakes, winds, or a collision. This illustrates the case of truncated harm (limited losses).For some structures with unbounded harm the effect is even stronger.

First Example: The Kerviel Rogue Trader Affair

The problem is summarized in *Antifragile* [111] as follows:

> On January 21, 2008, the Parisian bank Societé Générale rushed to sell in the market close to seventy billion dollars worth of stocks, a very large amount for any single "fire sale." Markets were not very active (called "thin"), as it was Martin Luther King Day in the United States, and markets worldwide dropped precipitously, close to 10 percent, costing the company close to six billion dollars in losses just from their fire sale. The entire point of the squeeze is that they couldn't wait, and they had no option but to turn a sale into a fire sale. For they had, over the weekend, uncovered a fraud. Jerome Kerviel, a rogue back office employee, was playing with humongous sums in the market and hiding these exposures from the main computer system. They had no choice but to sell, immediately, these stocks they didn't know they owned. Now, to see the effect of fragility from size (or concentration), consider losses as a function of quantity sold. A fire sale of $70 billion worth of stocks leads to a loss of $6 billion. But a fire sale a tenth of the size,$7 billion would result in no loss at all, as markets would absorb the quantities without panic, maybe without even noticing. So this tells us that if, instead of having one very large bank, with Monsieur Kerviel as a rogue trader, we had ten smaller units, each with a proportional Monsieur Micro- Kerviel, and each conducted his rogue trading independently and at random times, the total losses for the ten banks would be close to nothing.

Second Example: The Irish Potato Famine with a warning on GMOs

The same argument and derivations apply to concentration. Consider the tragedy of the Irish potato famine.

In the 19th Century, Ireland experienced a violent potato famine coming from concentration and lack of diversity. They concentrated their crops with the "lumper" potato variety. "Since potatoes can be propagated vegetatively, all of these lumpers were clones, genetically identical to one another."[2]

Now the case of genetically modified organism (GMOs) is rich in fragilities (and confusion about the "natural"): the fact that an error can spread beyond local spots bringing fat-tailedness, a direct result ofthe multiplication of large scale errors. But the mathematical framework here allows us to gauge its effect from loss of local diversity. The greater problem with GMOs is the risk of ecocide, examined in Chapter x.

Only Iatrogenics of Scale and Concentration

Note that, in this discussion, we only consider the harm, not the benefits of concentration under nonlinear (concave) response. Economies of scale (or savings from concentration and lack of diversity) are similar to short volatility exposures, with seen immediate benefits and unseen deferred losses.

2 the source is evolution.berkeley.edu/evolibrary but looking for author's name.

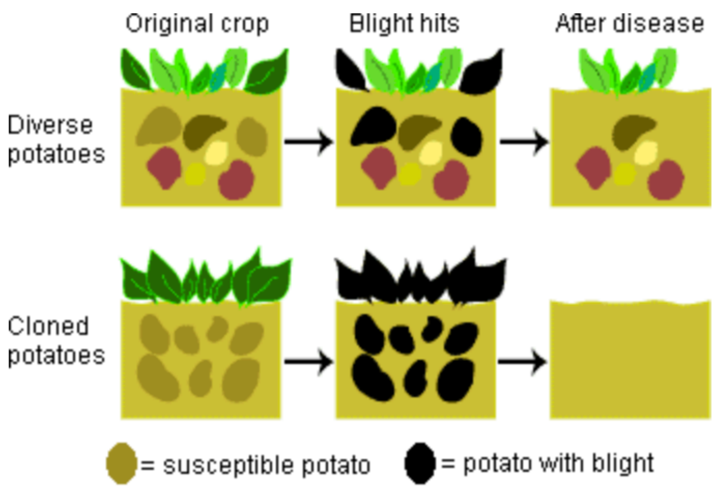


Figure 22.2: Integrating the evolutionary explanation of the Irish potato famine into our fragility framework, courtesy http://evolution.berkeley.edu/evolibrary .

The rest of the discussion is as follows. We will proceed, via convex transformation to show the effect of nonlinearity on the expectation. We start with open-ended harm, a monotone concave response, where regardless of probability distribution (satisfying some criteria), we can extract the harm from the second derivative of the exposure. Then we look at more natural settings represented by the "sigmoid" S-curve (or inverted S-curve) which offers more complex nonlinearities and spans a broader class of phenomena.

Unimodality as a general assumption Let the variable x, representing the stochastic stressor, follow a certain class of continuous probability distributions (unimodal), with the density $p(x)$ satisfying: $p(x) \geq p(x+\epsilon)$ for all $\epsilon > 0$, and $x > x^*$ and $p(x) \geq p(x-\epsilon)$ for all $x < x^*$ with $\{x^* : p(x^*) = \max_x p(x)\}$. The density $p(x)$ is Lipschitz. This condition will be maintained throughout the entire exercise.

22.2 UNBOUNDED CONVEXITY EFFECTS

In this section, we assume an unbounded harm function, where harm is a monotone (but nonlinear) function in C^2, with negative second derivative for all values of x in $\mathbb{R}^+$; so let $h(x)$, $\mathbb{R}^+ \to \mathbb{R}^-$ be the harm function. Let B be the size of the total unit subjected to stochastic stressor x, with $\theta(B) = B + h(x)$.

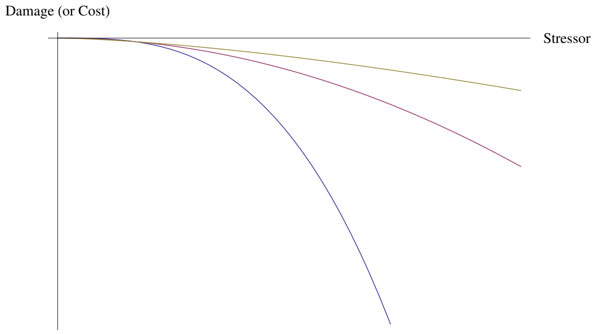


Figure 22.3: Simple Harm Functions, monotone: $k = 1$, $\beta = 3/2, 2, 3$.

We can prove by the inequalities from concave transformations that, the expectation of the large units is lower or equal to that of the sum of the parts. Because of the monotonocity and concavity of $h(x)$,

$$h\left(\sum_{i=1}^{N} \omega_i\, x\right) \leq \sum_{i=1}^{N} h(\omega_i\, x), \tag{22.1}$$

for all x in its domain ($\mathbb{R}^+$), where ω_i are nonnegative normalized weights, that is, $\sum_{i=1}^{N} \omega_i = 1$ and $0 \leq \omega_i \leq 1$.

And taking expectations on both sides, $\mathbb{E}(\theta(B)) \leq \mathbb{E}\left(\sum_{i=1}^{N} \theta(\omega_i\, B)\right)$: the mean of a large unit *under stochastic stressors* degrades compared to a series of small ones.

Application

Let $h(x)$ be the simplified harm function of the form

$$h(x) \equiv -k\, x^{\beta}, \tag{22.2}$$

$k \in (0, \infty)$, $\beta \in [0, \infty)$.

Table 25: Applications with unbounded convexity effects

Environment	Research	$h(x)$
Liquidation Costs	Toth et al.,[115],Bouchaud et al. [14]	$-kx^{\frac{3}{2}}$
Bridges	Flyvbjerg et al [42]	$-x(\frac{\log(x)+7.1}{10})$

Example 1: One-Tailed Standard Pareto Distribution Let the probability distribution of x (the harm) be a simple Pareto (which matters little for the exercise, as any one-tailed distribution does the job). The density:

$$p_{\alpha,L}(x) = \alpha\, L^{\alpha}\, x^{-\alpha-1} \text{ for } x \geq L \tag{22.3}$$

The distribution of the response to the stressor will have the distribution $g = (p \circ h)(x)$.

Given that k the stressor is strictly positive, $h(x)$ will be in the negative domain. Consider a second change of variable, dividing x in N equal fragments, so that the unit becomes $\xi = x/N$, $N \in \mathbb{N}_{\geq 1}$:

$$g_{\alpha,L,N}(\xi) = -\frac{\alpha^{\alpha} N^{-\alpha} \left(-\frac{\xi}{k}\right)^{-\alpha/\beta}}{\beta \, \xi}, \tag{22.4}$$

for $\xi \leq -k \left(\frac{L}{N}\right)^{\beta}$ and with $\alpha > 1 + \beta$. The expectation for a section x/N, $M_\beta(N)$:

$$M_\beta(N) = \int_{-\infty}^{-\frac{kL^\beta}{N}} \xi \quad g_{\alpha,L,N}(\xi) \, d\xi = -\frac{\alpha \, k \, L^\beta \, N^{\alpha\left(\frac{1}{\beta}-1\right)-1}}{\alpha - \beta} \tag{22.5}$$

which leads to a simple ratio of the mean of the total losses (or damage) compared to a κ number of its N fragments, allowing us to extract the "convexity effect" or the degradation of the mean coming from size (or concentration):

$$\frac{\kappa \, M_\beta(\kappa N)}{M_\beta(N)} = \kappa^{\alpha\left(\frac{1}{\beta}-1\right)} \tag{22.6}$$

With $\beta = 1$, the convexity effect =1. With $\beta = 3/2$ (what we observe in orderflow and many other domains related to planning, Bouchaud et al., 2012, Flyvbjerg et al, 2012), the convexity effect is shown in Figure 26.

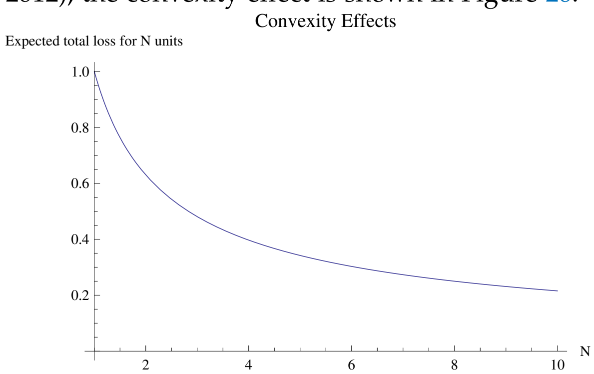


Table 26: The mean harm in total as a result of concentration. Degradation of the mean for N=1 compared to a large N, with $\beta = 3/2$

Unseen Harm The skewness of $g_{\alpha,L,N}(\xi)$ shows effectively how losses have properties that hide the mean in "small" samples (that is, large but insufficient number of observations), since, owing to skewness, the observed mean loss with tend to be lower than the true value. As with the classical Black Swan exposures, benefits are obvious and harm hidden.

22.3 A RICHER MODEL: THE GENERALIZED SIGMOID

Now the biological and physical domains (say animals, structures) do not incur unlimited harm, when taken as single units. The losses terminate somewhere: what is broken is broken. From the generalized sigmoid function of [**?**], where $S^M(x) = \sum_{k=1}^{M} \frac{a_k}{1+\exp(b_k(c_k-x))}$, a sum of single sigmoids. We assume as a special simplified case $M = 1$ and $a_1 = -1$ so we focus on a single stressor or source of harm $S(x), \mathbb{R}^+ \to [-1, 0]$ where x is a positive variable to simplify and the response a negative one. $S(0) = 0$, so $S(.)$ has the following form:

$$S(x) = \frac{-1}{1 + e^{\,b\,(c-x)}} + \frac{1}{1 + e^{b\,c}} \tag{22.7}$$

The second term is there to ensure that $S(0) = 0$. Figure 27 shows the different calibrations of b (c sets a displacement to the right).

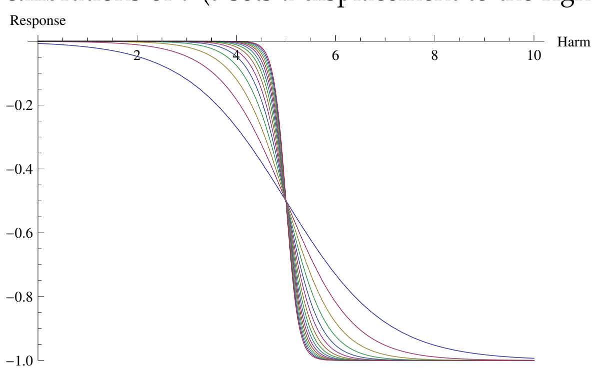
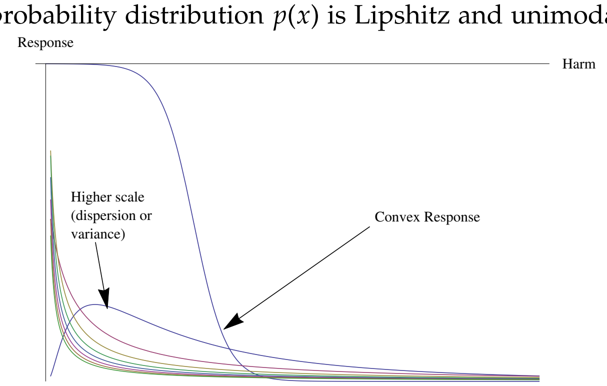


Table 27: Consider the object broken at -1 and in perfect condition at 0

[backgroundcolor=lightgray] The sigmoid, $S(x)$ in C^∞ is a class of generalized function (Sobolev, Schwartz [101]); it represents literally any object that has progressive positive or negative saturation; it is smooth and has derivatives of all order: simply anything bounded on the left and on the right has to necessarily have to have the sigmoid convex-concave (or mixed series of convex-concave) shape.

The idea is to measure the effect of the distribution, as in 3.14. Recall that the probability distribution $p(x)$ is Lipshitz and unimodal.



The second derivative $S''(x) = \frac{b^2 e^{b(c+x)}\left(e^{bx}-e^{bc}\right)}{\left(e^{bc}+e^{bx}\right)^3}$. Setting the point where $S''(x)$ becomes 0, at $x = c$, we get the following: S(x) is concave in the interval $x \in [0, c)$ and convex in the interval $x \in (c, \infty)$.

The result is mixed and depends necessarily on the parametrization of the sigmoids. We can thus break the probability distributions into two sections, the "concave" and "convex" parts: $\mathbb{E} = \mathbb{E}^- + \mathbb{E}^+$. Taking $\xi = x/N$, as we did earlier,

$$\mathbb{E}^- = N \int_0^c S(\xi)\ p(\xi)\,\mathrm{d}\xi,$$

and

$$\mathbb{E}^+ = N \int_c^\infty S(\xi)\ p(\xi)\,\mathrm{d}\xi$$

The convexity of $S(.)$ is symmetric around c,

$$S''(x)|_{x=c-u} = -2b^2 \sinh^4\left(\frac{b\,u}{2}\right)\operatorname{csch}^3(b\,u)$$

$$S''(x)|_{x=c+u} = 2b^2 \sinh^4\left(\frac{bu}{2}\right)\operatorname{csch}^3(b\,u)$$

We can therefore prove that the effect of the expectation for changes in N depends exactly on whether the mass to the left of a is greater than the mass to the right. Accordingly, if $\int_0^a p(\xi)\,\mathrm{d}\xi > \int_a^\infty p(\xi)\,\mathrm{d}\xi$, the effect of the concentration ratio will be positive, and negative otherwise.

Application

Example of a simple distribution: Exponential Using the same notations as 22.2, we look for the mean of the total (but without extracting the probability distribution of the transformed variable, as it is harder with a sigmoid). Assume x follows a standard exponential distribution with parameter λ, $p(x) \equiv \lambda e^{\lambda(-x)}$

$$M_\lambda(N) = \mathbb{E}\left(S(\xi)\right) = \int_0^\infty \lambda e^{\lambda(-x)}\left(-\frac{1}{e^{b\left(c-\frac{x}{N}\right)}+1} + \frac{1}{e^{\mathrm{bc}}+1}\right)\mathrm{d}x \tag{22.8}$$

$$M_\lambda(N) = \frac{1}{e^{bc}+1} - {}_2F_1\left(1, \frac{N\lambda}{b}; \frac{N\lambda}{b}+1; -e^{bc}\right)$$

where the Hypergeometric function ${}_2F_1(a,b;c;z) = \sum_{k=0}^\infty \frac{a_k b_k z^k}{k! c_k}$.

The ratio $\frac{\kappa\ M_\lambda(\kappa N)}{M_\lambda(N)}$ doesn't admit a reversal owing to the shape, as we can see in 22.4 but we can see that high variance reduces the effect of the concentration. However high variance increases the probability of breakage.

Example of a more complicated distribution: Pareto type IV Quasiconcave but neither convex nor concave PDF: The second derivative of the PDF for the Exponential doesn't change sign, $\frac{\partial^2}{\partial x^2}(\lambda \exp(-\lambda x)) = \lambda^3 e^{\lambda(-x)}$, so the distribution retains a convex shape. Further, it is not possible to move its mean beyond the point c where the sigmoid switches in the sign of the nonlinearity. So we elect a broader one, the Pareto Distibution of Type IV, which is extremely flexible because, unlike the simply convex shape (it has a skewed "bell" shape, mixed convex-concave-convex shape) and accommodates tail exponents, hence has power law properties for large

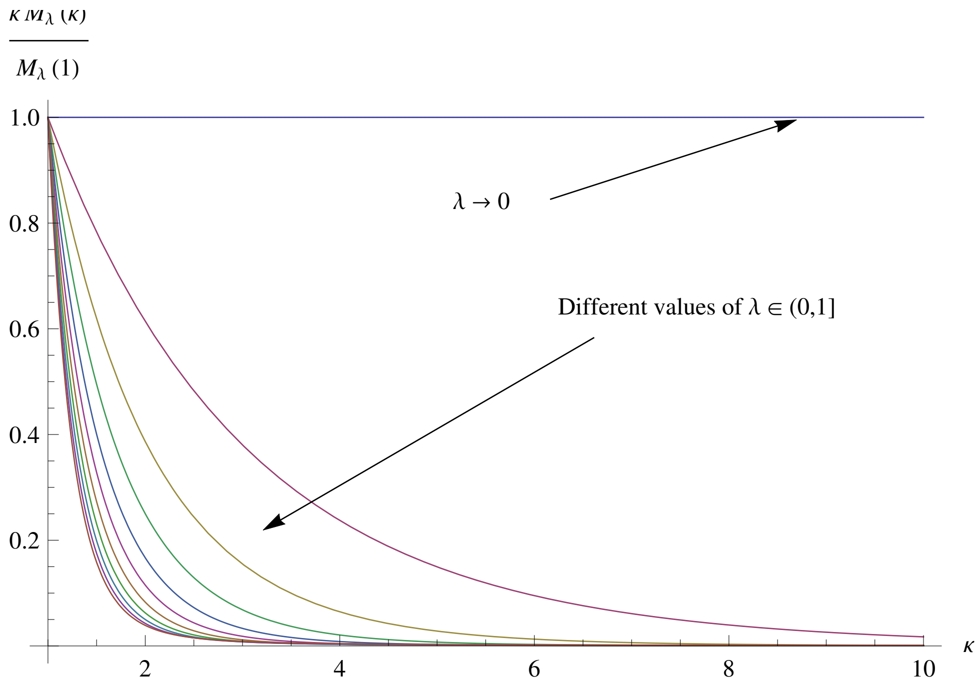


Figure 22.4: Exponential Distribution: The degradation coming from size at different values of λ.

deviations. It is quasiconcave but neither convex nor concave. A probability measure (hence PDF) $p : \mathfrak{D} \to [0, 1]$ is quasiconcave in domain $\mathfrak{D}$ if for all $x, y \in \mathfrak{D}$ and $\omega \in [0, 1]$ we have:

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

Where x is the same harm as in Equation 22.7:

$$p_{\alpha,\gamma,\mu,k}(x) = \frac{\alpha k^{-1/\gamma}(x-\mu)^{\frac{1}{\gamma}-1}\left(\left(\frac{k}{x-\mu}\right)^{-1/\gamma}+1\right)^{-\alpha-1}}{\gamma} \tag{22.9}$$

for $x \geq \mu$ and 0 elsewhere.

The Four figures in 3.14 shows the different effects of the parameters on the distribution.

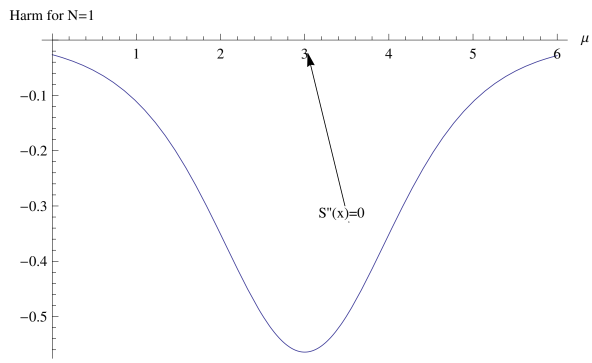
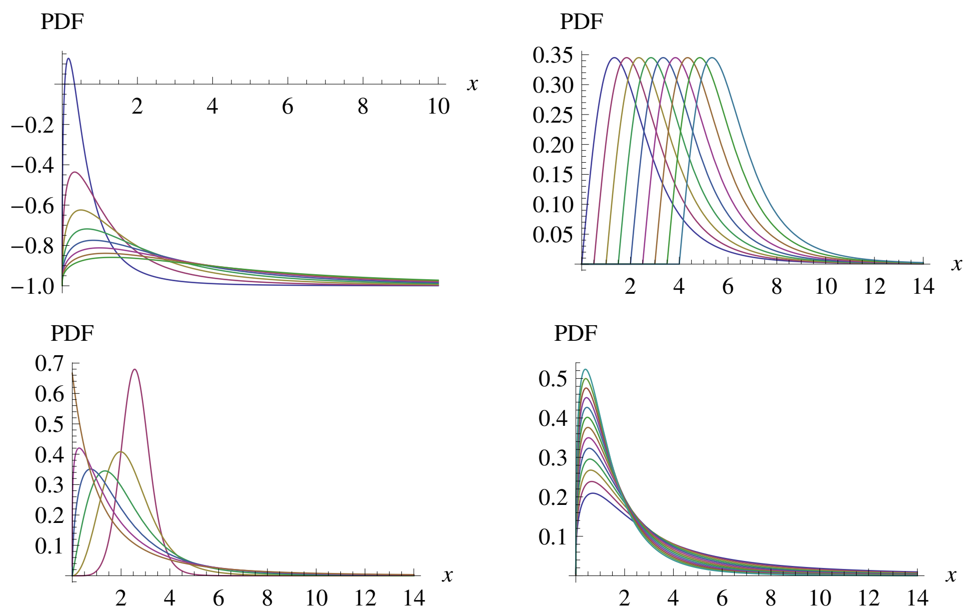


Figure 22.5: Harm increases as the mean of the probability distribution shifts to the right, to become maximal at c, the point where the sigmoid function S(.) switches from concave to convex.



The mean harm function, $M_{\alpha,\gamma,\mu,k}(N)$ becomes:

$$M_{\alpha,\gamma,\mu,k}(N) = \frac{\alpha k^{-1/\gamma}}{\gamma} \int_0^\infty (x-\mu)^{\frac{1}{\gamma}-1} \left(\frac{1}{e^{bc}+1} - \frac{1}{e^{b\left(c-\frac{x}{N}\right)}+1} \right) \left(\left(\frac{k}{x-\mu} \right)^{-1/\gamma} + 1 \right)^{-\alpha-1} \mathrm{d}x \quad (22.10)$$

M(.) needs to be evaluated numerically. Our concern is the "pathology" where the mixed convexities of the sigmoid and the probability distributions produce locally opposite results than 3.14 on the ratio $\frac{\kappa M_{\alpha,\gamma,\mu,k}(N)}{M_{\alpha,\gamma,\mu,k}(N)}$. We produce perturbations around zones where μ has maximal effects, as in 22.7. However as shown in Figure 22.5, the total expected harm is quite large under these conditions, and damage will be done regardless of the effect of scale.

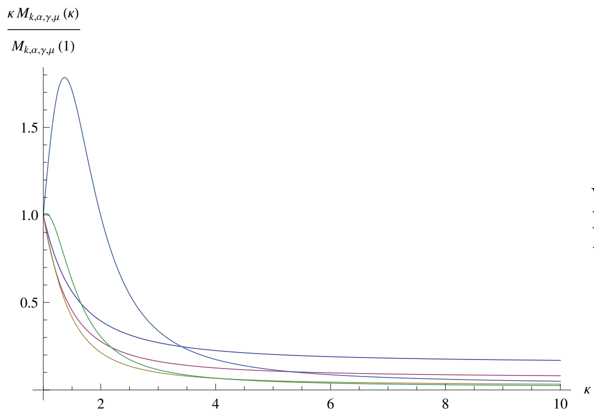


Figure 22.6: Different values of μ: we see the pathology where 2 M(2) is higher than M(1), for a value of $\mu = 4$ to the right of the point c.

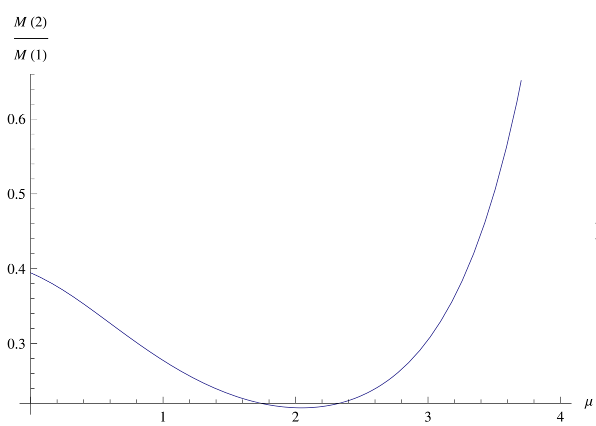


Figure 22.7: The effect of μ on the loss from scale.

Conclusion

This completes the math showing extracting the "small is beautiful" effect, as well as the effect of dose on harm in natural and biological settings where the Sigmoid is in use. More verbal discussions are in *Antifragile*.

Acknowledgments

Yaneer Bar-Yam, Jim Gatheral (naming such nonlinear fragility the "Tower of Babel effect"), Igor Bukanov, Edi Pigoni, Charles Tapiero.

23 | WHY IS THE FRAGILE NONLINEAR?

***Chapter Summary* 21: Explains why the fragilefragile is necessarily in the nonlinear.**

INCOMPLETE CHAPTER as of November 2014

The main framework of broken glass: very nonlinear in response. We replace the Heavyside with a continuous funtion in C^{∞}.

Imagine different classes of coffee cups or fragilefragile items that break as the dose increases, indexed by $\left\{\beta^i\right\}$ for their sigmoid of degree 1: the linearity in the left interval $\left(x_0, x_1\right]$, where $x$is the dose and $S(.)$ the response, $S : \mathbb{R}^+ \to [0,1]$. (Note that $\alpha = 1$; we keep a (which determines the height) constant so all start at the same point x_0 and end at the same one x_4. Note that c corresponds to the displacement to the right or the left on the dose-response line.

$$S_{a,\beta^i,\gamma}(x) \equiv \frac{a}{e^{\beta^i(-(\gamma+x))}+1}$$

The second derivative:

$$\frac{\partial^2 S_{a,\beta^i,\gamma}(x)}{\partial x^2} = -2a\beta^2 \sinh^4\left(\frac{1}{2}\beta(\gamma+x)\right)$$

$$\operatorname{csch}^3(\beta(\gamma+x)), \quad (23.1)$$

where sinh and csnh are the hyperbolic sine and cosine, respectively.

Next we subject all the families to a probability distribution of harm, $f(z)$ being a monomodal distribution with the expectation $\mathbb{E}(z) \in \left(x_0, x_1\right]$. We compose $f \circ S$to get $f\left(S_{\alpha,\beta^i,\gamma}(x)\right)$. In this case we pick a symmetric power law.

$f_{\alpha,\sigma}\left(S_{a,\beta,\gamma}(x)\right) =$,

with $\alpha \epsilon$ (1, ∞) and $\sigma \in$ (0, ∞)

The objects will produce a probability distribution around $[0,1]$ since $S_{a,\beta^i,\gamma}(x)$ is bounded at these levels; we can see to the right a Dirac mass concentrating observations at 1. Clearly what has survived is the nonlinear.

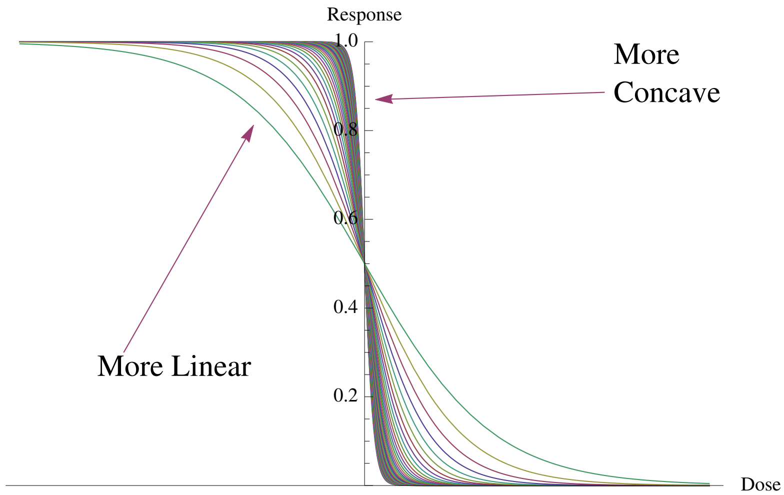
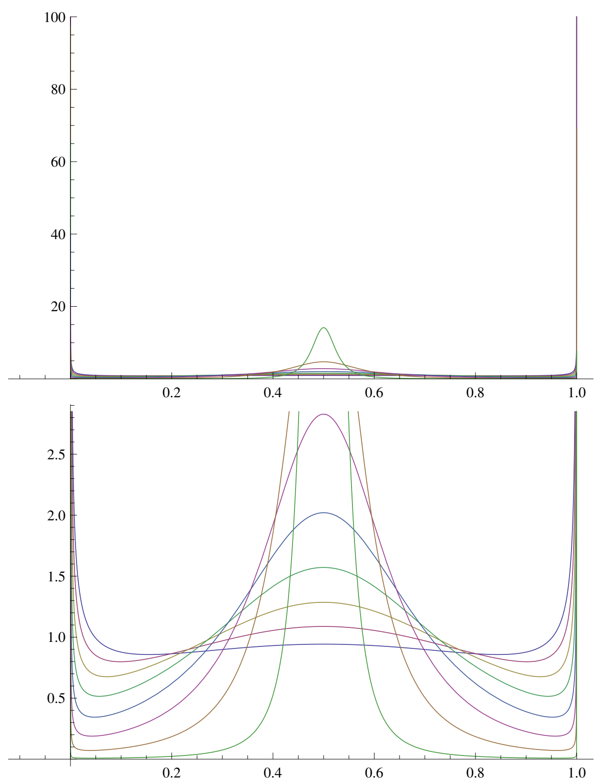


Figure 23.1: The different dose-response curves, at different values of $\left\{\beta^{i}\right\}$, corresponding to varying levels of concavity.



24 HOW THE WORLD WILL PROGRESSIVELY LOOK WEIRDER

Chapter Summary 22: **Information is convex to noise. 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 and the increase of variables available for epidemiological and other "empirical" research.**

24.1 HOW NOISE EXPLODES FASTER THAN DATA

To the observer, every day will seem weirder 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	Weirder and weirder events reported on the front pages
Epidemiological Studies, "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.

Figure 24.1: The picture of a "freak event" spreading on the web of a boa who ate a drunk person in Kerala, India, in November 2013. With 7 billion people on the planet and ease of communication the "tail" of daily freak events is dominated by such news. The make the point even more: it turned out to be false (thanks to Victor Soto).

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.

24.2 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", $\zeta = \frac{k}{m}$, 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:

$$F^{-1}(w) = \sqrt{2}\,\sigma\,\operatorname{erfc}^{-1}(2w) + \mu,$$

where $\operatorname{erfc}^{-1}$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}

$$\Phi^{-1}{}_{\sigma,\mu}(\zeta) = -\sqrt{2}\sigma\operatorname{erfc}^{-1}\left(2\frac{k}{m}\right) + \mu$$

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.

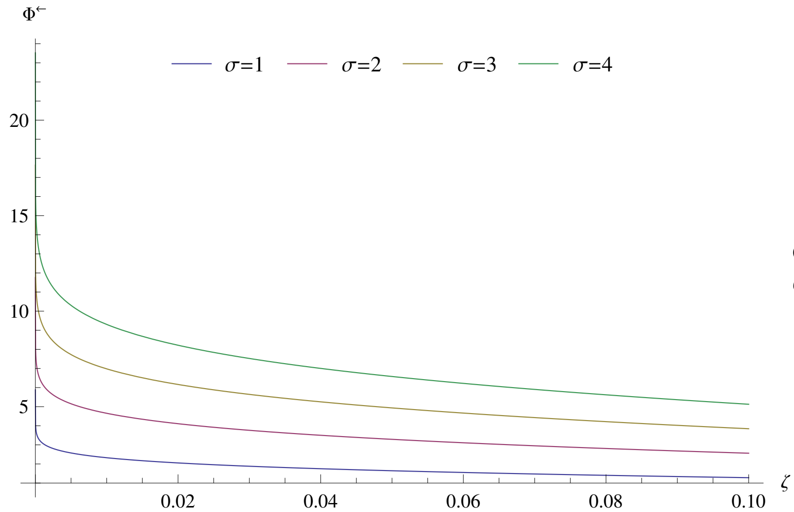


Figure 24.2: Gaussian, σ={1,2,3,4}

h!

$$f(x) \equiv \frac{\left(\frac{\alpha}{\alpha+\frac{(x-\mu)^2}{\sigma^2}}\right)^{\frac{\alpha+1}{2}}}{\sqrt{\alpha}\,\sigma\, B\left(\frac{\alpha}{2},\frac{1}{2}\right)} \tag{24.1}$$

Where we can get the inverse survival function.

$$\gamma^{-1}{}_{\sigma,\mu}(\zeta) = \mu + \sqrt{\alpha}\ \sigma\ \operatorname{sgn}(1-2\ \zeta)\sqrt{\frac{1}{I^{-1}_{(1,(2\zeta-1)\operatorname{sgn}(1-2\zeta))}\left(\frac{\alpha}{2},\frac{1}{2}\right)}-1} \tag{24.2}$$

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.

Fatter Tails: Alpha Stable Distribution

Part 2 of the discussion to come soon.

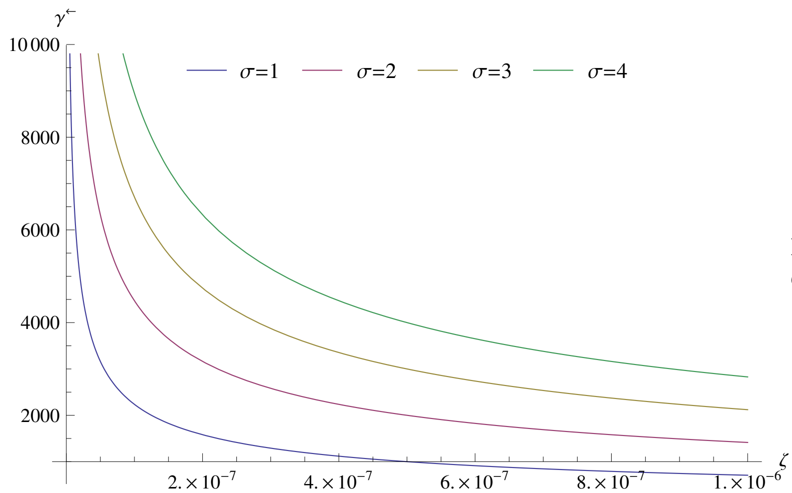


Figure 24.3: Power Law, σ={1,2,3,4}

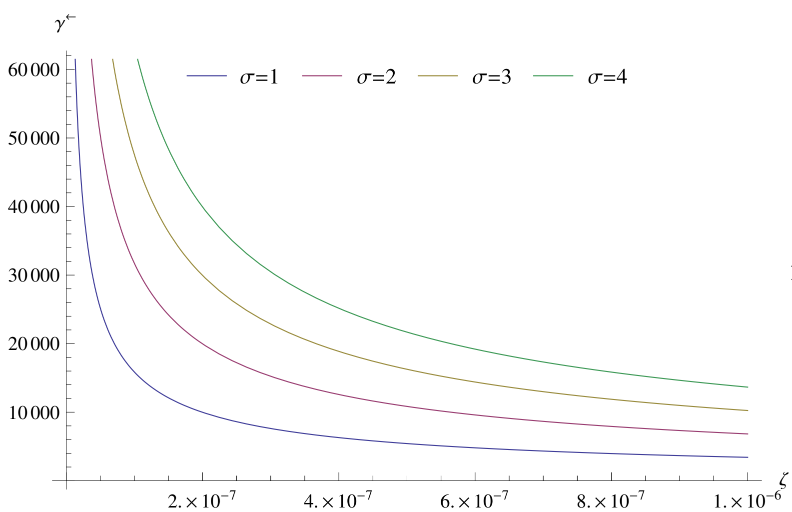


Figure 24.4: Alpha Stable Distribution

25 THE CONVEXITY OF WEALTH TO INEQUALITY

Chapter Summary 23: **The one percent of the one percent has tail properties such that the tail wealth (expectation $\int_K^\infty x\, p(x)\, \mathrm{d}x$) depends far more on inequality than wealth.**

25.1 THE ONE PERCENT OF THE ONE PERCENT ARE DIVORCED FROM THE REST

The one percent of the one percent of the population is vastly more sensitive to inequality than total GDP growth (which explains why the superrich are doing well now, and should do better under globalization, and why it is a segment that doesn't correlate well with the economy). For the super-rich, one point of GINI causes an increase equivalent to 6-10% increase in total income (say, GDP). More generally, the partial expectation in the tail is vastly more sensitive to changes in scale of the distribution than in its centering.

Sellers of luxury goods and products for the superwealthy profit from dispersion more than increase in total wealth or income. I looked at their case as a long optionality, benefit-from-volatility type of industry.

From textitAntifragile[111]:

> *Another business that does not care about the average but rather the dispersion around the average is the luxury goods industry—jewelry, watches, art, expensive apartments in fancy locations, expensive collec - tor wines, gourmet farm - raised probiotic dog food, etc. Such businesses only cares about the pool of funds available to the very rich. If the population in the Western world had an average income of fifty thousand dollars, with no inequality at all, the luxury goods sellers would not survive. But if the average stays the same, with a high degree of inequality, with some incomes higher than two million dollars, and potentially some incomes higher than ten million, then the business has plenty of customers—even if such high incomes were offset with masses of people with lower incomes. The "tails" of the distribution on the higher end of the income brackets, the extreme, are much more determined by changes in inequality than changes in the average. It gains from dispersion, hence is antifragile.*
>
> *This explains the bubble in real estate prices in Central London, determined by inequality in Russia and the Arabian Gulf and totally independent of the real estate dynamics in Britain. Some apartments, those for the very rich, sell for twenty times the average per square foot of a building a few blocks away.*
>
> *Harvard' s former president Larry Summers got in trouble explaining a version of the point and lost his job in the aftermath of the uproar. He was trying to say that males and females have equal intelligence, but the male population has more variations and dispersion (hence volatility), with more highly unintelligent men, and more highly intelligent ones. For Summers, this explained why men were overrepresented in the sci - entific and intellectual community (and also why men were overrepre - sented in jails or failures). The number of successful scientists depends on the "tails," the extremes,*

rather than the average. Just as an option does not care about the adverse outcomes, or an author does not care about the haters.

Derivations

Let the r.v. $x \in [x_{\min}, \infty)$ follow a Pareto distribution (type II), with expected return fixed at $\mathbb{E}(x) = m$, tail exponent $\alpha > 1$, the density function

$$p(x) = \frac{\alpha \left(\frac{(\alpha-1)(m-x_{\min})-x_{\min}+x}{(\alpha-1)(m-x_{\min})} \right)^{-\alpha-1}}{(\alpha-1)\left(m-x_{\min}\right)}$$

We are dealing with a three parameter function, as the fatness of the tails is determined by both α and $m - x_{\min}$, with $m - x_{\min} > 0$ (since $\alpha > 1$).

Note that with 7 billion humans, the one percent of the one percent represents 700,000 persons.

The same distribution applies to wealth and income (although with a different parametrization, including a lower α as wealth is more unevenly distributed than income.)

Note that this analysis does not take into account the dynamics (and doesn't need to): over time a different population will be at the top.

The Lorenz curve Where *F(x)*, short for $P(X < x)$ is the cumulative distribution function and inverse $F^{\leftarrow}(z) : [0,1] \rightarrow [x_{\min}, \infty)$, the Lorenz function for z $L(z)$:$[0,1] \rightarrow [0,1]$ is defined as:

$$L(z) \equiv \frac{\int_0^z F^{\leftarrow}(y)dy}{\int_0^1 F^{\leftarrow}(y)dy}$$

The distribution function

$$F(x) = 1 - \left(1 + \frac{x - x_{\min}}{(\alpha-1)\left(m - x_{\min}\right)}\right)^{-\alpha},$$

so its inverse becomes:

$$F^{\leftarrow}(y) = m(1-\alpha) + (1-y)^{-1/\alpha}(\alpha-1)\left(m - x_{\min}\right) + \alpha x_{\min}$$

Hence

$$L\left(z, \alpha, m, x_{\min}\right) = \frac{1}{m}(1-z)^{-1/\alpha}\left((z-1)\alpha\left(m - x_{\min}\right) + (z-1)^{\frac{1}{\alpha}}\left(m(z+\alpha-z\alpha) + (z-1)\alpha x_{\min}\right)\right) \quad (25.1)$$

Which gives us different combination of α and $m - x_{\min}$, producing different tail shapes: some can have a strong "middle class" (or equivalent) while being top-heavy; others can have more *equal* inequality throughout.

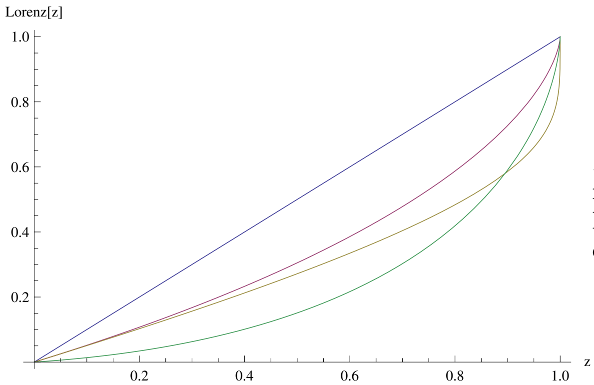


Figure 25.1: Different combinations $L(z,3,.2,.1), L(z,3,.95,.1), L(z,1.31,.2,.1)$ in addition to the perfect equality line L(z)= z. We see the criss-crossing at higher values of z.

Gini and Tail Expectation

The GINI Coefficient, ∈[0,1] is the difference between 1) the perfect equality,with a Lorenz $L(f) = f$ and 2) the observed $L\left(z, \alpha, m, x_{\min}\right)$

$$\text{GINI}\left(\alpha, m, x_{\min}\right) = \frac{\alpha}{(2\alpha - 1)} \frac{\left(m - x_{\min}\right)}{m}$$

Computing the tail mass above a threshold K, that is, the unconditional partial expectation $E_{>K} \equiv \int_K^\infty x p(x)\, dx$, which corresponds to the nominal share of the total pie for those with wealth above K,

$$E_{>K} = (\alpha - 1)^{\alpha - 1} \left(\alpha \left(K + m - x_{\min}\right) - m\right) \left(\frac{m - x_{\min}}{K + (\alpha - 1)m - \alpha x_{\min}}\right)^{\alpha}$$

The Probability of exceeding K, $P_{>K}$ (Short for $P(X > k)$)

$$P_{>K} = \left(1 + \frac{K - x_{\min}}{(\alpha - 1)\left(m - x_{\min}\right)}\right)^{-\alpha}$$

For the *One Percent of the One Percent* (or equivalent), we set the probability $P_{>K}$ and invert to K_P=$(\alpha - 1)\left(m - x_{\min}\right) p^{-1/\alpha} - \alpha\left(1 + m + x_{\min}\right)$,

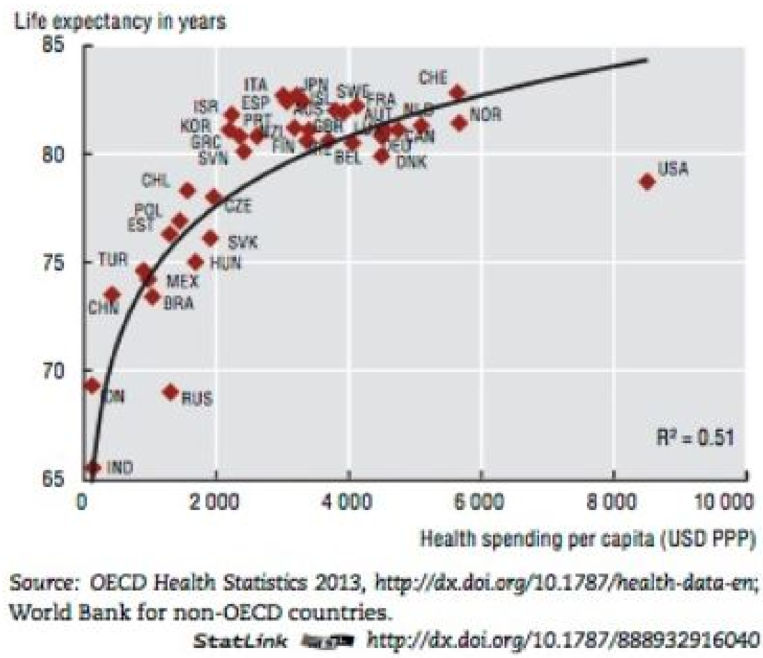
$$E_{>K} = \left(p^{\frac{\alpha-1}{\alpha}}\right) \left(\alpha\left(m - x_{\min}\right) + p^{\frac{1}{\alpha}} \left(m - m\alpha + \alpha x_{\min}\right)\right)$$

Now we can check the variations in GINI coefficient and the corresponding changes in $E_{>K}$for a constant m.

α	GINI	$E_{>K}$	$E_{>K}/\text{m}$
1.26	0.532895	0.33909	0.121103
1.23	0.541585	0.395617	0.141292
1.2	0.55102	0.465422	0.166222
1.17	0.561301	0.55248	0.197314
1.14	0.572545	0.662214	0.236505
1.11	0.584895	0.802126	0.286474
1.08	0.598522	0.982738	0.350978

26 NONLINEARITIES AND RISK IN MEDICINE

Chapter Summary 24: **Examines nonlinearities in medicine /iatrogenics as a risk management problem.**

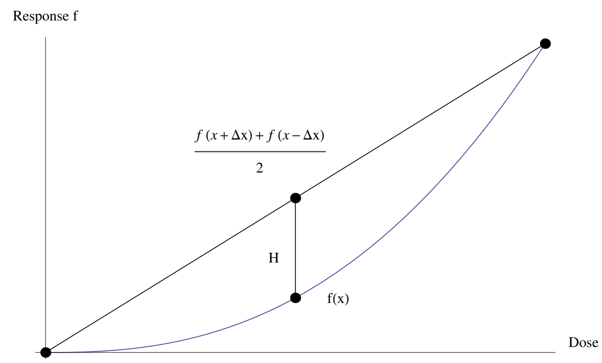


26.1 ANTIFRAGILITY FROM UNEVEN DISTRIBUTION

Take health effect a function "response" from a single parameter, $f: \mathfrak{R} \rightarrow \mathfrak{R}$ be a twice differentiable, the effect from dose x.

If over a range x ∈ [a,b], over a set time period Δt, $\frac{\partial^2 f(x)}{\partial x^2} > 0$ or more heuristically, $\frac{1}{2}(f(x+\Delta x) + f(x-\Delta x)) > f(x)$, with x+Δx and x-Δx ∈ [a,b] then there are benefits from unevenness of distribution: episodic deprivation, intermittent fasting, variable pulmonary ventilation, uneven distribution of proteins(autophagy), vitamins, high intensity training, etc.).

In other words, in place of a dose x, one can give 140% of x , then 60% of x, with a more favorable outcome.



Proof: Jensen's Inequality.

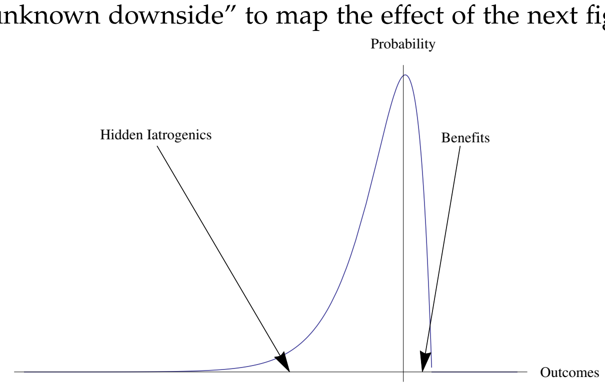
This is a simplification here since dose response is rarely monotone in its nonlinearity, as we will see further down.

Mixed Nonlinearities in Nature Nonlinearities are not monotone.

Nonlinearities in Biology- The shape convex-concave necessarily flows from anything increasing (monotone, i.e. never decreasing) and bounded, with a maximum and a minimum values, i.e. never reached infinity from either side. At low levels, the dose response is convex (gradually more and more effective). Additional doses tend to become gradually ineffective or hurt. The same can apply to anything consumed in too much regularity. This type of graph necessarily applies to any situation bounded on both sides, with a known minimum and maximum (saturation), which includes happiness.

For instance, If one considers that there exists a maximum level of happiness and unhappiness then the general shape of this curve with convexity on the left and concavity on the right has to hold for happiness (replace "dose" with wealth and "response" with happiness). Kahneman-Tversky Prospect theory models a similar one for "utility" of changes in wealth, which they discovered empirically.

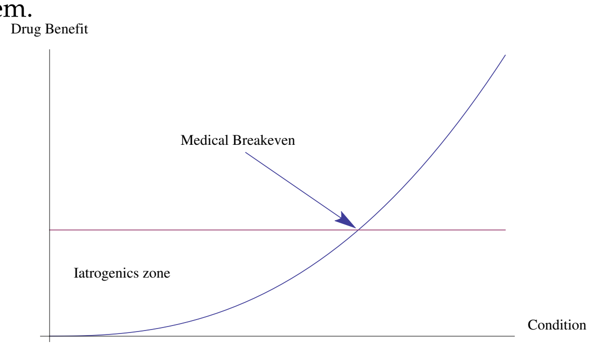
Iatrogenics If $\frac{\partial^2 f(x)}{\partial x^2} \leq 0$ for all x (to simplify), and x is symmetrically distributed, then the distribution of the "outcome" from administration of f (and only the effect of f) will be left-skewed as shown in Figure 1. Further "known limited upside, unknown downside" to map the effect of the next figure.



Medical Iatrogenics: Probability distribution of f. Case of small benefits and large Black Swan-style losses seen in probability space. Iatrogenics occur when we have small identifiable gains (say, avoidance of small discomfort or a minor infection) and exposure to Black Swans with delayed invisible large side effects (say, death). These concave benefits from medicine are just like selling a financial option (plenty of risk) against small tiny immediate gains while claiming "evidence of no harm".

In short, for a healthy person, there is a small probability of disastrous outcomes (discounted because unseen and not taken into account), and a high probability of mild benefits.

Proof: Convex transformation of a random variable, the Fragility Transfer Theorem.



In time series space:

Mother Nature v/s Medicine *The hypertension example. On the vertical axis, we have benefits of a treatment, on the horizontal, the severity of the condition. The arrow points at the level where probabilistic gains match probabilistic harm. Iatrogenics disappear nonlinearly as a function of the severity of the condition. This implies that when the patient is very ill, the distribution shifts to antifragile (thicker right tail), with large benefits from the treatment over possible iatrogenics, little to lose.*

Note that if you increase the treatment you hit concavity from maximum benefits, a zone not covered in the graph —seen more broadly, it would look like the graph of bounded upside

From *Antifragile*

Second principle of iatrogenics: it is not linear. We should not take risks with near-healthy people; but we should take a lot, a lot more risks with those deemed in danger.

Why do we need to focus treatment on more serious cases, not marginal ones? Take this example showing nonlinearity (convexity). When hypertension is mild, say marginally higher than the zone accepted as "normotensive," the chance of benefiting from a certain drug is close to 5.6 percent (only one person in eighteen benefit from the treatment). But when blood pressure is considered to be in the "high" or "severe" range, the chances of benefiting are now 26 and 72 percent, respectively (that is, one person in four and two persons out of three will benefit from the treatment). So the treatment benefits are convex to condition (the bene- fits rise disproportionally, in an accelerated manner). But consider that the iatrogenics should be constant for all categories! In the very ill condi- tion, the benefits are large relative to iatrogenics; in the borderline one, they are small. This means that we need to focus on high-symptom con- ditions and ignore, I mean really ignore, other situations in which the patient is not very ill.

The argument here is based on the structure of conditional survival probabilities, similar to the one that we used to prove that harm needs to be nonlinear for porcelain cups. Consider that Mother Nature had to have tinkered through selection in inverse proportion to the rarity of the condition. Of the hundred and twenty thousand drugs available today, I can hardly find a via positiva one that makes a healthy person uncondi- tionally "better" (and if someone shows me one, I will be skeptical of yet-unseen side effects). Once in a while we come up with drugs that enhance performance, such as, say, steroids, only to discover what peo- ple in finance have known for a while: in a "mature" market there is no free lunch anymore, and what appears as a free lunch has a hidden risk. When you think you have found a free lunch, say, steroids or trans fat, something that helps the healthy without visible downside, it is most likely that there is a concealed trap somewhere. Actually, my days in trading, it was called a "sucker's trade."

And there is a simple statistical reason that explains why we have not been able to find drugs that make us feel unconditionally better when we are well (or unconditionally stronger, etc.): nature would have been likely to find this magic pill by itself. But consider that illness is rare, and the more ill the person the less likely nature would have found the solu- tion by itself, in an accelerating way. A condition that is, say, three units of deviation away from the norm is more than three hundred times rarer than normal; an illness that is five units of deviation from the norm is more than a million times rarer!

The medical community has not modeled such nonlinearity of benefits to iatrogenics, and if they do so in words, I have not seen it in formal- ized in papers, hence into a decision-making methodology that takes probability into account (as we will see in the next section, there is little explicit use of convexity biases). Even risks seem to be linearly extrapo- lated, causing both underestimation and overestimation, most certainly miscalculation of degrees of harm—for instance, a paper on the effect of radiation states the following: "The standard model currently in use ap- plies a linear scale, extrapolating cancer risk from high doses to low doses of ionizing radiation." Further, pharmaceutical companies are under financial pressures to find diseases and satisfy the security ana- lysts. They have been scraping the bottom of the barrel, looking for disease among healthier and healthier people, lobbying for reclassifica- tions of conditions, and fine-tuning sales tricks to get doctors to overpre- scribe. Now, if your blood pressure is in the upper part of the range that used to be called "normal," you are no longer "normotensive" but "prehypertensive," even if there are no symptoms in view. There is nothing wrong with the classification if it leads to healthier lifestyle and robust via negativa measures—but what is behind such classification, often, is a drive for more medication.

27 AMERICAN OPTIONS AND HIDDEN CONVEXITY

Chapter Summary **25: American Options have hidden optionalities. Using a European option as a baseline we heuristically add the difference. We also show how these hidden options are extremely significant compared to the focus of the research literature on insignificant marginal improvements of the pricing equations but in the Black Scholes World.**

27.1 THIS NOTE

This is a paper in progress, not formatted for submission, but aiming at the development of ideas and mathematical results around the problem. We start with the math, and end with the explanations, much of which were aimed at updating *Dynamic Hedging*; the sequence will be somewhat reversed in the final paper, and some comments will be added.

27.2 THE GENERAL MATHEMATICAL RESULTS: PRICING SERIES OF HIDDEN OPTIONS "USE ONE LOSE ALL"

Define a probability triple $(\Omega, \mathcal{F}, \mathcal{P})$, with corresponding random variables indexed-ordered by size of maximal possible realizations so, with $X(\omega) : \Omega \to \mathbb{R}^n$ a measurable function, with $i \in \mathbb{N}^+, i \leq p$, we have a random vector $X \equiv (X_i)_{1 \leq i \leq p}$ with independent components but not the same probability distribution, as the realizations follow different Bernoulli distributions with varying probabilities and realizations $X_i(\omega)$:

$$X_i = \begin{cases} \lambda_1 & \text{w.p. } p_i \\ 0 & \text{w.p. } 1 - p_i \end{cases} \tag{27.1}$$

And of course we index the random variables by rank according to their maximum possible value λ_1, a rank that is retained for all realizations since the λ_i are constant, and in such a way that λ_p is the smallest :

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

These events are, casually, defined as a payoff (taking a single value) with its probability.

Optionality Only one payoff can be "accepted", which makes the maximal one the one of concern as we abandon the other, inferior ones. Define $\mathcal{A}$ as the new set of

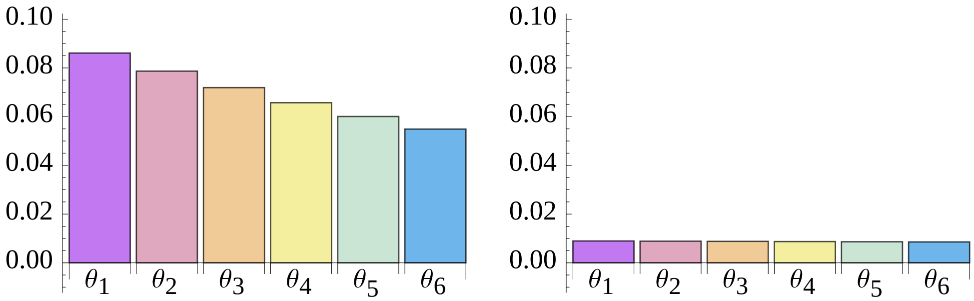


Figure 27.1: the vector of weights Θ under probabilities flat $p_1 = \cdots = p_6 = 10^2$ and 10^3 and $m = 90$ days. We can observe probabilities remain significant and cumulatively consequential. We also see how θ_i become equal, with a flat weights at small probabilities.

"events" ξ_i of concern and ξ_i^c the complentary event, that is that ξ_i does not take place, as follows:

$$\xi_i : \text{at least one } \lambda_i \text{ realization in } m,\ m \in \mathbb{N}^+, \text{ is} > 0$$
$$= \sup\left((X_{i,m})_{i \leq m}\right) > 0$$

$$\mathcal{A} = \left\{\xi_p\right\} \cup \left\{\xi_p^c \cap \xi_{n-1}\right\} \cup \left\{ \left\{\xi_p^c \cup \xi_{n-1}^c\right\} \cap \xi_{n-2}\right\} \cup \left\{ \left\{\xi_p^c \cup \xi_{n-1}^c \cup \xi_{n-2}^c\right\} \cap \xi_{n-3}\right\} \ldots \tag{27.2}$$

$$\mathcal{A} = \bigcup_{i=0}^{p} \left\{ \cup_{j=0}^{i} \xi_{n-j}^c \cap \xi_{n-i} \right\} \tag{27.3}$$

Now consider the weight vector Θ:

$$\mathbf{\Theta} \equiv \left((1-(1-p_n)^m) \left(\prod_{i=1}^{n-1} (1-p_{n-i})^m \right) \right)_{n \leq p}$$

$$\mathbf{\Theta} \equiv (\theta_1, \theta_2, \ldots, \theta_p)$$

$$\mathbf{\Lambda} \equiv (\lambda_1, \lambda_2, \ldots, \lambda_p)$$

We skip to the expectation without dealing with probability distribution:

$$\mathbb{E}[X|\mathcal{A}] = \mathbf{\Theta}.\mathbf{\Lambda}^{\mathbf{T}} \tag{27.4}$$

Given that the events are disjoint,the expected value of the option on n draws over a sequence of observations of length m (which could correspond to m time periods):

$$\Xi_m = \mathbb{E}[X|\mathcal{A}]$$
$$= \sum_{n=1}^{p} \lambda_n \left(1-(1-p_n)^m\right) \left(\prod_{i=1}^{n-1} (1-p_{n-i})^m\right) \quad (27.5)$$

Which gives the value of the hidden optionality.

What we have done here is find an exact representation of the expectation for the upper bound of a nonhomogeneous mixture of independent Bernouilli variables(or, rather, functions of Bernouilli variables). The result is generally applicable to many small things in life, so let us apply it to American options.

We can get further simplifications thanks to the effect that the options become additive as the probabilities get smaller, without necessarily becoming equal, so:

$$\mathbb{E}[X|\mathcal{A}] \approx \sum_{n=1}^{p} \lambda_n \left(1-(1-p_n)^m\right) \quad (27.6)$$

Note on the difference between the heuristic Bernouilli and a set of full distributions

It will take a few lines to show whether the tractability of the Bernouilli simplification causes losses in precision. We could use, say, the maximum of gamma/exponential family with different calibrations but Monte Carlo shows no difference.

27.3 MISPLACED PRECISION

So many "rigorous" research papers in derivatives have been involved in the "exact" pricing of American options, though *within* model when in fact their most interesting attribute is that they benefit from the breakdown of models, or they are **convex to model errors**.

Indeed an interesting test to see if someone has traded derivatives is to quiz him on American options. If he answers by providing a "pasting boundary" story but using a Black-Scholes type world, then you can safely make the conclusion that he has never gotten close to American options.

Furthermore, with faster computers, a faster pricing algorithm does not carry large advantages. The problem is in the hidden optionality... Major points:

An American option is always worth equally or more than the European option of the same nominal maturity.

An American option has always a shorter or equal expected life than a European option.

Rule 27.1.
The value of the difference between an American and European option of same strike and maturity increases with the following factors:

- *Higher volatility of interest rates.*
- *Higher volatility of volatility.*
- *Higher instability of the slope of the volatility curve.*

The major difference between an American and European option is that the holder of the American option has the right to decide on whether the option is worth more dead or alive. In other words is it worth more held to expiration or immediately exercised?

27.4 THE PRICING EQUATION

We can therefore show that, as of period t_0, for all periods to expiration t, where O_A is the "conventionally priced" American option (according to whichever method one choses), and O_E is the corresponding European option of the same maturity and strike,

$$O_A^* = O_E + \mathbb{E}\left((O_A - O_E) \vee \Xi_m\right) \tag{27.7}$$

the expectation of the maximum of two expectations, which allows the simplification:

$$O_A^* = O_E + \left((O_A - O_E) \vee \Xi_m\right) \tag{27.8}$$

We now need to define the components entering Ξ_m, namely the various probabilities p_i and associated payoff λ_i.

NOTE: This part will need some derivations, a bit more clarity about the derivations, etc. Also note that there is a need to prove iterated expectations...

27.5 WAR STORIES

War Story 1 : The Currency Interest rate Flip

I recall in the 1980s the German currency carried lower interest rates than the US. When rate 1 is lower than rate 2, then, on regular pricing systems, for vanilla currency options, the American Put is higher than the European Put, but American Call =European Call. At some point the rates started converging; they eventually flipped as the German rates rose a bit after the reunification of Deutschland. I recall the trade in which someone who understood model error trying to buy American Calls Selling European Calls and paying some trader who got an immediate marks-to-market P/L (from the mark-to-model). The systems gave an identical value to these -it looked like free money, until the trader blew up. Nobody could initially figure out why they were losing money after the flip –the systems were missing on the difference. There was no big liquidity but several billions went through. Eventually the payoff turned out to be big.

We repeated the game a few times around devaluations as interest rates would shoot up and there was always some sucker willing to do the trade.

War Story 2: The Stock Squeeze

Spitz called me once in during the 2000 Bachelier conference to tell me that we were in trouble. We were long listed American calls on some Argentinian stock and short the delta in stock. The stock was some strange ADR that got delisted and we had to cover our short ASAP. Somehow we could not find the stock, and begging Bear Stearns failed to help. The solution turned out to be trivial: exercise the calls, enough of them to get the stock. We were lucky that our calls were American, not European, otherwise we would have been squeezed to tears. Moral: an American call has hidden optionality on model error.

These hidden optionalities on model errors are more numerous than the ones in the two examples I just gave. I kept discovering new ones.

War Story 3: American Option and The Squeeze

I recall in the late 1990s seeing a strange situation: Long dated over-the-counter call options on a European Equity index were priced exceedingly below whatever measure of historical volatility one can think of. What happened was that traders were long the calls, short the future, and the market had been rallying slowly. They were losing on their future sales and had to pay for it -without collecting on their corresponding profits on the option side. The calls kept getting discounted; they were too long- dated and nobody wanted to toutch them. What does this mean? Consider that a long term European option can trade below intrinsic value! I mean intrinsic value by the forward! You may not have the funds to arb it... The market can become suddenly inefficient and bankrupt you on the marks as your options can be severely discounted. I recall seing the cash-future discount reach 10% during the crash of 1987. But with an American option you have a lower bound on how much you can be squeezed. Let us look for cases of differential valuation.

Case 1 (Simplest, the bang comes from the convexity to changes in the carry of the premium) Why do changes in interest rate carry always comparatively benefit the American option ? Take a 1 year European and American options on a forward trading at 100, i.e. with a spot at 100. The American option will be priced on the risk management system at exactly the same value as the European one. S=100, F=100, where S is the spot and F is the forward. Assume that the market rallies and the spot goes to 140. Both options will go to parity, and be worth $40.

Case 1 A Assume that interest rates are no longer 0, that both rates go to 10%. F stays equal to S. Suddenly the European option will go from $40 to the present value of $40 in one year using 10%, i.e. $36.36. The American option will stay at $40, like a rock.

Case 1 B Assume the domestic rate goes up to 10%, spot unchanged. F will be worth approximately of S. It will go from 140 to 126, but the P/L should be neutral if the option still has no gamma around 126 (i.e. the options trade at intrinsic value). The European option will still drop to the PV of 26, i.e. 23.636, while the American will be at 26.

We can thus see that the changes in carry always work to the advantage of the American option (assuming the trader is properly delta neutral in the forward). We saw in these two cases the outperformance of the American option. We know the rule that :

If in all scenarios option A is worth at least the same as option B and, in some scenarios can be worth more than option B, then it is not the greatest idea to sell option A and buy option B at the exact same price.

This tells us something but not too much: we know we need to pay more, but how much more?

Case 2 Sensitivity (more serious) to changes in the Dividend/Foreign rate

Another early exercise test needs to be in place, now. Say that we start with S = 140 and F = 140 and that we have both rates equal to 0. Let us compare a European and an American option on cash. As before, they will initially bear the same price on the risk management system.

Assume that that the foreign rate goes to 20%. F goes to approximately S, roughly 1.16. The European call option will be worth roughly $16 (assuming no time value), while the American option will be worth $40. Why ? because the American option being a very smart option, chooses whatever fits it better, between the cash and the future, and positions itself there.

Case 3: More Complex: Sensitivity to the Slope of the Yield Curve

Now let us assume that the yield curve has kinks it it, that it is not quite as linear as one would think. We often such niceties around year end events, when interest rates flip, etc.

As Figure TK shows the final forward might not be the most relevant item. Any bubbling on the intermediate date would affect the value of the American option. Remember that only using the final F is a recipe for being picked-on by a shrewd operator. A risk management and pricing system that uses no full term structure would be considered greatly defective, as it would price both options at the exact same price when clearly the American put is worth more because one can lock-in the forward to the exact point in the middle – where the synthetic underlying is worth the most. Thus using the final interest rate differential would be totally wrong.

To conclude from these examples, the American option is extremely sensitive to the interest rates and their volatility. The higher that volatility the higher the difference between the American and the European. Pricing Problems

It is not possible to price American options using a conventional Monte Carlo simulator. We can, however, try to price them using a more advanced version -or a combination between Monte Carlo and an analytical method. But the knowledge thus gained would be simply comparative.

Further results will follow. It would be great knowledge to quantify their difference, but we have nothing in the present time other than an ordinal relationship.

27.6 THE STOPPING TIME PROBLEM

Another non-trivial problem with American options lies in the fact that the forward hedge is unknown. It resembles the problem with a barrier option except that the conditions of termination are unknown and depend on many parameters (such as volatility, base interest rate, interest rate differential). The intuition of the stopping time problem is as follows: the smart option will position itself on the point on the curve that fits it the best.

Note that the forward maturity ladder in a pricing and risk management system that puts the forward delta in the terminal bucket is WRONG.

27.7 EXPRESSING THE VARIOUS SUB-OPTIONS

27.8 CONCLUSION

BIBLIOGRAPHY

[1] Martha Amram and Nalin Kulatilaka. Real options:: Managing strategic investment in an uncertain world. *OUP Catalogue*, 1998.

[2] Kenneth J Arrow. Aspects of the theory of risk-bearing (yrjo jahnsson lectures). *Yrjo Jahnssonin Saatio, Helsinki*, 1965.

[3] Kenneth Joseph Arrow. *Essays in the theory of risk-bearing*, volume 1. Markham Publishing Company Chicago, 1971.

[4] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] Louis Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 1900.

[6] Kevin P Balanda and HL MacGillivray. Kurtosis: a critical review. *The American Statistician*, 42(2):111–119, 1988.

[7] Nicholas Barberis. The psychology of tail events: Progress and challenges. *American Economic Review*, 103(3):611–16, 2013.

[8] Shlomo Benartzi and Richard H Thaler. Myopic loss aversion and the equity premium puzzle. *The quarterly journal of Economics*, 110(1):73–92, 1995.

[9] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[10] Serge Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97(1):1–59, 1927.

[11] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[12] Marvin Blum. On the sums of independently distributed pareto variates. *SIAM Journal on Applied Mathematics*, 19(1):191–198, 1970.

[13] Émile Borel. *Les probabilités et la vie*, volume 91. Presses universitaires de France, 1943.

[14] Jean-Philippe Bouchaud, J Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand. *(September 11, 2008)*, 2008.

[15] Leo Breiman. Probability, classics in applied mathematics, vol. 7. *Society for Industrial and Applied Mathematics (SIAM), Pennsylvania*, 1992.

[16] L Brennan, I Reed, and William Sollfrey. A comparison of average-likelihood and maximum-likelihood ratio tests for detecting radar targets of unknown doppler frequency. *Information Theory, IEEE Transactions on*, 14(1):104–110, 1968.

[17] VV Buldygin and Yu V Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980.

[18] Rémy Chicheportiche and Jean-Philippe Bouchaud. The joint distribution of stock returns is not elliptical. *International Journal of Theoretical and Applied Finance*, 15(03), 2012.

[19] VP Chistyakov. A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probability & Its Applications*, 9(4):640–648, 1964.

[20] George M Constantinides and Anastasios G Malliaris. Portfolio theory. *Handbooks in operations research and management science*, 9:1–30, 1995.

[21] DA Darling. The influence of the maximum term in the addition of independent random variables. *Transactions of the American Mathematical Society*, 73(1):95–107, 1952.

[22] Bruno De Finetti. *Theory of Probability, volumes I and 2*. Wiley, 1977.

[23] Emanuel Derman and Paul Wilmott. The financial modelers' manifesto. In *SSRN: http://ssrn. com/abstract*, volume 1324878, 2009.

[24] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[25] Wolfgang Doeblin. Sur certains mouvements aléatoires discontinus. *Scandinavian Actuarial Journal*, 1939(1):211–222, 1939.

[26] Wolfgang Doeblin. Sur les sommes dÂŠun grand nombre de variables aléatoires indépendantes. *Bull. Sci. Math*, 63(2):23–32, 1939.

[27] Joseph L Doob. Heuristic approach to the kolmogorov-smirnov theorems. *The Annals of Mathematical Statistics*, 20(3):393–403, 1949.

[28] Bradley Efron. Bayes' theorem in the 21st century. *Science*, 340(6137):1177–1178, 2013.

[29] Jon Elster. Hard and soft obscurantism in the humanities and social sciences. *Diogenes*, 58(1-2):159–170, 2011.

[30] Edwin J Elton and Martin J Gruber. Modern portfolio theory, 1950 to date. *Journal of Banking & Finance*, 21(11):1743–1759, 1997.

[31] Paul Embrechts. *Modelling extremal events: for insurance and finance*, volume 33. Springer, 1997.

[32] Paul Embrechts and Charles M Goldie. On convolution tails. *Stochastic Processes and their Applications*, 13(3):263–278, 1982.

[33] Paul Embrechts, Charles M Goldie, and Noël Veraverbeke. Subexponentiality and infinite divisibility. *Probability Theory and Related Fields*, 49(3):335–347, 1979.

[34] M Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):247–271, 1909.

[35] Robert Engle. Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, pages 157–168, 2001.

[36] CG Esseen. On the concentration function of a sum of independent random variables. *Probability Theory and Related Fields*, 9(4):290–308, 1968.

[37] Kai-Tai Fang. Elliptically contoured distributions. *Encyclopedia of Statistical Sciences*, 2006.

[38] Doyne James Farmer and John Geanakoplos. Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates. 2009.

[39] William Feller. 1971an introduction to probability theory and its applications, vol. 2.

[40] William Feller. An introduction to probability theory. 1968.

[41] Bent Flyvbjerg. Phronetic planning research: theoretical and methodological reflections. *Planning Theory & Practice*, 5(3):283–306, 2004.

[42] Bent Flyvbjerg. From nobel prize to project management: getting risks right. *arXiv preprint arXiv:1302.3642*, 2013.

[43] Shane Frederick, George Loewenstein, and Ted O'donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

[44] DA Freedman and PB Stark. What is the chance of an earthquake? *NATO Science Series IV: Earth and Environmental Sciences*, 32:201–213, 2003.

[45] David Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.

[46] Rainer Froese. Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations. *Journal of Applied Ichthyology*, 22(4):241–253, 2006.

[47] Xavier Gabaix. Power laws in economics and finance. Technical report, National Bureau of Economic Research, 2008.

[48] Gerd Gigerenzer. *Adaptive thinking: rationality in the real world*. Oxford University Press, New York, 2000.

[49] Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009.

[50] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

[51] BV Gnedenko and AN Kolmogorov. Limit distributions for sums of independent random variables (1954). *Cambridge, Mass.*

[52] Charles M Goldie. Subexponential distributions and dominated-variation tails. *Journal of Applied Probability*, pages 440–442, 1978.

[53] Daniel Goldstein and Nassim Taleb. We don't quite know what we are talking about when we talk about volatility. *Journal of Portfolio Management*, 33(4), 2007.

[54] Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2008.

[55] Clyde D Hardin Jr. Skewed stable variables and processes. Technical report, DTIC Document, 1984.

[56] Espen Gaarder Haug and Nassim Nicholas Taleb. Option traders use (very) sophisticated heuristics, never the black–scholes–merton formula. *Journal of Economic Behavior & Organization*, 77(2):97–106, 2011.

[57] Martin Hlusek. On distribution of absolute values. 2011.

[58] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[59] Chi-fu Huang and Robert H Litzenberger. *Foundations for financial economics*. Prentice Hall, 1988.

[60] IA Ibragimov and KE Chernin. On the unimodality of geometric stable laws. *Theory of Probability & Its Applications*, 4(4):417–419, 1959.

[61] Jean-Pierre Kahane. *Some random series of functions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2nd edition, 1993.

[62] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[63] Harry Kesten. A sharper form of the doeblin-lévy-kolmogorov-rogozin inequality for concentration functions. *Mathematica Scandinavica*, 25:133–144, 1969.

[64] John M Keynes. A treatise on probability. 1921.

[65] Leopold Kohr. Leopold kohr on the desirable scale of states. *Population and Development Review*, 18(4):745–750, 1992.

[66] A.N. Kolmogorov. *Selected Works of AN Kolmogorov: Probability theory and mathematical statistics*, volume 26. Springer, 1992.

[67] David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

[68] Paul Lévy and M Émile Borel. *Théorie de l'addition des variables aléatoires*, volume 1. Gauthier-Villars Paris, 1954.

[69] Andrew Lo and Mark Mueller. Warning: physics envy may be hazardous to your wealth! 2010.

[70] Michel Loève. *Probability Theory. Foundations. Random Sequences*. New York: D. Van Nostrand Company, 1955.

[71] Michel Loeve. Probability theory, vol. ii. *Graduate texts in mathematics*, 46:0–387, 1978.

[72] HL MacGillivray and Kevin P Balanda. Mixtures, myths and kurtosis. *Communications in Statistics-Simulation and Computation*, 17(3):789–802, 1988.

[73] M Machina and M Rothschild. Risk.[in:] utility and probability. the new pal-grave. red. j. eatwell, m. milgate, p. newman, 1987.

[74] Mark Machina and Michael Rothschild. Risk. in the new palgrave dictionary of economics, edited by steven n. durlauf and lawrence e. blume, 2008.

[75] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

[76] Harry M Markowitz. *Portfolio selection: efficient diversification of investments*, volume 16. Wiley, 1959.

[77] T Mikosch and AV Nagaev. Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, 1(1):81–110, 1998.

[78] Frederick Mosteller and John W Tukey. Data analysis and regression. a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley, 1977*, 1, 1977.

[79] Aleksandr Viktorovich Nagaev. Integral limit theorems taking into account large deviations when cramér's condition does not hold. ii. *Teoriya Veroyatnostei i ee Primeneniya*, 14(2):203–216, 1969.

[80] Sergey V Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.

[81] Sergey Victorovich Nagaev. Some limit theorems for large deviations. *Theory of Probability & Its Applications*, 10(2):214–235, 1965.

[82] SV Nagaev and IF Pinelis. Some inequalities for the distribution of sums of independent random variables. *Theory of Probability & Its Applications*, 22(2):248–256, 1978.

[83] Gloria Origgi. Is trust an epistemological notion? *Episteme*, 1(01):61–72, 2004.

[84] Athanasios Papoulis. Probability, random variables, and stochastic processes, 1991.

[85] Giovanni Peccati and Murad S Taqqu. *Wiener Chaos: Moments, Cumulants and Diagrams, a Survey with Computer Implementation*, volume 1. Springer, 2011.

[86] Valentin V Petrov. Limit theorems of probability theory. 1995.

[87] Iosif Pinelis. On the characteristic function of the positive part of a random variable. *arXiv preprint arXiv:1309.5928*, 2013.

[88] Steven Pinker. *The better angels of our nature: Why violence has declined*. Penguin, 2011.

[89] EJG Pitman. Subexponential distribution functions. *J. Austral. Math. Soc. Ser. A*, 29(3):337–347, 1980.

[90] John W Pratt. Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society*, pages 122–136, 1964.

[91] Yu V Prokhorov. An extremal problem in probability theory. *Theory of Probability & Its Applications*, 4(2):201–203, 1959.

[92] Yu V Prokhorov. Some remarks on the strong law of large numbers. *Theory of Probability & Its Applications*, 4(2):204–208, 1959.

[93] Colin M Ramsay. The distribution of sums of certain iid pareto variates. *Communications in StatisticsÂŮTheory and Methods*, 35(3):395–405, 2006.

[94] BA Rogozin. An estimate for concentration functions. *Theory of Probability & Its Applications*, 6(1):94–97, 1961.

[95] BA Rogozin. The concentration functions of sums of independent random variables. In *Proceedings of the Second Japan-USSR Symposium on Probability Theory*, pages 370–376. Springer, 1973.

[96] Michael Rothschild and Joseph E Stiglitz. Increasing risk: I. a definition. *Journal of Economic theory*, 2(3):225–243, 1970.

[97] Michael Rothschild and Joseph E Stiglitz. Increasing risk ii: Its economic consequences. *Journal of Economic Theory*, 3(1):66–84, 1971.

[98] Gennady Samorodnitsky and Murad S Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[99] Leonard J Savage. *The foundations of statistics*. Courier Dover Publications, 1954.

[100] Mr Christian Schmieder, Mr Tidiane Kinda, Mr Nassim N Taleb, Elena Loukoianova, and Mr Elie Canetti. *A new heuristic measure of fragility and tail risks: application to stress testing*. Number 12-216. Andrews McMeel Publishing, 2012.

[101] Laurent Schwartz. Théorie des distributions. *Bull. Amer. Math. Soc. 58 (1952), 78-85 DOI: http://dx. doi. org/10.1090/S0002-9904-1952-09555-0 PII*, pages 0002–9904, 1952.

[102] William F Sharpe. Mutual fund performance. *Journal of business*, pages 119–138, 1966.

[103] Vernon L Smith. *Rationality in economics: constructivist and ecological forms*. Cambridge University Press, Cambridge, 2008.

[104] Emre Soyer and Robin M Hogarth. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3):695–711, 2012.

[105] N N Taleb and R Douady. Mathematical definition, mapping, and detection of (anti) fragility. *Quantitative Finance*, 2013.

[106] Nassim Taleb. *Fooled by randomness: The hidden role of chance in life and in the markets*. Random House Trade Paperbacks, 2001/2005.

[107] Nassim N Taleb and Daniel G Goldstein. The problem is beyond psychology: The real world is more random than regression analyses. *International Journal of Forecasting*, 28(3):715–716, 2012.

[108] Nassim Nicholas Taleb. *Dynamic Hedging: Managing Vanilla and Exotic Options*. John Wiley & Sons (Wiley Series in Financial Engineering), 1997.

[109] Nassim Nicholas Taleb. Errors, robustness, and the fourth quadrant. *International Journal of Forecasting*, 25(4):744–759, 2009.

[110] Nassim Nicholas Taleb. *The Black Swan:: The Impact of the Highly Improbable Fragility*. Random House Digital, Inc., 2010.

[111] Nassim Nicholas Taleb. *Antifragile: things that gain from disorder*. Random House and Penguin, 2012.

[112] Albert Tarantola. *Inverse problem theory: Methods for data fitting and model parameter estimation*. Elsevier Science, 2002.

[113] Jozef L Teugels. The class of subexponential distributions. *The Annals of Probability*, 3(6):1000–1011, 1975.

[114] Peter M Todd and Gerd Gigerenzer. *Ecological rationality: intelligence in the world*. Evolution and cognition series. Oxford University Press, Oxford, 2012.

[115] Bence Toth, Yves Lemperiere, Cyril Deremble, Joachim De Lataillade, Julien Kockelkoren, and J-P Bouchaud. Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006, 2011.

[116] Jack L Treynor. How to rate management of investment funds. *Harvard business review*, 43(1):63–75, 1965.

[117] Lenos Trigeorgis. *Real options: Managerial flexibility and strategy in resource allocation*. MIT press, 1996.

[118] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

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

[120] Willem Rutger van Zwet. *Convex transformations of random variables*, volume 7. Mathematisch centrum, 1964.

[121] Rafał Weron. Levy-stable distributions revisited: tail index> 2 does not exclude the levy-stable regime. *International Journal of Modern Physics C*, 12(02):209–223, 2001.

[122] Stephen J Wolfe. On the local behavior of characteristic functions. *The Annals of Probability*, pages 862–866, 1973.

[123] Stephen James Wolfe. On the unimodality of infinitely divisible distribution functions. *Probability Theory and Related Fields*, 45(4):329–335, 1978.

[124] IV Zaliapin, Yan Y Kagan, and Federic P Schoenberg. Approximating the distribution of pareto sums. *Pure and Applied geophysics*, 162(6-7):1187–1228, 2005.

[125] Vladimir M Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Soc., 1986.

INDEX

Skin in the Game, 59

Antifragility, 277, 281, 313, 337, 343

Black Swan, 19, 21, 47, 51, 53, 54, 65, 73, 94, 95, 116, 200, 219, 237, 264, 322, 342, 343

Citation Rings, 58
Concavity/Convexity, iii, iv, xvi, 5, 45, 47, 48, 74, 76, 78, 91, 93, 118, 235, 264–269, 278, 279, 281, 287–289, 297, 303, 313, 315, 319–321, 323–326, 342, 343

Econometrics, 54, 94, 192, 196, 200, 201, 297
Egomaniac as an rebuttal, 200
Empiricism (Naive Empiricism), 47

Finance Artists, 81
Fragility, 46, 49, 129, 264, 266, 287, 329

Law of Large Numbers, xv, 19, 22, 41, 51, 65, 95, 119, 123, 124, 133, 137, 139, 153, 155, 156, 181, 183
Loss Function, 66, 67

Measure Theory, 52, 53, 164

Pinker Problem, 186
Pseudo-biases, 131

Skin in the Game, xv, xvi, 7, 35, 60, 61, 201, 249, 258

Triffat Fallacy, 59
Turkey Problem, 65, 95, 184

Via Negativa, 36