The Technical Incerto

Statistical Consequences of Fat Tails

REAL WORLD PREASYMPTOTICS, EPISTEMOLOGY, AND APPLICATIONS

Papers and Commentary

Nassim Nicholas Taleb

THE TECHNICAL INCERTO COLLECTION

STATISTICAL CONSEQUENCES OF FAT TAILS

Real World Preasymptotics, Epistemology, and Applications

Papers and Commentary

NASSIM NICHOLAS TALEB

STEM
ACADEMIC PRESS

2019

COAUTHORS [1]

Pasquale Cirillo (Chapters 13, 15, and 16)
Raphael Douady (Chapter 14)
Andrea Fontanari (Chapter 13)
Hélyette Geman (Chapter 25)
Donald Geman (Chapter 25)
Espen Haug (Chapter 22)
The Universa Investments team (Chapter 23)

1 Papers relied upon here are [45, 46, 47, 94, 105, 124, 163, 221, 224, 225, 226, 228, 229, 230, 231, 239, 240, 241]

CONTENTS

*Nontechnical chapters are indicated with a star *; Discussion chapters are indicated with a †; adaptation from published ("peer-reviewed") papers with a ‡.*
While chapters are indexed by Arabic numerals, expository and very brief mini-chapters (half way between appendices and full chapters) use letters such as A, B, etc.

1 PROLOGUE*,† 1

2 GLOSSARY, DEFINITIONS, AND NOTATIONS 5

2.1 General Notations and Frequently Used Symbols 5

2.2 Catalogue Raisonné of General & Idiosyncratic concepts 7

2.2.1 Power Law Class $\mathfrak{P}$ 7

2.2.2 Law of Large Numbers (Weak) 8

2.2.3 The Central Limit Theorem (CLT) 8

2.2.4 Law of Medium Numbers or Preasymptotics 8

2.2.5 Kappa Metric 8

2.2.6 Elliptical distribution 9

2.2.7 Statistical independence 9

2.2.8 Stable (Lévy stable) Distribution 9

2.2.9 Multivariate Stable Distribution 10

2.2.10 Karamata point 10

2.2.11 Subexponentiality 10

2.2.12 Student T as proxy 11

2.2.13 Citation ring 11

2.2.14 Rent seeking in academia 12

2.2.15 Pseudo-empiricism or Pinker Problem 12

2.2.16 Preasymptotics 12

2.2.17 Stochasticizing 12

2.2.18 Value at Risk, Conditional VaR 13

2.2.19 Skin in the game 13

2.2.20 MS Plot 13

2.2.21 Maximum domain of attraction, MDA 14

2.2.22 Substitution of Integral in the psychology literature 14

2.2.23 Inseparability of probability (another common error) 14

2.2.24 Wittgenstein's Ruler 15

2.2.25 Black swan 15

2.2.26 The empirical distribution is not empirical 15

2.2.27 The hidden tail 16

2.2.28 Shadow moment 16
2.2.29 Tail dependence 17
2.2.30 Metaprobability 17
2.2.31 Dynamic hedging 17

I FAT TAILS AND THEIR EFFECTS, AN INTRODUCTION 19

3 A NON-TECHNICAL OVERVIEW – THE DARWIN COLLEGE LECTURE *,‡ 21
3.1 On the Difference Between Thin and Thick Tails 21
3.2 A (More Advanced) Categorization and Its Consequences 23
3.3 The Main Consequences and How They Link to the Book 28
3.3.1 Forecasting 37
3.3.2 The Law of Large Numbers 40
3.4 Epistemology and Inferential Asymmetry 40
3.5 Naive Empiricism: Ebola Should Not Be Compared to Falls from Ladders 46
3.6 Primer on Power Laws (almost without mathematics) 49
3.7 Where Are the Hidden Properties? 52
3.8 Bayesian Schmayesian 55
3.9 X vs $F(X)$, exposures to X confused with knowledge about X 56
3.10 Ruin and Path Dependence 58
3.11 What To Do? 61

4 UNIVARIATE FAT TAILS, LEVEL 1, FINITE MOMENTS† 63
4.1 A Simple Heuristic to Create Mildly Fat Tails 63
4.1.1 A Variance-preserving heuristic 65
4.1.2 Fattening of Tails With Skewed Variance 66
4.2 Does Stochastic Volatility Generate Power Laws? 68
4.3 The Body, The Shoulders, and The Tails 69
4.3.1 The Crossovers and Tunnel Effect. 70
4.4 Fat Tails, Mean Deviation and the Rising Norms 73
4.4.1 The Common Errors 73
4.4.2 Some Analytics 74
4.4.3 Effect of Fatter Tails on the "efficiency" of STD vs MD 76
4.4.4 Moments and The Power Mean Inequality 77
4.4.5 Comment: Why we should retire standard deviation, now! 80
4.5 Visualizing the effect of rising p on iso-norms 84

5 LEVEL 2: SUBEXPONENTIALS AND POWER LAWS 87
5.0.1 Revisiting the Rankings 87
5.0.2 What is a borderline probability distribution? 89
5.0.3 Let Us Invent a Distribution 90
5.1 Level 3: Scalability and Power Laws 91
5.1.1 Scalable and Nonscalable, A Deeper View of Fat Tails 91
5.1.2 Grey Swans 93
5.2 Some Properties of Power Laws 94
5.2.1 Sums of variables 94
5.2.2 Transformations 95
5.3 Bell Shaped vs Non Bell Shaped Power Laws 96
5.4 Super-Fat Tails: The Log-Pareto Distribution 97

5.5 Pseudo-Stochastic Volatility: An investigation 98
6 THICK TAILS IN HIGHER DIMENSIONS† 101
6.1 Thick Tails in Higher Dimension, Finite Moments 102
6.2 Joint Fat-Tailedness and Ellipticality of Distributions 104
6.3 Multivariate Student T 106
6.3.1 Ellipticality and Independence under Thick Tails 106
6.4 Fat Tails and Mutual Information 108
6.5 Fat Tails and Random Matrices, a Rapid Interlude 108
6.6 Correlation and Undefined Variance 109
6.7 Fat Tailed Residuals in Linear Regression Models 110
A SPECIAL CASES OF THICK TAILS 115
A.1 Multimodality and Thick Tails, or the War and Peace Model 115
A.2 Transition Probabilities: What Can Break Will Break 119

II THE LAW OF MEDIUM NUMBERS 121
7 LIMIT DISTRIBUTIONS, A CONSOLIDATION*,† 123
7.1 Refresher: The Weak and Strong LLN 123
7.2 Central Limit in Action 125
7.2.1 The Stable Distribution 125
7.2.2 The Law of Large Numbers for the Stable Distribution 126
7.3 Speed of Convergence of CLT: Visual Explorations 127
7.3.1 Fast Convergence: the Uniform Dist. 127
7.3.2 Semi-slow convergence: the exponential 128
7.3.3 The slow Pareto 129
7.3.4 The half-cubic Pareto and its basin of convergence 131
7.4 Cumulants and Convergence 131
7.5 Technical Refresher: Traditional Versions of CLT 133
7.6 The Law of Large Numbers for Higher Moments 134
7.6.1 Higher Moments 134
7.7 Mean deviation for a Stable Distributions 137
8 HOW MUCH DATA DO YOU NEED? AN OPERATIONAL METRIC FOR FAT-TAILEDNESS‡ 139
8.1 Introduction and Definitions 140
8.2 The Metric 142
8.3 Stable Basin of Convergence as Benchmark 144
8.3.1 Equivalence for Stable distributions 145
8.3.2 Practical significance for sample sufficiency 145
8.4 Technical Consequences 147
8.4.1 Some Oddities With Asymmetric Distributions 147
8.4.2 Rate of Convergence of a Student T Distribution to the Gaussian Basin 147
8.4.3 The Lognormal is Neither Thin Nor Fat Tailed 148
8.4.4 Can Kappa Be Negative? 148
8.5 Conclusion and Consequences 148
8.5.1 Portfolio Pseudo-Stabilization 149
8.5.2 Other Aspects of Statistical Inference 150
8.5.3 Final comment 150
8.6 Appendix, Derivations, and Proofs 150

8.6.1 Cubic Student T (Gaussian Basin) 150
8.6.2 Lognormal Sums 152
8.6.3 Exponential 154
8.6.4 Negative Kappa, Negative Kurtosis 155
9 EXTREME VALUES AND HIDDEN RISK *,† 157
9.1 Preliminary Introduction to EVT 157
9.1.1 How Any Power Law Tail Leads to Fréchet 159
9.1.2 Gaussian Case 160
9.1.3 The Picklands-Balkema-de Haan Theorem 162
9.2 The Invisible Tail for a Power Law 163
9.2.1 Comparison with the Normal Distribution 166
9.3 Appendix: The Empirical Distribution is Not Empirical 166
B THE LARGE DEVIATION PRINCIPLE, IN BRIEF 169
C CALIBRATING UNDER PARETIANITY 171
C.1 Distribution of the sample tail Exponent 173
10 "IT IS WHAT IT IS": DIAGNOSING THE SP500† 175
10.1 Paretianity and Moments 175
10.2 Convergence Tests 177
10.2.1 Test 1: Kurtosis under Aggregation 177
10.2.2 Maximum Drawdowns 178
10.2.3 Empirical Kappa 179
10.2.4 Test 2: Excess Conditional Expectation 180
10.2.5 Test 3- Instability of 4^{th} moment 182
10.2.6 Test 4: MS Plot 182
10.2.7 Records and Extrema 184
10.2.8 Asymmetry right-left tail 187
10.3 Conclusion: It is what it is 187
D THE PROBLEM WITH ECONOMETRICS 189
D.1 Performance of Standard Parametric Risk Estimators 190
D.2 Performance of Standard NonParametric Risk Estimators 191
E MACHINE LEARNING CONSIDERATIONS 195
E.0.1 Calibration via Angles 197

III PREDICTIONS, FORECASTING, AND UNCERTAINTY 199

11 PROBABILITY CALIBRATION CALIBRATION UNDER FAT TAILS ‡ 201
11.1 Continuous vs. Discrete Payoffs: Definitions and Comments 202
11.1.1 Away from the Verbalistic 203
11.1.2 There is no defined "collapse", "disaster", or "success" under fat tails 206
11.2 Spurious overestimation of tail probability in psychology 208
11.2.1 Thin tails 208
11.2.2 Fat tails 209
11.2.3 Conflations 209
11.2.4 Distributional Uncertainty 212
11.3 Calibration and Miscalibration 213
11.4 Scoring Metrics 213
11.4.1 Deriving Distributions 216

11.5 Non-Verbalistic Payoff Functions and The Good News from Machine Learning 217
11.6 Conclusion: 219
11.7 Appendix: Proofs and Derivations 220
11.7.1 Distribution of Binary Tally $P^{(p)}(n)$ 220
11.7.2 Distribution of the Brier Score 220
12 ELECTION PREDICTIONS AS MARTINGALES: AN ARBITRAGE APPROACH‡ 223
12.0.1 Main results 225
12.0.2 Organization 226
12.0.3 A Discussion on Risk Neutrality 228
12.1 The Bachelier-Style valuation 228
12.2 Bounded Dual Martingale Process 230
12.3 Relation to De Finetti's Probability Assessor 231
12.4 Conclusion and Comments 233

IV INEQUALITY ESTIMATORS UNDER FAT TAILS 237
13 GINI ESTIMATION UNDER INFINITE VARIANCE ‡ 239
13.1 Introduction 239
13.2 Asymptotics of the Nonparametric Estimator under Infinite Variance 243
13.2.1 A Quick Recap on α-Stable Random Variables 244
13.2.2 The α-Stable Asymptotic Limit of the Gini Index 245
13.3 The Maximum Likelihood Estimator 246
13.4 A Paretian illustration 247
13.5 Small Sample Correction 250
13.6 Conclusions 253
14 ON THE SUPER-ADDITIVITY AND ESTIMATION BIASES OF QUANTILE CONTRIBUTIONS ‡ 259
14.1 Introduction 259
14.2 Estimation For Unmixed Pareto-Tailed Distributions 261
14.2.1 Bias and Convergence 261
14.3 An Inequality About Aggregating Inequality 264
14.4 Mixed Distributions For The Tail Exponent 267
14.5 A Larger Total Sum is Accompanied by Increases in $\widehat{\kappa}_q$ 270
14.6 Conclusion and Proper Estimation of Concentration 270
14.6.1 Robust methods and use of exhaustive data 271
14.6.2 How Should We Measure Concentration? 271

V SHADOW MOMENTS PAPERS 273
15 ON THE SHADOW MOMENTS OF APPARENTLY INFINITE-MEAN PHENOMENA (WITH P. CIRILLO)‡ 275
15.1 Introduction 275
15.2 The dual Distribution 276
15.3 Back to Y: the shadow mean (or population mean) 278
15.4 Comparison to other methods 281
15.5 Applications 282

16 ON THE TAIL RISK OF VIOLENT CONFLICT AND ITS UNDERESTIMATION (WITH P. CIRILLO)‡ 285
16.1 Introduction/Summary 285
16.2 Summary statistical discussion 288
16.2.1 Results 288
16.2.2 Conclusion 289
16.3 Methodological Discussion 289
16.3.1 Rescaling method 289
16.3.2 Expectation by Conditioning (less rigorous) 291
16.3.3 Reliability of data and effect on tail estimates 292
16.3.4 Definition of an "event" 293
16.3.5 Missing events 294
16.3.6 Survivorship Bias 294
16.4 Data analysis 294
16.4.1 Peaks over Threshold 295
16.4.2 Gaps in Series and Autocorrelation 296
16.4.3 Tail Analysis 297
16.4.4 An Alternative View on Maxima 299
16.4.5 Full Data Analysis 299
16.5 Additional robustness and reliability tests 300
16.5.1 Bootstrap for the GPD 300
16.5.2 Perturbation Across Bounds of Estimates 301
16.6 Conclusion: is the world more unsafe than it seems? 302
16.7 Acknowledgments 304
F WHAT ARE THE CHANCES OF A THIRD WORLD WAR?*,† 305

VI METAPROBABILITY PAPERS 309
17 HOW THICK TAILS EMERGE FROM RECURSIVE EPISTEMIC UNCERTAINTY† 311
17.1 Methods and Derivations 312
17.1.1 Layering Uncertainties 312
17.1.2 Higher Order Integrals in the Standard Gaussian Case 313
17.1.3 Effect on Small Probabilities 317
17.2 Regime 2: Cases of decaying parameters $a(n)$ 319
17.2.1 Regime 2-a;"Bleed" of Higher Order Error 319
17.2.2 Regime 2-b; Second Method, a Non Multiplicative Error Rate 320
17.3 Limit Distribution 321
18 STOCHASTIC TAIL EXPONENT FOR ASYMMETRIC POWER LAWS† 323
18.1 Background 324
18.2 One Tailed Distributions with Stochastic Alpha 324
18.2.1 General Cases 324
18.2.2 Stochastic Alpha Inequality 325
18.2.3 Approximations for the Class $\mathfrak{P}$ 326
18.3 Sums of Power Laws 327
18.4 Asymmetric Stable Distributions 328
18.5 Pareto Distribution with lognormally distributed α 329
18.6 Pareto Distribution with Gamma distributed Alpha 330
18.7 The Bounded Power Law in Cirillo and Taleb (2016) 330

18.8 Additional Comments 331
18.9 Acknowledgments 331
19 META-DISTRIBUTION OF P-VALUES AND P-HACKING‡ 333
19.1 Proofs and derivations 335
19.2 Inverse Power of Test 339
19.3 Application and Conclusion 340
G SOME CONFUSIONS IN BEHAVIORAL ECONOMICS 343
G.1 Case Study: How the myopic loss aversion is misspecified 343

VII OPTION TRADING AND PRICING UNDER FAT TAILS 349
20 FINANCIAL THEORY'S FAILURES WITH OPTION PRICING† 351
20.1 Bachelier not Black-Scholes 351
20.1.1 Distortion from Idealization 352
20.1.2 The Actual Replication Process: 354
20.1.3 Failure: How Hedging Errors Can Be Prohibitive. 354
21 UNIQUE OPTION PRICING MEASURE WITH NEITHER DYNAMIC HEDGING NOR COMPLETE MARKETS‡ 355
21.1 Background 355
21.2 Proof 357
21.2.1 Case 1: Forward as risk-neutral measure 357
21.2.2 Derivations 357
21.3 Case where the Forward is not risk neutral 361
21.4 comment 361
22 OPTION TRADERS NEVER USE THE BLACK-SCHOLES-MERTON FORMULA*,‡ 363
22.1 Breaking the Chain of Transmission 363
22.2 Introduction/Summary 364
22.2.1 Black-Scholes was an argument 364
22.3 Myth 1: Traders did not price options before BSM 367
22.4 Methods and Derivations 368
22.4.1 Option formulas and Delta Hedging 371
22.5 Myth 2: Traders Today use Black-Scholes 372
22.5.1 When do we value? 373
22.6 On the Mathematical Impossibility of Dynamic Hedging 373
22.6.1 The (confusing) Robustness of the Gaussian 375
22.6.2 Order Flow and Options 376
22.6.3 Bachelier-Thorp 376
23 OPTION PRICING UNDER POWER LAWS: A ROBUST HEURISTIC*,‡ 379
23.1 Introduction 380
23.2 Call Pricing beyond the Karamata constant 380
23.2.1 First approach, S is in the regular variation class 381
23.2.2 Second approach, S has geometric returns in the regular variation class 382
23.3 Put Pricing 384
23.4 Arbitrage Boundaries 385
23.5 Comments 386
24 FOUR MISTAKES IN QUANTITATIVE FINANCE*,‡ 387
24.1 Conflation of Second and Fourth Moments 387

24.2 Missing Jensen's Inequality in Analyzing Option Returns 388
24.3 The Inseparability of Insurance and Insured 389
24.4 The Necessity of a Numéraire in Finance 390
24.5 Appendix (Betting on Tails of Distribution) 390
25 TAIL RISK CONSTRAINTS AND MAXIMUM ENTROPY (W. D.& H. GEMAN)‡ 393
25.1 Left Tail Risk as the Central Portfolio Constraint 393
25.1.1 The Barbell as seen by E.T. Jaynes 396
25.2 Revisiting the Mean Variance Setting 396
25.2.1 Analyzing the Constraints 397
25.3 Revisiting the Gaussian Case 398
25.3.1 A Mixture of Two Normals 399
25.4 Maximum Entropy 400
25.4.1 Case A: Constraining the Global Mean 401
25.4.2 Case B: Constraining the Absolute Mean 401
25.4.3 Case C: Power Laws for the Right Tail 402
25.4.4 Extension to a Multi-Period Setting: A Comment 403
25.5 Comments and Conclusion 405
25.6 Appendix/Proofs 405

Bibliography and Index 407

1 PROLOGUE*,†

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

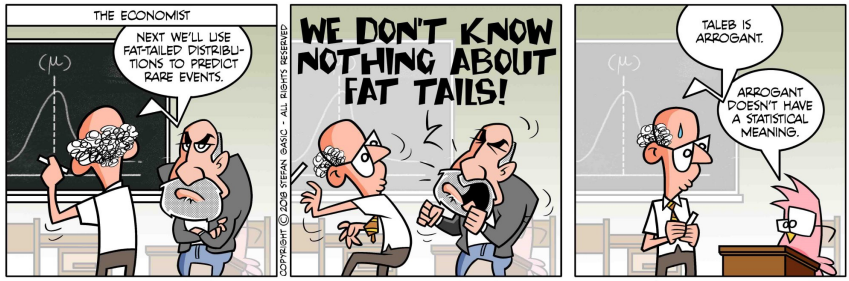


Figure 1.1: *The problem is not awareness of "fat tails", but the lack of understanding of their consequences. Saying "it is fat tailed" implies much more than changing the name of the distribution, but a general overhaul of the statistical tools and types of decisions made. Credit Stefan Gasic.*

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

THIS BOOK consists in 1) published papers and 2) (uncensored) commentary, about classes of statistical distributions that deliver extreme events, and how we should deal with them for both statistical inference and decision making. Most "standard" statistics come from theorems designed for thin tails:

* Discussion chapter.

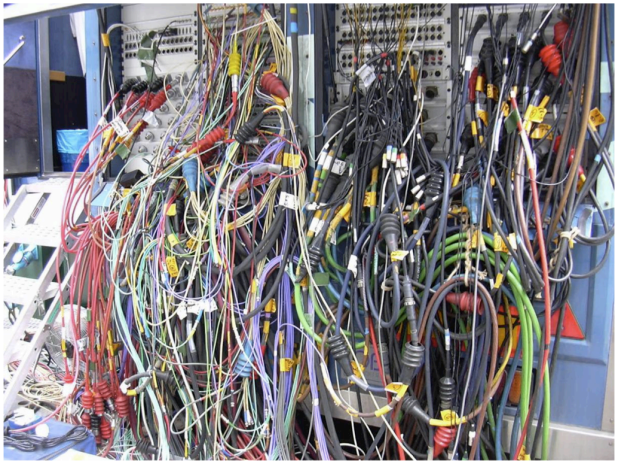


Figure 1.2: *Complication without insight: the clarity of mind of many professionals using statistics and data science without an understanding of the core concepts, what it is fundamentally about.* *Credit: Wikimedia.*

they need to be adapted preasymptotically to fat tails, which is not trivial –or abandoned altogether.

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

More generally, this book draws on the author's multi-volume series, *Incerto* [223] and associated technical research program, which is about how to live in the real world, a world with a structure of uncertainty that is too complicated for us.

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

A WORD ON TERMINOLOGY

"Thick tails" is often used in academic contexts. For us, here, it maps to much "higher kurtosis than the Gaussian" –to conform to the finance practitioner's lingo. As to "Fat Tails", we prefer to reserve it both extreme thick tails or membership in the power law class (which we show in Chapter 8 cannot be disentangled). For many it is meant to be a narrower definition, limited to "power laws" or "regular variations" – but we prefer to call "power laws" "power laws" (when we are quite

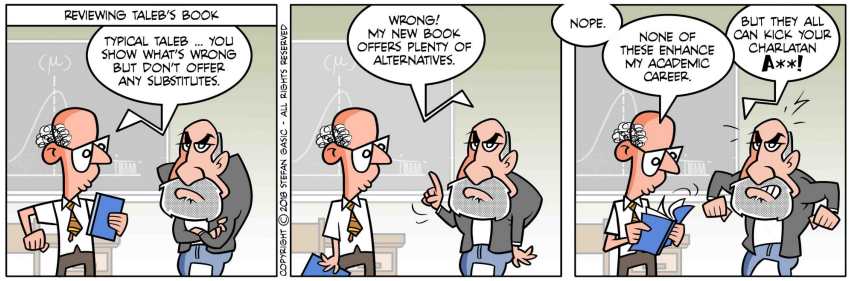


Figure 1.3: *Credit: Stefan Gasic.*

certain about the process), so what we call "fat tails" may sometimes be more technically "extremely thick tails" for many.

To avoid ambiguity, we stay away from designations such as "heavy tails" or "long tails".

The next two chapters will clarify.

ACKNOWLEDGMENTS

In addition to coauthors mentioned earlier, the author is indebted to Zhuo Xi, Jean-Philippe Bouchaud, Robert Frey, Spyros Makridakis, Mark Spitznagel, Brandon Yarkin, Raphael Douady, Peter Carr, Marco Avellaneda, Didier Sornette, Paul Embrechts, Bruno Dupire, Jamil Baz, Damir Delic, Yaneer Bar-Yam, Diego Zviovich, Joseph Norman, Ole Peters, Robert Frey, Chitpuneet Mann, Harry Crane –and of course endless, really endless discussions with the great Benoit Mandelbrot.

Twitter volunteer editors such as Maxime Biette cleared many typos.

Some of the papers that turned into chapters have been presented at conferences; the author thanks Lauren de Haan, Bert Zwart, and others for comments on extreme value related problems. More specific acknowledgements will be made within individual chapters. As usual, the author would like to express his gratitude to the staff at Naya restaurant in NY.

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

Clearly, in business or in anything subjected to the rigors of the real world, he would have been terminated. People who never had any skin in the game [233] cannot figure out the necessity of circumstantial suspension of belief and the informational value of unreliability for decision making: *don't give a pilot a faulty metric, learn to provide only reliable information; letting the pilot know that the plane is defective saves lives*. Nor can they get the outperformance of *via negativa* –Popperian science works by removal. The late David Freedman had tried unsuccessfully to tame vapid and misleading statistical modeling vastly *outperformed* by "nothing".

But it is the case that the various chapters and papers here *do* offer solutions and alternatives, except that these aren't the most comfortable for some as they require some mathematical work for re-derivations for fat tailed conditions.

2 GLOSSARY, DEFINITIONS, AND NOTATIONS

THIS IS A catalogue raisonné of the main topics and notations. Notations are redefined in the text every time; this is an aid for the random peruser. Some chapters extracted from papers will have specific notations, as specified. Note that while our terminology may be at variance with that of some research groups, it aims at remaining consistent.

2.1 GENERAL NOTATIONS AND FREQUENTLY USED SYMBOLS

$\mathbb{P}$ is the probability symbol; typically in $\mathbb{P}(X > x)$, X is the random variable, x is the realization. More formal measure-theoretic definitions of events and other French matters are in Chapter 11 and other places where such formalism makes sense.

$\mathbb{E}$ is the expectation operator.

$\mathbb{V}$ is the Variance operator.

$\mathbb{M}$ is the mean absolute deviation which is, when centered, centered around the mean (rather than the median).

$\varphi(.)$ and $f(.)$ are usually reserved for the PDF (probability density function) of a pre-specified distribution. In some chapters, a distinction is made for $f_x(x)$ and $f_y(y)$, particularly when X and Y follow two separate distributions.

n is usually reserved for the number of summands.

p is usually reserved for the moment order.

r.v. is short for a random variable.

$F(.)$ is reserved for the CDF (cumulative distribution function $\mathbb{P}(X < x)$, $\overline{F}(.)$, or S is the survival function $\mathbb{P}(X > x)$.

$\sim$ indicates that a random variable is distributed according to a certain specified law.

$\chi(t) = \mathbb{E}(e^{itX_s})$ is the characteristic function of a distribution. In some discussions, the argument $t \in \mathbb{R}$ is represented as ω. Sometimes Ψ is used.

$\xrightarrow{D}$ denotes convergence in distribution, as follows. Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables; $X_n \xrightarrow{D} X$ means the CDF F_n for X_n has the following limit:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every real x for which F is continuous.

$\xrightarrow{P}$ denotes convergence in probability, that is for $\varepsilon > 0$, we have, using the same sequence as before

$$\lim_{n\to\infty} \Pr\left(|X_n - X| > \varepsilon\right) = 0.$$

$\xrightarrow{a.s.}$ denotes almost sure convergence, the stronger form:

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

S_n is typically a sum for n summands.

α and α_s: we shall typically try to use $\alpha_s \in (0, 2]$ to denote the tail exponent of the limiting and Platonic stable distribution and $\alpha_p \in (0, \infty)$ the corresponding Paretian (preasymptotic) equivalent but only in situations where there could be some ambiguity. Plain α should be understood in context.

$\mathcal{N}(\mu_1, \sigma_1)$ the Gaussian distribution with mean μ_1 and variance σ_1^2.

$\mathcal{L}(.,.)$ or $\mathcal{LN}(.,.)$ is the Lognormal distribution, with pdf $f^{(L)}(.)$ typically parametrized here as $\mathcal{L}(X_0 - \frac{1}{\sigma^2}, \sigma)$ to get a mean X_0, and variance $\left(e^{\sigma^2} - 1\right) X_0^2$.

$\mathcal{S}(\alpha_S, \beta, \mu, \sigma)$ is the stable distribution with tail index α_s in $(0, 2]$, symmetry index β in $-1, 1)$, centrality parameter μ in $\mathbb{R}$ and scale $\sigma > 0$.

$\mathfrak{P}$ is the power law class (see below).

$\mathfrak{S}$ is the subexponential class (see below).

$\delta(.)$ is the Dirac delta function.

$\theta(.)$ is the Heaviside theta function.

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

$\|.\|_p$ is a norm defined for (here a real vector) $\mathbf{X} = (X_1, \ldots, X_n)^T$,

$\|X\|_p \triangleq \left(\frac{1}{n} \sum_{i=1}^n |x_i|^p\right)^{1/p}$. Note the absolute value in this text.

${}_1F_1(.;.;.)$ is the Kummer confluent hypergeometric function: ${}_1F_1(a;b;z) = \sum_{k=0}^{\infty} \frac{a_k \frac{z^k}{k}!}{b_k}$.

${}_2\tilde{F}_2$ is the generalized hypergeometric function regularized: ${}_2\tilde{F}_2(.,.;.,.;.) = \frac{{}_2F_2(a;b;z)}{\left(\Gamma(b_1)...\Gamma(b_q)\right)}$ and ${}_pF_q(a;b;z)$ has series expansion $\sum_{k=0}^{\infty} \frac{(a_1)_k...(a_p)_k}{(b_1)_k...(b_p)_k} z^k/k!$, were $(a_q)_{(.)}$ is the Pockhammer symbol.

$(a_q)_{(.)}$ is the Q-Pochhammer symbol $(a_q)_n = \prod_{i=1}^{n-1} \left(1 - aq^i\right)$.

2.2 CATALOGUE RAISONNÉ OF GENERAL & IDIOSYNCRATIC CONCEPTS

Next is the duplication of the definition of some central topics.

2.2.1 Power Law Class $\mathfrak{P}$

The power law class is conventionally defined by the property of the survival function, as follows. Let X be a random variable belonging to the class of distributions with a "power law" right tail, that is:

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

where $L : [x_{\min}, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as

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

for any $k > 0$ [22].

The survival function of X is called to belong to the "regular variation" class RV_α. More specifically, a function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is index varying at infinity with index ρ ($f \in RV_\rho$) when

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^\rho.$$

More practically, there is a point where $L(x)$ approaches its limit, l, becoming a constant –which we call the "Karamata constant" and the point is dubbed the "Karamata point". Beyond such value the tails for power laws are calibrated using such standard techniques as the Hill estimator. The distribution in that zone is dubbed the strong Pareto law by B. Mandelbrot [160],[74].

The same applies, when specified, to the left tail.

2.2.2 Law of Large Numbers (Weak)

The standard presentation is as follows. Let $X_1, X_2, \dots X_n$ be an infinite sequence of independent and identically distributed (Lebesgue integrable) random variables with expected value $\mathbb{E}(X_n) = \mu$ (though one can somewhat relax the i.i.d. assumptions). The sample average $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ converges to the expected value, $\overline{X}_n \to \mu$, for $n \to \infty$.

Finiteness of variance is not necessary (though of course the finite higher moments accelerate the convergence).

The strong law is discussed where needed.

2.2.3 The Central Limit Theorem (CLT)

The Standard (Lindeberg-Lévy) version of CLT is as follows. Suppose a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2 < +\infty$, and $\overline{X}_n$ the sample average for n. Then as n approaches infinity, the sum of the random variables $\sqrt{n}(\overline{X}_n \mu)$ converges in distribution to the Gaussian [20] [21]:

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

Convergence in distribution here means that the CDF (cumulative distribution function) of $\sqrt{n}$ converges pointwise to the CDF of $\mathcal{N}(0, \sigma)$ for every real z,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}(\overline{X}_n - \mu) \leq z\right) = \lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq \frac{z}{\sigma}\right] = \Phi\left(\frac{z}{\sigma}\right), \ \sigma > 0$$

where $\Phi(z)$ is the standard normal CDF evaluated at z.

There are many other versions of the CLT, presented as needed.

2.2.4 Law of Medium Numbers or Preasymptotics

This is pretty much the central topic of this book. We are interested in the behavior of the random variable for n large but not too large or asymptotic. While it is not a big deal for the Gaussian owing to extremely rapid convergence (by both LLN and CLT), this is not the case for other random variables.

See **Kappa** next.

2.2.5 Kappa Metric

Metric here should not be interpreted in the mathematical sense of a distance function, but rather in its engineering sense, as a quantitative measurement.

Kappa, in [0, 1], developed by this author here, in Chapter 8, and in a paper [232], gauges the preasymptotic behavior or a random variable; it is 0 for the Gaussian considered as benchmark, and 1 for a Cauchy or a r.v. that has no mean.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean, that is $\mathbb{E}(X) < +\infty$. Let $S_n = X_1 + X_2 + \ldots + X_n$ be a partial sum. Let $\mathbb{M}(n) = \mathbb{E}(|S_n - \mathbb{E}(S_n)|)$ be the expected mean absolute deviation from the mean for n summands (recall we do not use the median but center around the mean). Define the "rate" of convergence for n additional summands starting with n_0:

$$\kappa_{n_0,n} = \min\left\{\kappa_{n_0,n} : \frac{\mathbb{M}(n)}{\mathbb{M}(n_0)} = \left(\frac{n}{n_0}\right)^{\frac{1}{2-\kappa_{n_0,n}}}, n_0 = 1, 2, \ldots\right\}, \tag{2.2}$$

$n > n_0 \geq 1$, hence

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\log\left(\frac{\mathbb{M}(n)}{\mathbb{M}(n_0)}\right)}. \tag{2.3}$$

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

2.2.6 Elliptical distribution

X, a $p \times 1$ random vector is said to have an elliptical (or elliptical contoured) distribution with location parameters μ, a non-negative matrix Σ and some scalar function Ψ if its characteristic function is of the form $exp(it'\mu)\Psi(t\Sigma t')$.

2.2.7 Statistical independence

Independence between two variables X and Y with marginal PDFs $f(x)$ and $f(y)$ and joint PDF $f(x, y)$ is defined by the identity:

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

regardless of the correlation coefficient. In the class of elliptical distributions, the bivariate Gaussian with coefficient 0 is both independent and uncorrelated. This does not apply to the Student T or the Cauchy in their multivariate forms.

2.2.8 Stable (Lévy stable) Distribution

This is a generalization of the CLT.

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

$$\frac{Sn - a_n}{b_n} \xrightarrow{D} X_s, \tag{2.4}$$

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

$$\chi(t) = e^{\left(i\mu t - |t\sigma|_s^{\alpha}\left(1 - i\beta\tan\left(\frac{\pi\alpha_s}{2}\right)\mathrm{sgn}(t)\right)\right)} \text{ when } \alpha_s \neq 1. \tag{2.5}$$

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

2.2.9 Multivariate Stable Distribution

A random vector $\mathbf{X} = (X_1, \ldots, X_k)^T$ is said to have the multivariate stable distribution if every linear combination of its components $\mathbf{Y} = a_1X_1 + \cdots + a_kX_k$ has a stable distribution. That is, for any constant vector $\mathbf{a} \in \mathbb{R}^k$, the random variable $\mathbf{Y} = a^T\mathbf{X}$ should have a univariate stable distribution.

2.2.10 Karamata point

See **Power Law Class**

2.2.11 Subexponentiality

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

Let $\mathbf{X} = X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with support in $(\mathbb{R}^+)$, with cumulative distribution function F. The subexponential class of distributions is defined by (see [243], [193]):

$$\lim_{x \to +\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2 \tag{2.6}$$

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

More generally, it can be shown that the sum of n variables is dominated by the maximum of the values over those variables in the same way. Formally, the following two properties are equivalent to the subexponential condition [43],[83]. For a given $n \geq 2$, let $S_n = \Sigma_{i=1}^{n} x_i$ and $M_n = \max_{1 \leq i \leq n} x_i$

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

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

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

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

$$\int_0^\infty \mathbf{e}^{\epsilon x}\, dF(x) = +\infty \tag{2.7}$$

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

2.2.12 Student T as proxy

We use the student T with α degrees of freedom as a convenient two-tailed power law distribution. For $\alpha = 1$ it becomes a Cauchy, and of course Gaussian for $\alpha \to \infty$.

The student T is the main bell-shaped power law, that is, the PDF is continuous and smooth, asymptotically approaching zero for large negative/positive x, and with a single, unimodal maximum (further, the PDF is quasiconcave but nor concave).

2.2.13 Citation ring

A highly circular mechanism by which academic prominence is reached thanks to discussions where papers are considered prominent because other people are citing them, with no external filtering, thus causing research to concentrate and get stuck around "corners", focal areas of no real significance. This is linked to the operation of the academic system in the absence of adult supervision or the filtering of skin in the game.

XAMPLE of fields that are, practically, frauds in the sense that their results are not portable to reality and only serve to feed additional papers that, in turn, will produce more papers: Modern Financial Theory, econometrics (particularly for macro variables), GARCH processes, psychometrics, stochastic control models in finance, behavioral economics and finance, decision making under uncertainty, macroeconomics, and a bit more.

2.2.14 Rent seeking in academia

There is a conflict of interest between a given researcher and the subject under consideration. The objective function of an academic department (and person) becomes collecting citations, honors, etc. at the expense of the purity of the subject: for instance many people get stuck in research corners because it is more beneficial to their careers and to their department.

2.2.15 Pseudo-empiricism or Pinker Problem

Discussion of "evidence" that is not statistically significant, or use of metrics that are uninformative because they do not apply to the random variables under consideration –like for instance making inferences from the means and correlations for fat tailed variables. This is the result of:

i) the focus in statistical education on Gaussian or thin-tailed variables,

ii) the absence of probabilistic knowledge combined with memorization of statistical terms,

iii) complete cluelessness about dimensionality,

all of which are prevalent among social scientists.

Example of pseudo-empiricism: comparing death from terrorist actions or epidemics such as ebola (fat tailed) to falls from ladders (thin tailed).

This confirmatory "positivism" is a disease of modern science; it breaks down under both dimensionality and fat-tailedness.

Actually one does not need to distinguish between fat tailed and Gaussian variables to get the lack of rigor in these activities: simple criteria of statistical significance are not met –not do these operators grasp the notion of such a concept as significance.

2.2.16 Preasymptotics

Mathematical statistics is largely concerned with what happens with $n = 1$ (where n is the number of summands) and $n = \infty$. What happens in between is what we call the real world –and the major focus of this book. Some distributions (say those with finite variance) are Gaussian in behavior asymptotically, for $n = \infty$, but not for extremely large but not infinite n.

2.2.17 Stochasticizing

Making a deterministic parameter stochastic, (i) in a simple way, or (ii) via a more complex continuous or discrete distribution.

(i) Let s be the deterministic parameter; we stochasticize (entry-level style) by creating a two-state Bernouilli with p probability of taking a value s_1, $1 - p$ of

taking value s_2. A transformation is mean-preserving when $ps_1 + (1-p)s_2 = s$, that is, preserves the mean. More generally, it can be in a similar manner variance preserving, etc.

(ii) We can use a full probability distribution, typically a Gaussian if the variable is two-tailed, and the Lognormal or the exponential if the variable is one-tailed (rarely a power law). When s is standard deviation, one can stochasticize s^2, where it becomes "stochastic volatility", with a variance or standard deviation typically dubbed "Vvol".

2.2.18 Value at Risk, Conditional VaR

The mathematical expression of the Value at Risk, VaR, for a random variable X with distribution function F and threshold $\lambda \in [0,1]$

$$\mathrm{VaR}_\lambda(X) = -\inf\{x \in \mathbb{R} : F(x) > \lambda\},$$

and the corresponding CVar or Expected Shortfall ES at threshold λ:

$$\mathrm{ES}_\lambda(X) = \mathbb{E}\left(-X \mid_{X \leq -\mathrm{VaR}_\lambda(X)}\right)$$

or, in the positive domain, by considering the tail for X instead of that of $-X$.

More generally the expected shortfall for threshold K is $\mathbb{E}(X|_{X>K})$.

2.2.19 Skin in the game

A filtering mechanism that forces cooks to eat their own cooking and be exposed to harm in the event of failure, thus throws dangerous people out of the system. Fields that have skin in the game: plumbing, dentistry, surgery, engineering, activities where operators are evaluated by tangible results or subjected to ruin and bankruptcy. Fields where people have no skin in the game: circular academic fields where people rely on peer assessment rather than survival pressures from reality.

2.2.20 MS Plot

The MS plot, "maximum to sum", allows us to see the behavior of the LLN for a given moment consider the contribution of the maximum observation to the total, and see how it behaves as n grows larger. For a r.v. X, an approach to detect if $\mathbb{E}(X^p)$ exists consists in examining convergence according to the law of large numbers (or, rather, absence of), by looking the behavior of higher moments in a given sample. One convenient approach is the Maximum-to-Sum plot, or MS plot as shown in Figure 10.3.

The MS Plot relies on a consequence of the law of large numbers [181] when it comes to the maximum of a variable. For a sequence $X_1, X_2, ..., X_n$ of nonnegative i.i.d. random variables, if for $p = 1, 2, 3, \dots$, $\mathbb{E}[X^p] < \infty$, then

$$R_n^p = M_n^p / S_n^p \rightarrow^{a.s.} 0$$

as $n \rightarrow \infty$, where $S_n^p = \sum_{i=1}^{n} X_i^p$ is the partial sum, and $M_n^p = \max(X_1^p, ..., X_n^p)$ the partial maximum. (Note that we can have X the absolute value of the random variable in case the r.v. can be negative to allow the approach to apply to odd moments.)

2.2.21 Maximum domain of attraction, MDA

The extreme value distribution concerns that of the maximum r.v., when $x \rightarrow x^*$, where $x^* = \sup\{x : F(x) < 1\}$ (the right "endpoint" of the distribution) is in the maximum domain of attraction, MDA [115]. In other words,

$$\max(X_1, X_2, \dots X_n) \xrightarrow{P} x^*.$$

2.2.22 Substitution of Integral in the psychology literature

The verbalistic literature makes the following conflation. Let $K \in \mathbb{R}^+$ be a threshold, $f(.)$ a density function and $p_K \in [0, 1]$ the probability of exceeding it, and $g(x)$ an impact function. Let I_1 be the expected payoff above K:

$$I_1 = \int_K^\infty g(x) f(x) \, \mathrm{d}x,$$

and Let I_2 be the impact at K multiplied by the probability of exceeding K:

$$I_2 = g(K) \int_K^\infty f(x) \, dx = g(K) p_K.$$

The substitution comes from conflating I_1 and I_2, which becomes an identity if and only if $g(.)$ is constant above K (say $g(x) = \theta_K(x)$, the Heaviside theta function). For $g(.)$ a variable function with positive first derivative, I_1 can be close to I_2 only under thin-tailed distributions, not under the fat tailed ones.

2.2.23 Inseparability of probability (another common error)

Let $F : \mathcal{A} \rightarrow [0, 1]$ be a probability distribution (with derivative f) and $g : \mathbb{R} \rightarrow \mathbb{R}$ a measurable function, the "payoff"". Clearly, for $\mathcal{A}'$ a subset of $\mathcal{A}$:

$$\int_{\mathcal{A}'} g(x)\mathrm{d}F(x) = \int_{\mathcal{A}'} f(x)g(x)\mathrm{d}x \neq \int_{\mathcal{A}'} f(x)\mathrm{d}x \; g\left(\int_{\mathcal{A}'} \mathrm{d}x\right)$$

In discrete terms, with $\pi(.)$ a probability mass function:

$$\sum_{x \in \mathcal{A}'} \pi(x)g(x) \neq \sum_{x \in \mathcal{A}'} \pi(x) \; g\left(\frac{1}{n}\sum_{x \in \mathcal{A}'} x\right) \tag{2.8}$$

= probability of event $\times$ payoff of average event

The general idea is that probability is a kernel into an equation not an end product by itself outside of explicit bets.

2.2.24 Wittgenstein's Ruler

"Wittgenstein's ruler" is the following riddle: are you using the ruler to measure the table or using the table to measure the ruler? Well, it depends on the results. Assume there are only two alternatives: a Gaussian distribution and a Power Law one. We show that a large deviation, say a "six sigma" indicates the distribution is power law.

2.2.25 Black swan

Basically, things that fall outside what you can model, and carry large consequences. The idea is to not predict them, but be convex (or at least not concave) to them. And no, it is not about fat tails; it is just that fat tails make them worse. It is hard to explain to modelers that we need to learn to work with things we have never seen before.

2.2.26 The empirical distribution is not empirical

The empirical distribution, or survival function $\widehat{F}(t)$ is as follows: Let $X_1, \dots X_n$ be independent, identically distributed real random variables with the common cumulative distribution function $F(t)$.

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \geq t},$$

where $\mathbb{1}_A$ is the indicator function.

By the Glivenko-Cantelli theorem, we have uniform convergence of the max norm to a specific distribution –the Kolmogorov-Smirnoff –regardless of the initial distribution. We have:

$$\sup_{t\in\mathbb{R}} \left|\widehat{F}_n(t) - F(t)\right| \xrightarrow{\text{as.}} 0; \tag{2.9}$$

this distribution-independent convergence concerns probabilities of course, not moments –a result this author has worked on and generalized for the "hidden moment" above the maximum.

We note the main result (further generalized by Donsker into a Brownian Bridge since we know the extremes are 0 and 1)

$$\sqrt{n}\left(\widehat{F}_n(t) - F(t)\right) \xrightarrow{D} \mathcal{N}\left(0, F(t)(1-F(t))\right) \tag{2.10}$$

"The empirical distribution is not empirical" means that since the empirical distributions are necessarily censured on the interval $[x_{min}, x_{max}]$, for fat tails this can carry huge consequences because we cannot analyze fat tails in probability space but in payoff space.

Further see the entry on **the hidden tail**

2.2.27 The hidden tail

Consider K_n the maximum of a sample of n independent identically distributed variables; $K_n = \max(X_1, X_2, \ldots, X_n)$. Let $\phi(.)$ be the density of the underlying distribution. We can decompose the moments in two parts, with the "hidden" moment above K_0.

$$\mathbb{E}(X^p) = \underbrace{\int_L^{K_n} x^p \phi(x)\, dx}_{\mu_{L,p}} + \underbrace{\int_{K_n}^{\infty} x^p \phi(x)\, dx}_{\mu_{K,p}}$$

where μ_L is the observed part of the distribution and μ_K the hidden one (above K). By Glivenko-Cantelli the distribution of $\mu_{K,0}$ should be independent of the initial distribution of X, but higher moments do not, hence there is a bit of a problem with Kolmogorov-Smirnoff-style tests.

2.2.28 Shadow moment

This is called in this book "plug-in" estimation. It is not done by measuring the directly observable sample mean which is biased under fat-tailed distributions, but by using maximum likelihood parameters, say the tail exponent α, and deriving the shadow mean or higher moments.

2.2.29 Tail dependence

Let X_1 and X_2 be two random variables not necessarily in the same distribution class. Let $F^{\leftarrow}(q)$ be the inverse CDF for probability q, that is $F^{\leftarrow}(q) = \inf\{x \in \mathbb{R} : F(x) \geq q\}$, λ_u the upper tail dependence is defined as

$$\lambda_u = \lim_{q \to 1} \mathrm{P}\left(X_2 > F_2^{\leftarrow}(q) | X_1 > F_1^{\leftarrow}(q)\right) \tag{2.11}$$

Likewise for the lower tail dependence index.

2.2.30 Metaprobability

Comparing two probability distributions via some tricks which includes stochasticizing parameters. Or stochasticize a parameter to get the distribution of a call price, a risk metric such as VaR (see entry), CVaR, etc., and check the robustness or convexity of the resulting distribution.

2.2.31 Dynamic hedging

The payoff of a European call option C on an underlying S with expiration time indexed at T should be replicated with the following stream of dynamic hedges, the limit of which can be seen here, between present time t and T:

$$\lim_{\Delta t \to 0} \left(\sum_{i=1}^{n=T/\Delta t} \frac{\partial C}{\partial S} |_{S=S_{t+(i-1)\Delta \mathrm{t}}, t=t+(i-1)\Delta \mathrm{t}, \left(S_{t+i\Delta \mathrm{t}} - S_{t+(i-1)\Delta \mathrm{t}}\right)} \right) \tag{2.12}$$

We break up the period into n increments Δt. Here the hedge ratio $\frac{\partial C}{\partial S}$ is computed as of time $t + (i\text{-}1)\ \Delta t$, but we get the nonanticipating difference between the price at the time the hedge was initiatied and the resulting price at $t + i\ \Delta t$.

This is supposed to make the payoff deterministic *at the limit of* $\Delta t \to 0$. In the Gaussian world, this would be an Ito-McKean integral.

We show where this replication is never possible in a fat-tailed environment, owing to presamptotics.

Part I

FAT TAILS AND THEIR EFFECTS, AN INTRODUCTION

3 A NON-TECHNICAL OVERVIEW - THE DARWIN COLLEGE LECTURE *,‡

Abyssus abyssum invocat
תהום אל־תהום כורא
Psalms

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

3.1 ON THE DIFFERENCE BETWEEN THIN AND THICK TAILS

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

Research and discussion chapter.

A shorter version of this chapter was given at Darwin College, Cambridge (UK) on January 27 2017, as part of the Darwin College Lecture Series on Extremes. The author extends the warmest thanks to D.J. Needham and Julius Weitzdörfer, as well as their invisible assistants who patiently and accurately transcribed the lecture into a coherent text. The author is also grateful towards Susan Pfannenschmidt and Ole Peters who corrected some mistakes. Jamil Baz prevailed upon me to add more commentary to the chapter to accommodate economists and econometricians who, one never knows, may eventually identify with some of it.

- In Mediocristan, when a sample under consideration gets large, no single observation can really modify the statistical properties.
- In Extremistan, the tails (the rare events) play a disproportionately large role in determining the properties.

Another way to view it:

Assume a large deviation X.

- In Mediocristan, the probability of sampling higher than X twice in a row is greater than sampling higher than $2X$ once.
- In Extremistan, the probability of sampling higher than $2X$ once is greater than the probability of sampling higher than X twice in a row.

Let us randomly select two people in Mediocristan; assume we obtain a (very unlikely) combined height of 4.1 meters – a tail event. According to the Gaussian distribution (or, rather its one-tailed siblings), the most likely combination of the two heights is 2.05 meters and 2.05 meters. Not 10 centimeters and 4 meters.

Simply, the probability of exceeding 3 sigmas is 0.00135. The probability of exceeding 6 sigmas, twice as much, is 9.86×10^{-10}. The probability of two 3-sigma events occurring is 1.8×10^{-6}. Therefore the probability of two 3-sigma events occurring is considerably higher than the probability of one single 6-sigma event. This is using a class of distribution that is not fat-tailed.

Figure 3.1 shows that as we extend the ratio from the probability of two 3-sigma events divided by the probability of a 6-sigma event, to the probability of two 4-sigma events divided by the probability of an 8-sigma event, i.e., the further we go into the tail, we see that a large deviation can only occur via a combination (a sum) of a large number of intermediate deviations: the right side of Figure 3.1. In other words, for something bad to happen, it needs to come from a series of very unlikely events, not a single one. This is the logic of Mediocristan.

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

This highlights the crisp distinction between the two domains; for the class of subexponential distributions, ruin is more likely to come from a single extreme event than from a series of bad episodes. This logic underpins classical risk theory as outlined by the actuary Filip Lundberg early in the 20^{th} Century [153] and formalized in the 1930s by Harald Cramer [50], but forgotten by economists in recent times. For insurability, losses need to be more likely to come from many events than a single one, thus allowing for diversification,

This indicates that insurance can only work in Mediocristan; you should never write an uncapped insurance contract if there is a risk of catastrophe. The point is called the catastrophe principle.

As we saw earlier, with thick tailed distributions, extreme events away from the centre of the distribution play a very large role. Black Swans are not "more fre-

quent" (as it is commonly misinterpreted), they are more consequential. The fattest tail distribution has just one very large extreme deviation, rather than many departures form the norm. Figure 4.4 shows that if we take a distribution such as the Gaussian and start fattening its tails, then the number of departures away from one standard deviation drops. The probability of an event staying within one standard deviation of the mean is 68 percent. As the tails fatten, to mimic what happens in financial markets for example, the probability of an event staying within one standard deviation of the mean rises to between 75 and 95 percent. So note that as we fatten the tails we get higher peaks, smaller shoulders, and a higher incidence of a very large deviation. Because probabilities need to add up to 1 (even in France) increasing mass in one area leads to decreasing it in another.

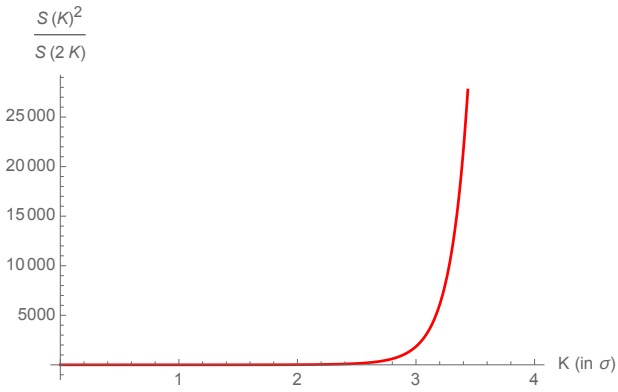


Figure 3.1: *Ratio of two occurrences of size K by one of 2K for a Gaussian distribution*. The larger the K, that is, the more we are in the tails, the more likely the event is to come from 2 independent realizations of K (hence $P(K)^2$, and the less likely from a single event of magnitude 2K.*
*This is fudging for pedagogical simplicity. The more rigorous approach would be to compare 2 occurrences of size K to 1 occurrence of size $2K$ plus 1 regular deviation – but the end graph would not change at all.

3.2 A (MORE ADVANCED) CATEGORIZATION AND ITS CONSEQUENCES

Let us now consider the degrees of thick tailedness in a casual way for now (we will get deeper and deeper later in the book). The ranking is by severity.

Distributions:

Thick Tailed ⊃ Subexponential ⊃ Power Law (Paretian)

First there are entry level thick tails. This is any distribution with fatter tails than the Gaussian i.e. with *more* observations within ±1 standard deviation than $\text{erf}\left(\frac{1}{\sqrt{2}}\right) \approx 68.2\%$[3] and with kurtosis (a function of the fourth central moment) higher than 3 [4].

Second, there are subexponential distributions satisfying our thought experiment earlier (the one illustrating the catastrophe principle). Unless they enter the class

3 The error function erf is the integral of the Gaussian distribution $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z dt e^{-t^2}$.

4 The moment of order p for a random variable X is the expectation of a p power of X, $\mathbb{E}(X^p)$.

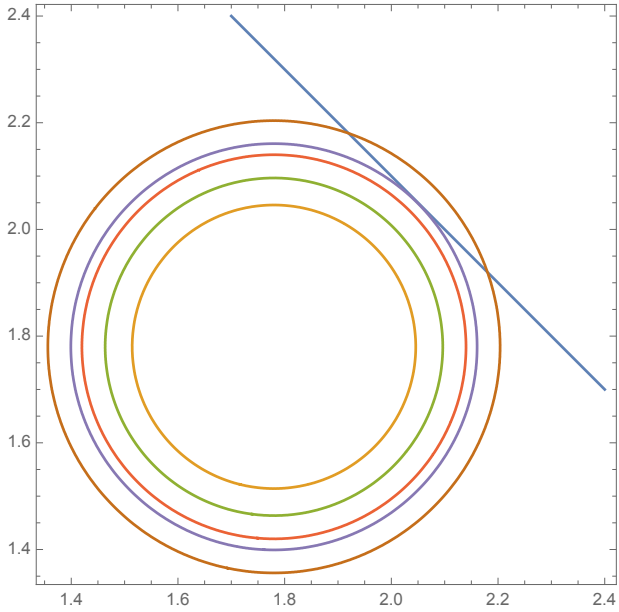


Figure 3.2: *Iso-densities for two independent Gaussian distributions. The line shows* $x + y = 4.1$. *Visibly the maximal probability is for* $x = y = 2.05$.

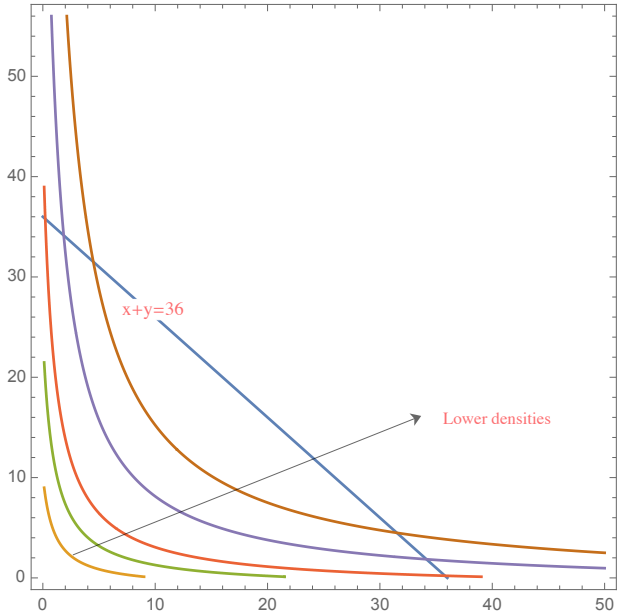


Figure 3.3: *Iso-densities for two independent thick tailed distributions (in the power law class). The line shows* $x + y = 36$. *Visibly the maximal probability is for either* $x = 36 - \epsilon$ *or* $y = 36 - \epsilon$, *with* ϵ *going to* 0 *as the sum* $x + y$ *becomes larger.*

of power laws, distributions are not really thick tailed because they do not have monstrous impacts from rare events. In other words, they can have all the moments

.

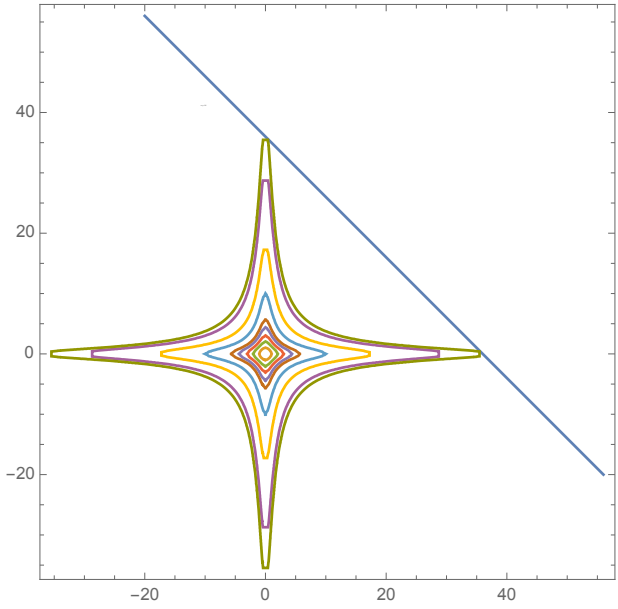


Figure 3.4: *Same representation as in Figure 3.1, but concerning power law distributions with support on the real line; we can see the isodensities looking more and more like a cross for lower and lower probabilities.*

Level three, what is called by a variety of names, power law, or member of the regular varying class, or the "Pareto tails" class; these correspond to real thick tails but the fattailedness depends on the parametrization of their tail index. Without getting into a tail index for now, consider that there will be *some* moment that is infinite, and moments higher than that one will also be infinite.

Let us now work our way from the bottom to the top of the central tableau in Figure 3.7. At the bottom left we have the degenerate distribution where there is only one possible outcome i.e. no randomness and no variation. Then, above it, there is the Bernoulli distribution which has two possible outcomes, not more. Then above it there are the two Gaussians. There is the natural Gaussian (with support on minus and plus infinity), and Gaussians that are reached by adding random walks (with compact support, sort of, unless we have infinite summands)[5]. These are completely different animals since one can deliver infinity and the other cannot (except asymptotically). Then above the Gaussians sit the distributions in the subexponential class that are not members of the power law class. These members have all moments. The subexponential class includes the lognormal, which is one of the strangest things in statistics because sometimes it cheats and fools us. At low variance, it is thin-tailed; at high variance, it behaves like the very thick tailed. Some people take it as good news that the data is not Paretian but lognormal; it is not necessarily so. Chapter 8 gets into the weird properties of the lognormal.

5 Compact support means the real-valued random variable X takes realizations in a bounded interval, say $[a, b]$,$(a, b]$, $[a, b)$, etc. The Gaussian has an exponential decline e^{-x^2} that accelerates with deviations, so some people such as Adrien Douady consider it effectively of compact support.

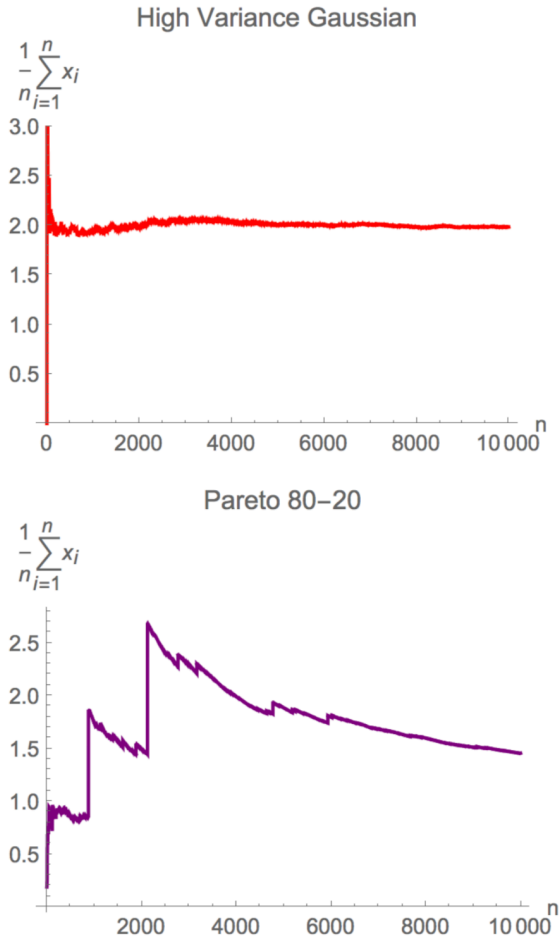


Figure 3.5: *The law of large numbers, that is how long it takes for the sample mean to stabilize, works much more slowly in Extremistan (here a Pareto distribution with 1.13 tail exponent , corresponding to the "Pareto 80-20". Both have the same mean absolute deviation. Note that the same applies to other forms of sampling, such as portfolio theory.*

Membership in the subexponential class does not satisfy the so-called Cramer condition , allowing insurability, as we illustrated in Figure 3.1, recall out thought experiment in the beginning of the chapter. More technically, the Cramer condition means that the expectation of the exponential of the random variable exists.[6]

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

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

where $\mathbb{E}$ is the expectation operator.

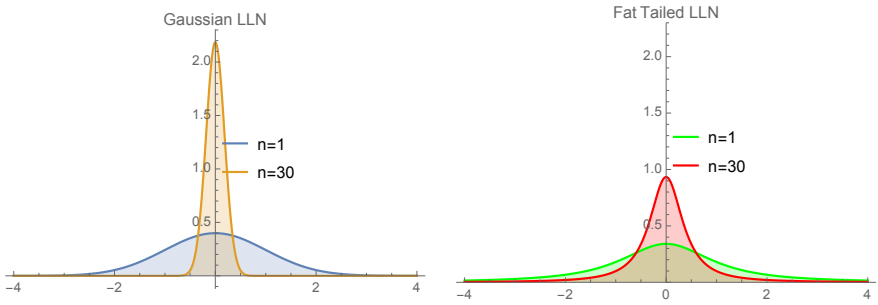


Figure 3.6: *What happens to the distribution of an average as the number of observations n increases? This is the same representation as in Fig. 3.5 seen in distribution/probability space. The fat tailed distribution does not compress as easily as the Gaussian. You need a much, much larger sample. It is what it is.*

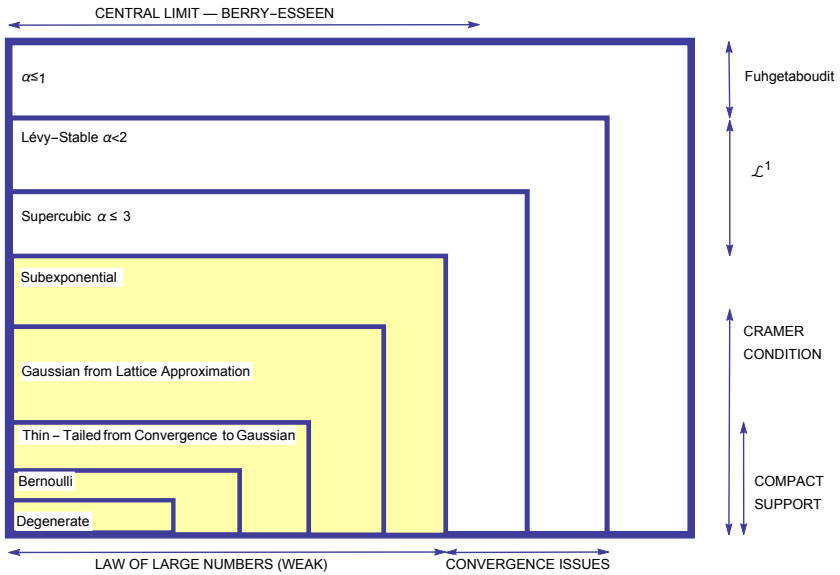


Figure 3.7: *The tableau of thick tails, along the various classifications for convergence purposes (i.e., convergence to the law of large numbers, etc.) and gravity of inferential problems. Power Laws are in white, the rest in yellow. See Embrechts et al [81].*

Once we leave the yellow zone, where the law of large numbers (LLN) largely works[7], and the central limit theorem (CLT) eventually ends up working[8], then

7 Take for now the following definition for the law of large numbers: it roughly states that if a distribution has a finite mean, and you add independent random variables drawn from it —that is, your sample gets

we encounter convergence problems. So here we have what are called power laws. We rank them by their tail index α, on which later; take for now that the lower the tail index, the fatter the tails. When the tail index is $\alpha \leq 3$ we call it supercubic ($\alpha = 3$ is cubic). That's an informal borderline: the distribution has no moment other than the first and second, meaning both the laws of large number and the central limit theorem apply in theory.

Then there is a class with $\alpha \leq 2$ we call the Levy-Stable to simplify (though it includes similar power law distributions with α less than 2 not explicitly in that class; but in theory, as we add add up variables, the sum ends up in that class rather than in the Gaussian one thanks to something called the generalized central limit theorem, GCLT). From here up we are increasingly in trouble because there is no variance. For $1 \leq \alpha \leq 2$ there is no variance, but mean absolute deviation (that is, the average variations taken in absolute value) exists.

Further up, in the top segment, there is no mean. We call it the Fuhgetaboudit. If you see something in that category, you go home and you don't talk about it.

The traditional statisticians approach to thick tails has been to claim to assume a different distribution but keep doing business as usual, using same metrics, tests, and statements of significance. Once we leave the yellow zone, for which statistical techniques were designed (even then), things no longer work as planned. The next section presents a dozen issues, almost all terminal. We will get a bit more technical and use some jargon.

3.3 THE MAIN CONSEQUENCES AND HOW THEY LINK TO THE BOOK

Here are some consequences of moving out of the yellow zone, the statistical comfort zone:

Consequence 1

The law of large numbers, when it works, works too slowly in the real world.

This is more shocking than you think as it cancels most statistical estimators. See Figure 3.5 in this chapter for an illustration. The subject is treated in Chapter 8 and distributions are classified accordingly.[9]

larger— you eventually converge to the mean. How quickly? that is the question and the topic of this book.

8 We will address ad nauseam the central limit theorem but here is the initial intuition. It states that n-summed independent random variables with finite second moment end up looking like a Gaussian distribution. Nice story, but how fast? Power laws on paper need an infinity of such summands, meaning they never really reach the Gaussian. Chapter 7 deals with the limiting distributions and answers the central question: "how fast?" both for CLT and LLN. How fast is a big deal because in the real world we have something different from n equals infinity.

9 What we call preasymptotics is the behavior of a sum or sequence when n is large but not infinite. This is (sort of) the focus of this book.

Summary of the problem with overstandardized statistics

STATISTICAL ESTIMATION is based on two elements: the central limit theorem (which is assumed to work for "large" sums, thus making about everything conveniently normal) and that of the law of large numbers, which reduces the variance of the estimation as one increases the sample size. However, things are not so simple; there are caveats. In Chapter 8, we show how sampling is distribution dependent, and varies greatly within the same class. As shown by Bouchaud and Potters in [27] and Sornette in [211], the tails for some finite variance but infinite higher moments can, under summation, converge to the Gaussian within $\pm\sqrt{n \log n}$, meaning the center of the distribution inside such band becomes Gaussian, but remote parts, those tails, don't –and the remote parts determine so much of the properties.

Life happens in the preasymptotics.

Sadly, in the entry on estimators in the monumental *Encyclopedia of Statistical Science* [145], W. Hoeffding writes:

> "The exact distribution of a statistic is usually highly complicated and difficult to work with. Hence the need to approximate the exact distribution by a distribution of a simpler form whose properties are more transparent. The limit theorems of probability theory provide an important tool for such approximations. In particular, the classical central limit theorems state that the sum of a large number of independent random variables is approximately normally distributed under general conditions. In fact, the normal distribution plays a dominating role among the possible limit distributions. To quote from Gnedenko and Kolmogorov's text [[110], Chap. 5]:
>
> > *"Whereas for the convergence of distribution functions of sums of independent variables to the normal law only restrictions of a very general kind, apart from that of being infinitesimal (or asymptotically constant), have to be imposed on the summands, for the convergence to another limit law some very special properties are required of the summands"*.
>
> Moreover, many statistics behave asymptotically like sums of independent random variables. All of this helps to explain the importance of the normal distribution as an asymptotic distribution."

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

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

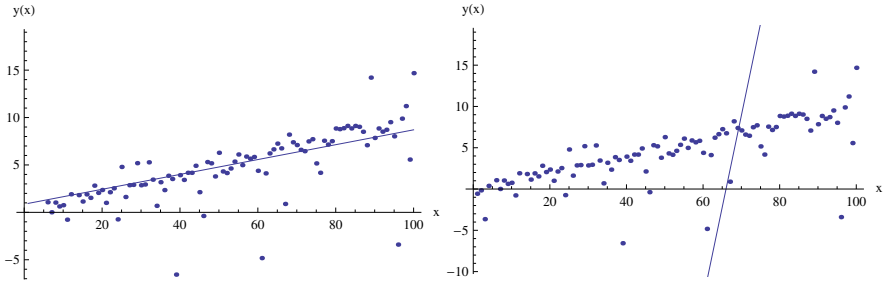


Figure 3.8: *In the presence of thick tails, we can fit markedly different regression lines to the same story (the Gauss-Markov theorem —necessary to allow for linear regression methods —doesn't apply anymore). Left: a regular (naïve) regression. Right: a regression line that tries to accommodate the large deviation —a "hedge ratio" so to speak, one that protects the agent from a large deviation, but mistracks small ones. Missing the largest deviation can be fatal. Note that the sample doesn't include the critical observation, but it has been guessed using "shadow mean" methods.*

Figure 3.9: *Inequality measures such as the Gini coefficient require completely different methods of estimation under thick tails, as we will see in Part III. Science is hard.*

Consequence 2

The mean of the distribution will rarely correspond to the sample mean; it will have a persistent small sample effect (downward or upward) particularly when the distribution is skewed (or one-tailed).

This is another problem of sample insufficiency. In fact, there is no very thick tailed- one tailed distribution in which the population mean can be properly estimated directly from the sample mean –rare events determine the mean, and these, *being rare*, take a lot of data to show up[10]. Consider that some power laws (like the one described as the "80/20" in common parlance have 92 percent of the observations falling below the true mean). For the sample average to be informative, we need orders of magnitude more data than we do (people in economics still do not understand this, though traders have an intuitive grasp of the point). The problem is discussed briefly further down in 3.7, and more formally in the "shadow mean" chapters, Chapters 15 and 16. Further, we will introduce the notion of hidden properties are in 3.7. Clearly by the same toke, variance will be likely to be underestimatwd.

Consequence 3

Metrics such as standard deviation and variance are not useable.

They fail out of sample –even when they exist; even when all moments exist. Discussed in ample details in Chapter 4.1. It is a scientific error that the notion of standard deviation (often mistaken for average deviation by its users) found its way as a measure of variation as it is very narrowly accurate in what it purports to do, in the best of circumstances.

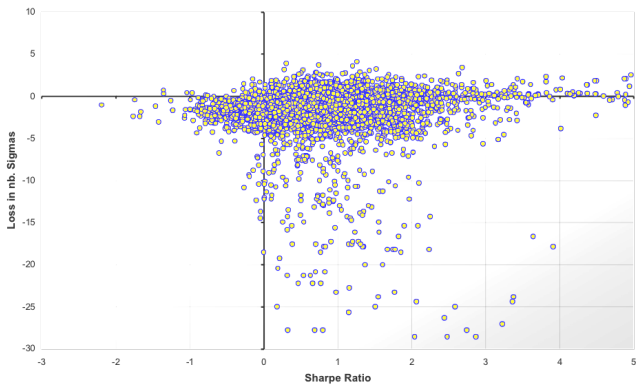


Figure 3.10: *We plot the Sharpe ratio of hedge funds on the horizontal axis as computed up to crisis of 2008 and their subsequent losses expressed in standard deviation during the crisis. Not only does the Sharpe ratio completely fail to predict out of sample performance, but, if anything, it can be seen as a weak predictor of failure. Courtesy Raphael Douady.*

Consequence 4

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

10 The population mean is the average if we sampled the entire population. The sample mean is, obviously, what we have in front of us. Sometimes, as with wealth or war casualties, we can have the entire population, yet the population mean isn't that of the sample. In these situations we use the concept of "shadow mean", which is the expectation as determined by the data generating process or mechanism.

This is a simple consequence of the previous point.[11] Either they require much more data, many more orders of magnitude, or some different model than the one being used, of which we are not yet aware. Figure 3.3 show the Sharpe ratio, supposed to predict performance, fails out of sample — it acts in exact reverse of the intention. Yet it is still used because people can be suckers for numbers.

> Practically every single economic variable and financial security is thick tailed. Of the 40,000 securities examined, not one appeared to be thin-tailed. This is the main source of failure in finance and economics.

Financial theorists claim something like that if the first two moments exist, then the so-called portfolio theory works, even if the distribution has thick tails (they add some conditions of ellipticality we will discuss later). The main problem is that even if variance exists, we don't know what it can be with acceptable precision; it obeys a slow law of large numbers because the second moment of a random variable is necessarily more thick tailed than the variable itself. Further, stochastic correlations or covariances also represent a form of thick tails, which invalidates these metrics.

> Practically any paper in economics using covariance matrices is suspicious.

Details are in Chapters 4 for the univariate case and 6.1 for multivariate situations.

> **Consequence 5**
>
> *Robust statistics is not robust and the empirical distribution is not empirical.*

The story of my life. Much like the Soviet official journal was named *Pravda* which means "truth" in Russian, almost as a joke, robust statistics are like a type of prank, except that most professionals aren't aware of it.

First, robust statistics shoots for measures that can handle tail events —large observations —without changing much. This the wrong idea of robustness: a metric that doesn't change in response of a tail event may be doing so precisely because it is uninformative. Further, these measures do not help with expected payoffs. Second, robust statistics are usually associated with a branch called "nonparametric" statistics, under the impression that the absence of parameters will make the analysis less distribution dependent. This book shows all across that it makes things worse.

The winsorization of the data, by removing outliers, distort the expectation operation and actually reduce information –though it would be a good idea to check if the outlier is real or a fake outlier we call in finance a "bad print" (some clerical error or computer glitch).

11 Roughly, Beta is a metric showing how much an asset A is expected to move in response to a move in the general market (or a given benchmark or index), expressed as the ratio of the covariance between A and the market over the variance of the market.
The Sharpe ratio expresses the average return (or excess return) of an asset or a strategy divided by its standard deviation.

The so-called (nonparametric) "empirical distribution" is not empirical at all (as it misrepresents the expected payoffs in the tails), as we will show in Chapter 10 —this is at least the case for the way it is used in finance and risk management. Take for now the following explanation: future maxima are poorly tracked by past data without some intelligent extrapolation.

Consider someone looking at building a flood protection system with levees. The naively obtained "empirical" distribution will show the worst past flood level, the past maxima. Any worse level will have zero probability (or so). But by definition, if it was a past maxima, it had to have exceeded what was a past maxima before it to become one, and the empirical distribution would have missed it. For thick tails, the difference between past maxima and future expected maxima is much larger than thin tails.

Consequence 6

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

See Figure 3.8 and the commentary. The logic behind the least-square minimization method is the Gauss-Markov theorem which explicitly requires a thin-tailed distribution to allow the line going through the data points to be unique. So either we need a lot, a lot of data to minimize the squared deviations (in other words, the Gauss-Markov theorem applies, but not for our preasymptotic situations as the real world has finite, not infinite data), or we can't because the second moment does not exist. In the latter case, if we minimize mean absolute deviations (MAD), as we see in 4.1, not only we may still be facing an insufficiency of data for proper convergence, but the deviation slope may not be unique.

We discuss the point in some details in 6.7 and show how thick tails produce an in-sample higher coefficient of determination (R^2) than the real one because of the small sample effect of thick tails. When variance is infinite, R^2 should be 0. But because samples are necessarily finite, it will show, deceivingly, higher numbers than 0. Effectively, to conclude, under thick tails, R^2 is useless, uninformative, and often (as with IQ studies) downright fraudulent.

Consequence 7

Maximum likelihood methods can work well for some parameters of the distribution (good news).

Take a power law. We may estimate a parameter for its shape, the tail exponent (for which we use the symbol α in this book[12]), which, adding some other parameter (the scale) connects us back to its mean considerably better than going about it directly by sampling the mean.

Example: The mean of a simple Pareto distribution with minimum value L and tail exponent α and PDF $\alpha L^{\alpha} x^{-\alpha-1}$ is $L\frac{\alpha}{\alpha-1}$, a function of α. So we can get it

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

from these two parameters, one of which may already be known. This is what we call "plug-in" estimator. One can estimate α with a low error with visual aid (or using maximum likelihood methods with low variance — it is inverse-gamma distributed), then get the mean. It beats the direct observation of the mean.

The logic is worth emphasizing:

> The tail exponent α captures, by extrapolation, the low-probability deviation not seen in the data, but that plays a disproportionately large share in determining the mean.

This generalized approach to estimators is also applied to Gini and other inequality estimators.

So we can produce more reliable (or at least less unreliable) estimators for, say, a function of the tail exponent in *some* situations. But, of course, not all.

Now a real-world question is warranted: what do we do when we do not have a reliable estimator? Better stay home. We must not expose ourselves to harm in the presence of fragility, but can still take risky decisions if we are bounded for maximum losses (Figure 3.3).

Consequence 8

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

From a controversy the author had with the cognitive linguist and science writer Steven Pinker: making pronouncements (and generating theories) from recent variations in data is not easily possible, unless one meets some standards of significance, which requires more data under thick tails (the same logic of the slow LLN). Stating "violence has dropped" because the number of people killed in wars has declined from the previous year or decade is not a scientific statement: a scientific claim distinguishes itself from an anecdote as it aims at affecting what happens out of sample, hence the concept of statistical significance.

Let us repeat that nonstatistically significant statements are not the realm of science. However, saying violence has risen upon a single observation may be a rigorously scientific claim. The practice of reading into descriptive statistics may be acceptable under thin tails (as sample sizes do not have to be large), but never so under thick tails, except, to repeat, in the presence of a large deviation.

Consequence 9

Principal component analysis (PCA) and factor analysis are likely to produce spurious factors and loads.

This point is a bit technical; it adapts the notion of sample insufficiency to large random vectors seen via the dimension reduction technique called principal component analysis (PCA) . The issue a higher dimensional version of our law of large

number complications. The story is best explained in Figure 3.26, which shows the accentuation of what is called the "Wigner Effect", from insufficiency of data for the PCA. Also, to be technical, note that the Marchenko-Pastur distribution is not applicable in the absence of a finite fourth moment (or, has been shown in [23], for tail exponent in excess of 4).[13]

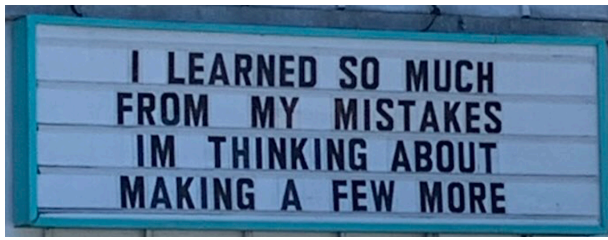


Figure 3.11: *Under thick tails (to the left), mistakes are terminal. Under thin tails (to the left) they can be great learning experiences. Source: You had one Job.*

Consequence 10

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

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

Simply, higher moments for thick tailed distributions are explosive. Particularly in economics.

Consequence 11

There is no such thing as a typical large deviation.

Conditional on having a "large" move, the magnitude of such a move is not defined, especially under serious thick tails (the Power Law tails class). This is associated with the catastrophe principle we saw earlier. In the Gaussian world, the expectation of a movement, conditional that the movement exceeds 4 standard deviations, is about 4 standard deviations. For a Power Law it will be a multiple of that. We call this the Lindy property and it is discussed in 5 and particularly in Chapter 11.2.

Consequence 12

The Gini coefficient ceases to be additive..

13 To be even more technical, principal components are independent when correlations are 0. However, for fat tailed distributions, as we will see more technically in 6.3.1, absence of correlation does not imply independence.

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

It is not just Gini, but other measures of concentration such as the top 1% owns x% of the total wealth, etc. The derivations are in Chapters 13 and 14.

Consequence 13

Large deviation theory fails to apply to thick tails. I mean, it really doesn't apply.

I really mean it doesn't apply. The methods behind the large deviation principle (Varadan [255] , Dembo and Zeituni [58], etc.) will be very useful in the thin-tailed world. And there only. See discussion and derivations in Appendix B as well as the limit theorem chapters, particularly Chapter 7.

Consequence 14

Option risks are never mitigated by dynamic hedging.

This might be technical and uninteresting for nonfinance people but the entire basis of financial hedging behind Black-Scholes rests on the possibility and necessity of dynamic hedging, both of which will be shown to be erroneous in Chapters 20 21 and 22. The required exponential decline of deviates away from the center requires the probability distribution to be outside the sub-exponential class. Again, we are talking about something related the Cramer condition –it all boils down to that exponential moment.

Recall the author has been an option trader and to option traders dynamic hedging is not the way prices are derived —and it has been so, as shown by Haug and the author, for centuries.

Consequence 15

Forecasting in frequency space diverges from expected payoff.

This point is explored in the next section here and in an entire chapter (Chapter 11.1): the foolish notion of focus on frequency rather than expectation can carry a mild effect under thin tails; not under thick tails. Figures 3.12 and 3.13 show the effect.

Consequence 16

Ruin problems are more acute and ergodicity is required under thick tails.

This is a bit technical but explained in the end of this chapter.

Let us discuss some of the points.

3.3.1 Forecasting

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

In the real world one is not paid in probability, but in dollars (or in survival, etc.). The fatter the tails, the more one needs to worry about payoff space – the saying goes: "payoff swamps probability" (see box). One can be wrong very frequently if the cost is low, so long as one is convex to payoff (i.e. make large gains when one is right). Further, one can be forecasting with 99.99% accuracy and still go bust (in fact, more likely to go bust: funds with impeccable track records were those that went bust during the 2008-2009 rout [14]). A point that may be technical for those outside quantitative finance: it is the difference between a vanilla option and a corresponding binary of the same strike, as discussed in *Dynamic Hedging* [222]: counterintuitively, thick tailedness *lowers* the value of the binary and raise that of the vanilla. This is expressed by the author's adage: "I've never seem a rich forecaster." We will examine in depth in 4.3.1 where we show that fattening the tails cause the probability of events higher than 1 standard deviations to drop –but the consequences to rise (in term of contribution to moments, say effect on the mean or other metrics).

Figure 11.1 shows the extent of the problem.

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

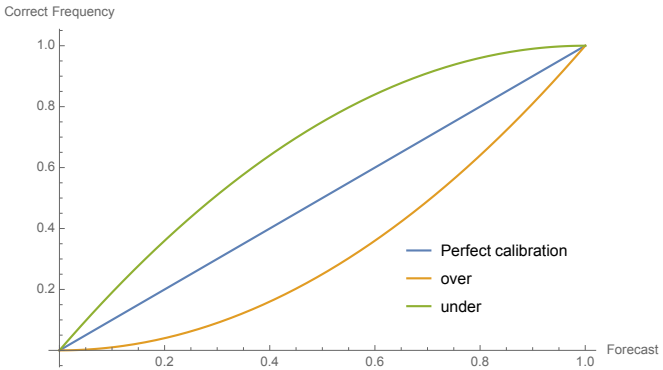


Figure 3.12: *Probabilistic calibration as seen in the psychology literature. The x axis shows the estimated probability produced by the forecaster, the y axis the actual realizations, so if a weather forecaster predicts* 30% *chance of rain, and rain occurs* 30% *of the time, they are deemed "calibrated". We hold that calibration in frequency (probability) space is an academic exercise (in the bad sense of the word) that mistracks real life outcomes outside narrow binary bets. It is particularly fallacious under thick tails.*

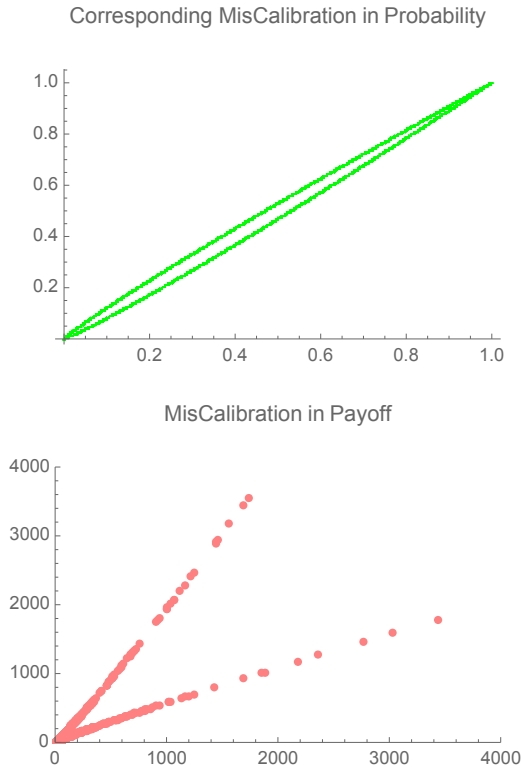


Figure 3.13: *How miscalibration in probability corresponds to miscalibration in payoff under power laws. The distribution under consideration is Pareto with tail index* $\alpha = 1.15$

Remark 1

Probabilistic forecast errors ("calibration") are in a different probability class from that true real-world P/L variations (or true payoffs).

"Calibration", which is a measure of how accurate one's predictions, lies in probability space –between 0 and 1. Any standard measure of such calibration will necessarily be thin-tailed (and, if anything, extra-thin tailed since it is bounded) – whether the random variable under such prediction is thick tailed or not. On the other hand, payoffs in the real world can be thick tailed, hence the distribution of such "calibration" will follow the property of the random variable.

We show full derivations and proofs in Chapter 11.

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

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

- For the first type, management consists in reducing the probability –the frequency – of such events. Remember that we count events and aim at reducing their counts.
- For the second type, it consists in reducing the effect should such an event take place. We do not count events, we measure impact.

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

But banks are only about money; consider that for wars we cannot afford the naive focus on event frequency without taking into account the magnitude, as done by the science writer Steven Pinker in [191], discussed in Chapter 16. This is without even examining the ruin problems (and nonergodicity) in the end of this section. More technically, one needs to meet the Cramer condition of non-subexponentiality for a tally of events (taken at face value) for raw probability to have any meaning at all. The plane analogy was proposed by the insightful Russ Robert during one of his econtalk podcasts with the author.

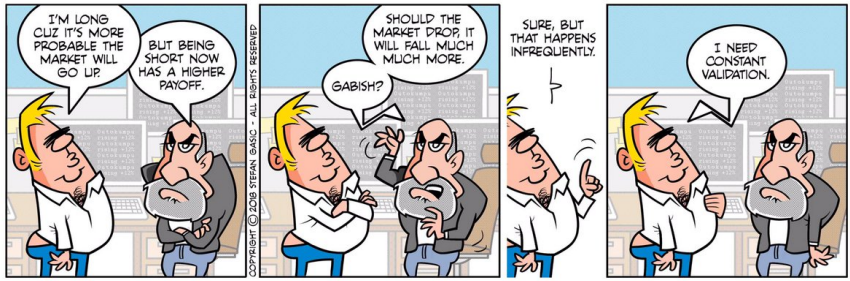


Figure 3.14: *Life is about payoffs, not forecasting, and the difference increases in Extremistan. (Why "Gabish" rather than "capisce"? Gabish is the recreated pronunciation of Siculo-Galabrez (Calabrese); the "p" used to sound like a "b" and the "g" like a Semitic kof, a hard K, from Punic. Much like capicoli is "gabagool".)*

3.3.2 The Law of Large Numbers

Let us now discuss the law of large numbers which is the basis of much of statistics. The law of large numbers tells us that as we add observations the mean becomes more stable, the rate being around $\sqrt{n}$. Figure 3.5 shows that it takes many more observations under a fat-tailed distribution (on the right hand side) for the mean to stabilize.

The "equivalence" is not straightforward.

One of the best known statistical phenomena is Pareto's 80/20 e.g. twenty percent of Italians own 80 percent of the land. Table 3.1 shows that while it takes 30 observations in the Gaussian to stabilize the mean up to a given level, it takes 10^{11} observations in the Pareto to bring the sample error down by the same amount (assuming the mean exists).

Despite this being trivial to compute, few people compute it. You cannot make claims about the stability of the sample mean with a thick tailed distribution. There are other ways to do this, but not from observations on the sample mean.

3.4 EPISTEMOLOGY AND INFERENTIAL ASYMMETRY

Table 3.1: *Corresponding n_α, or how many observations to get a drop in the error around the mean for an equivalent α-stable distribution (the measure is discussed in more details in Chapter 8). The Gaussian case is the $\alpha = 2$. For the case with equivalent tails to the 80/20 one needs at least 10^{11} more data than the Gaussian.*

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

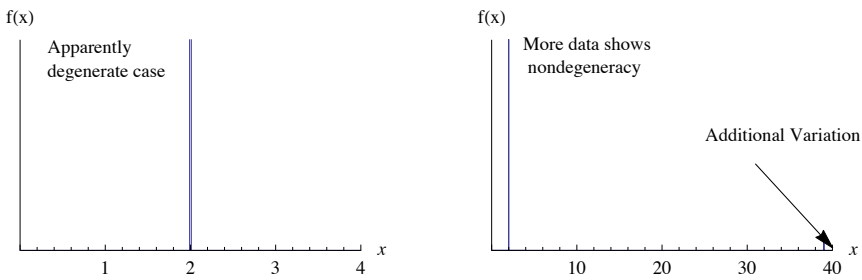


Figure 3.15: ***The Masquerade Problem (or Central Asymmetry in Inference).*** *To the left, a degenerate random variable taking seemingly constant values, with a histogram producing a Dirac stick. One cannot rule out nondegeneracy. But the right plot exhibits more than one realization. Here one can 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 can use the asymmetry to produce rigorous rules.*

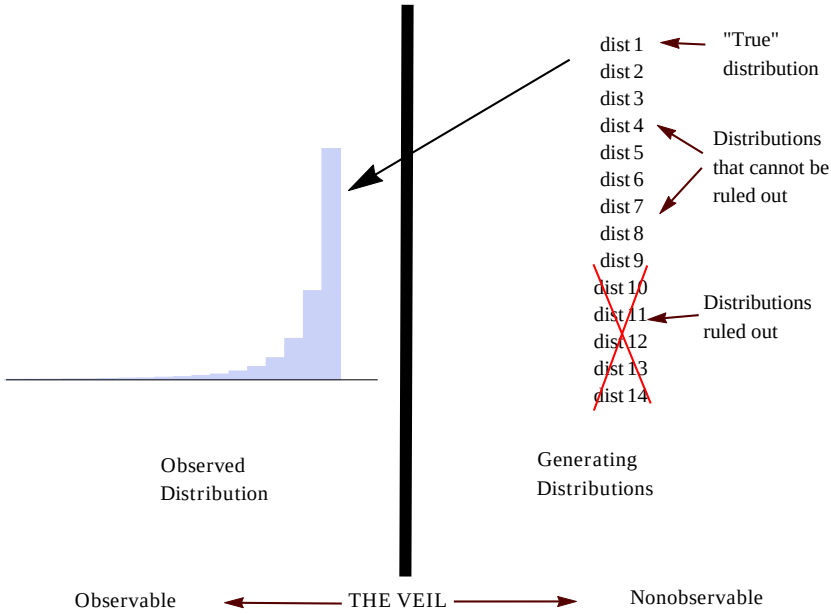


Figure 3.16: ***"The probabilistic veil"***. *Taleb and Pilpel [241] cover the point from an epistemological standpoint with the "veil" thought experiment by which an observer is supplied with data (generated by someone with "perfect statistical information", that is, producing it from a generator of 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.*

Principle 3.1 (Epistemology: the invisibility of the generator.)

- *We do not observe probability distributions, just realizations.*
- *A probability distribution cannot tell you if the realization belongs to it.*
- *You need a meta-probability distribution to discuss tail events (i.e., the conditional probability for the variable to belong to a certain distributions vs. others).*

Definition 3.1 (Asymmetry in distributions)
It is much easier for a criminal to fake being an honest person than for an honest person to fake being a criminal. Likewise it is easier for a fat-tailed distribution to fake being thin tailed than for a thin-tailed distributions to fake being thick tailed.

Let us now examine the epistemological consequences. Figure 3.15 illustrates the Masquerade Problem (or Central Asymmetry in Inference). On the left is a

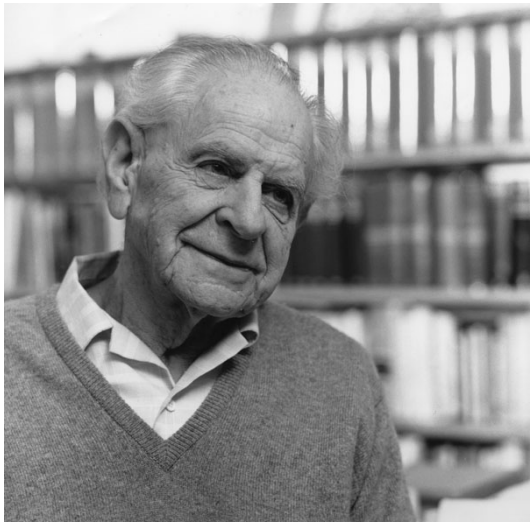


Figure 3.17: *Popper's solution of the problem of induction is in asymmetry: relying on confirmatory empiricism, that is focus on "ruling out" what fails to work, via negativa style. We extend this approach to statistical inference with the probabilistic veil by progressively ruling out entire classes of distributions.*

Scientific Rigor and Asymmetries by The Russian School of Probability

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

Members across three generations: P.L. Chebyshev, A.A. Markov, A.M. Lyapunov, S.N. Bernshtein (ie. Bernstein), E.E. Slutskii, N.V. Smirnov, L.N. Bol'shev, V.I. Romanovskii, A.N. Kolmogorov, Yu.V. Linnik, and the new generation: V. Petrov, A.N. Nagaev, A. 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 asymmetry to robustness is as follows. Is robust what does not produce variability across perturbation of parameters of the probability distribution. If there is change, but with an asymmetry, i.e. a concave or convex response to such perturbations, the classification is fragility and antifragility, respectively, see [220].

Figure 3.18: *The Problem of Induction. The philosophical problem of enumerative induction, expressed in the question:* ***"How many white swans do you need to count before ruling out the future occurrence of a black one?"*** *maps surprisingly perfectly to our problem of the working of the law of large numbers:* ***"How much data do you need before making a certain claim with an acceptable error rate?"*** *It turns out that the very nature of statistical inference reposes on a clear definition and quantitative measure of the inductive mechanism. It happens that, under thick tails, we need considerably more data; as we will see in Chapters 7 and 8 there is a way to gauge the relative speed of the inductive mechanism, even if ultimately the problem of induction cannot be perfectly solved. The problem said of induction is generally misattributed to Hume, [224] .*

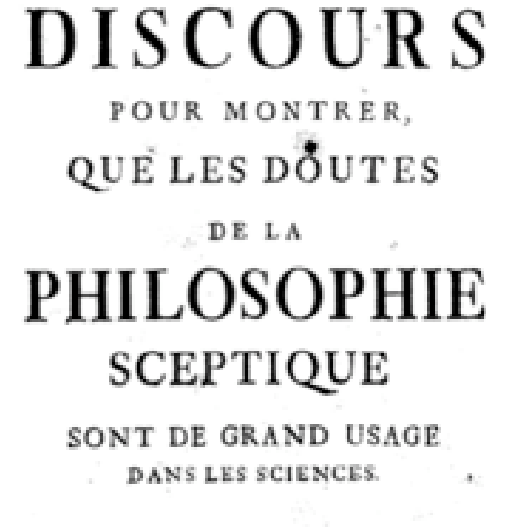

DISCOURS
POUR MONTRER,
QUE LES DOUTES
DE LA
PHILOSOPHIE
SCEPTIQUE
SONT DE GRAND USAGE
DANS LES SCIENCES.

Figure 3.19: *A Discourse to Show that Skeptical Philosophy is of Great Use in Science by François de La Mothe Le Vayer (1588-1672), apparently Bishop Huet's source. Every time I find a IJoriginal thinkerÌ who figured out the skeptical solution to the Black Swan problem, it turns out that he may just be cribbing a predecessor –not maliciously, but we forget to dig to the roots. As we insist, "Hume's problem" has little to do with Hume, who carried the heavy multi-volume Dictionary of Pierre Bayle (his predecessors) across Europe. I thought it was Huet who was as one digs, new predecessors crop up*

.

degenerate random variable taking seemingly constant values with a histogram producing a Dirac stick.

We have known at least since Sextus Empiricus that we cannot rule out degeneracy but there are situations in which we can rule out non-degeneracy. If I see a distribution that has no randomness, I cannot say it is not random. That is, we

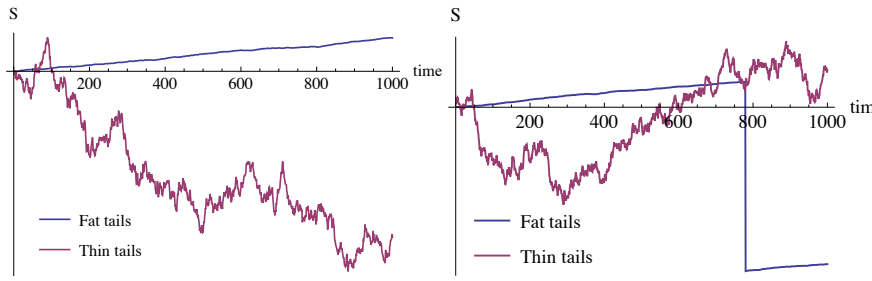


Figure 3.20: *It is not possible to "accept" thin tails, very easy to reject thintailedness. One distribution can produce jumps and quiet days will not help rule out their occurrence.*

cannot say there are no Black Swans . Let us now add one observation. I can now see it is random, and I can rule out degeneracy. I can say it is not "not random". On the right hand side we have seen a Black Swan , therefore the statement that there are no Black Swans is wrong. This is the negative empiricism that underpins Western science. As we gather information, we can rule things out. The distribution on the right can hide as the distribution on the left, but the distribution on the right cannot hide as the distribution on the left (check). This gives us a very easy way to deal with randomness. Figure 3.16 generalizes the problem to how we can eliminate distributions.

If we see a 20 sigma event, we can rule out that the distribution is thin-tailed. If we see no large deviation, we can not rule out that it is not thick tailed unless we understand the process very well. This is how we can rank distributions. If we reconsider Figure 3.7 we can start seeing deviations and ruling out progressively from the bottom. These ranks are based on how distributions can deliver tail events. Ranking distributions (by order or priority for the sake of inference) becomes very simple. Consider the logic: if someone tells you there is a ten-sigma event, it is much more likely that they have the wrong distribution than it is that you really have ten-sigma event (we will refine the argument later in this chapter). Likewise, as we saw, thick tailed distributions do not deliver a lot of deviation from the mean. But once in a while you get a big deviation. So we can now rule out what is not mediocristan. We can rule out where we are not Ş we can rule out mediocristan. I can say this distribution is thick tailed by elimination. But I can not certify that it is thin tailed. This is the Black Swan problem.

Application of the Maquerade Problem: Argentina's stock market before and after Aug 12, 2019 For an illustration of the asymmetry of inference applied to parameters of a distribution, or how a distribution can masquerade as having thinner tails than it actually has, consider what we knew about the Argentinian market before and after the large drop of Aug 12, 2019 (shown in Fig. 3.21). Using this reasoning, any future parameter uncertainty should make tails fatter, not thinner. Rafal Weron, in [259], showed how we are more likely to overestimate the tail index when fitting a stable distribution (lower means fatter tails).

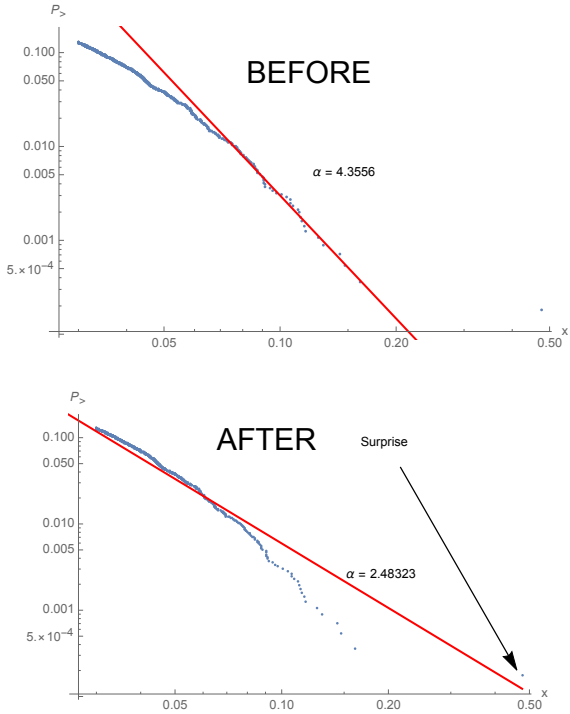


Figure 3.21: *A single day reveals the true tails of a distribution. Argentina's stock market before and after Aug 12, 2019. You may suddenly revise the tails as thicker (lower parameter α), never the reverse –it would take a long, long time for that to happen. Data obtained thanks to Diego Zviovich.*

3.5 NAIVE EMPIRICISM: EBOLA SHOULD NOT BE COMPARED TO FALLS FROM LADDERS

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

Principle 3.2
Thou shalt not compare a multiplicative fat-tailed process in Extremistan in the subexponential class to a thin-tailed process, particularly one that has Chernoff bounds from Mediocristan.

This is a simple consequence of the catastrophe principle we saw earlier, as illustrated in Figure 3.1. It is naïve empiricism to compare these processes, to suggest that we worry too much about ebola and too little about diabetes. In fact it is the other way round. We worry too much about diabetes and too little about ebola and other ailments with multiplicative effects. This is an error of reasoning that comes from not understanding thick tails –sadly it is more and more common. What is

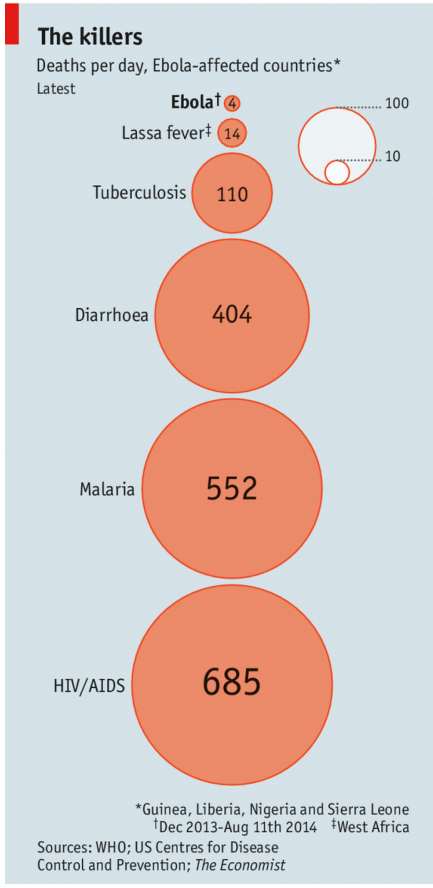


Figure 3.22: *Naive empiricism: never compare thick tailed variables to thin tailed ones, since the means do not belong to the same class of distributions. This is a generalized mistake made by The Economist, but very common in the so-called learned discourse. Even the Royal Statistical Society fell for it once they hired a "risk communication" person with a sociology or journalism background to run it.*

worse, such errors of reasoning are promoted by **empirical psychology** which does not appear to be empirical. It is also used by shills for industry passing for "risk communicators" selling us pesticides and telling us not to worry because harm appears to be minimal in past data (see Figure).

The correct reasoning is generally absent in decision theory and risk circles outside of the branches of extreme value theory and the works of the ABC group in Berlin's Max Planck directed by Gerd Gigerenzer [107] which tells you that your grandmother's instincts and teachings are not to be ignored and, when her recommendations clash with psychologists and decision theorists, it is usually the psychologists and decision theorists who are unrigorous. A simple look at the summary by "most cited author" Baruch Fishhoff's in *Risk: a Very Short Introduction*[92] shows no effort to disentangle the two classes of distribution. The problem linked to the "risk calibration" and "probabilistic calibration" misunderstood by psychologists and discussed more technically in Chapter 11 discussing expert calibration under thick tails.

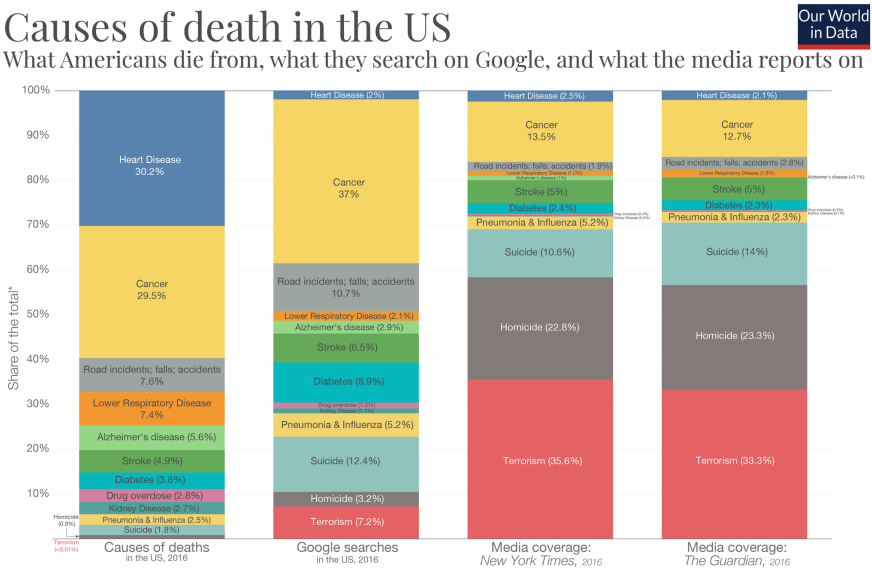


Figure 3.23: *Bill Gates's Naive (Non-Statistical) Empiricism: the founder of Microsoft*[1] *is promoting and financing the development of the above graph, yet at the same time claiming that the climate is causing an existential risk, not realizing that his arguments conflict since existential risks are necessarily absent in past data. Furthermore, a closer reading of the graphs shows that cancer, heart disease, and Alzheimer, being ailments of age, do not require the attention on the part of young adults and middle-aged people something terrorism and epidemics warrant.*

Another logical flaw is that terrorism is precisely low because of the attention it commands. Relax your vigilance and it may go out of control. The same applies to homicide: fears lead to safety.

If this map shows something, it is the rationality of common people with a good tail risk detector, compared to the ignorance of "experts". People are more calibrated to consequences and properties of distributions than psychologists claim.

[1] *Microsoft is a technology company still in existence at the time of writing.*

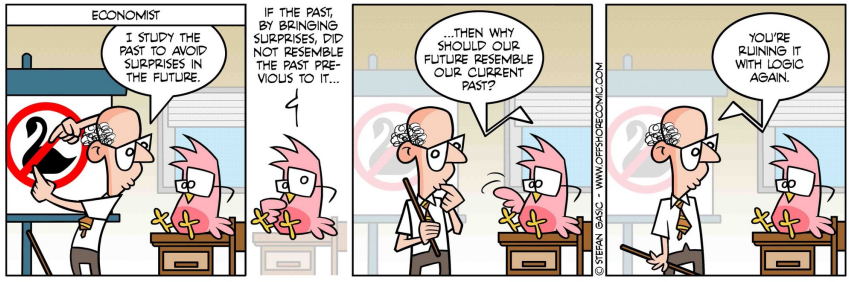


Figure 3.24: *Because of the slowness of the law of large numbers, under thick tails, the past's past doesn't resemble the past's future; accordingly, today's past will not resemble today's future. Things are easier under thin tails. Credit Stefan Gasic.*

Figure 3.25: *Beware the lobbyist using pseudo-empirical arguments. "Risk communications" shills such as the fellow here, with a journalism background, are hired by firms such as Monsanto (and cars and Tobacco companies) to engage in smear campaigns on their behalf using "science", "empirical arguments" and "evidence", and downplay "public fears" they deem irrational. Lobbying organizations penetrate such centers as "Harvard Center for Risk Analysis" with a fancy scholarly name that helps convince the layperson. The shills' line of argument, commonly, revolves around "no evidence of harm" and "rationality". Other journalists, in turn, espouse such arguments owing to their ability to sway the statistically naive. Probabilistic and risk literacy, statistical knowledge and journalism have suffered greatly from the spreading of misconceptions by nonscientists, or, worse, non-statisticians.*

3.6 PRIMER ON POWER LAWS (ALMOST WITHOUT MATHEMATICS)

Let us now discuss the intuition behind the Pareto Law. It is simply defined as: say X is a random variable. For x sufficiently large, the probability of exceeding $2x$ divided by the probability of exceeding x is no different from the probability of exceeding $4x$ divided by the probability of exceeding $2x$, and so forth. This property is called "scalability".[15]

Table 3.2: *An example of a power law*

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

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

15 To put some minimum mathematics: let X be a random variable belonging to the class of distributions with a "power law" right tail:

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{3.1}$$

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

Table 3.3: *Kurtosis from a single observation for financial data* $\frac{Max\left(X^4_{t-\Delta t_i}\right)_{i=0}^{n}}{\sum_{i=0}^{n} X^4_{t-\Delta t_i}}$

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

butions. But because it has no mean we have to ditch the statistical textbooks and do something more solid, more rigorous, even if it seems less mathematical.

A Pareto distribution has no higher moments: moments either do not exist or become statistically more and more unstable. So next we move on to a problem with economics and econometrics. In 2009 I took 55 years of data and looked at how much of the kurtosis (a function of the fourth moment) came from the largest observation –see Table 3.3. For a Gaussian the maximum contribution over the same time span should be around .008 ± .0028. For the S&P 500 it was about 80 percent. This tells us that we don't know anything about the kurtosis of these securities. Its sample error is huge; or it may not exist so the measurement is heavily sample dependent. If we don't know anything about the fourth moment, we know nothing about the stability of the second moment. It means we are not in a class of distribution that allows us to work with the variance, even if it exists. Science is hard; quantitative finance is hard too.

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

Table 3.3, for financial data, debunks all the college textbooks we are currently using. A lot of econometrics that deals with squares goes out of the window. This explains why economists cannot forecast what is going on –they are using the wrong methods and building the wrong confidence intervals. It will work within

the sample, but it will not work outside the sample –and samples are by definition finite and will always have finite moments. If we say that variance (or kurtosis) is infinite, we are not going to observe anything that is infinite within a sample.

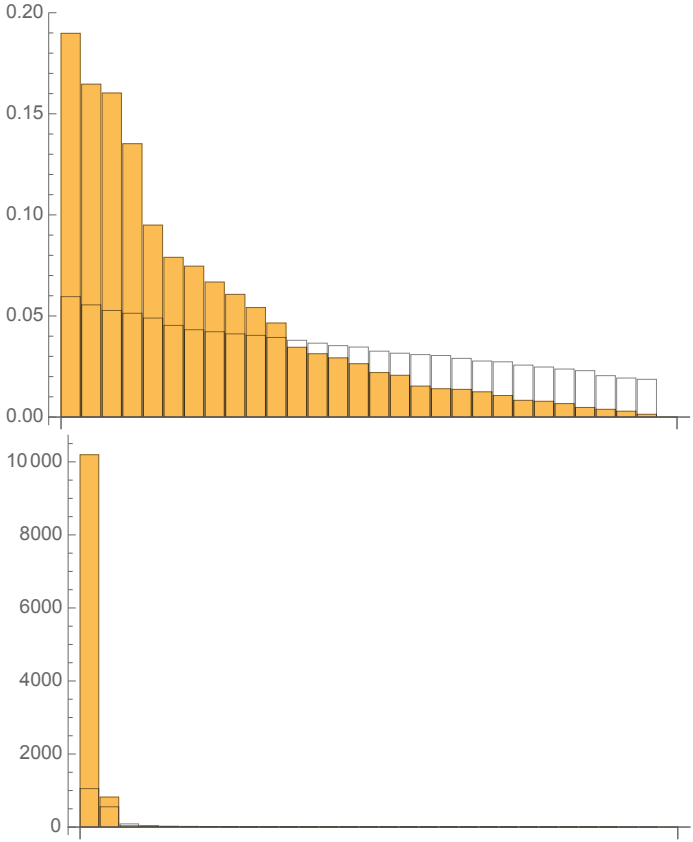


Figure 3.26: ***Spurious PCAs Under Thick Tails:*** *A Monte Carlo experiment that shows how spurious correlations and covariances are more acute under thick tails. Principal Components ranked by variance for 30 Gaussian uncorrelated variables, n=100 (above) and 1000 data points, and principal components ranked by variance for 30 Stable Distributed (with tail* $\alpha = \frac{3}{2}$ *, symmetry* $\beta = 1$*, centrality* $\mu = 0$*, scale* $\sigma = 1$*) (below). Both are "uncorrelated" identically distributed variables. We can see the "flatter" PCA structure with the Gaussian as n increases (the difference between PCAs shrinks). Such flattening does not occur in reasonable time under fatter tails.*

Principal component analysis, PCA (see Figure 3.26) is a dimension reduction method for big data and it works beautifully with thin tails (at least sometimes). But if there is not enough data there is an illusion of what the structure is. As we increase the data (the n variables), the structure becomes flat (something called in some circles the "Wigner effect" for random matrices, after Eugene Wigner — do not confuse with Wigner's discoveries about the dislocation of atoms under radiation). In the simulation, the data that has absolutely no structure: principal components (PCs) should be all equal (asymptotically, as data becomes large); but

the small sample effect causes the ordered PCs to show a declining slope. We have zero correlation on the matrix. For a thick tailed distribution (the lower section), we need a lot more data for the spurious correlation to wash out i.e., dimension reduction does not work with thick tails.

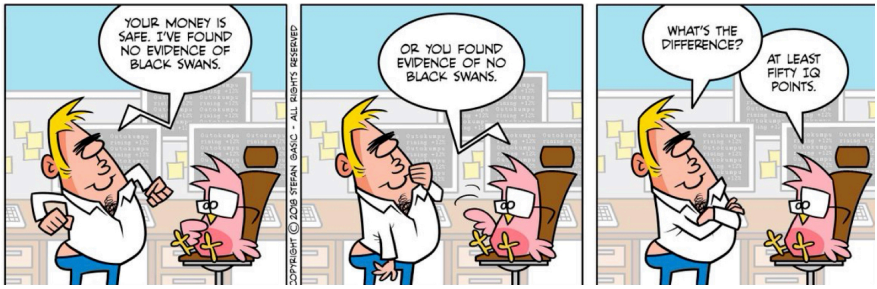


Figure 3.27: ***A central asymmetry****: the difference between absence of evidence and evidence of absence is compounded by thick tails. It requires a more elaborate understanding of random events —or a more naturalistic one. (Please do not attribute IQ points here as equivalent to the ones used in common psychometrics, as the suspicion is high scoring people fail to get the asymmetry.) Courtesy Stefan Gasic.*

3.7 WHERE ARE THE HIDDEN PROPERTIES?

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

See Figure 3.28 for the intuition: if it is thick tailed and we look at the sample mean, we observe fewer tail events. The common mistake is to think that we can naively derive the mean in the presence of one-tailed distributions. But there are unseen rare events and with time these will fill in. But by definition, they are low probability events. The trick is to estimate the distribution and then derive the mean. This is called in this book "plug-in" estimation, see Table 3.4. It is not done by measuring the directly observable sample mean which is biased under fat-tailed distributions. This is why, outside a crisis, banks seem to make large profits. Then once in a while they lose everything and more and have to be bailed out by the taxpayer. The way we handle this is by differentiating the true mean (which I call "shadow") from the realized mean, as in the Tableau in Table 3.4.

We can also do that for the Gini coefficient to estimate the "shadow" one rather than the naively observed one.

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

WITTGENSTEIN'S RULER: WAS IT REALLY A "10 SIGMA EVENT"?

In the summer of 1998, the hedge fund called "Long Term Capital Management" (LTCM) proved to have a very short life; it went bust from some deviations in the markets –those "of an unexpected nature". The loss was a *yuuuge deal* because two of the partners received the Swedish Riksbank Prize, marketed as the "Nobel" in economics. More significantly, the fund harbored a large number of finance professors; LTCM had imitators among professors (at least sixty finance PhDs blew up during that period from trades similar to LTCM's, and owing to risk management methods that were identical). At least two of the partners made the statement that it was a "10 sigma" event (10 standard deviations), hence they should be absolved of all accusations of incompetence (I was first hand witness of two such statements).

Let us apply what the author calls "Wittgenstein's ruler": are you using the ruler to measure the table or using the table to measure the ruler?

Assume to simplify there are only two alternatives: a Gaussian distribution and a Power Law one. For the Gaussian, the "event" we define as the survival function of 10 standard deviations is 1 in 1.31×10^{-23}. For the Power law of the same scale, a student T distribution with tail exponent 2, the survival function is 1 in 203.

What is the probability of the data being Gaussian conditional on a 10 sigma event, compared to that alternative?

We start with Bayes' rule. $\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}$. Replace $\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\overline{A})\mathbb{P}(B|\overline{A})$ and apply to our case.

$$P(\text{Gaussian}|\text{Event}) = \frac{\mathbb{P}(\text{Gaussian})P(\text{Event}|\text{Gaussian})}{(1-\mathbb{P}(\text{Gaussian}))P(\text{Event}|\text{NonGaussian}) + \mathbb{P}(\text{Gaussian})P(\text{Event}|\text{Gaussian})}$$

$\mathbb{P}(\text{Gaussian})$	$P(\text{Gaussian}\|\text{Event})$
0.5	2×10^{-21}
0.999	2×10^{-18}
0.9999	2×10^{-17}
0.99999	2×10^{-16}
0.999999	2×10^{-15}
1	1

Moral: If there is a tiny probability, $< 10^{-10}$ that the data might not be Gaussian, one can firmly reject Gaussianity in favor of the thick tailed distribution. The heuristic is to reject Gaussianity in the presence of any event > 4 or > 5 STDs –we will see throughout the book why patches such as conditional variance are inadequate and can be downright fraudulent.[a]

[a] The great Benoit Mandelbrot used to be extremely critical of methods that relied on a Gaussian and added jumps or other ad hoc trick to explain what happened in the data (say Merton's jump diffusion process [171]) –one can always fit back jumps ex post. He used to cite the saying attributed to John von Neumann: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

from a broader phenomenon (the past is in sample; inference is what works out of sample).

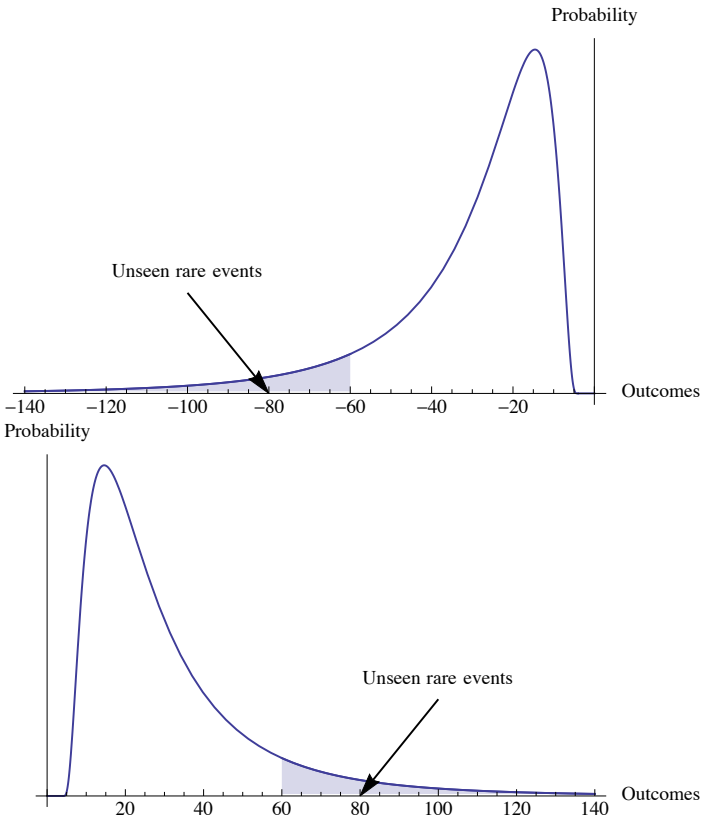


Figure 3.28: *Shadow Mean at work: Below: Inverse Turkey Problem – The unseen rare event is positive. When you look at a positively skewed (antifragile) time series and make (nonparametric) inferences about the unseen, you miss the good stuff an underestimate the benefits. Above: The opposite problem. The filled area corresponds to what we do not tend to see in small samples, from insufficiency of data points. Interestingly the shaded area increases with model error.*

Once we have figured out the distribution, we can estimate the statistical mean. This works much better than directly measuring the sample mean. For a Pareto distribution, for instance, 98% of observations are below the mean. There is a bias in the observed mean. But once we know that we have a Pareto distribution, we should ignore the sample mean and look elsewhere. Chapters 13 and 15 discuss the techniques.

Note that the field of Extreme Value Theory [114] [81] [115] focuses on tail properties, not the mean or statistical inference.

Table 3.4: *Shadow mean, sample mean, maximum likelihood mean (from plug-in estimators) and their ratio for different minimum thresholds. In bold the values for the 145k threshold. Rescaled data. From Cirillo and Taleb [46]. Details are explained in Chapters 16 and 13.*

L	Sample Mean	ML Mean	Ratio
10K	9.079×10^6	3.11×10^7	3.43
25K	9.82×10^6	3.62×10^7	3.69
50K	1.12×10^7	4.11×10^7	3.67
100K	1.34×10^7	4.74×10^7	3.53
200K	1.66×10^7	6.31×10^7	3.79
500K	2.48×10^7	8.26×10^7	3.31

3.8 BAYESIAN SCHMAYESIAN

In the absence of reliable information, Bayesian methods can be of little help. This author has faced since the publication of *The Black Swan* numerous questions concerning the use of something vaguely Bayesian to solve problems about the unknown under thick tails. Since one cannot manufacture information beyond what's available, no technique, Bayesian nor Schmayesian can help. The key is that one needs a reliable prior, something not readily observable (see Diaconis and Friedman [65] for the difficulty for an agent in formulating a prior).

A problem is the speed of updating, as we will cover in Chapter 7, which is highly distribution dependent. The mistake in the rational expectation literature is to believe that two observers supplied with the same information would necessarily converge to the same view. Unfortunately, the conditions for that to happen in real time or to happen at all are quite specific.

One of course can use Bayesian methods (under adequate priors) for the estimation of parameters if 1) one has a clear idea about the range of values (say from universality classes or other stable basins) and 2) these parameters follow a tractable distribution with low variance such as, say, the tail exponent of a Pareto distribution (which is inverse-gamma distributed), [11].

MORAL HAZARD AND RENT SEEKING in financial education: One of the most depressing experience this author had was when teaching a course on Fat Tails at the University of Massachusetts Amherst, at the business school, during a very brief stint there. One PhD student in finance bluntly said that he liked the ideas but that a financial education career commanded "the highest salary in the land" (that is, among all other specialties in education). He preferred to use Markowitz methods (even if they failed in fat-tailed domains) as these were used by other professors, hence allowed him to get his papers published, and get a high paying job.
I was disgusted, but predicted he would subsequently have a very successful career writing non-papers. He did.

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

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

Practitioners and risk takers often observe the following disconnect: people (non-practitioners) talking X (with the implication that practitioners should care about X in running their affairs) while practitioners think about $F(X)$, nothing but $F(X)$. And the straight confusion since Aristotle between X and $F(X)$ has been chronic as discussed in *Antifragile* [227] which is written around that theme. Sometimes people mention $F(X)$ as utility but miss the full payoff. And the confusion is at two level: one, simple confusion; second, in the decision-science literature, seeing the difference and not realizing that action on $F(X)$ is easier than action on X.

- The variable X can be unemployment in Senegal, $F_1(X)$ is the effect on the bottom line of the IMF, and $F_2(X)$ is the effect on your grandmother (which I assume is minimal).
- X can be a stock price, but you own an option on it, so $F(X)$ is your exposure an option value for X, or, even more complicated the utility of the exposure to the option value.
- X can be changes in wealth, $F(X)$ the convex-concave way it affects your well-being. One can see that $F(X)$ is vastly more stable or robust than X (it has thinner tails).

Convex vs. linear functions of a variable X Consider Fig. 3.30; 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 inequality. But beyond Jensen's inequality, the difference in risks between the two will be more and more considerable. When it comes to probability, the more nonlinear F, the less the probabilities of X matters compared to that of F. Moral of the story: focus on F, which we can alter, rather than on the measurement of the elusive properties of X.

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

The danger with the treatment of the Black Swan problem is as follows: people focus on X ("predicting X"). My point is that, although we do not understand X, we can deal with it by working on F which we can understand, while others work on predicting X which we can't because small probabilities are incomputable, particularly in thick 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 distribu-

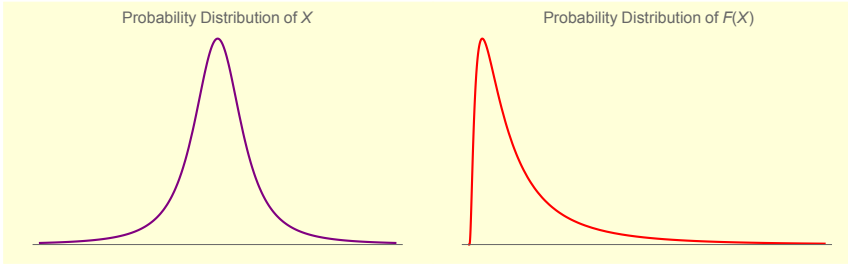


Figure 3.29: *The Conflation Problem X (random variable) and F(X) a function of it (or payoff). If F(X) is convex we don't need to know much about it –it becomes an academic problem. And it is safer to focus on transforming F(X) than X.*

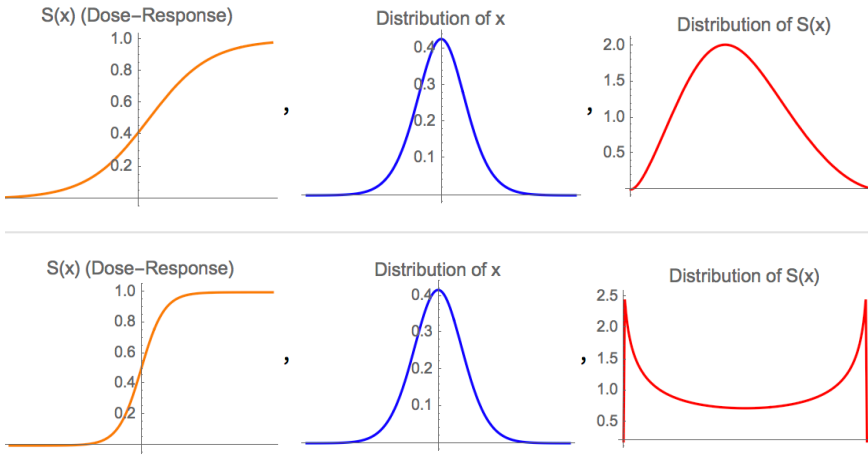


Figure 3.30: *The Conflation Problem: a convex-concave transformation of a thick tailed X produces a thin tailed distribution (above). A sigmoidal transformation (below) that is bounded on a distribution in* $(-\infty, \infty)$ *produces an ArcSine distribution, with compact support.*

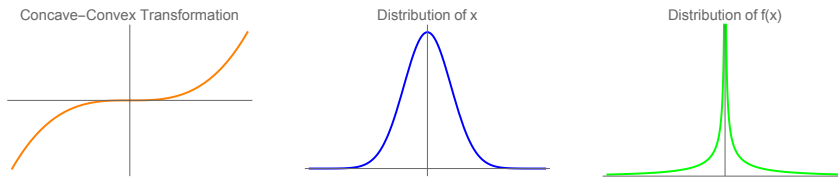


Figure 3.31: *A concave-convex transformation (of the style of a probit –an inverse CDF for the Gaussian– or of a logit) makes the tails of the distribution of* $f(x)$ *thicker*

tion of X to get $F(X)$. We had to wait until 1964 to start a discussion on "convex transformations of random variables", Van Zwet (1964)[254] –as the topic didn't seem important before.

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

Fragility and Antifragility When $F(X)$ is concave (fragile), errors about X can translate into extreme negative values for $F(X)$. When $F(X)$ is convex, one is largely immune from severe negative variations. In situations of trial and error, or with an option, we do not need to understand X as much as our exposure to the risks. Simply the statistical properties of X are swamped by those of H. The point of *Antifragile* is that exposure is more important than the naive notion of "knowledge", that is, understanding X.
The more nonlinear F the less the probabilities of X matters in the probability distribution of the final package F.
Many people confuse the probabilites of X with those of F. I am serious: the *entire* literature reposes largely on this mistake. For Baal's sake, focus on F, not X.

3.10 RUIN AND PATH DEPENDENCE

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

Path dependence is as follows. If I iron my shirts and then wash them, I get vastly different results compared to when I wash my shirts and then iron them. My first work, *Dynamic Hedging* [222], was about how traders avoid the "absorbing barrier" since once you are bust, you can no longer continue: anything that will eventually go bust will lose *all* past profits.

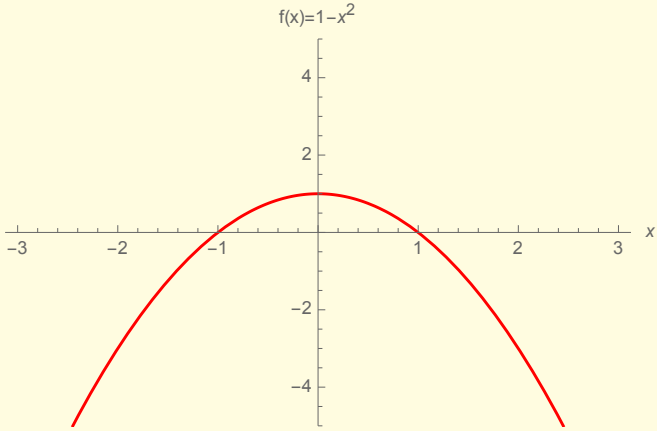
The physicists Ole Peters and Murray Gell-Mann [183] shed new light on this point, and revolutionized decision theory showing that a key belief since the development of applied probability theory in economics was wrong. They pointed out that all economics textbooks make this mistake; the only exception are by information theorists such as Kelly and Thorp.

Let us explain ensemble probabilities.

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

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

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



Now consider the following two types successions of deviations of x for 7 days (expressed in standard deviations).

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

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

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

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

Ole Peters has observed, is running his life across branching parallel universes). We can only get the return on the market under strict conditions.

Time probability and ensemble probability are not the same. This only works if the risk takers has an allocation policy compatible with the Kelly criterion [140],[245] using logs. Peters wrote three papers on time probability (one with Murray Gell-Mann) and showed that a lot of paradoxes disappeared.

Let us see how we can work with these and what is wrong with the literature. If we visibly incur a tiny risk of ruin, but have a frequent exposure, it will go to

probability one over time. If we ride a motorcycle we have a small risk of ruin, but if we ride that motorcycle a lot then we will reduce our life expectancy. The way to measure this is:

Principle 3.3 (Repetition of exposures)
Focus only on the reduction of life expectancy of the unit assuming repeated exposure at a certain density or frequency.

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

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

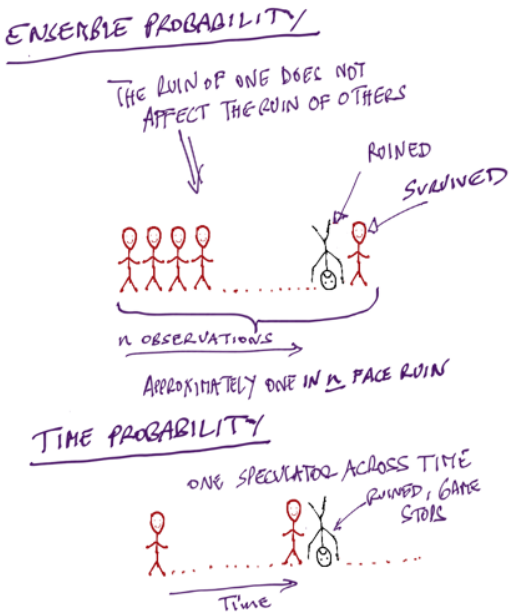


Figure 3.32: *Ensemble probability vs. time probability. The treatment by option traders is done via the absorbing barrier. I have traditionally treated this in Dynamic Hedging [222] and Antifragile[220] as the conflation between X (a random variable) and $f(X)$ a function of said r.v., which may include an absorbing state.*

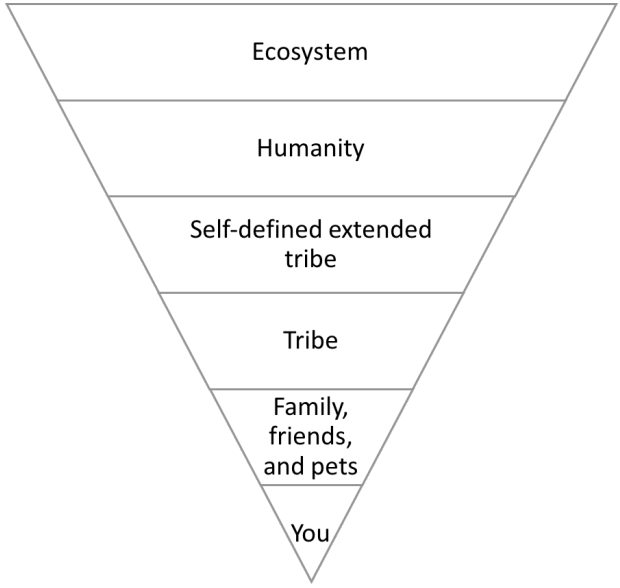


Figure 3.33: *A hierarchy for survival. Higher entities have a longer life expectancy, hence tail risk matters more for these. Lower entities such as you and I are renewable.*

wish it to have and see by how much the life expectancy is reduced by *repeated* exposure.

Remark 2: Psychology of decision making

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

Next let us consider layering, why systemic risks are in a different category from individual, idiosyncratic ones. Look at the (inverted) pyramid in Fig. 3.33: the worst-case scenario is not that an individual dies. It is worse if your family, friends and pets die. It is worse if you die and your arch enemy survives. They collectively have more life expectancy lost from a terminal tail event.

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

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

3.11 WHAT TO DO?

To summarize, we first need to make a distinction between mediocristan and Extremistan, two separate domains that about never overlap with one another. If

we fail to make that distinction, we don't have any valid analysis. Second, if we don't make the distinction between time probability (path dependent) and ensemble probability (path independent), we don't have a valid analysis.

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

The beautiful thing we discovered is that everything that is fragile has to present a concave exposure [220] similar –if not identical –to the payoff of a short option, that is, a negative exposure to volatility. It is nonlinear, necessarily. It has to have harm that accelerates with intensity, up to the point of breaking. If I jump 10m I am harmed more than 10 times than if I jump one meter. That is a necessary property of fragility. We just need to look at acceleration in the tails. We have built effective stress testing heuristics based on such an option-like property [237].

In the real world we want simple things that work [108]; we want to impress our accountant and not our peers. (My argument in the latest instalment of the *Incerto*, *Skin in the Game* is that systems judged by peers and not evolution rot from overcomplication). To survive we need to have clear techniques that map to our procedural intuitions.

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

NEXT

The next three chapters will examine the technical intuitions behind thick tails in discussion form, in not too formal a language. Derivations and formal proofs come later with the adaptations of the journal articles.

4 UNIVARIATE FAT TAILS, LEVEL 1, FINITE MOMENTS†

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

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

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

4.1 A SIMPLE HEURISTIC TO CREATE MILDLY FAT TAILS

A couple of reminders about convexity and Jensen's inequality:

Let $\mathscr{A}$ be a convex set in a vector space in $\mathbb{R}$, and let $\varphi : \mathscr{A} \to \mathbb{R}$ be a function; φ is called convex if $\forall x_1, x_2 \in \mathscr{A}, \forall t \in [0,1]$:

$$\varphi\left(t x_1 + (1-t) x_2\right) \leq t\varphi\left(x_1\right) + (1-t)\varphi\left(x_2\right)$$

† Discussion chapter.

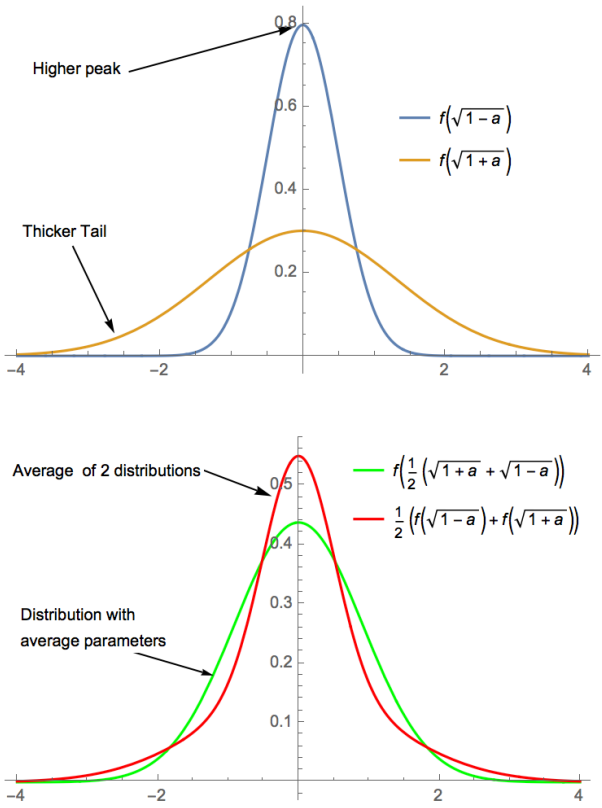


Figure 4.1: *How random volatility creates fatter tails owing to the convexity of some parts of the density to the scale of the distribution.*

For a random variable X and $\varphi(.)$ a convex function, by Jensen's inequality [133]:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

Remark 3: Fat Tails and Jensen's inequality

For a Gaussian distribution (and, members of the location-scale family of distributions), tail probabilities are convex to the scale of the distribution, here the standard deviation σ (and to the variance σ^2). This allows us to fatten the tails by "stochasticizing" either the standard deviation or the variance, hence checking the effect of Jensen's inequality on the probability distribution.

Heteroskedasticity is the general technical term often used in time series analysis to characterize a process with fluctuating scale. Our method "stochasticizes", that is, perturbates the variance or the standard deviation[2] of the distribution under the constraint of conservation of the mean.

2 "Volatility" in the quant language means standard deviation, but "stochastic volatility" is usually stochastic variance.

But note that *any* heavy tailed process, even a power law, can be described *in sample* (that is finite number of observations necessarily discretized) by a simple Gaussian process with changing variance, a regime switching process, or a combination of Gaussian plus a series of variable jumps (though not one where jumps are of equal size, see the summary in [172]).[3]

This method will also allow us to answer the great question: "where do the tails start?" in 4.3.

Let $f(\sqrt{a}, x)$ be the density of the normal distribution (with mean 0) as a function of the variance for a given point x of the distribution. Compare $f\left(\frac{1}{2}\left(\sqrt{1-a}+\sqrt{a+1}\right), x\right)$ to $\frac{1}{2}\left(f\left(\sqrt{1-a}, x\right)+f\left(\sqrt{a+1}, x\right)\right)$; the difference between the two will be owed to Jensen's inequality. We assume the average σ^2 constant, but the discussion works just as well if we just assumed σ constant —it is a long debate whether one should put a constraint on the average variance or on that of the standard deviation, but 1) doesn't matter much so long as one remains consistent, 2) for our illustrative purposes here there is no real fundamental difference.

Since higher moments increase under fat tails, though not necessarily lower ones, it should be possible to simply increase fat tailedness (via the fourth moment) while keeping lower moments (the first two or three) invariant. [4]

4.1.1 A Variance-preserving heuristic

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

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

The characteristic function[5] is

$$\phi(t, a) = \frac{1}{2} e^{-\frac{1}{2}(1+a)t^2\sigma^2}\left(1 + e^{at^2\sigma^2}\right) \tag{4.1}$$

3 The jumps for such a process can be simply modeled as a regime that is characterized by a Gaussian with low variance and extremely large mean (and a low-probability of occurrence), so, technically, Poisson jumps are mixed Gaussians.

4 To repeat what we stated in the previous chapter, 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). Fughedaboutdit. 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.

5 Note there is no difference between characteristic and moment generating functions when the mean is 0, a property that will be useful in later, more technical chapters.

Odd moments are nil. The second moment is preserved since

$$M(2) = (-i)^2 \partial^{t,2} \phi(t)|_0 \ = \sigma^2 \tag{4.2}$$

and the fourth moment

$$M(4) = (-i)^4 \partial^{t,4} \phi|_0 = 3\left(a^2+1\right)\sigma^4 \tag{4.3}$$

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

Limitations of the simple heuristic This heuristic, while useful for intuition building, is of limited powers as it can only raise kurtosis to twice that of a Gaussian, so it should be used only pedagogically, to get some intuition about the effects of the convexity. Section 4.1.2 will present a more involved technique.

Remark 4: Peaks

As Figure 4.4 shows: fat tails manifests themselves with higher peaks, a concentration of observations around the center of the distribution.

This is usually misunderstood.

4.1.2 Fattening of Tails With Skewed Variance

We can improve on the fat-tail heuristic in 4.1, (which limited the kurtosis to twice the Gaussian) as follows. We Switch between Gaussians with variance:

$$\begin{cases} \sigma^2(1+a), & \text{with probability } p \\ \sigma^2(1+b), & \text{with probability } 1-p \end{cases} \tag{4.4}$$

with $p \in [0,1)$ 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.

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

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

$$\phi(t,V) = \int_0^\infty \frac{e^{-\frac{t^2 v}{2} - \frac{\left(\log(v) - v0 + \frac{Vv^2}{2}\right)^2}{2Vv^2}}}{\sqrt{2\pi} v Vv} dv \tag{4.5}$$

where v is the variance and Vv is the second order variance, often called volatility of volatility. Thanks to integration by parts we can use the Fourier transform to obtain all varieties of payoffs (see Gatheral [101]). But the absence of a closed-form distribution can be remedied as follows, with the use of distributions for the variance that are analytically more tractable.

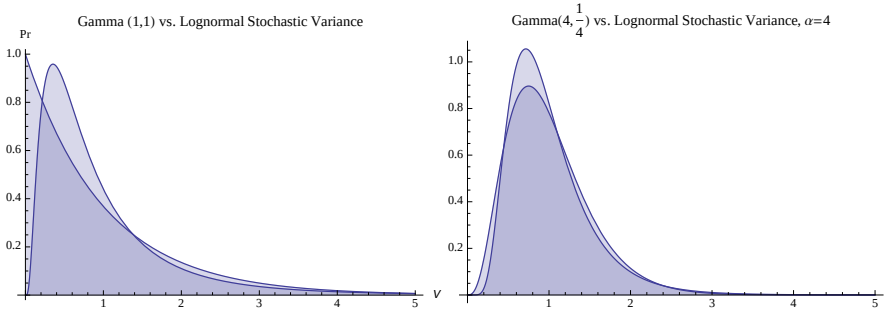


Figure 4.2: *Stochastic Variance: Gamma distribution and Lognormal of same mean and variance.*

Gamma Variance The gamma distribution applied to the variance of a Gaussian is is a useful shortcut for a full distribution of the variance, which allows us to go beyond the narrow scope of heuristics [36]. It is easier to manipulate analytically than the Lognormal.

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

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

with mean V and variance $\frac{V}{\sqrt{a}}$. Figure 4.2 shows the matching to a lognormal with same first two moments where we calibrate the lognormal to mean $\frac{1}{2}\log\left(\frac{aV^3}{aV+1}\right)$ and standard deviation $\sqrt{-\log\left(\frac{aV}{aV+1}\right)}$. The final distribution becomes (once again, assuming the same mean as a fixed volatility situation:

$$f_{a,V}(x) = \int_0^\infty \frac{e^{-\frac{(x-\mu)^2}{2v}}}{\sqrt{2\pi}\sqrt{v}} \Gamma_a(v) \mathrm{d}v, \tag{4.6}$$

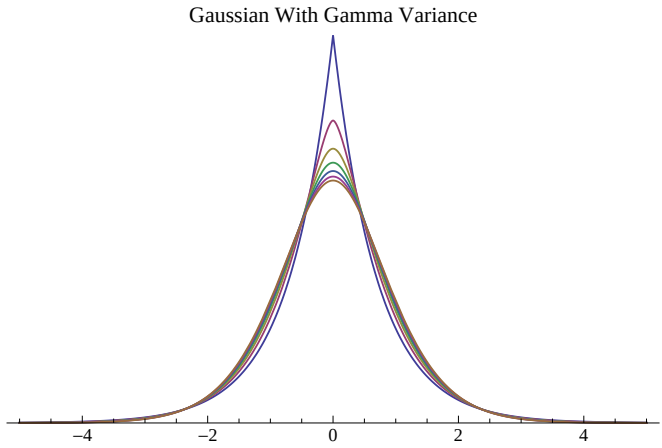


Figure 4.3: *Stochastic Variance using Gamma distribution by perturbating α in equation 4.7.*

allora:

$$f_{\alpha,V}(x) = \frac{2^{\frac{3}{4}-\frac{a}{2}} a^{\frac{a}{2}+\frac{1}{4}} V^{-\frac{a}{2}-\frac{1}{4}} |x-\mu|^{a-\frac{1}{2}} K_{a-\frac{1}{2}}\left(\frac{\sqrt{2}\sqrt{a}|x-\mu|}{\sqrt{V}}\right)}{\sqrt{\pi}\Gamma(a)}. \tag{4.7}$$

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

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

4.2 DOES STOCHASTIC VOLATILITY GENERATE POWER LAWS?

We have not yet defined power laws; take for now the condition that least one of the moments is infinite.

And the answer: depend on whether we are stochasticizing σ or σ^2 on one hand, or $\frac{1}{\sigma}$ or $\frac{1}{\sigma^2}$ on the other.

Assume the base distribution is the Gaussian, the random variable $X \sim \mathcal{N}(\mu, \sigma)$. Now there are different ways to make σ, the scale, stochastic. Note that since σ is nonnegative, we need it to follow some one-tailed distribution.

- We can make σ^2 (or, possibly σ) follow a Lognormal distribution. It does not yield closed form solutions, but we can get the moments and verify it is not a power law.
- We can make σ^2 (or σ) follow a gamma distribution. It does yield closed form solutions, as we saw in the example above, in Eq. 4.7.
- We can make $\frac{1}{\sigma^2}$ —the precision parameter—follow a gamma distribution.
- We can make $\frac{1}{\sigma^2}$ follow a lognormal distribution.

The results shown in Table 4.1 come from the following simple properties of density functions and expectation operators. Let X be any random variable with

Table 4.1: *Transformations for stochastic volatility. We can see from the density of the transformations* $\frac{1}{x}$ *or* $\frac{1}{\sqrt{x}}$ *if we have a power law on hand.* $\mathcal{LN}$, $\mathcal{N}$,$\mathcal{G}$ *and* $\mathcal{P}$ *are the Lognormal, Normal, Gamma, and Pareto distributions, respectively.*

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

Table 4.2: *The p-moments of possible distributions for variance*

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

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

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

and

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

(via the Mellin transform).

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

4.3 THE BODY, THE SHOULDERS, AND THE TAILS

Where do the tails start?

We assume the tails start at the level of convexity of the segment of the probability distribution to the scale of the distribution –in other words, affected by the stochastic volatility effect.

4.3.1 The Crossovers and Tunnel Effect.

Notice in Figure 4.4 a series of crossover zones, invariant to a. Distributions called "bell shape" have a convex-concave-convex shape (or quasi-concave shape).

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

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

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

1. There exist a "high peak" inner tunnel, $A_T = (a_2, a_3)$ for which the δ-perturbed σ of the probability distribution $p^\delta(x) \geq p(x)$ if $x \in (a_2, a_3)$
2. There exists outer tunnels, the "tails", for which $p^\delta(x) \geq p(x)$ if $x \in (-\infty, a_1)$ or $x \in (a_4, \infty)$

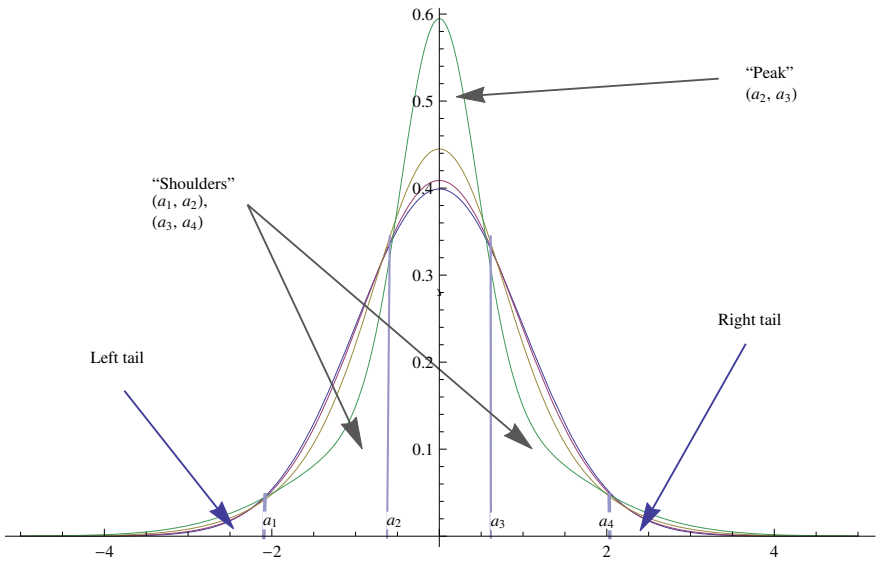


Figure 4.4: *Where do the tails start? Fatter and fatter fails through perturbation of the scale parameter σ for a Gaussian, made more stochastic (instead of being fixed). Some parts of the probability distribution gain in density, others lose. Intermediate events are less likely, tails events and moderate deviations are more likely. We can spot the crossovers a_1 through a_4. The "tails" proper start at a_4 on the right and $a_1$on the left.*

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 (easily) computable, not reliable for any effective use. The implication is that Black Swans do not necessarily come from fat tails; le problem can result from an incomplete assessment of tail events.

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

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

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

$$\begin{aligned}\{a_1, a_2, a_3, a_4\} = \Bigg\{&\mu - \sqrt{\frac{1}{2}\left(5+\sqrt{17}\right)}\sigma, \mu - \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma, \\ &\mu + \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma, \mu + \sqrt{\frac{1}{2}\left(5+\sqrt{17}\right)}\sigma\Bigg\}\end{aligned} \tag{4.8}$$

In figure 4.4, the crossovers for the intervals are numerically $\{-2.13\sigma, -.66\sigma, .66\sigma, 2.13\sigma\}$.

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

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

$$\begin{aligned}\{a_1, a_2, a_3, a_4\} = \Bigg\{&-\frac{\sqrt{\frac{5\alpha - \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, -\frac{\sqrt{\frac{5\alpha + \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, \\ &\frac{\sqrt{\frac{5\alpha - \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}, \frac{\sqrt{\frac{5\alpha + \sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}\Bigg\}\end{aligned}$$

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

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

In Summary, Where Does the Tail Start?

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

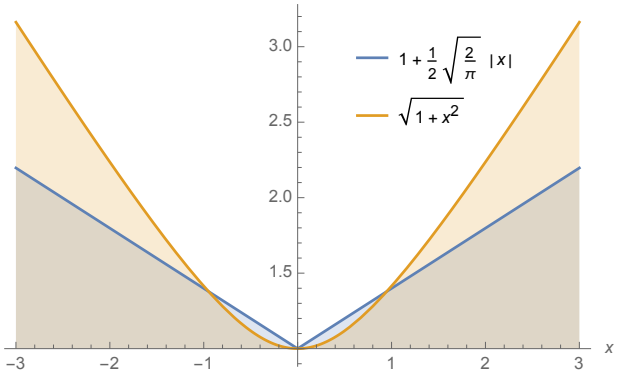


Figure 4.5: *We compare the behavior of $\sqrt{K+x^2}$ and $K+|x|$. The difference between the two weighting functions increases for large values of the random variable x, which explains the divergence of the two (and, more generally, higher moments) under fat tails.*

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

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 Lognormal is a special case.

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

Stochastic Parameters The problem of elliptical distributions is that they do not map the return of securities, owing to the absence of a single variance at any point in time, see Bouchaud and Chicheportiche (2010) [42]. When the scales of the distributions of the individuals move but not in tandem, the distribution ceases to be elliptical. Figure 6.2 shows the effect of applying the equivalent of stochastic volatility methods: the more annoying stochastic correlation. Instead of perturbating the correlation matrix Σ as a unit as in section 6.1, we perturbate the correlations with surprising effect.

4.4 FAT TAILS, MEAN DEVIATION AND THE RISING NORMS

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

4.4.1 The Common Errors

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

The difference between standard deviation (assuming mean and median of 0 to simplify) $\sigma = \sqrt{\frac{1}{n}\sum x_i^2}$ and mean absolute deviation $MAD = \frac{1}{n}\sum|x_i|$ increases under fat tails, as one can see in Figure 4.5 . This can provide a conceptual approach to the notion.

Dan Goldstein and the author [112] put the following question to investment professionals and graduate students in financial engineering –people who work with risk and deviations all day long.

> A stock (or a fund) has an average return of 0%. It moves on average 1% a day in absolute value; the average up move is 1% and the average down move is 1%. It does not mean that all up moves are 1% –some are .6%, others 1.45%, and so forth.

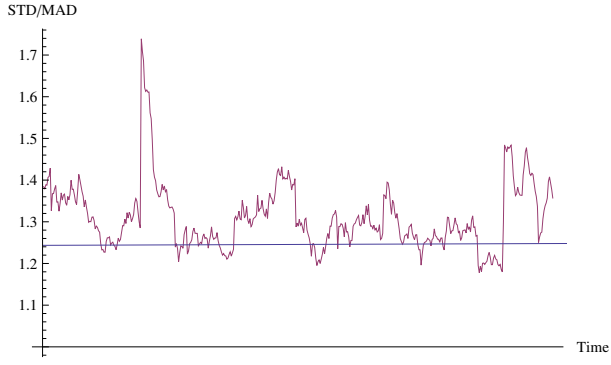


Figure 4.6: *The Ratio STD/MAD for the daily returns of the SP500 over the past 47 years, seen with a monthly rolling window. We can consider the level* $\sqrt{\frac{\pi}{2}} \approx 1.253$ *(as approximately the value for Gaussian deviations), as the cut point for fat tailedness.*

> Assume that we live in the Gaussian world in which the returns (or daily percentage moves) can be safely modeled using a Normal Distribution. Assume that a year has 256 business days. What is its standard deviation of returns (that is, of the percentage moves), the âĂIJsigmaâĂİ that is used for volatility in financial applications?
>
> What is the daily standard deviation?
>
> What is the yearly standard deviation?

As the reader can see, the question described mean deviation. And the answers were overwhelmingly wrong. For the daily question, almost all answered 1%. Yet a Gaussian random variable that has a daily percentage move in absolute terms of 1% has a standard deviation that is higher than that, about 1.25%. It should be up to 1.7% in empirical distributions. The most common answer for the yearly question was about 16%, which is about 80% of what would be the true answer. The professionals were scaling daily volatility to yearly volatility by multiplying by $\sqrt{256}$ which is correct provided one had the correct daily volatility.

So subjects tended to provide MAD as their intuition for STD. When professionals involved in financial markets and continuously exposed to notions of volatility talk about standard deviation, they use the wrong measure, mean absolute deviation (MAD) instead of standard deviation (STD), causing an average underestimation of between 20 and 40%. In some markets it can be up to 90%. Further, responders rarely seemed to immediately understand the error when it was pointed out to them. However when asked to present the equation for standard deviation they effectively expressed it as the mean root mean square deviation. Some were puzzled as they were not aware of the existence of MAD.

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

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

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

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

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

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

4.4.2 Some Analytics

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

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

where $\mathbb{E}$ is the expectation operator (under the probability measure of concern and X is a centered variable such $\mathbb{E}(x) = 0$); the ratio increases with the fat tailedness of the distribution; (The general case corresponds to $\frac{(\mathbb{E}(x^p))^{\frac{1}{p}}}{\mathbb{E}(|x|)}$, $p > 1$, under the condition that the distribution has finite moments up to n, and the special case here $n = 2$).[6]

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

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

Mean Deviation vs Standard Deviation, more technical Why the [REDACTED] did statistical science pick STD over Mean Deviation? Here is the story, with analytical derivations not seemingly available in the literature. In Huber [129]:

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

Let us rederive and see what Fisher meant.

Let n be the number of summands:

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

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

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

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

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

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

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

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

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

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

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

Finalmente, the Asymptotic Relative Efficiency For a Gaussian

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

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

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

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

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

For ease, a simple Monte Carlo simulation would do. Using $p = .01$ and $n = 1000$... Figure 4.8 shows how a=2 causes degradation. A minute presence of outliers makes MAD more "efficient" than STD. Small "outliers" of 5 standard deviations cause MAD to be five times more efficient.[7]

7 The natural way is to center MAD around the median; we find it more informative for many of our purposes here (and decision theory) to center it around the mean. We will make note when the centering is around the mean.

Figure 4.7: *Harald Cramér, of the Cramer condition, and the ruin problem.*

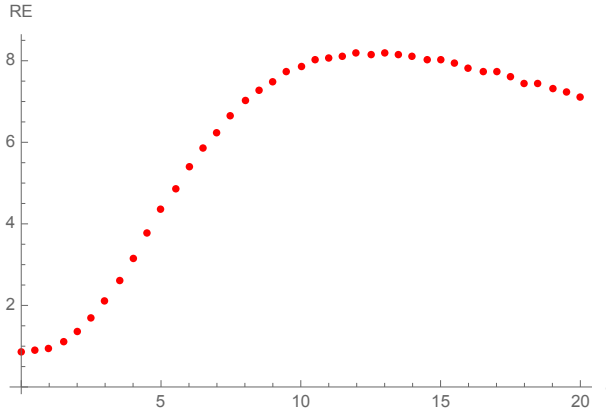


Figure 4.8: *A simulation of the Relative Efficiency ratio of Standard deviation over Mean deviation when injecting a jump size $\sqrt{(1+a)} \times \sigma$, as a multiple of σ the standard deviation.*

4.4.4 Moments and The Power Mean Inequality

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

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

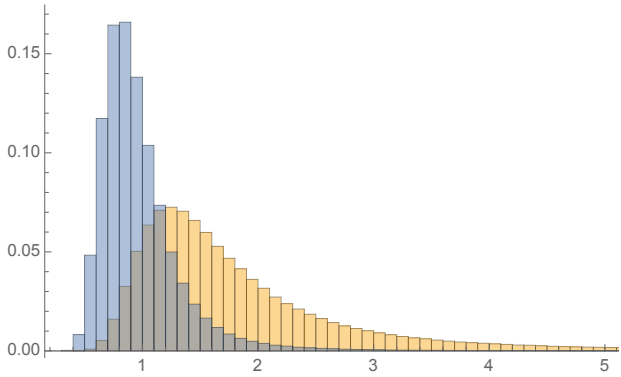


Figure 4.9: *Mean deviation vs standard deviation for a finite variance power law. The result is expected (MD is the thinner distribution), complicated by the fact that standard deviation has an infinite variance since the square of a Paretian random variable with exponent α is Paretian with an exponent of $\frac{1}{2}\alpha$. In this example the mean deviation of standard deviation is 5 times higher.*

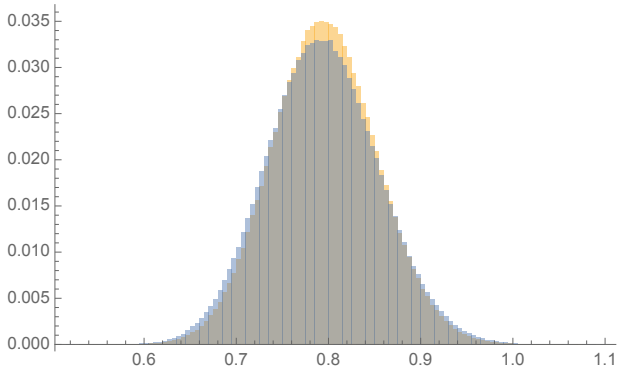


Figure 4.10: *For a Gaussian, there is small difference between MD and STD.*

For any $1 \leq p < q$ the following inequality holds:

$$\sqrt[p]{\sum_{i=1}^{n} w_i |x_i|^p} \leq \sqrt[q]{\sum_{i=1}^{n} w_i |x_i|^q} \tag{4.9}$$

where the positive weights w_i sum to unity. (Note that we avoid $p < 1$ because it does not satisfy the triangle inequality).

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

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

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

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

$$\{\|X\|_n\}_{n=1}^{5} = \left\{1, \frac{\sqrt{5}}{2}, \frac{\sqrt[3]{7}}{2^{2/3}}, \frac{\sqrt[4]{41}}{2}, \frac{\sqrt[5]{61}}{2^{4/5}}\right\}. \tag{4.10}$$

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

$$\{||X||_n\}_{n=1}^{5} = \left\{1, \frac{5}{4}, \frac{\sqrt[3]{\frac{43}{2}}}{2}, \frac{\sqrt[4]{1201}}{4}, \frac{\sqrt[5]{2101}}{2 \times 2^{3/5}}\right\}. \tag{4.11}$$

So we can see it becomes rapidly explosive.

One property quite useful with power laws with infinite moment:

$$\|X\|_\infty = \sup\left(\frac{1}{n}|x_i|\right)_{i=1}^{n} \tag{4.12}$$

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

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

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

or, alternatively

$$\frac{\mathbb{E}(X^p)}{\mathbb{E}(|X|)} = 2^{\frac{1}{2}(p-3)}\left(1 + (-1)^p\right)\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}-\frac{p}{2}}\Gamma\left(\frac{p+1}{2}\right) \tag{4.13}$$

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

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

hence

$$\frac{\sqrt{\mathbb{E}(X^2)}}{\mathbb{E}(|X|)} = \frac{MD}{STD} = \sqrt{\frac{\pi}{2}}$$

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

For a Power Law distribution with tail exponent α=3, say a Student T

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

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.

Pareto Case For a standard Pareto distribution with minimum value (and scale) L, PDF $f(x) = \alpha L^\alpha x^{-\alpha-1}$ and standard deviation $\frac{\sqrt{\frac{\alpha}{\alpha-2}}L}{\alpha-1}$, we have

$$\frac{MD}{STD} = \frac{\left(2^{1/\alpha}-1\right)(\alpha-2)(\alpha-1)}{L}, \tag{4.14}$$

by centering around the median.

"Infinite" moments Infinite moments, say infinite variance, always manifest themselves as computable numbers in observed sample, yielding finite moments of all orders, simply because the sample is finite. A distribution, say, Cauchy, with undefined means will always deliver a measurable mean in finite samples; but different samples will deliver completely different means. Figures 4.11 and 4.12 illustrate the "drifting" effect of the moments with increasing information.

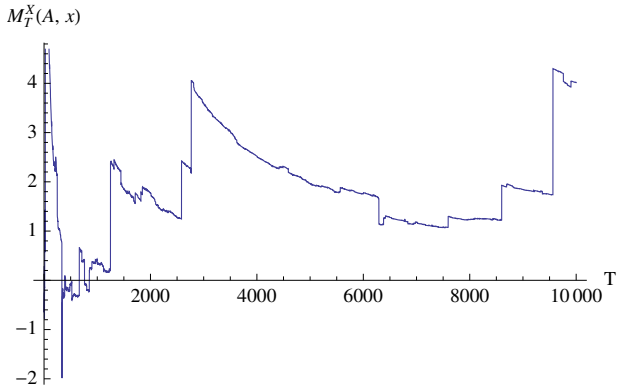


Figure 4.11: *The mean of a series with undefined mean (Cauchy).*

4.4.5 Comment: Why we should retire standard deviation, now!

The notion of standard deviation has confused hordes of scientists; it is time to retire it from common use and replace it with the more effective one of mean deviation. Standard deviation, STD, should be left to mathematicians, physicists and

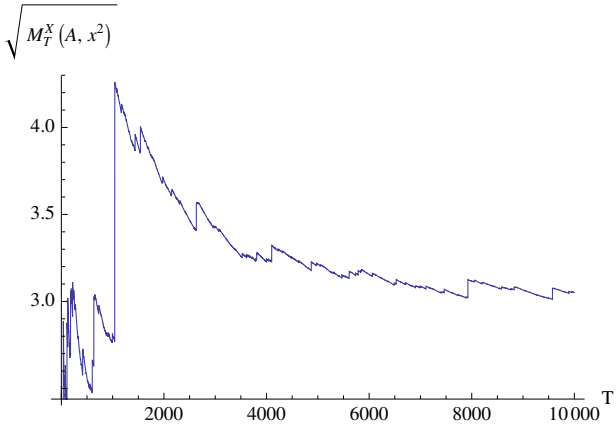


Figure 4.12: *The square root of the second moment of a series with infinite variance. We observe pseudo-convergence before a jump.*

mathematical statisticians deriving limit theorems. There is no scientific reason to use it in statistical investigations in the age of the computer, as it does more harm than good-particularly with the growing class of people in social science mechanistically applying statistical tools to scientific problems.

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

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

It is all due to a historical accident: in 1893, the great Karl Pearson introduced the term "standard deviation" for what had been known as "root mean square error". The confusion started then: people thought it meant mean deviation. The idea stuck: every time a newspaper has attempted to clarify the concept of market "volatility", it defined it verbally as mean deviation yet produced the numerical measure of the (higher) standard deviation.

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

It all comes from bad terminology for something non-intuitive. By a psychological phenomenon called attribute substitution, some people mistake MAD for STD

because the former is easier to come to mind – this is "Lindy"[8] as it is well known by cheaters and illusionists.

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

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

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

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

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

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

$$\mathbb{V}(X_1 + \cdots + X_n) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n).$$

by the linearity of the variance. But then it assumes that one cannot use another metric then by simple transformation make it additive[9]. As we will see, for the Gaussian $\mathrm{md}(X) = \sqrt{\frac{2}{\pi}}\sigma$ —for the Student T with 3 degrees of freedom, the factor is $\frac{2}{\pi}$, etc.

8 See a definition of "Lindy" in 5.0.2

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

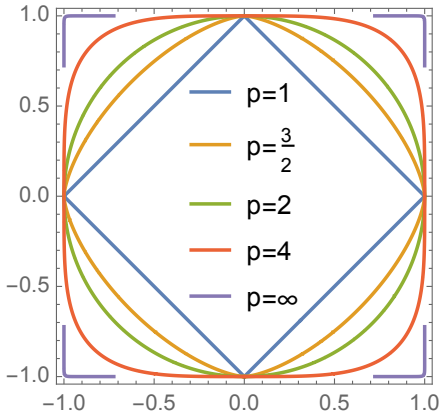


Figure 4.13: *Rising norms and the unit circle/square: values of the iso-norm* $\left(|x_1|^p + |x_2|^p\right)^{1/p} = 1$. *We notice the area inside the norm (i.e. satisfying norm* ≤ 1*),* $v(p) = \frac{4\Gamma\left(\frac{p+1}{p}\right)^2}{\Gamma\left(\frac{p+2}{p}\right)}$, *with* $v(1) = 2$ *and* $v(\infty) = 4$.

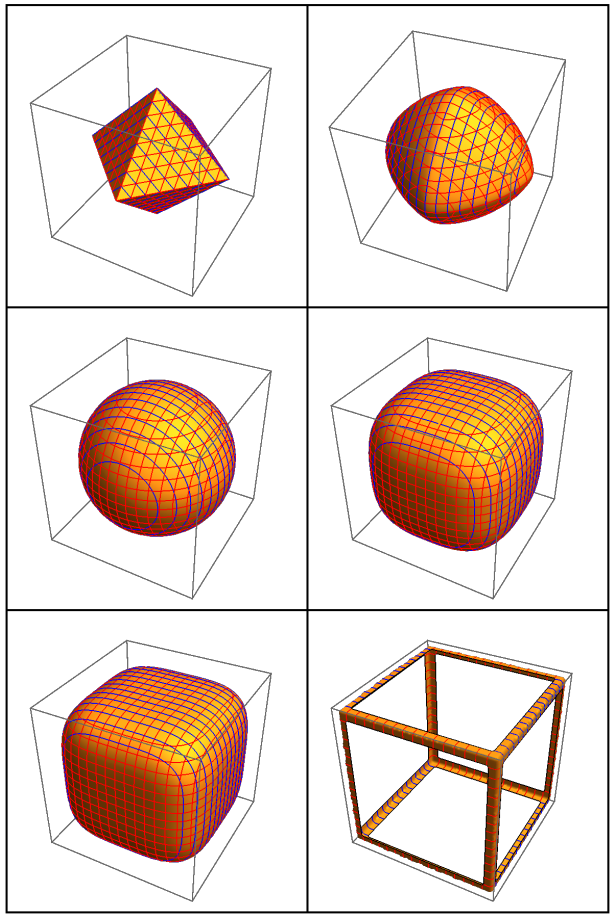


Figure 4.14: *Rising norms and the unit cube: values of the iso-norm* $\left(|x_1|^p + |x_2|^p + |x_3|^p\right)^{1/p} = 1$ *for* $p = 1, \frac{3}{2}, 2, 3, 4$, *and* ∞. *The volume satisfying the inequality norm* ≤ 1 *increases for* $\frac{4}{3}$ *for* $p = 1$, $\frac{4\pi}{3}$ *for* $p = 2$ *(the unit sphere), to* 2^3 *for* $p = \infty$ *(the unit cube), a much higher increase than in Fig. 4.13 . We can see the operation of the curse of dimensionality in the smaller and smaller volume for* $p = 1$ *relative to the maximum when* $p = \infty$

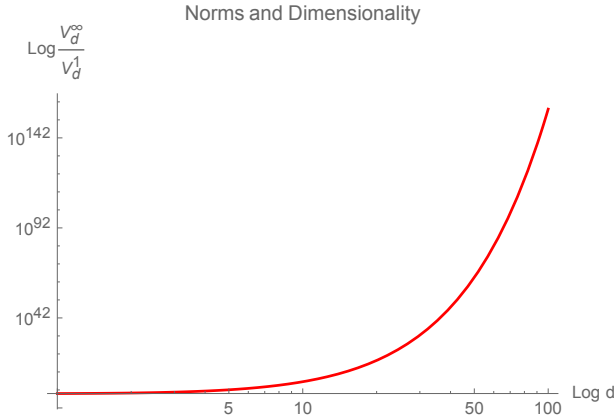


Figure 4.15: *The curse of dimensionality, with yuuuge applications across statistical areas, particularly model error in higher dimentions. As d increases, the ratio of V^1 over V^∞ blows up. If for $d = 2$, it is 2 it is already six figures for $d = 9$.*

4.5 VISUALIZING THE EFFECT OF RISING p ON ISO-NORMS

Consider the region $\mathscr{R}_{(p)}^{(n)}$ defined as $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n}) :\in \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i^p}\right)^{\mathbf{1/p}} \leq \mathbf{1}$, with the border defined by the identity. As the norm rises, we calculate the following measure of the ball:

$$V_n^p = \int \ldots \int_{\mathbf{X} \in \mathscr{R}_{(\mathbf{n})}^{(\mathbf{n})}} 1 d\mathbf{X} = \frac{\left(\mathbf{4\Gamma}\left(\mathbf{1} + \frac{\mathbf{1}}{\mathbf{p}}\right)\right)^{\mathbf{n}}}{\mathbf{\Gamma}\left(\frac{\mathbf{n}}{\mathbf{p}} + \mathbf{1}\right)}$$

Figures 4.13 and 4.14 show two effects.

The first is how rising norms occupy a larger share of the space.

The second gives us a hint of the curse of dimensionality, useful in many circumstances (and, centrally, for model error). Compare figures 4.13 and 4.14: you will notice that in the first case, for $d = 2$, $p = 1$, m occupies half the area of the square, with $p = \infty$ all of it. The ratio of norms is $\frac{1}{2}$. But for $d = 3$, $p = 1$ occupies $\frac{4/3}{2^3} = \frac{1}{6}$ of the space (again, $p = \infty$ occupies all of it). The ratio of higher moments to lower moments increases with dimensionality, as seen in Fig. 4.15.

URTHER READING: We stop here and present probability books in general. For more general intuition about probability, the indispensable Borel's [84]. Kolmogorov [143], Loeve [152], Feller [91],[90]. For measure theory, Billingsley [20].

For subexponentiality Pitman [193], Embrechts and Goldie (1982) [82], Embrechts (1979, which seems to be close to his doctoral thesis) [83], Chistyakov (1964) [43], Goldie (1978) [111], and Teugels [243].

For extreme value distributions Embrechts et al [81], De Haan and Ferreira [115].

For stable distributions Uchaikin and Zolotarev [252], Zolotarev [266], Samorindsky and Taqqu [206].

Stochastic processes Karatsas and Shreve [139], Oksendal [180], Varadhan [256].

5 LEVEL 2: SUBEXPONENTIALS AND POWER LAWS

THIS CHAPTER briefly presents the exponential and power law classes as "true fat tails" (already defined in Chapter 3) and presents some wrinkles associated with them. Subexponentiality (without scalability), that is membership in the subexponential but not power law class is a small category (of the common distributions, only the borderline exponential –and gamma associated distributions such as the Laplace – and the the lognormal fall in that class).

5.0.1 Revisiting the Rankings

Table 5.1 reviews the rankings of Chapter 3. Recall that probability distributions range between extreme thin-tailed (Bernoulli) and extreme fat tailed. Among the categories of distributions that are often distinguished due to the convergence properties of moments are:

1. Having a support that is compact but not degenerate
2. Subgaussian
3. Subexponential
4. Power Law with exponent greater than 2
5. Power Law with exponent less than or equal to 2. In particular, Power Law distributions have a finite mean only if the exponent is greater than 1, and have a finite variance only if the exponent exceeds 2.
6. Power Law with exponent less than 1.

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

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

Table 5.1: *Ranking distributions*

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

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

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

Let $\mathbf{X} = X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with support in $(\mathbb{R}^+)$, with cumulative distribution function F. The subexponential class of distributions is defined by (see [243], [193]):

$$\lim_{x \to +\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2 \tag{5.1}$$

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

More generally, it can be shown that the sum of n variables is dominated by the maximum of the values over those variables in the same way. Formally, the following two properties are equivalent to the subexponential condition [43],[83]. For a given $n \geq 2$, let $S_n = \Sigma_{i=1}^{n} x_i$ and $M_n = \max_{1 \leq i \leq n} x_i$

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

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

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

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

$$\int_0^\infty \mathbf{e}^{\epsilon x}\, dF(x) = +\infty \tag{5.2}$$

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

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

5.0.2 What is a borderline probability distribution?

The best way to figure out a probability distribution is to... invent one. In fact in the next section, 5.0.3, we will build one that is the exact borderline between thin and fat tails *by construction*. Consider for now that the properties are as follows:

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

$$\lim_{x\to+\infty} \frac{\overline{F}(x)^n}{\overline{F}(nx)} = 1, \tag{5.3}$$

and

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

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

Note : another property of the demarcation is the absence of Lucretius fallacy from *The Black Swan*, mentioned earlier (i.e. future extremes will not be similar to past extremes under fat tails, and such dissimilarity increases with fat tailedness):

Let us look at the demarcation properties for now. Let X be a random variable that lives in either $(0,\infty)$ or $(-\infty,\infty)$ and $\mathbb{E}$ the expectation operator under "real world" (physical) distribution. By classical results [81]:

$$\lim_{K\to\infty} \frac{1}{K}\mathbb{E}(X|_{X>K}) = \lambda \tag{5.4}$$

- If $\lambda = 1$, X is said to be in the thin tailed class $\mathcal{D}_1$ and has a characteristic scale
- If $\lambda > 1$, X is said to be in the fat tailed regular variation class $\mathcal{D}_2$ and has no characteristic scale
- If
$$\lim_{K\to\infty} \mathbb{E}(X|_{X>K}) - K = \mu$$
where $\mu > 0$, then X is in the borderline exponential class

The first case is called the "Lindy effect" when the random variable X is time survived. The subject is examined outside of this fat-tails project. See Iddo eliazar's exposition [76].

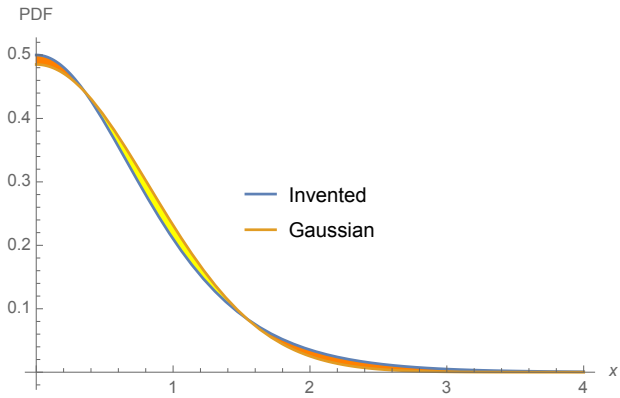


Figure 5.1: *Comparing the invented distribution (at the cusp of subexponentiality) to the Gaussian of the same variance ($k = 1$). It does not take much to switch from Gaussian to subexponential properties.*

5.0.3 Let Us Invent a Distribution

While the exponential distribution is at the cusp of the subexponential class but with support in $[0, \infty)$, we can construct a borderline distribution with support in $(-\infty, \infty)$, as follows [1]. Find survival functions $\overline{F} : \mathbb{R} \to [0, 1]$ that satisfy:

$$\forall x \geq 0, \lim_{x\to+\infty} \frac{\overline{F}(x)^2}{\overline{F}(2x)} = 1, \ \overline{F}'(x) \leq 0$$

and

$$\lim_{x\to+\infty} \overline{F} = 0.$$

$$\lim_{x\to-\infty} \overline{F} = 1.$$

1 The Laplace distribution, which doubles the exponential on both sides, does not fit the property as the ratio of the square to the double is $\frac{1}{2}$.

Let us assume a candidate function a sigmoid, using the hyperbolic tangent

$$\overline{F}^{\kappa}(x) = \frac{1}{2}\left(1 - \tanh(kx)\right) \ , \kappa > 0.$$

We can use this as a kernel distribution (we mix later to modify the kurtosis).

Let $f(.)$ be the density function:

$$f(x) = -\frac{\partial \overline{F}(x)}{\partial x} = \frac{1}{2} k \operatorname{sech}^2(kx). \tag{5.5}$$

The characteristic function:

$$\phi(t) = \frac{\pi t \operatorname{csch}\left(\frac{\pi t}{2k}\right)}{2k}. \tag{5.6}$$

Given that it is all real, we can guess that the mean is 0 –so are all odd moments.

The second moment will be $\lim_{t\to 0} (-i)^2 \frac{\partial^2}{\partial t^2} \frac{\pi t \operatorname{csch}\left(\frac{\pi t}{2k}\right)}{2k} = \frac{\pi^2}{12k^2}$ And the fourth moment: $\lim_{t\to 0} (-i)^4 \frac{\partial^4}{\partial t^4} \frac{\pi t \operatorname{csch}\left(\frac{\pi t}{2k}\right)}{2k} = \frac{7\pi^4}{240k^4}$, hence the Kurtosis will be $\frac{21}{5}$. The distribution we invented has slightly fatter tails than the Gaussian.

5.1 LEVEL 3: SCALABILITY AND POWER LAWS

Now we get into the serious business.

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

5.1.1 Scalable and Nonscalable, A Deeper View of Fat Tails

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

Note: We can see from the scaling difference between the Student and the Pareto the conventional definition of a Power Law tailed distribution is expressed more

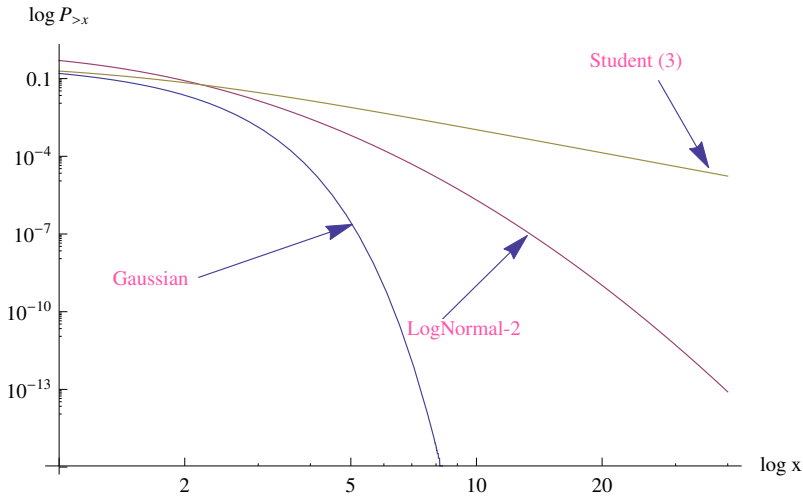


Figure 5.2: *Three Types of Distributions. As we hit the tails, the Student remains scalable while the Standard Lognormal shows an intermediate position before eventually ending up getting an infinite slope on a log-log plot. But beware the lognormal as it may have some surprises (Chapter 8)*

.

Table 5.2: *Scalability, comparing regularly varying functions/powerlaws to other distributions*

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

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 as deviations get larger.

Definition 5.1 (the class $\mathfrak{P}$)
The $\mathfrak{P}$ class of power laws (regular variation) is defined for r.v. X as follows:

$$\mathfrak{P} = \{X : \mathbb{P}(X > x) \sim L(x)\, x^{-\alpha}\} \tag{5.7}$$

5.1.2 Grey Swans

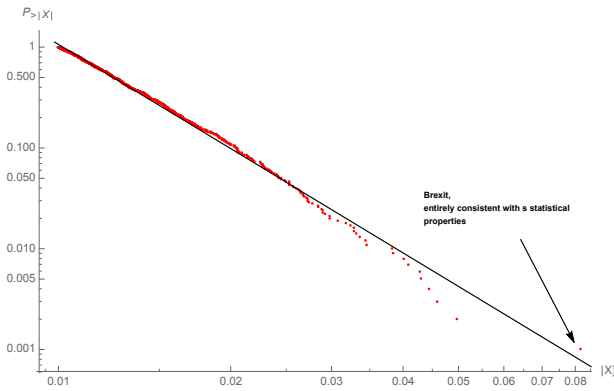


Figure 5.3: *The Grey Swan of Brexit when seen using a power law.*

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

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

More advanced methods involving the use of Levy laws may help in the event of asymmetry, but the use of two different Pareto distributions with two different

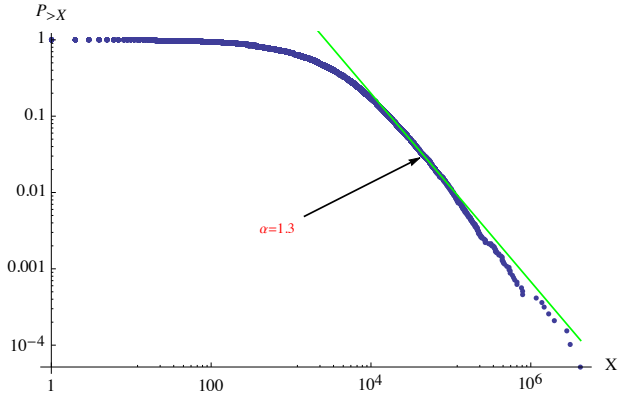


Figure 5.4: *Book Sales: the near tail can be robust for estimation of sales from rank and vice versa –it works well and shows robustness so long as one doesn't compute general expectations or higher non-truncated moments.*

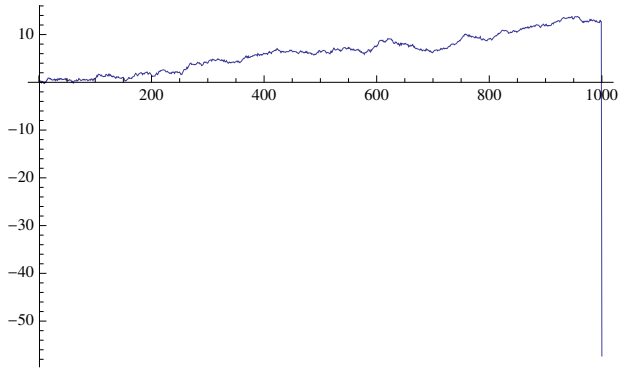


Figure 5.5: *The Turkey Problem, where nothing in the past properties seems to indicate the possibility of the jump.*

exponents, one for the left tail and the other for the right one would do the job (without unnecessary complications).

Estimation issues Note that there are many methods to estimate the tail exponent α from data, what is called a "calibration. However, we will see, the tail exponent is rather hard to guess, and its calibration marred with errors, owing to the insufficiency of data in the tails. In general, the data will show thinner tail than it should.

We will return to the issue in more depth in later chapters.

5.2 SOME PROPERTIES OF POWER LAWS

Two central properties.

5.2.1 Sums of variables

Property 1: Tail exponent of a sum

Let $X_1, X_2, \dots X_n$ be random variables neither independent nor identically distributed, each X_i following a distribution with a different asymptotic tail exponent α_i (we assume that random variables outside the power law class will have an asymptotic alpha = $+\infty$). Assume further we are concerned with the right tail of the distribution (the argument remains identical when we apply it to the left tail). See [98] for further details.

Consider the weighted sum $S_n = \sum_{i=1}^{n} w_i X_i$, with all weights w_i strictly positive. Consider α_s the tail exponent for the sum.

For all $w_i > 0$,

$$\alpha_s = \min(\alpha_i).$$

Clearly, if $\alpha_2 \leq \alpha_1$ and $w_2 > 0$,

$$\lim_{x \to \infty} \frac{\log\left(w_1 x^{-\alpha_1} + w_2 x^{-\alpha_2}\right)}{\log(x)} = \alpha_2.$$

The implication is that adding a single summand with undefined (or infinite) mean, variance, or higher moments leads to the total sum to have undefined (or infinite) mean, variance, or higher moments.

Principle 5.1 (Power Laws + Thin Tails = Power Laws)
Mixing power law distributed and thin tailed variables results in power laws, no matter the composition.

5.2.2 Transformations

The second property while appearing benign, can be vastly more annoying:

Property 2

Let X be a random variable with tail exponent α. The tail exponent of X^p is $\frac{\alpha}{p}$.

This tells us that the variance of a finite variance random variable with tail exponent < 4 will be infinite. In fact we will see it does cause problems for stochastic volatility models, when the real process can actually be of infinite variance.

This gives us a hint, without too much technical effort, on how a convex transformation of a random variable thickens the tail.

Proof. The general approach is as follows. Let $p(.)$ be a probability density function and $\phi(.)$ a transformation (with some restrictions). We have the distribution of the transformed variable (assuming the support is conserved –stays the same):

$$p\left(\phi(x)\right) = \frac{p\left(\phi^{(-1)}(x)\right)}{\phi'\left(\phi^{(-1)}(x)\right)}. \tag{5.8}$$

Assume that $x > l$ and l is large (i.e. a point where the slowly varying function "ceases to vary" within some order of x). The PDF for these values of x can be written as $p(x) \propto Kx^{-\alpha-1}$. Consider $y = \phi(x) = x^p$: the inverse function of $y = x^p$ is $x = y^{\frac{1}{p}}$. Applying to the denominator in Eq. 5.8, we get $\frac{1}{p} x^{\frac{1-p}{p}}$. □

Integrating above l, the survival function will be: $\mathbb{P}(Y > y) \propto y^{-\frac{\alpha}{p}}$.

5.3 BELL SHAPED VS NON BELL SHAPED POWER LAWS

The slowly varying function effect, a case study The fatter the tails, the less the "body" matters for the moments (which become infinite, eventually). But for power laws with thinner tails, the zone that is not power law (the slowly moving part) plays a role –"slowly varying" is more or less formally defined in 5.1.1,18.2.2 and 5.1.1. This section will show how apparently equal distributions can have different shapes.

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

$$f_P(x) = \begin{cases} \alpha(1+x)^{-\alpha-1} & x \geq 0 \\ \\ \alpha(1-x)^{-\alpha-1} & x < 0 \end{cases}$$

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

We have two ways to compare distributions.

- Equalizing by tail ratio: setting $\lim_{x\to\infty} \frac{f_P(x)}{f_s(x)} = 1$ to get the same tail ratio, we get the equivalent "tail" distribution with $s = \left(\alpha^{1-\frac{\alpha}{2}} B\left(\frac{\alpha}{2},\frac{1}{2}\right)\right)^{1/\alpha}$.
- Equalizing by standard deviations (when finite): we have, with $\alpha > 2$, $\mathbb{E}(X_P^2) = \frac{2}{\alpha^2-3\alpha+2}$ and $\mathbb{E}(X_S^2) = \frac{\alpha\left(\alpha^{1-\frac{\alpha}{2}} B\left(\frac{\alpha}{2},\frac{1}{2}\right)\right)^{2/\alpha}}{\alpha-2}$.

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

Finally, we have the comparison "bell shape" semi-concave vs the angular double-convex one as seen in Fig. 5.6.

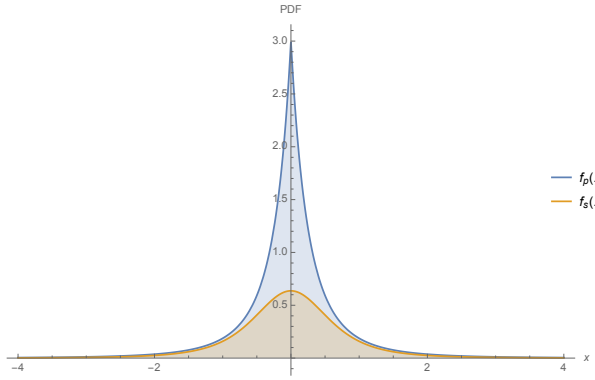


Figure 5.6: *Comparing two symmetric power laws of same exponent, one with a brief slowly varying function, the other with an extended one. All moments eventually become the same in spite of the central differences in their shape for small deviations.*

5.4 SUPER-FAT TAILS: THE LOG-PARETO DISTRIBUTION

The mother of all fat tails, the log-Pareto distribution, is not present in common lists of distributions but we can rederive it here. The log-Pareto is the Paretian analog of the lognormal distribution.

Remark 5: Rediscovering the log-Pareto distribution

If $X \sim \mathcal{P}(L, \alpha)$ *the Pareto distribution with PDF* $f^{(P)}(x) = \alpha L^\alpha x^{-\alpha-1}$, $x \geq L$ *and survival function* $S^{(P)}(x) = L^\alpha x^{-\alpha}$, *then:*

$e^X \sim \mathcal{LP}(L, \alpha)$ *the log-Pareto distribution with PDF*

$$f^{(LP)}(x) = \frac{\alpha L^\alpha \log^{-\alpha-1}(x)}{x} \, , x \geq e^L$$

and survival function

$$S^{(LP)}(x) = L^\alpha \log^{-\alpha}(x)$$

While for a regular power law, we have an asymptotic linear slope on the log-log plot, i.e.,

$$\lim_{x \to \infty} \frac{\log\left(L^\alpha x^{-\alpha}\right)}{\log(x)} = -\alpha,$$

the slope for a log-Pareto goes to 0:

$$\lim_{x \to \infty} \frac{\log\left(L^\alpha \log(x)^{-\alpha}\right)}{\log(x)} = 0,$$

and clearly no moment can exist regardless of the value of the tail parameter α. The difference between asymptotic behaviors is visible is Fig 5.7.

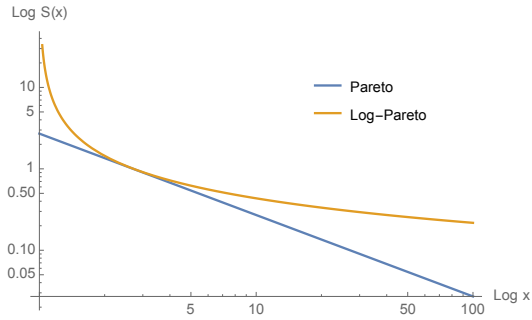


Figure 5.7: *Comparing log-log plots for the survival functions of the Pareto and log-Pareto*

5.5 PSEUDO-STOCHASTIC VOLATILITY: AN INVESTIGATION

We mentioned earlier in Chapter 3 that a "10 sigma" statement means we are not in the Gaussian world. We also discussed the problem of nonobservability of probability distributions: we observe data, not generating processes.

It is therefore easy to be fooled by a power law by mistaking it for a heteroskedastic process. In hindsight, we can always say: "conditional volatility was high, at such standard deviation it is no longer a 10 sigma, but a mere 3 sigma deviation".

The way to debunk these claims is to reason with the aid of an inverse problem: how a power law with a constant scale can masquerade as a heteroskedastic process. We will see in Appendix how econometrics' reliance on heteroskedasticity (i.e. moving variance) has severe defects since the variance of that variance doesn't have a structure.

Fig. 5.8 shows the volatility of returns of a market that greatly resemble ones should one use a standard simple stochastic volatility process. By stochastic volatility we assume the variance is distributed randomly [2].

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

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

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

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

2 One can have models with either stochastic variance or stochastic standard deviation. The two have different expectations.

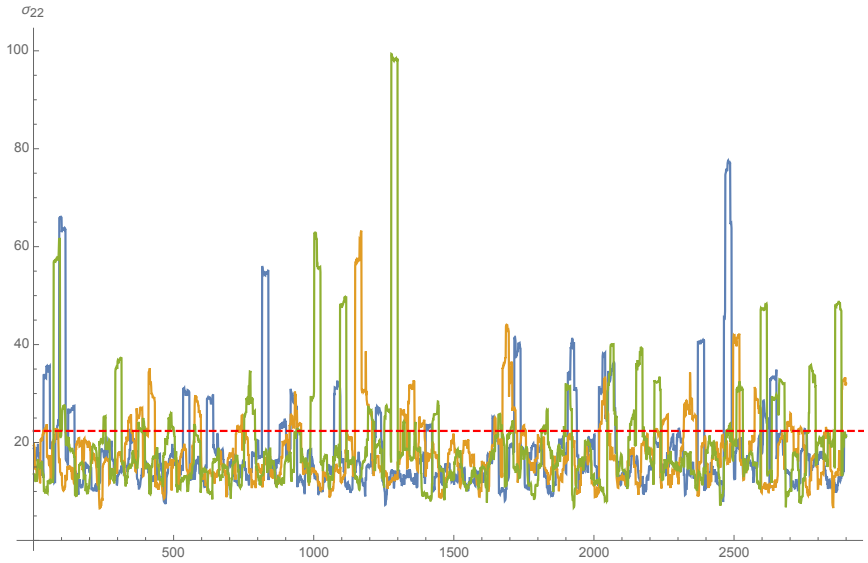


Figure 5.8: *Running 22-day (i.e., corresponding to monthly) realized volatility (standard deviation) for a Student T distributed returns sampled daily. It gives the impression of stochastic volatility when in fact the scale of the distribution is constant.*

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

$$\chi_y(\omega) \tag{5.9}$$

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

From which we get the mean deviation of the second moment[3].

3 As customary, we do not use standard deviation as a metric owing to its instability and its lack of information, but prefer mean deviation.

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

NEXT

The next chapter will venture in higher dimensions. Some consequences are obvious, others less so –say correlations exist even when covariances do not.

6 THICK TAILS IN HIGHER DIMENSIONS†

THIS DISCUSSION is about as simplified as possible handling of higher dimensions. We will look at 1) the simple effect of fat-tailedness for multiple random variables, 2) Ellipticality and distributions, 3) random matrices and the associated distribution of eigenvalues , 4) How we can look at covariance and correlations when moments don't exist (say, as in the Cauchy case).

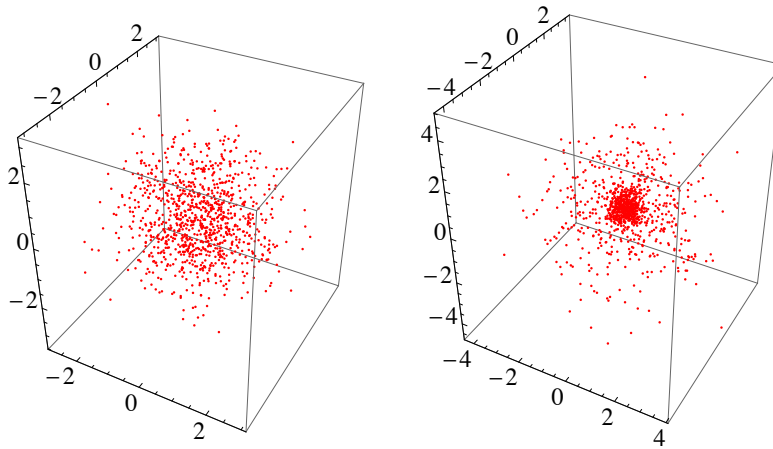


Figure 6.1: *Thick tails in higher dimensions: For a 3 dimentional vector, thin tails (left) and thick tails (right) of the same variance. In place of a bell curve with higher peak (the "tunnel") of the univariate case, we see an increased density of points towards the center.*

† Discussion chapter.

6.1 THICK TAILS IN HIGHER DIMENSION, FINITE MOMENTS

We will build the intuitions of thick tails from convexity to scale as we did in the previous chapter, but using higher dimensions.

Let $\vec{X} = (X_1, X_2, \ldots, X_m)$ be a $p \times 1$ random vector with the variables assumed to be drawn from a multivariate Gaussian. Consider the joint probability distribution $f(x_1, \ldots, x_m)$. We denote the m-variate multivariate Normal distribution by $\mathcal{N}(\vec{\mu}, \Sigma)$, with mean vector $\vec{\mu}$, variance-covariance matrix Σ, and joint pdf,

$$f\left(\vec{x}\right) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \tag{6.1}$$

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

We can apply the same simplied variance preserving heuristic as in 4.1 to fatten the tails:

$$\begin{aligned} f_a\left(\vec{x}\right) &= \frac{1}{2}(2\pi)^{-m/2}|\Sigma_1|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T {\Sigma_1}^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \\ &+ \frac{1}{2}(2\pi)^{-m/2}|\Sigma_2|^{-1/2}\exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T {\Sigma_2}^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \end{aligned} \tag{6.2}$$

where a is a scalar that determines the intensity of stochastic volatility, $\Sigma_1 = \Sigma(1+a)$ and $\Sigma_2 = \Sigma(1-a)$.[2]

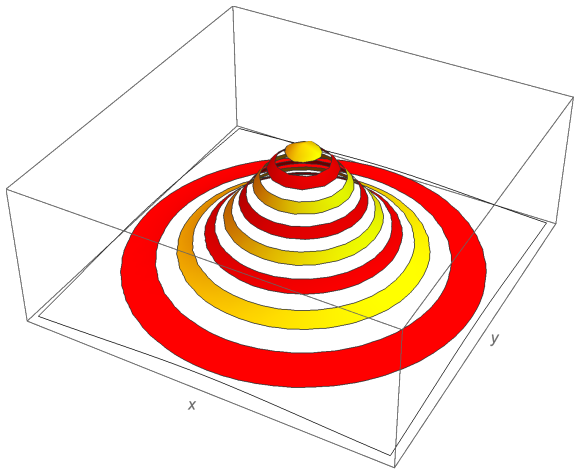


Figure 6.2: *Elliptically Contoured Joint Returns of Powerlaw (Student T)*

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

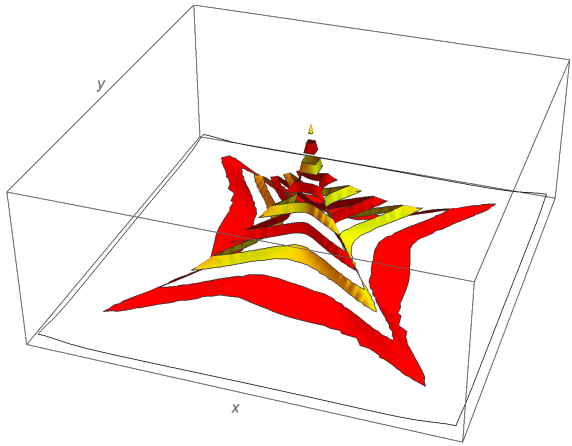


Figure 6.3: *NonElliptical Joint Returns, from stochastic correlations*

Notice in Figure 6.1, as with the one-dimensional case, a concentration in the middle part of the distribution.[3]

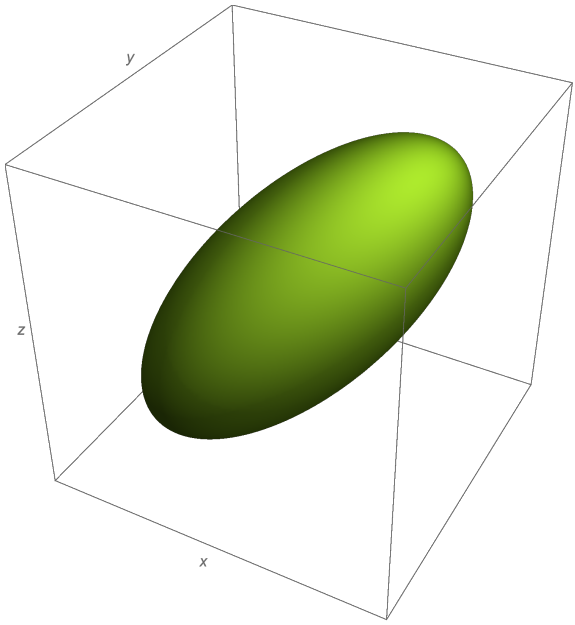


Figure 6.4: *Elliptically Contoured Joint Returns for for a multivariate distribution* (x, y, z) *solving to the same density.*

3 We created thick tails making the variances stochastic while keeping the correlations constant; this is to preserve the positive definite character of the matrix.

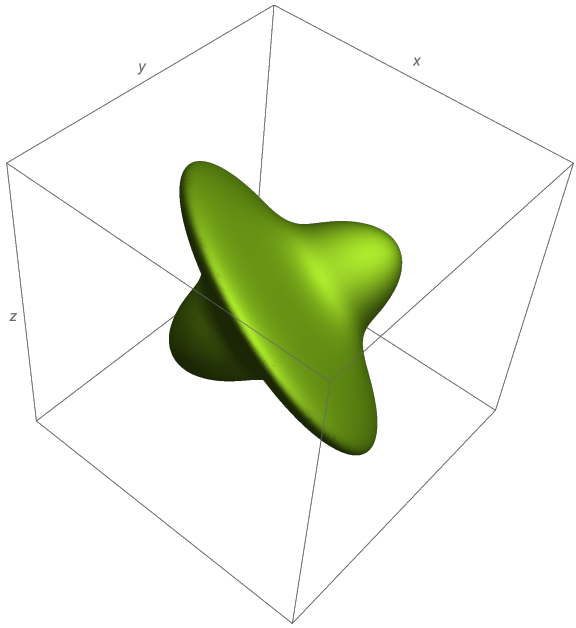


Figure 6.5: *NonElliptical Joint Returns, from stochastic correlations, for a multivariate distribution* (x, y, z)*, solving to the same density.*

6.2 JOINT FAT-TAILEDNESS AND ELLIPTICALITY OF DISTRIBUTIONS

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

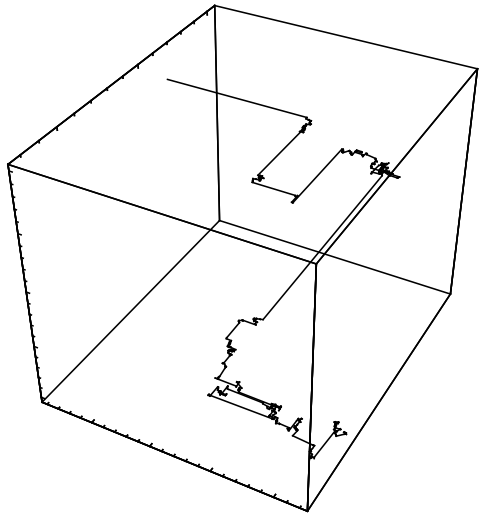


Figure 6.6: ***History moves by jumps****: A thick tailed historical process, in which events are distributed according to a power law that corresponds to the "80/20", with* $\alpha \simeq 1.13$*, represented as a 3-D Levy process.*

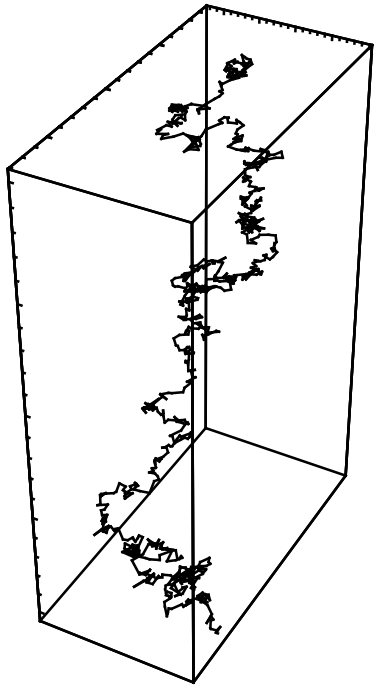


Figure 6.7: *What the proponents of "great moderation" or "long peace" have in mind: history as a thin-tailed process.*

What is an Elliptical Distribution? From the definition in [87], $\mathbf{X}$, a $p \times 1$ random vector is said to have an elliptical (or elliptical contoured) distribution with location parameters μ, a non-negative matrix Σ and some scalar function Ψ if its characteristic function is of the form $exp(it'\mu)\Psi(t\Sigma t')$.

Intuitively an elliptical distribution should show an ellipse for iso-density plots when represented in 2-D (for a bivariate) and 3-D (for a trivariate) as in Figures 6.2 and 6.4; a noneliptical would violate the shape as in Figures 6.3 and 6.5.

The main property of the class of elliptical distribution is that it is closed under linear transformation. This leads to attractive properties in the building of portfolios, and in the results of portfolio theory (in fact one cannot have portfolio theory without ellitipcality of distributions). Under ellipticality, all portfolios can be characterized completely by their location and scale and any two portfolios with identical location and scale (in return space) have identical distributions returns.

Note that (ironically) Lévy-Stable distributions are elliptical –but only in the way they are defined.

6.3 MULTIVARIATE STUDENT T

The multivariate Student T is a convenient way to model, as it collapses to the Cauchy for $\alpha = 1$. The alternative would be the multivariate stable, which, we will see, is devoid of density.

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

$$f(\mathbf{X}) = \left(\frac{(\mathbf{X} - \mathbf{M}).\mathbf{\Sigma}^{-1}.(\mathbf{X} - \mathbf{M})}{\nu} + 1 \right)^{-\frac{1}{2}(\nu+p)}. \tag{6.3}$$

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

$$f(x_1, x_2) = \frac{\nu\sqrt{1-\rho^2}\left(\frac{-\nu\rho^2+\nu-2\rho x_1 x_2+x_1^2+x_2^2}{\nu-\nu\rho^2}\right)^{-\frac{\nu}{2}-1}}{2\pi\left(\nu-\nu\rho^2\right)}. \tag{6.4}$$

6.3.1 Ellipticality and Independence under Thick Tails

Take the product of two Cauchy densities for x and y (what we used in Fig. 3.1):

$$f(x)f(y) = \frac{1}{\pi^2\left(x^2+1\right)\left(y^2+1\right)} \tag{6.5}$$

which, patently, as we saw in Chapter 3 (with the example of the two randomly selected persons with a total net worth of $36 million), is not elliptical. Compare to the joint distribution $f_\rho(x, y)$:

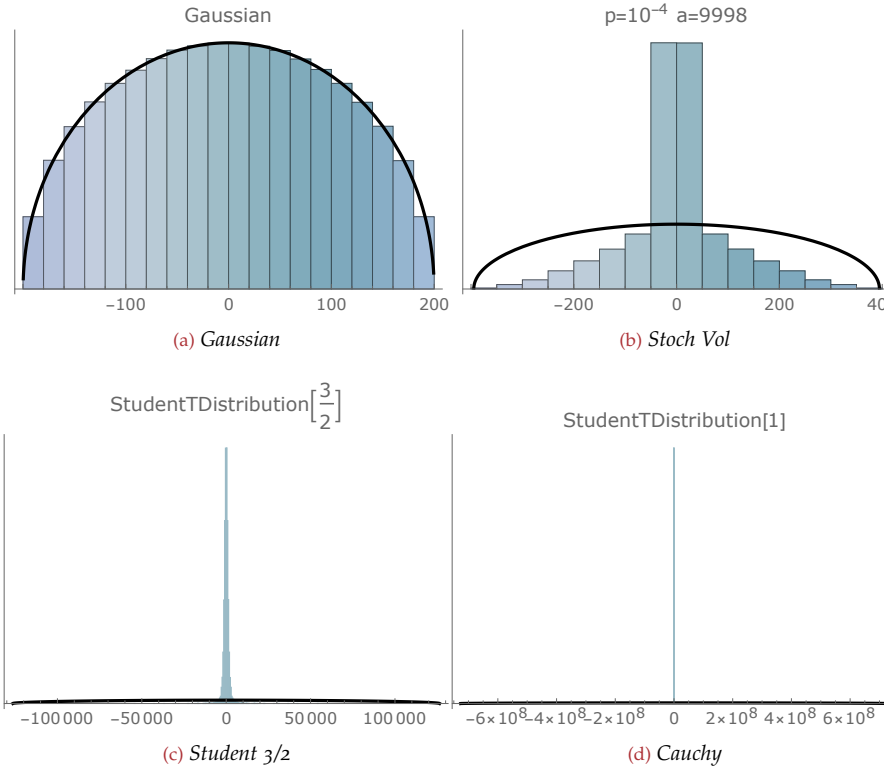
$$f_\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\left(y\left(\frac{y}{1-\rho^2} - \frac{\rho x}{1-\rho^2}\right) + x\left(\frac{x}{1-\rho^2} - \frac{\rho y}{1-\rho^2}\right) + 1\right)^{3/2}}, \tag{6.6}$$

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

$$f_0(x, y) = \frac{1}{2\pi\left(x^2+y^2+1\right)^{3/2}} \tag{6.7}$$

which is elliptical. This shows that absence of correlation is not independence as:

> Independence between two variables X and Y is defined by the identity:
>
> $$\frac{f(x, y)}{f(x)f(y)} = 1,$$
>
> regardless of the correlation coefficient. In the class of elliptical distributions, the bivariate Gaussian with coefficient 0 is both independent and uncorrelated. This does not apply to the Student T or the Cauchy.



(a) *Gaussian* (b) *Stoch Vol*

(c) *Student 3/2* (d) *Cauchy*

Figure 6.8: *The various shapes of the distribution of the eigenvalues for random matrices, which in the Gaussian case follow the Wigner semicircle distribution. The Cauchy case corresponds to the Student parametrized to have 1 degrees of freedom.*

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

A random vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is said to have the multivariate stable distribution if every linear combination of its components $\mathbf{Y} = a_1 X_1 + \cdots + a_k X_k$ has a stable distribution. That is, for any constant vector $\mathbf{a} \in \mathbb{R}^k$, the random variable $\mathbf{Y} = a^T \mathbf{X}$ should have a univariate stable distribution. And to have a linear combination remain within the same class requires ellipticality. Hence *by construction*, $f_0(x, y)$ is not necessarily equal to $f(x)f(y)$. Consider the Cauchy case that has an explicit density function. The denominator of the product of densities includes an additional term, x^2y^2, which pushes the iso-densities in one direction or another, as we saw in the introductory examples of Chapter 3.

6.4 FAT TAILS AND MUTUAL INFORMATION

We notice that because of the artificiality in constructing multivariate distributions, mutual information is not 0 in the presence of independence, since the ratio of joint densities/product of densities $\neq 1$ under 0 "correlation" ρ.

What is the mutual information of a Student T (which includes the Cauchy)?

$$\mathbb{I}(X,Y) = \mathbb{E}\log\left(\frac{f(x,y)}{f(x)f(y)}\right)$$

where the expectation is taken under the joint distribution for X and y. The mutual information thanks to the log becomes additive (Note that one can use any logarithmic base and translate by dividing by log(2)).

So $\mathbb{I}(X,Y) = \mathbb{E}\left(logf(x,y)\right) - \mathbb{E}log\left(f(x)\right) - \mathbb{E}log\left(f(y)\right)$ or $\mathbb{H}(X)$+$\mathbb{H}(Y)$ -$\mathbb{H}(X,Y)$ where $\mathbb{H}$ is the entropy and $\mathbb{H}(X,Y)$ the joint entropy.

We note that $-\frac{1}{2}\log(1-\rho^2)$ is the mutual information of a Gaussian regardless of parametrization. So for $X, Y \sim$ Multivariate Student T (α, ρ), the mutual information $\mathbb{I}_\alpha(X,Y)$:

$$\mathbb{I}_\alpha(X,Y) = -\frac{1}{2}\log\left(1-\rho^2\right) + \lambda_\alpha \tag{6.8}$$

where

$$\lambda_\alpha = -\frac{2}{\alpha} + \log(\alpha) + 2\pi(\alpha+1)\csc(\pi\alpha) + 2\log\left(B\left(\frac{\alpha}{2}, \frac{1}{2}\right)\right) - (\alpha+1)H_{-\frac{\alpha}{2}} + (\alpha+1)H_{-\frac{\alpha}{2}-\frac{1}{2}} - 1 - \log(2\pi) \tag{6.9}$$

where csc(.) is the cosecant of the argument, $B(.,.)$ is the beta function and $H(.)^{(r)}$ is the harmonic number $H_n^r = \sum_{i=1}^{n} \frac{1}{i^r}$ with $H_n = H_n^{(1)}$. We note that $\lambda_\alpha \underset{\alpha\to\infty}{\to} 0$.

To conclude this brief section metrics linked to entropy such as mutual information are vastly more potent than correlation; mutual information can detect nonlinearities.

6.5 FAT TAILS AND RANDOM MATRICES, A RAPID INTERLUDE

The eigenvalues of matrices themselves have an analog to Gaussian convergence: the semi-circle distribution, as shown in Figure 6.8.

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

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

$$f(\lambda) = \frac{2}{\pi R^2}\sqrt{R^2-\lambda^2} \text{ for } -R \leq \lambda \leq R. \tag{6.10}$$

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

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

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

6.6 CORRELATION AND UNDEFINED VARIANCE

Next we examine a paradox: while covariances can be infinite, correlation is finite. However, it will have a huge sampling error to be informative –same problem we discussed with PCA in Chapter 3.

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

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

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

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

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

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

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

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

$$\lim_{\substack{x_n \to +\infty \\ y_n \to +\infty}} \frac{x_n y_n + \kappa_1}{\sqrt{x_n^2 + \kappa_2}\sqrt{y_n^2 + \kappa_3}} = 1$$

$$\lim_{\substack{x_n \to +\infty \\ y_n \to -\infty}} \frac{x_n y_n + \kappa_1}{\sqrt{x_n^2 + \kappa_2}\sqrt{y_n^2 + \kappa_3}} = -1$$

and

$$\lim_{\substack{x_n \to -\infty \\ y_n \to +\infty}} \frac{x_n y_n + \kappa_1}{\sqrt{x_n^2 + \kappa_2}\sqrt{y_n^2 + \kappa_3}} = -1$$

for all values of $n \geq 2$.

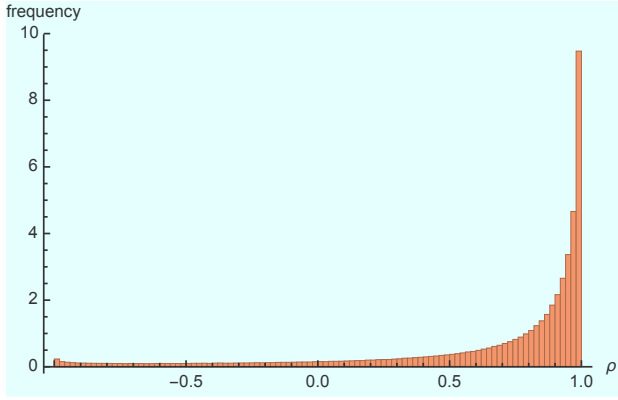


Figure 6.9: *Sample distribution of correlation for a sample of* 10^3. *The correlation exists for a bivariate T distribution (exponent* $\frac{2}{3}$, *correlation* $\frac{3}{4}$ *) but... not useable*

An example of the distribution of correlation is shown in Fig. 6.9. Finite correlation doesn't mean low variance: it exists, but may not be useful for statistical purpose owing to the noise and slow convergence.

6.7 FAT TAILED RESIDUALS IN LINEAR REGRESSION MODELS

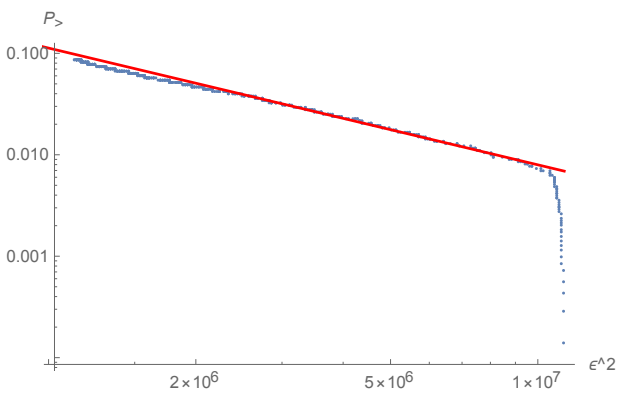


Figure 6.10: *The log-log-plot of the survival function of thee squared residuals* ϵ^2 *for the IQ-income linear regression using the standard Winsconsin Longitudinal Studies (WLS) data. We notice that the income variables are winsorized. Clipping the tails creates the illusion of a high* R^2. *Actually, even without clipping the tail, the coefficient of determination will show much higher values owing to the small sample properties for the variance of a power law.*

We mentioned in Chapter 3 that linear regression fails to inform under fat tails. Yet it is practiced. For instance, it is patent that income and wealth variables are power law distributed (with a spate of problems, see our Gini discussions in 13). However IQ scores are Gaussian (seemingly by design). Yet people regress one on the other failing to see that it is improper.

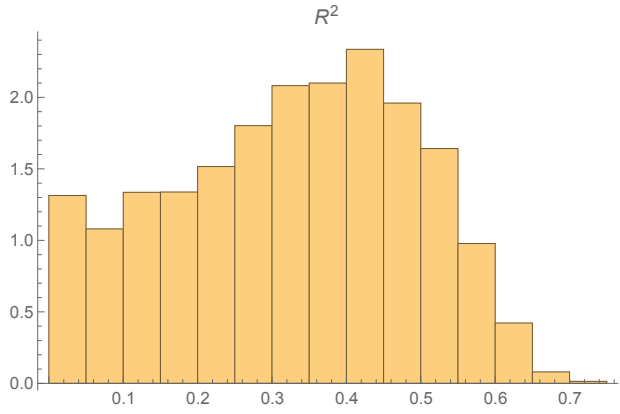


Figure 6.11: *An infinite variance case that shows a high R^2 in sample; but it ultimately has a value of 0. Remember that R^2 is stochastic. The problem greatly resembles that of P values in Chapter 19 owing to the complication of a metadistribution in* $[0, 1]$

Consider the following linear regression in which the independent and indepen-dent are of different classes:

$$Y = aX + b + \epsilon,$$

where X is standard Gaussian ($\mathcal{N}(0, 1)$) and ϵ is power law distributed, with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) < +\infty$. There are no restrictions on the parameters.

Clearly we can compute the coefficient of determination R^2 as 1 minus the ratio of the expectation of the sum of residuals over the total squared variations, so we get the more general answer to our idiosyncratic model. Since $X \sim \mathcal{N}(0, 1)$, $aX + b \sim \mathcal{N}(b, |a|)$, we have

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - (ax_i + b + \epsilon_i)\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}.$$

We can show that, for large n

$$R^2 = \frac{a^2}{a^2 + \mathbb{E}(\epsilon_i^2)} + O\left(\frac{1}{n^2}\right). \tag{6.11}$$

And of course, for infinite variance:

$$\lim_{E(\epsilon^2) \to +\infty} \mathbb{E}(R^2) = 0.$$

When ϵ is T-distributed with α degrees of freedom, clearly ϵ^2 will follow an FRatio distribution $(1, \alpha)$ –a power law with exponent $\frac{\alpha}{2}$.

Note that we can also compute the same "expectation" by taking, simply, the square of the correlation between X and Y. For instance, assume the distribution for ϵ is the Student T distribution with zero mean, scale σ and tail exponent $\alpha > 2$ (as we saw earlier, we get identical results with other ones so long as we constrain the mean to be 0). Let's start by computing the correlation: the numerator is the co-

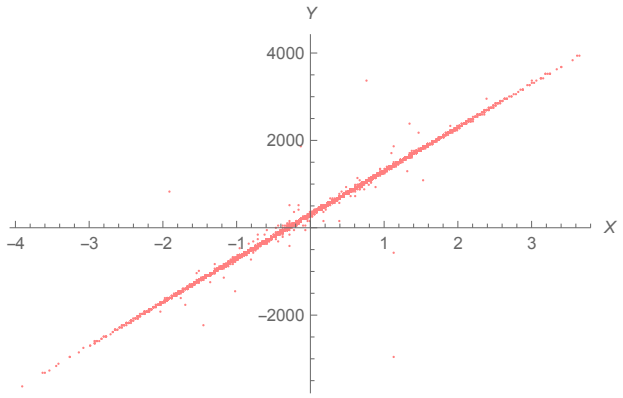


Figure 6.12: *A Cauchy regression with an expected $R^2 = 0$, faking it but showing higher values in small samples (here .985).*

variance $Cov(X, Y) = \mathbb{E}\left((aX + b + \epsilon)X\right) = a$. The denominator (standard deviation for Y) becomes $\sqrt{\mathbb{E}\left(((aX + \epsilon) - a)^2\right)} = \sqrt{\frac{2\alpha a^2 - 4a^2 + \alpha\sigma^2}{\alpha - 2}}$. So

$$\mathbb{E}(R^2) = \frac{a^2(\alpha - 2)}{2(\alpha - 2)a^2 + \alpha\sigma^2} \tag{6.12}$$

And the limit from above:

$$\lim_{\alpha \to 2^+} \mathbb{E}(R^2) = 0.$$

We are careful here to use $\mathbb{E}(R^2)$ rather than the seemingly deterministic R^2 because it is a stochastic variable that will be extremely sample dependent, and only stabilize for large n, perhaps even astronomically large n. Indeed, recall that *in sample* the expectation will always be finite, even if the ϵ are Cauchy! The point is illustrated in Figures 6.11 and 6.12. Actually, when one uses the maximum likelihood estimation of R^2 via $\mathbb{E}\left(\epsilon^2\right)$ using α, (the "shadow mean" method in Chapters 13 and 14, among others) we notice that in the IQ example used in the graph, the mean of the sample residuals are about half of the maximum likelihood one, making R^2 even lower (that is, virtually 0)[4].

The point invalidates much studies of the relations IQ-wealth and IQ-income of the kind [263]; we can see the striking effect in Fig. 6.10. Given that R is bounded in $[0, 1]$, it will reach its true value very slowly – see the P-Value problem in Chapter 19.

Property 3

When a fat tailed random variable is regressed against a thin tailed one, the coefficient of determination R^2 will be biased higher, and requires a much larger sample size to converge (if it ever does).

Note that sometimes people try to solve the problem by some nonlinear transformation of a random variable (say, the logarithm) to try to establish a linear

4 $2.2\ 10^9$ vs $1.24\ 10^9$.

relationship. If the required transformation is exact, things will be fine –but only if exact. Errors can arise from the discrepancy. For correlation is extremely delicate and unlike mutual information, non-additive and often uninformative. The point has been explored by this author in [235].

NEXT

We will examine in chapter 8 the slow convergence of power laws distributed variables under the law of large numbers (LLN): it can be as much as 10^{13} times slower than the Gaussian.

A SPECIAL CASES OF THICK TAILS

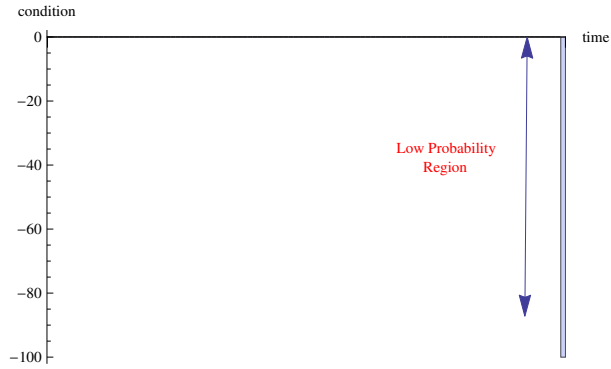


Figure A.1: *A coffee cup is less likely to incur "small" than large harm. It shatters, hence is exposed to (almost) everything or nothing. The same type of payoff is prevalent in markets with, say, (reval)devaluations, where small movements beyond a barrier are less likely than larger ones.*

FOR UNIMODAL distributions, thick 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.

A.1 MULTIMODALITY AND THICK TAILS, OR THE WAR AND PEACE MODEL

We noted earlier in 4.1 that stochasticizing (that is, making a deterministic variable stochastic), ever so mildly, variances, the distribution gains in thick tailedness (as expressed by kurtosis). But we maintained the same mean.

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

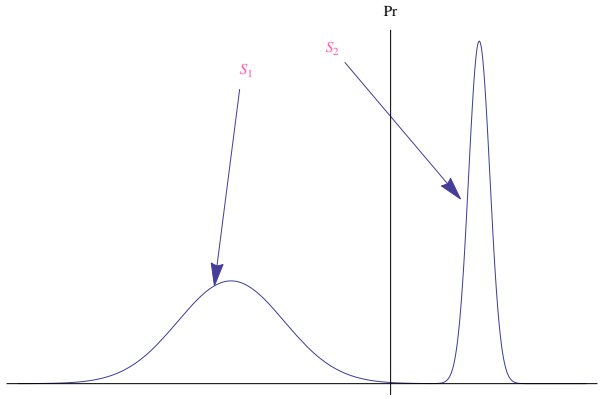


Figure A.2: *The War and peace model. Kurtosis =1.7, much lower than the Gaussian.*

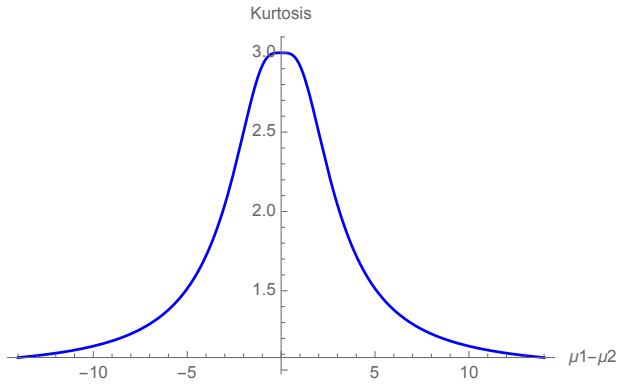


Figure A.3: *Negative (relative) kurtosis and bimodality (3 is the Gaussian).*

Either the very meaning of "thick tails" loses its significance under multimodality, or takes on a new one where the "middle", around the expectation ceases to matter.[7, 154].

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

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

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

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

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

the stochasticity of the mean offsets the stochasticity of volatility.

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

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

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

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

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

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

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

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

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

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

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 PDF 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.

The same idea applies to currency pegs, as devaluations cannot be "mild", with all-or- nothing type of volatility and low density in the "valley" between the two distinct regimes.

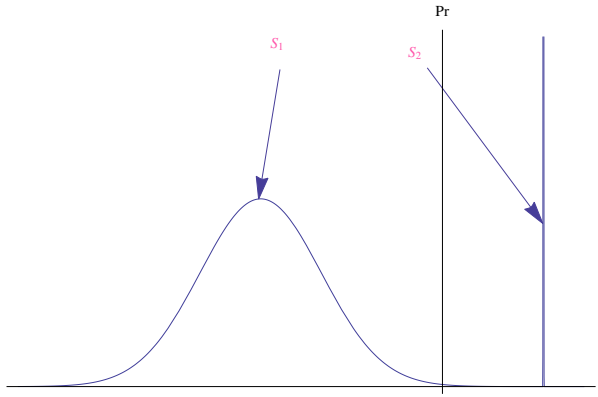


Figure A.4: *The Bond payoff/Currency peg model. Absence of volatility stuck at the peg, deterministic payoff in regime 2, mayhem in regime 1. Here the kurtosis K=2.5. Note that the coffee cup is a special case of both regimes 1 and 2 being degenerate .*

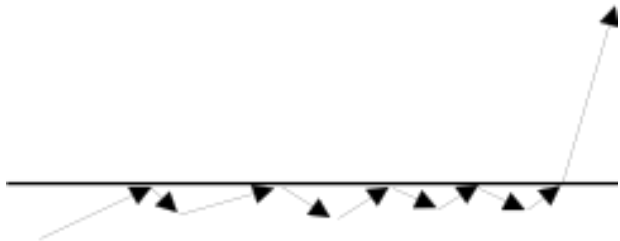


Figure A.5: *Pressure on the peg which may give a Dirac PDF in the "no devaluation" regime (or, equivalently,low volatility). It is typical for finance imbeciles to mistake regime S_2 for low volatility.*

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

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

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

A brief list of other situations where bimodality is encountered:

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

A.2 TRANSITION PROBABILITIES: WHAT CAN BREAK WILL BREAK

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

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

After n periods, that is, n steps,

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

Where

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

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

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

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

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

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

and the "ergodic" probabilities:

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

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

With the "ergodic" matrix,

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

where $\mathbb{1}^{\mathsf{T}}$ is the transpose of unitary vector $\{1,1\}$, π the matrix of eigenvectors. The eigenvalues become $\lambda = \begin{pmatrix} 1 \\ p_{1,1} + p_{2,2} - 1 \end{pmatrix}$ and associated eigenvectors $\pi = \begin{pmatrix} 1 & 1 \\ \frac{1-p_{1,1}}{1-p_{2,2}} & 1 \end{pmatrix}$.

Part II

THE LAW OF MEDIUM NUMBERS

7 LIMIT DISTRIBUTIONS, A CONSOLIDATION*,†

In this expository chapter we proceed to consolidate the literature on limit distributions seen from our purpose, with some shortcuts where indicated. After introducing the law of large numbers, we show the intuition behind the central limit theorem and illustrate how it varies preasymptotically across distributions. Then we discuss the law of large numbers as applied to higher moments. A more formal and deeper approach will be presented in the next chapter.

Both the law of large numbers and the central limit theorem are partial answers to a general problem: "What is the limiting behavior of a sum (or average) of random variables as the number of summands approaches infinity?". And our law of medium numbers (or preasymptotics) is: now what when the number of summands doesn't reach infinity?

7.1 REFRESHER: THE WEAK AND STRONG LLN

The standard presentation is as follows. Let $X_1, X_2, \ldots$ be an infinite sequence of independent and identically distributed (Lebesgue integrable) random variables with expected value $\mathbb{E}(X_n) = \mu$ (we will see further down one can somewhat relax the i.i.d. assumptions). For all n, the sample average $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ converges to the expected value, $\overline{X}_n \to \mu$,for $n \to \infty$.

Finiteness of variance is not necessary (though of course the finite higher moments accelerate the convergence).

There are two modes of convergence: convergence in probability $\xrightarrow{P}$ (which implies convergence in distribution, though not always the reverse), and the stronger $\xrightarrow{a.s.}$ almost sure convergence (similar to pointwise convergence) (or almost every-

* Discussion chapter (with some research).

where or almost always). Applied here the distinction corresponds to the weak and strong LLN respectively.

The weak LLN The weak law of large numbers (or Kinchin's law, or sometimes called Bernouilli's law) can be summarized as follows: the probability of a variation in excess of some threshold for the average becomes progressively smaller as the sequence progresses. In estimation theory, an estimator is called consistent if it thus converges in probability to the quantity being estimated.

$$\overline{X}_n \xrightarrow{P} \mu \text{ when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| > \varepsilon \right) = 0.$$

Note that standard proofs are based on Chebyshev's inequality: if X has a finite non-zero variance σ^2. Then for any real number $k > 0$,

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

The strong LLN The strong law of large numbers states that, as the number of summands n goes to infinity, the probability that the average converges to the expectation equals 1.

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu \text{ when } n \to \infty.$$

That is,

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu \right) = 1.$$

Relaxations of i.i.d. Now one can relax the identically distributed assumption under some conditions: Kolmogorov's proved that non identical distributions for the summands X_i require for each summand the existence of a finite second moment.

As to independence, some weak dependence is allowed. Traditionally the conditions are, again, the usual finite variance 1) $\mathbb{V}(X_i) \leq c$ and some structure on the covariance matrix, 2) $\lim_{|i-j|\to+\infty} \operatorname{Cov}(X_i, X_j) = 0$.

However it turns out 1) can be weakened to $\sum_{i=1}^{n} \mathbf{V}[X_i] = o(n^2)$, and 2) $|\operatorname{Cov}(X_i, X_j)| \leq \varphi(|i-j|)$, where $\frac{1}{n}\sum_{i=1}^{n} \varphi(i) \to 0$. See Bernstein [19] and Kozlov [146] (in Russian).[2]

2 Thanking "romanoved", a mysterious Russian speaking helper on Mathematics Stack Exchange.

Our Interest Our concern in this chapter and the next one is clearly to look at the "speed" of such convergence. Note that under the stronger assumption of i.i.d. we do not need variance to be finite, so we can focus on mean absolute deviation as a metric for divergence.

7.2 CENTRAL LIMIT IN ACTION

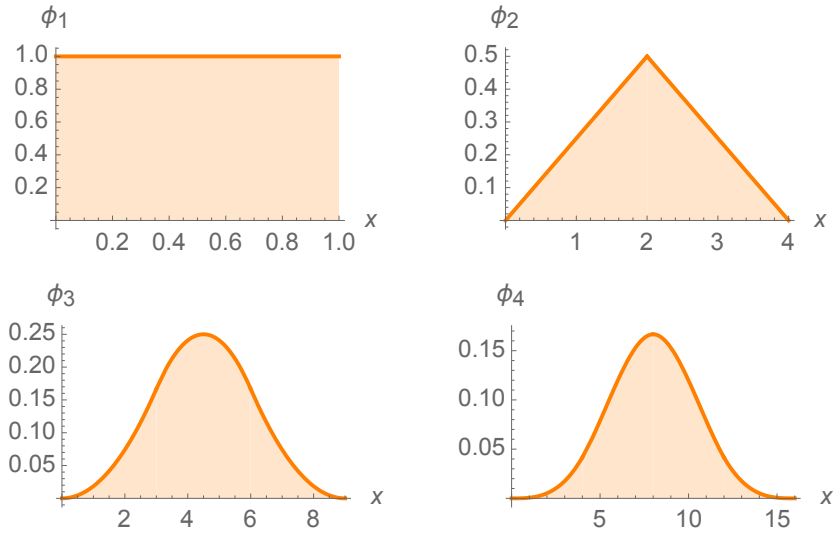


Figure 7.1: *The fastest CLT: the Uniform becomes Gaussian in a few steps. We have, successively, 1, 2, 3, and 4 summands. With 3 summands we see a well formed bell shape.*

We will start with a simplification of the generalized central limit theorem (GCLT), as formulated by Paul Lévy (the traditional approaches to CLT as well as the technical backbone will be presented later):

7.2.1 The Stable Distribution

Using the same notation as above, let $X_1, \ldots, X_n$ be independent and identically distributed random variables. Consider their sum S_n. We have

$$\frac{Sn - a_n}{b_n} \xrightarrow{D} X_s, \tag{7.1}$$

where X_s follows a stable distribution $\mathcal{S}$, a_n and b_n are norming constants, and, to repeat, $\xrightarrow{D}$ denotes convergence in distribution (the distribution of X as $n \to \infty$). The properties of $\mathcal{S}$ will be more properly defined and explored in the next chapter. Take it for now that a random variable X_s follows a stable (or α-stable) distribution,

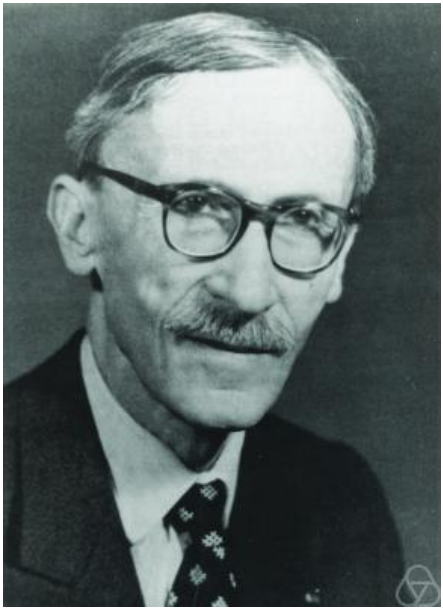


Figure 7.2: *Paul Lévy, 1886-1971, formulated the generalized central limit theorem.*

symbolically $X_s \sim S(\alpha_s, \beta, \mu, \sigma)$, if its characteristic function $\chi(t) = \mathbb{E}(e^{itX_s})$ is of the form:

$$\chi(t) = e^{\left(i\mu t - |t\sigma|_s^{\alpha}\left(1 - i\beta\tan\left(\frac{\pi\alpha_s}{2}\right)\text{sgn}(t)\right)\right)} \text{ when } \alpha_s \neq 1. \tag{7.2}$$

The constraints are $-1 \leq \beta \leq 1$ and $0 < \alpha_s \leq 2$.[3]

The designation stable distribution implies that the distribution (or class) is stable under summation: you sum up random variables following any the various distributions that are members of the class $\mathfrak{S}$ explained next chapter (actually the same distribution with different parametrizations of the characteristic function), and you stay within the same distribution. Intuitively, $\chi(t)^n$ is the same form as $\chi(t)$, with $\mu \to n\mu$, and $\sigma \to n^{\frac{1}{\alpha}}\sigma$. The well known distributions in the class (or some people call it a "basin") are: the Gaussian, the Cauchy and the Lévy with $\alpha = 2, 1$, and $\frac{1}{2}$, respectively. Other distributions have no closed form density.[4]

7.2.2 The Law of Large Numbers for the Stable Distribution

Let us return to the law of large numbers.

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

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

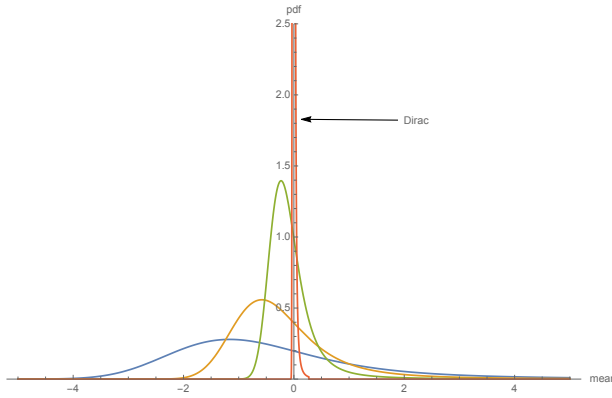


Figure 7.3: *The law of large numbers show a tightening distribution around the mean leading to degeneracy converging to a Dirac stick at the exact mean.*

By standard results, we can observe the law of large numbers at work for the stable distribution, as illustrated in Fig. 7.3:

$$\lim_{n\to+\infty} \chi\left(\frac{t}{n}\right)^n = e^{i\mu t},\ 1 < \alpha_s \leq 2 \tag{7.3}$$

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

$$\frac{1}{\sqrt{2\pi}}\mathcal{F}_t\left(e^{i\mu t}\right)(x) = \delta(\mu + x). \tag{7.4}$$

Further, we can observe the "real-time" operation for all $1 < n < +\infty$ in the following ways, as we will explore in the next sections.

7.3 SPEED OF CONVERGENCE OF CLT: VISUAL EXPLORATIONS

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

7.3.1 Fast Convergence: the Uniform Dist.

Consider a uniform distribution –the simplest of all. If its support is in $[0, 1]$, it will simply have a density of $\phi(x_1) = 1$ for $0 \leq x_1 \leq 1$ and integrates to 1. Now add another variable, x_2, identically distributed and independent. The sum $x_1 + x_2$ immediately changed in shape! Look at $\phi_2(.)$, the density of the sum in Fig. 7.1. It is now a triangle. Add one variable and now consider the density ϕ_3 of the distribution of $X_1 + X_2 + X_3$. It is already almost bell shaped, with $n = 3$ summands.

The uniform sum distribution

$$\phi_n(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(\frac{x-L}{H-L} - k \right)^{n-1} \operatorname{sgn}\left(\frac{x-L}{H-L} - k \right) \text{ for } nL \leq x \leq nH$$

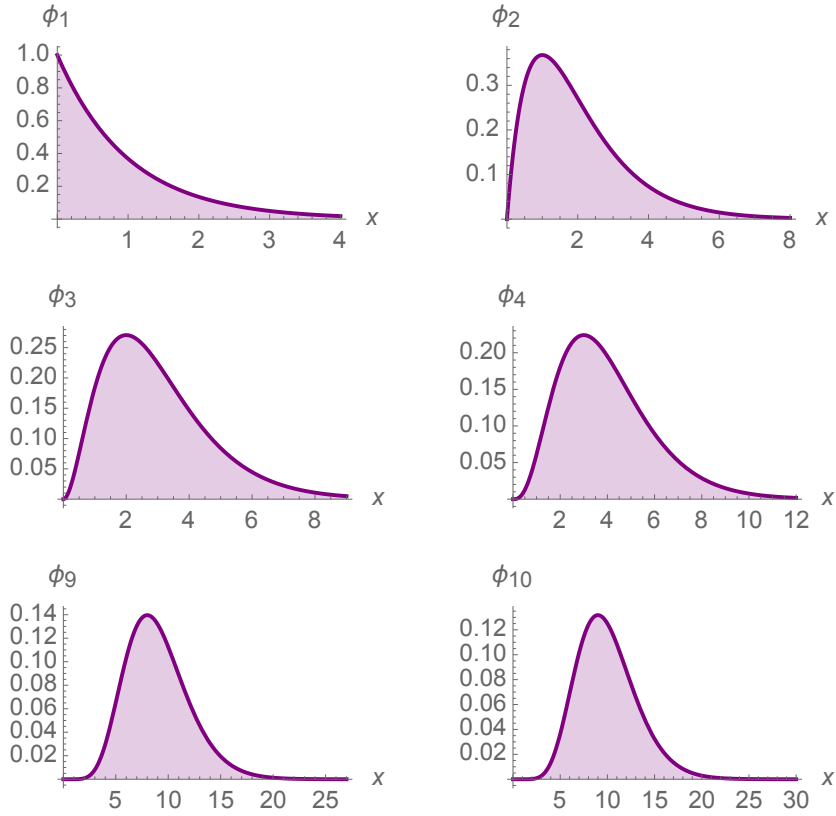


Figure 7.4: *The exponential distribution,ϕ indexed by the number of summands. Slower than the uniform, but good enough.*

7.3.2 Semi-slow convergence: the exponential

Let us consider a sum of exponential random variables.

We have for initial density

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

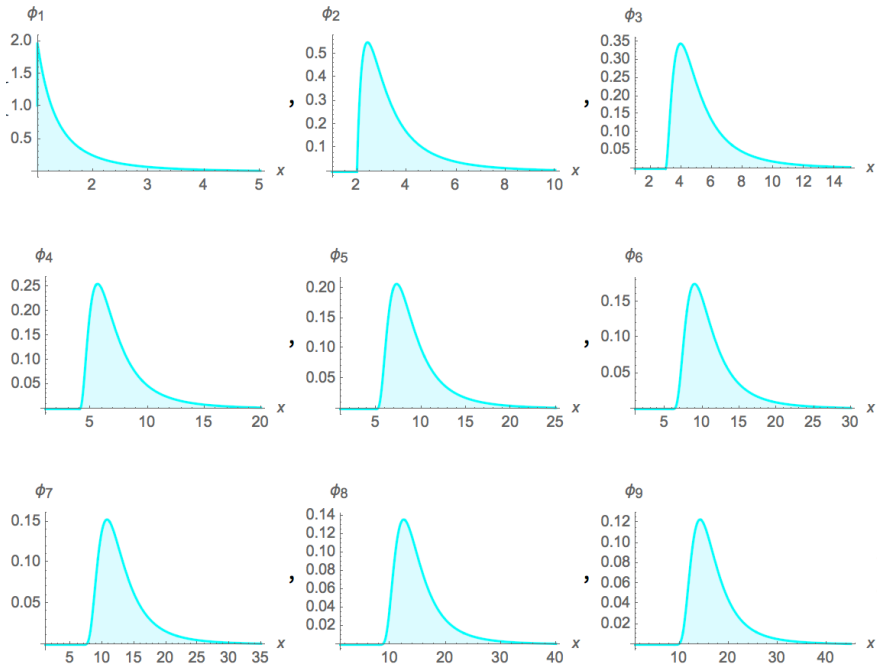


Figure 7.5: *The Pareto distribution. Doesn't want to lose its skewness, although in this case it should converge to the Gaussian... eventually.*

and for n summands[5]

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

We have, replacing x by n/λ (and later in the illustrations in Fig. 7.4 $\lambda = 1$),

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

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

We can see how we get more slowly to the Gaussian, as shown in Fig. 7.4, mostly on account of its skewness. Getting to the Gaussian requires symmetry.

7.3.3 The slow Pareto

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

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

5 We derive the density of sums either by convolving, easy in this case, or as we will see with the Pareto, via characteristic functions.

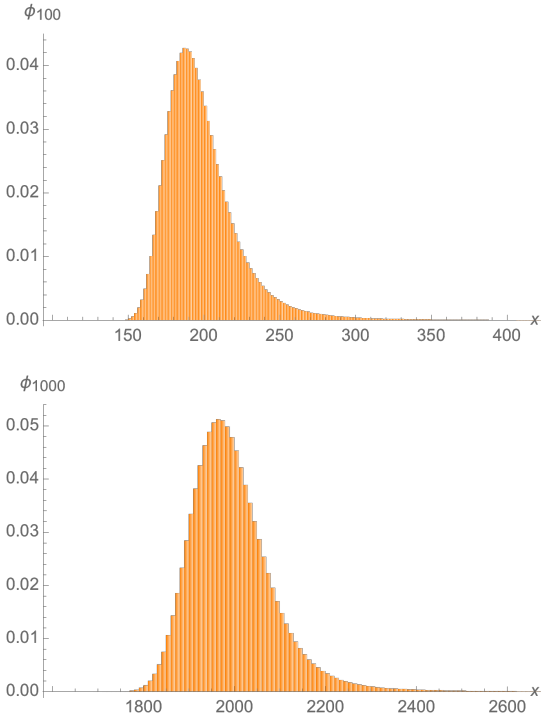


Figure 7.6: *The Pareto distribution, ϕ_{100} and ϕ_{1000}, not much improvement towards Gaussianity, but an $\alpha = 2$ will eventually get you there if you are patient and have a long, very long, life.*

and inverting the characteristic function,

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

Where $E_{(.)}(.)$ is the exponential integral $E_n(z) = \int_1^\infty \frac{dte^{t(-z)}}{t^n}$. Clearly, the integration is done numerically (so far nobody has managed to pull out the distribution of a Pareto sum). It can be exponentially slow (up to 24 hours for $n = 50$ vs. 45 seconds for $n = 2$), so we have used Monte Carlo simulations for Figs. 7.3.1.

Recall from Eq. 7.1 that the convergence requires norming constants a_n and b_n. From Uchaikin and Zolotarev [252], we have (narrowing the situation for $1 < \alpha_p \leq 2$):

$$\mathbb{P}(X > x) = cx^{-\alpha_p}$$

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

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

as $x \to \infty$. The norming constants become $a_n = n\ \mathbb{E}(X)$ for $\alpha_p > 1$ (for other cases, consult [252] as these are not likely to occur in practice), and

$$b_n = \begin{cases} \pi n^{1/\alpha_p} \left(2 \sin\left(\frac{\pi\alpha_p}{2}\right) \Gamma(\alpha_p)\right)^{-\frac{1}{\alpha_p}} (c+d)^{1/\alpha_p} & \text{for } 1 < \alpha_p < 2 \\ \sqrt{c+d}\sqrt{n \log(n)} & \text{for } \alpha_p = 2 \end{cases}. \tag{7.5}$$

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

7.3.4 The half-cubic Pareto and its basin of convergence

Of interest is the case of $\alpha = \frac{3}{2}$. Unlike the situations where as in Fig. 7.3.1, the distribution ends up slowly being symmetric. But, as we will cover in the next chapter, it is erroneous to conflate its properties with those of a stable. It is, in a sense, more fat-tailed.

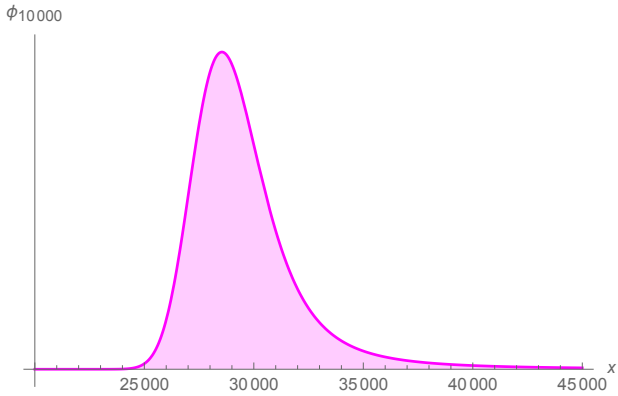


Figure 7.7: *The half-cubic Pareto distribution never becomes symmetric in real life. Here* $n = 10^4$

7.4 CUMULANTS AND CONVERGENCE

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

Definition 7.1 (Excess p-cumulants)
Let $\chi(\omega)$ *be characteristic functionof a given distribution,* n *the number of summands (for independent random variables),* p *the order of the moment. We define the ratio of cumulants for the corresponding* p^{th} *moment:*

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

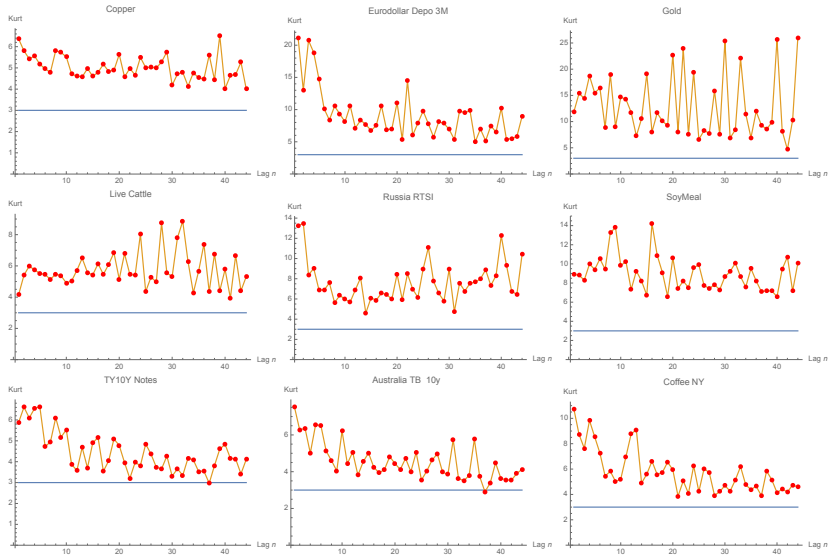


Figure 7.8: *Behavior of the 4^{th} moment under aggregation for a few financial securities deemed to converge to the Gaussian but in fact do not converge (backup data for [225]). There is no conceivable way to claim convergence to Gaussian for data sampled at a lower frequency.*

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

Remark 6

We note that

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

for all probability distributions outside the Power Law class.

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

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

A sketch of the proof can be done using the stable distribution as the limiting basin and the nonderivability at order p greater than its tail index, using Eq. 8.4.

Table 7.1 shows what happens to the cumulants $K(.)$ for n-summed variables.

We would expect a drop at a rate $\frac{1}{N^2}$ for stochastic volatility (gamma variance wlog). However, figure 10.2 shows the drop does not take place at any such speed. Visibly we are not in the basin. As seen in [225] there is an absence of convergence of kurtosis under summation across economic variables.

Table 7.1: *Table of Normalized Cumulants For Thin Tailed Distributions Speed of Convergence for N Independent Summands*

-

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

7.5 TECHNICAL REFRESHER: TRADITIONAL VERSIONS OF CLT

This is a refresher of the various approaches bundled under the designation CLT.

The Standard (Lindeberg-Lévy) version of CLT Suppose as before a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2 < +\infty$, and $\overline{X}_n$ the sample average for n. Then as n approaches infinity, the sum of the random variables $\sqrt{n}(\overline{X}_n\mu)$ converges in distribution to a Gaussian [20] [21]:

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

Convergence in distribution means that the CDF (cumulative distribution function) of $\sqrt{n}$ converges pointwise to the CDF of $\mathcal{N}(0, \sigma)$ for every real z,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}(\overline{X}_n - \mu) \leq z\right) = \lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq \frac{z}{\sigma}\right] = \Phi\left(\frac{z}{\sigma}\right),\ \sigma > 0$$

where $\Phi(z)$ is the standard normal cdf evaluated ar z. Note that the convergence is uniform in z in the sense that

$$\lim_{n\to\infty} \sup_{z\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}(\overline{X}_n - \mu) \leq z\right) - \Phi\left(\frac{z}{\sigma}\right)\right| = 0,$$

where sup denotes the least upper bound, that is, the supremum of the set.

Lyapunov's CLT In Lyapunov's derivation, summands have to be independent, but not necessarily identically distributed. The theorem also requires that random variables $|'X_i|$ have moments of some order $(2 + \delta$, and that the rate of growth of these moments is limited by the Lyapunov condition given below.

The condition is as follows. Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

If for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left(|X_i - \mu_i|^{2+\delta}\right) = 0,$$

then a sum of $\frac{X_i-\mu_i}{s_i}$ converges in distribution to a standard normal random variable, as n goes to infinity:

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{D} N(0,1).$$

If a sequence of random variables satisfies Lyapunov's condition, then it also satisfies Lindeberg's condition that we cover next. The converse implication, however, does not hold.

Lindeberg's condition Lindeberg allows to reach CLT under weaker assumptions. With the same notations as earlier:

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left((X_i - \mu_i)^2 \cdot \mathbb{1}_{\{|X_i-\mu_i|>\varepsilon s_n\}}\right) = 0$$

for all $\varepsilon > 0$, where $\mathbb{1}$ indicator function, then the random variable $Z_n = \frac{\sum_{i=1}^{n}(X_i-\mu_i)}{s_n}$ converges in distribution] to a Gaussian as $n \to \infty$.

Lindeberg's condition is sufficient, but not in general necessary except if the sequence under consideration satisfies:

$$\max_{1\le k\le n} \frac{\sigma_i^2}{s_n^2} \to 0, \quad \text{as } n \to \infty,$$

then Lindeberg's condition is both sufficient and necessary, i.e. it holds if and only if the result of central limit theorem holds.

7.6 THE LAW OF LARGE NUMBERS FOR HIGHER MOMENTS

7.6.1 Higher Moments

A test of fat tailedness can be seen by applying the law of large number to higher moments and see how they converge. A visual examination of the behavior of the cumulative mean of the moment can be done in a similar way to the standard visual tests of LLN we saw in Chapter 3– except that it applies to X^p (raw or centered) rather than X. We check the functioning of the law of large numbers by seeing if adding observations causes a reduction of the variability of the average (or its variance if it exists). Moments that do not exist will show occasional jumps – or, equivalently, large subsamples will produce different averages. When moments exist, adding observations eventually prevents further jumps.

Another visual technique is to consider the contribution of the maximum observation to the total, and see how it behaves as n grows larger. It is called the MS plot [114], "maximum to sum", and shown in Figure 7.9.

Table 7.2: *Kurtosis K(t) for t daily, 10-day, and 66-day windows for the random variables*

	K(1)	K(10)	K(66)	Max Quartic	Years
Australian Dollar/USD	6.3	3.8	2.9	0.12	22.
Australia TB 10y	7.5	6.2	3.5	0.08	25.
Australia TB 3y	7.5	5.4	4.2	0.06	21.
BeanOil	5.5	7.0	4.9	0.11	47.
Bonds 30Y	5.6	4.7	3.9	0.02	32.
Bovespa	24.9	5.0	2.3	0.27	16.
British Pound/USD	6.9	7.4	5.3	0.05	38.
CAC40	6.5	4.7	3.6	0.05	20.
Canadian Dollar	7.4	4.1	3.9	0.06	38.
Cocoa NY	4.9	4.0	5.2	0.04	47.
Coffee NY	10.7	5.2	5.3	0.13	37.
Copper	6.4	5.5	4.5	0.05	48.
Corn	9.4	8.0	5.0	0.18	49.
Crude Oil	29.0	4.7	5.1	0.79	26.
CT	7.8	4.8	3.7	0.25	48.
DAX	8.0	6.5	3.7	0.20	18.
Euro Bund	4.9	3.2	3.3	0.06	18.
Euro Currency/DEM previously	5.5	3.8	2.8	0.06	38.
Eurodollar Depo 1M	41.5	28.0	6.0	0.31	19.
Eurodollar Depo 3M	21.1	8.1	7.0	0.25	28.
FTSE	15.2	27.4	6.5	0.54	25.
Gold	11.9	14.5	16.6	0.04	35.
Heating Oil	20.0	4.1	4.4	0.74	31.
Hogs	4.5	4.6	4.8	0.05	43.
Jakarta Stock Index	40.5	6.2	4.2	0.19	16.
Japanese Gov Bonds	17.2	16.9	4.3	0.48	24.
Live Cattle	4.2	4.9	5.6	0.04	44.
Nasdaq Index	11.4	9.3	5.0	0.13	21.
Natural Gas	6.0	3.9	3.8	0.06	19.
Nikkei	52.6	4.0	2.9	0.72	23.
Notes 5Y	5.1	3.2	2.5	0.06	21.
Russia RTSI	13.3	6.0	7.3	0.13	17.
Short Sterling	851.8	93.0	3.0	0.75	17.
Silver	160.3	22.6	10.2	0.94	46.
Smallcap	6.1	5.7	6.8	0.06	17.
SoyBeans	7.1	8.8	6.7	0.17	47.
SoyMeal	8.9	9.8	8.5	0.09	48.
Sp500	38.2	7.7	5.1	0.79	56.
Sugar #11	9.4	6.4	3.8	0.30	48.

Table 7.2: *(continued from previous page)*

	K(1)	K(10)	K(66)	Max Quartic	Years
SwissFranc	5.1	3.8	2.6	0.05	38.
TY10Y Notes	5.9	5.5	4.9	0.10	27.
Wheat	5.6	6.0	6.9	0.02	49.
Yen/USD	9.7	6.1	2.5	0.27	38.

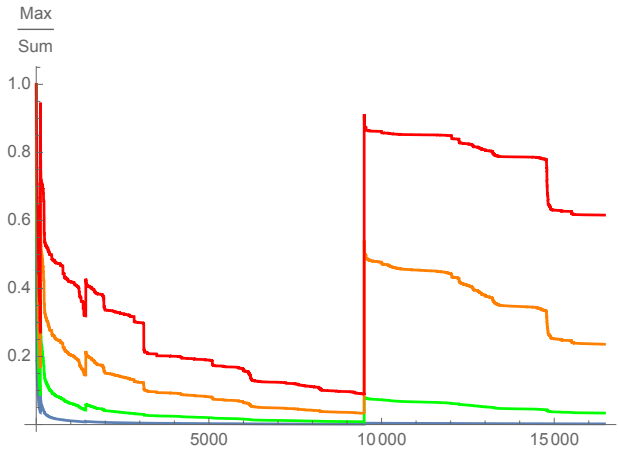


Figure 7.9: *MS Plot showing the behavior of cumulative moments $p = 1, 2, 3, 4$ for the SP500 over the 60 years ending in 2018. The MS plot (Maximum to sum) will be presented in 10.2.6.*

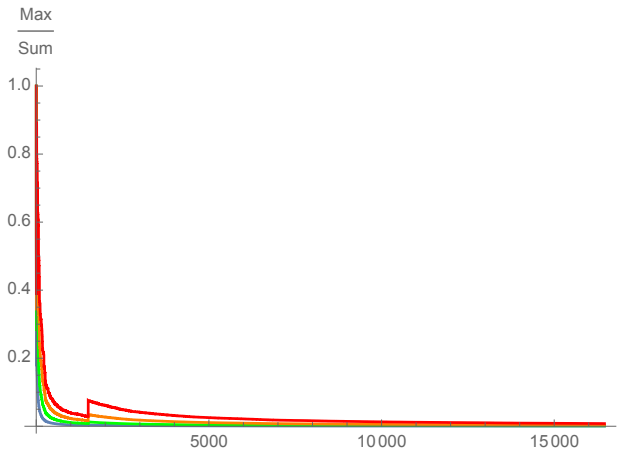


Figure 7.10: *Gaussian Control for the data in Figure* **??**.

7.7 MEAN DEVIATION FOR A STABLE DISTRIBUTIONS

Let us prepare a result for the next chapter using the norm L^1 for situations of finite mean but infinite variance.[6] Clearly we have no way to measure the compression of the distribution around the mean within the norm L^2.

The error of a sum in the norm L^1 is as follows. Let $\theta(x)$ be the Heaviside function (whose value is zero for negative arguments and one for positive arguments). Since $\text{sgn}(x) = 2\theta(x) - 1$, its characteristic function will be:

$$\chi^{\text{sgn}(x)}(t) = \frac{2i}{t}. \tag{7.6}$$

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

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

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

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

which is the Hilbert transform of χ when $\int$ is taken in the p.v. sense (Pinelis, 2015)[190]. In our situation, given that all independents summands are copies from the same distribution, we can replace the product $\chi(t)^n$ with $\chi_s(t)$ which is the same characteristic function with $\sigma_s = n^{1/\alpha}\sigma$, β remaining the same:

$$\mathbb{E}(|X|) = 2i\frac{\partial}{\partial u}\text{p.v.}\int_{-\infty}^{\infty} \frac{\chi_s(t-u)}{t}\, dt|_{t=0}. \tag{7.7}$$

Now, [190] the Hilbert transform H,

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

can be rewritten as

$$(Hf)(t) = -i\frac{\partial}{\partial u}\left(1 + \chi_s(u) + \frac{1}{\pi i}\int_0^{\infty-} \chi_s(u+t) - \chi_s(u-t) - \chi_s(t) + \chi_s(-t)\,\frac{dt}{t}\right). \tag{7.8}$$

Consider the stable distribution defined in 7.2.1.

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

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

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

which then integrates nicely to:

$$\mathbb{E}|X|_{(\alpha_s,\beta,\sigma_s,0)} = \frac{\sigma_s}{2\pi}\Gamma\left(\frac{\alpha_s-1}{\alpha_s}\right)\left(\left(1+i\beta\tan\left(\frac{\pi\alpha_s}{2}\right)\right)^{1/\alpha_s}+\left(1-i\beta\tan\left(\frac{\pi\alpha_s}{2}\right)\right)^{1/\alpha_s}\right). \quad (7.9)$$

NEXT

The next chapter presents a central concept: how to work with the law of middle numbers? How can we translate between distributions?

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

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

Applications: Among others, it

- helps assess the sample size n needed for statistical significance outside the Gaussian,
- helps measure the speed of convergence to the Gaussian (or stable basin),
- allows practical comparisons across classes of fat-tailed distributions,
- allows the assessment of the number of securities needed in portfolio construction to achieve a certain level of stability from diversification,
- helps understand some inconsistent attributes of the lognormal, pending on the parametrization of its variance.

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

Background : Conventional measures of fat-tailedness, namely 1) the tail index for the Power Law class, and 2) Kurtosis for finite moment distributions fail to apply to some distributions, and do not allow comparisons across classes and

Research chapter.

The author owes the most to the focused comments by Michail Loulakis who, in addition, provided the rigorous derivations for the limits of the κ for the Student T and lognormal distributions, as well as to the patience and wisdom of Spyros Makridakis. The paper was initially presented at *Extremes and Risks in Higher Dimensions*, Sept 12-16 2016, at the Lorentz Center, Leiden and at Jim Gatheral's Festschrift at the Courant Institute, in October 2017. The author thanks Jean-Philippe Bouchaud, John Einmahl, Pasquale Cirillo, and others. Laurens de Haan suggested changing the name of the metric from "gamma" to "kappa" to avoid confusion. Additional thanks to Colman Humphrey, Michael Lawler, Daniel Dufresne and others for discussions and insights with derivations.

parametrization, that is between power laws outside the Levy-Stable basin, or power laws to distributions in other classes, or power laws for different number of summands. How can one compare a sum of 100 Student T distributed random variables with 3 degrees of freedom to one in a Levy-Stable or a Lognormal class? How can one compare a sum of 100 Student T with 3 degrees of freedom to a single Student T with 2 degrees of freedom?

We propose an operational and heuristic metric that allows us to compare n-summed independent variables under all distributions with finite first moment. The method is based on the rate of convergence of the law of large numbers for finite sums, n-summands specifically.

We get either explicit expressions or simulation results and bounds for the lognormal, exponential, Pareto, and the Student T distributions in their various calibrations –in addition to the general Pearson classes.

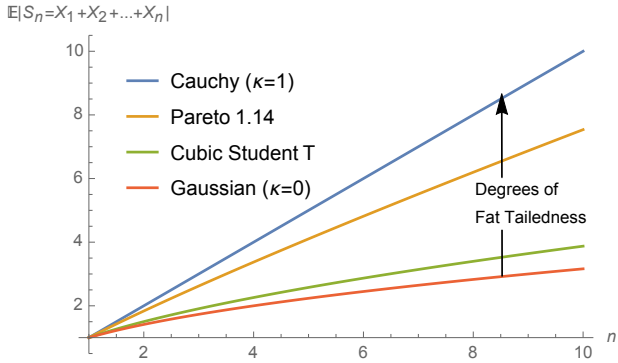


Figure 8.1: *The intuition of what κ is measuring: how the mean deviation of the sum of identical copies of a r.v. $S_n = X_1 + X_2 + \dots X_n$ grows as the sample increases and how we can compare preasymptotically distributions from different classes.*

8.1 INTRODUCTION AND DEFINITIONS

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

Criterion for "fat-tailedness" There are various ways to "define" Fat Tails and rank distributions according to each definition. In the narrow class of distributions having all moments finite, it is the kurtosis, which allows simple comparisons and

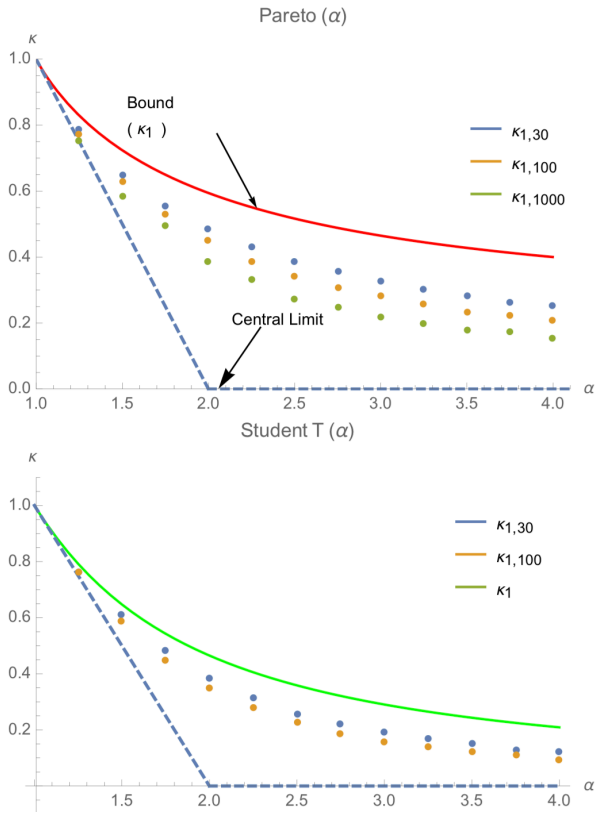


Figure 8.2: *Watching the effect of the Generalized Central Limit Theorem: Pareto and Student T Distribution, in the $\mathfrak{P}$ class, with α exponent, κ converge to $2 - (\mathbb{1}_{\alpha<2}\alpha + \mathbb{1}_{\alpha\geq 2}2)$, or the Stable $\mathfrak{S}$ class. We observe how slow the convergence, even after 1000 summands. This discounts Mandelbrot's assertion that an infinite variance Pareto can be subsumed into a stable distribution.*

measure departures from the Gaussian, which is used as a norm. For the Power Law class, it can be the tail exponent . One can also use extremal values, taking the probability of exceeding a maximum value, adjusted by the scale (as practiced in extreme value theory). For operational uses, practitioners' fat-tailedness is a degree of concentration, such as "how much of the statistical properties will be attributable to a single observation?", or, appropriately adjusted by the scale (or the mean dispersion), "how much is the total wealth of a country in the hands of the richest individual?"

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

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

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

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

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

The state of statistical inference The last point, the "speed", appears to have been ignored (see earlier comments in Chapter 3 about the 9,400 pages of the *Encyclopedia of Statistical Science* [145]). It is very rare to find a discussion about how long it takes to reach the asymptote, or how to deal with n summands that are large but perhaps not sufficiently so for the so-called "normal approximation".

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

8.2 THE METRIC

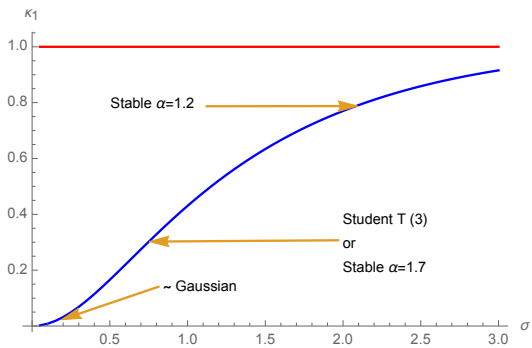


Figure 8.3: *The lognormal distribution behaves like a Gaussian for low values of σ, but becomes rapidly equivalent to a power law. This illustrates why, operationally, the debate on whether the distribution of wealth was lognormal (Gibrat) or Pareto (Zipf) doesn't carry much operational significance.*

Definition 8.1 (the κ metric)
Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean, that is $\mathbb{E}(X) < +\infty$. Let $S_n = X_1 + X_2 + \ldots + X_n$ be a partial sum. Let $\mathbb{M}(n) = \mathbb{E}(|S_n - \mathbb{E}(S_n)|)$ be the expected mean absolute deviation from the mean for n summands. Define the "rate" of convergence for n additional summands starting with n_0:

Table 8.1: *Kappa for 2 summands, κ_1.*

Distribution	κ_1
Student T (α)	$2-\frac{2\log(2)}{2\log\left(\frac{2^{2-\alpha}\Gamma\left(\alpha-\frac{1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)^2}\right)+\log(\pi)}$
Exponential/Gamma	$2-\frac{\log(2)}{2\log(2)-1}\approx .21$
Pareto (α)	$2-\frac{\log(2)}{\log\left((\alpha-1)^{2-\alpha}\alpha^{\alpha-1}\int_0^{\frac{2}{\alpha-1}}-2\alpha^2(y+2)^{-2\alpha-1}\left(\frac{2}{\alpha-1}-y\right)\left(B_{\frac{1}{y+2}}(-\alpha,1-\alpha)-B_{\frac{y+1}{y+2}}(-\alpha,1-\alpha)\right)dy\right)}$ [3]
Normal (μ,σ) with switching variance $\sigma^2 a$ w.p p[4].	$2-\frac{\log(2)}{\log\left(\frac{\sqrt{2}\left(\sqrt{\frac{ap}{p-1}+\sigma^2}+p\left(-2\sqrt{\frac{ap}{p-1}+\sigma^2}+p\left(\sqrt{\frac{ap}{p-1}+\sigma^2}-\sqrt{2a\left(\frac{1}{p-1}+2\right)+4\sigma^2+\sqrt{a+\sigma^2}}\right)+\sqrt{2a\left(\frac{1}{p-1}+2\right)+4\sigma^2}\right)\right)}{p\sqrt{a+\sigma^2}-(p-1)\sqrt{\frac{ap}{p-1}+\sigma^2}}\right)}$
Lognormal (μ,σ)	$\approx 2-\frac{\log(2)}{\log\left(\frac{2\,\mathrm{erf}\left(\frac{\sqrt{\log\left(\frac{1}{2}\left(e^{\sigma^2}+1\right)\right)}}{2\sqrt{2}}\right)}{\mathrm{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)}\right)}$.

Table 8.2: *Summary of main results*

Distribution	κ_n
Exponential/Gamma	Explicit
Lognormal (μ,σ)	No explicit κ_n but explicit lower and higher bounds (low or high σ or n). Approximated with Pearson IV for σ in between.
Pareto (α) (Constant)	Explicit for κ_2 (lower bound for all α).
Student T(α) (slowly varying function)	Explicit for κ_1 , $\alpha = 3$.

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

Table 8.3: *Comparing Pareto to Student T (Same tail exponent α)*

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

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

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\log\left(\frac{\mathbb{M}(n)}{\mathbb{M}(n_0)}\right)}. \tag{8.1}$$

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

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

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\sum_{i=0}^{n} \frac{\log(i+1) - \log(i)}{2 - \kappa(i, i+1)}}. \tag{8.2}$$

Use of Mean Deviation Note that we use for measure of dispersion around the mean the mean absolute deviation, to stay in norm L^1 in the absence of finite variance –actually, even in the presence of finite variance, under Power Law regimes, distributions deliver an unstable and uninformative second moment. Mean deviation proves far more robust there. (Mean absolute deviation can be shown to be more "efficient" except in the narrow case of kurtosis equals 3 (the Gaussian), see a longer discussion in [234]; for other advantages, see [184].)

8.3 STABLE BASIN OF CONVERGENCE AS BENCHMARK

Definition 8.2 (the class $\mathfrak{P}$)
The $\mathfrak{P}$ class of power laws (regular variation) is defined for r.v. X as follows:

$$\mathfrak{P} = \{X : \mathbb{P}(X > x) \sim L(x)\, x^{-\alpha}\} \tag{8.3}$$

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

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

Definition 8.3
(stable $\mathfrak{S}$ *class) A random variable* X *follows a stable (or* α*-stable) distribution, symbolically* $X \sim S(\tilde{\alpha}, \beta, \mu, \sigma)$, *if its characteristic function*$\chi(t) = \mathbb{E}(e^{itX})$ *is of the form:*

$$\chi(t) = \begin{cases} e^{\left(i\mu t - |t\sigma|^{\tilde{\alpha}}\left(1 - i\beta\tan\left(\frac{\pi\tilde{\alpha}}{2}\right)sgn(t)\right)\right)} & \tilde{\alpha} \neq 1 \\ e^{it\left(\frac{2\beta\sigma\log(\sigma)}{\pi} + \mu\right) - |t\sigma|\left(1 + \frac{2i\beta sgn(t)\log(|t\sigma|)}{\pi}\right)} & \tilde{\alpha} = 1 \end{cases}, \tag{8.4}$$

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

$$\tilde{\alpha} \triangleq \begin{cases} \alpha\, \mathbb{1}_{\alpha<2} + 2\, \mathbb{1}_{\alpha\geq 2} & \text{if } X \text{ is in } \mathfrak{P} \\ 2 & \text{otherwise.} \end{cases} \tag{8.5}$$

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

8.3.1 Equivalence for Stable distributions

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

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

simply from the property that

$$\mathbb{M}(n) = n^{\frac{1}{\tilde{\alpha}}}\mathbb{M}(1) \tag{8.6}$$

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

The problem of the preasymptotics for n summands reduces to:

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

8.3.2 Practical significance for sample sufficiency

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

Computations of the sort done Table 8.2 for instance allows us to compare various distributions under various parametriazation. (comparing various Pareto distributions to symmetric Student T and, of course the Gaussian which has a flat kappa of 0)

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

Many papers, when discussing financial matters, say [98] use finite variance as a binary classification for fat tailedness: power laws with a tail exponent greater than 2 are therefore classified as part of the "Gaussian basin", hence allowing the use of variance and other such metrics for financial applications. A much more natural boundary is finiteness of expectation for financial applications [226]. Our metric can thus be useful as follows:

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

κ_n is indicative of both the rate of convergence under the law of large number, and for $\kappa_n \to 0$, for rate of convergence of summands to the Gaussian under the central limit, as illustrated in Figure 8.2.

$$n_{\min} = \inf\left\{ n_v : \mathbb{E}\left(\left|\sum_{i=1}^{n_v} \frac{X_{v,i} - m_p}{n_v}\right|\right) \leq \mathbb{E}\left(\left|\sum_{i=1}^{n_g} \frac{X_{g,i} - m_g}{n_g}\right|\right), n_v > 0 \right\} \tag{8.7}$$

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

$$n_v = n_g^{-\frac{1}{\kappa_{1,n_g}{}^{-1}}} \approx n_g^{-\frac{1}{\kappa_1{}^{-1}}}, n_g > 1 \tag{8.8}$$

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

We can also elicit quite counterintuitive results. From Eq. 8.8, the "Pareto 80/20" in the popular mind, which maps to a tail exponent around $\alpha \approx 1.14$, requires $> 10^9$ more observations than the Gaussian.

8.4 TECHNICAL CONSEQUENCES

8.4.1 Some Oddities With Asymmetric Distributions

The stable distribution, when skewed, has the same κ index as a symmetric one (in other words, κ is invariant to the β parameter in Eq. 8.4, which conserves under summation). But a one-tailed simple Pareto distribution is fatter tailed (for our purpose here) than an equivalent symmetric one.

This is relevant because the stable is never really observed in practice and used as some limiting mathematical object, while the Pareto is more commonly seen. The point is not well grasped in the literature. Consider the following use of the substitution of a stable for a Pareto. In Uchaikin and Zolotarev [252]:

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

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

8.4.2 Rate of Convergence of a Student T Distribution to the Gaussian Basin

We show in the appendix –thanks to the explicit derivation of κ for the sum of students with $\alpha = 3$, the "cubic" commonly noticed in finance –that the rate of convergence of κ to 0 under summation is $\frac{1}{\log(n)}$. This (and the semi-closed form for the density of an n-summed cubic Student) complements the result in Bouchaud and Potters [28] (see also [211]), which is as follows. Their approach is to separate the "Gaussian zone" where the density is approximated by that of a Gaussian, and a "Power Law zone" in the tails which retains the original distribution with Power Law decline. The "crossover" between the two moves right and left of the center at a rate of $\sqrt{n \log(n)}$ standard deviations) which is excruciatingly slow. Indeed, one can note that more summands fall at the center of the distribution, and fewer outside of it, hence the speed of convergence according to the central limit theorem will differ according to whether the density concerns the center or the tails.

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

8.4.3 The Lognormal is Neither Thin Nor Fat Tailed

Naively, as we can see in Figure 8.2, at low values of the parameter σ, the lognormal behaves like a Gaussian, and, at high σ, it appears to have the behavior of a Cauchy of sorts (a one-tailed Cauchy, rather a stable distribution with $\alpha = 1$, $\beta = 1$), as κ gets closer and closer to 1. This gives us an idea about some aspects of the debates as to whether some variable is Pareto or lognormally distributed, such as, say, the debates about wealth [160], [52], [53]. Indeed, such debates can be irrelevant to the real world. As P. Cirillo [44] observed, many cases of Paretianity are effectively lognormal situations with high variance; the practical statistical consequences, however, are smaller than imagined.

8.4.4 Can Kappa Be Negative?

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

Details and derivations are presented in the appendix.

8.5 CONCLUSION AND CONSEQUENCES

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

We may draw a few operational consequences:

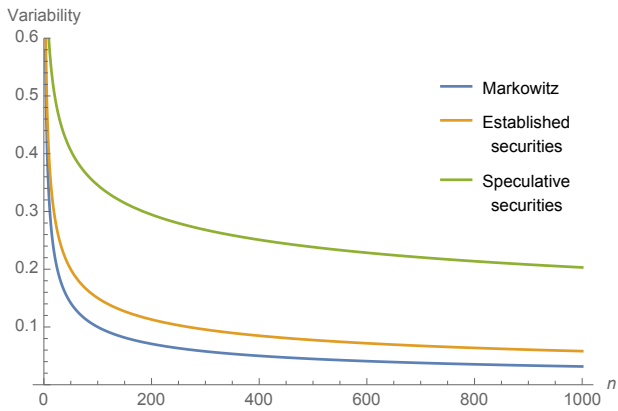


Figure 8.4: *In short, why the 1/n heuristic works: it takes many, many more securities to get the same risk reduction as via portfolio allocation according to Markowitz . We assume to simplify that the securities are independent, which they are not, something that compounds the effect.*

8.5.1 Portfolio Pseudo-Stabilization

Our method can also naturally and immediately apply to portfolio construction and the effect of diversification since adding a security to a portfolio has the same "stabilizing" effect as adding an additional observation for the purpose of statistical significance. "How much data do you need?" translates into "How many securities do you need?". Clearly, the Markowicz allocation method in modern finance [164] (which seems to not be used by Markowitz himself for his own portfolio [176]) applies only for κ near 0; people use convex heuristics, otherwise they will underestimate tail risks and "blow up" the way the famed portfolio-theory oriented hedge fund Long Term Management did in 1998 [233] [245].)

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

It has also been noted that there is practically no financial security that is not fatter tailed than the Gaussian, from the simple criterion of kurtosis [225], meaning Markowitz portfolio allocation is *never* the best solution. It happens that agents wisely apply a noisy approximation to the $\frac{1}{n}$ heuristic which has been classified as one of those biases by behavioral scientists but has in fact been debunked as false (a false bias is one in which, while the observed phenomenon is there, it does not constitute a "bias" in the bad sense of the word; rather it is the researcher who is mistaken owing to using the wrong tools instead of the decision-maker). This tendency to "overdiversify" has been deemed a departure from optimal investment behavior by Benartzi and Thaler [18], explained in [16] "when faced with n options, divide assets evenly across the options. We have dubbed this heuristic the "$1/n$ rule."" However, broadening one's diversification is effectively as least as optimal as standard allocation(see critique by Windcliff and Boyle [260] and [61]). In short, an equally weighted portfolio outperforms the SP500 across a broad range range of metrics. But even the latter two papers didn't conceive of the full effect and properties of fat tails, which we can see here with some precision. Fig. 8.5 shows the effect for securities compared to Markowitz.

This false bias is one in many examples of policy makers "nudging" people into the wrong rationality [233] and driving them to increase their portfolio risk many folds.

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

Likewise the metric gives us some guidance in the treatment of data for forecasting, by establishing sample sufficiency, to state such matters as how many years of data do we need before stating whether climate conditions "have changed", see [158].

8.5.2 Other Aspects of Statistical Inference

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

8.5.3 Final comment

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

8.6 APPENDIX, DERIVATIONS, AND PROOFS

We show here some derivations

8.6.1 Cubic Student T (Gaussian Basin)

The Student T with 3 degrees of freedom is of special interest in the literature owing to its prevalence in finance [98]. It is often mistakenly approximated to be Gaussian owing to the finiteness of its variance. Asymptotically, we end up with a Gaussian, but this doesn't tell us anything about the rate of convergence. Mandelbrot and Taleb [163] remarks that the cubic acts more like a powerlaw in the distribution of the extremes, which we will elaborate here thanks to an explicit PDF for the sum.

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

$$p(x) = \frac{6\sqrt{3}}{\pi \left(x^2+3\right)^2}\,, x \in (-\infty, \infty) \tag{8.9}$$

Proposition 8.1
Let Y be a sum of $X_1, \ldots, X_n$, n identical copies of X. Let $\mathbb{M}(n)$ be the mean absolute deviation from the mean for n summands. The "rate" of convergence $\kappa_{1,n} = \left\{\kappa : \frac{\mathbb{M}(n)}{\mathbb{M}(1)} = n^{\frac{1}{2-\kappa}}\right\}$ is:

$$\kappa_{1,n} = 2 - \frac{\log(n)}{\log\left(e^n n^{-n}\Gamma(n+1,n) - 1\right)} \tag{8.10}$$

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

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

$$\mathbb{M}(n) = \begin{cases} \frac{2\sqrt{3}}{\pi} & \text{for } n = 1 \\ \frac{2\sqrt{3}}{\pi}\left(e^n n^{-n}\Gamma(n+1,n) - 1\right) & \text{for } n > 1 \end{cases} \tag{8.11}$$

The derivations are as follows. For the pdf and the MAD we followed different routes.

We have the characteristic functionfor n summands:

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

The pdf of Y is given by:

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

After arduous integration we get the result in 8.11. Further, since the following result does not appear to be found in the literature, we have a side useful result: the PDF of Y can be written as

$$p(y) = \frac{e^{n-\frac{iy}{\sqrt{3}}} \left(e^{\frac{2iy}{\sqrt{3}}} E_{-n}\left(n + \frac{iy}{\sqrt{3}}\right) + E_{-n}\left(n - \frac{iy}{\sqrt{3}}\right)\right)}{2\sqrt{3}\pi} \tag{8.12}$$

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

Note the following identities (from the updating of Abramowitz and Stegun) [68]

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

As to the asymptotics, we have the following result (proposed by Michail Loulakis): Reexpressing Eq. 8.11:

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

Further,

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

(From the behavior of the sum of Poisson variables as they converge to a Gaussian by the central limit theorem: $e^{-n}\sum_{m=0}^{n-1}\frac{n^m}{m!} = \mathbb{P}(X_n < n)$ where X_n is a Poisson random variable with parameter n. Since the sum of n independent Poisson random variables with parameter 1 is Poisson with parameter n, the Central Limit Theorem says the probability distribution of $Z_n = (X_n - n)/\sqrt{n}$ approaches a standard normal distribution. Thus $\mathbb{P}(X_n < n) = \mathbb{P}(Z_n < 0) \to 1/2$ as $n \to \infty$.[5] For another approach, see [177] for proof that $1 + \frac{n}{1!} + \frac{n^2}{2!} + \cdots + \frac{n^{n-1}}{(n-1)!} \sim \frac{e^n}{2}$.)

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

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

5 Robert Israel on Math Stack Exchange

thus κ goes to 0 (i.e, the average becomes Gaussian) at speed $\frac{1}{\log(n)}$, which is excruciatingly slow. In other words, even with 10^6 summands, the behavior cannot be summarized as that of a Gaussian, an intuition often expressed by B. Mandelbrot [163].

8.6.2 Lognormal Sums

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

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

Cumulants:

$$\begin{aligned} K_1^n &= ne^{\mu+\frac{\sigma^2}{2}} \\ K_2^n &= n\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2} \\ K_3^n &= n\left(e^{\sigma^2}-1\right)^2\left(e^{\sigma^2}+2\right)e^{3\mu+\frac{3\sigma^2}{2}} \\ K_4^n &= \dots \end{aligned}$$

Which allow us to compute: $Skewness = \frac{\sqrt{e^{\sigma^2}-1}\left(e^{\sigma^2}+2\right)e^{\frac{1}{2}\left(2\mu+\sigma^2\right)-\mu-\frac{\sigma^2}{2}}}{\sqrt{n}}$ and $Kurtosis = 3+\frac{e^{2\sigma^2}\left(e^{\sigma^2}\left(e^{\sigma^2}+2\right)+3\right)-6}{n}$

We can immediately prove from the cumulants/moments that:

$$\lim_{n\to+\infty}\kappa_{1,n}=0,\ \lim_{\sigma\to 0}\kappa_{1,n}=0$$

and our bound on κ becomes explicit:

Let $\kappa_{1,n}^*$ be the situation under which the sums of lognormal conserve the lognormal density, with the same first two moments. We have

$$0\le\kappa_{1,n}^*\le 1,$$

$$\kappa_{1,n}^*=2-\frac{\log(n)}{\log\left(\frac{n\,\mathrm{erf}\left(\frac{\sqrt{\log\left(\frac{n+e^{\sigma^2}-1}{n}\right)}}{2\sqrt{2}}\right)}{\mathrm{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)}\right)}$$

Heuristic attempt Among other heuristic approaches, we can see in two steps how 1) under high values of σ, $\kappa_{1,n} \to \kappa^*_{1,n}$, since the law of large numbers slows down, and 2) $\kappa^*_{1,n} \overset{\sigma\to\infty}{\to} 1$.

Loulakis' Proof Proving the upper bound, that for high variance $\kappa_{1,n}$ approaches 1 has been shown formally my Michail Loulakis[6] which we summarize as follows. We start with the identify $\mathbb{E}\left(|X-m|\right) = 2\int_m^\infty (x-m)f(x)dx = 2\int_m^\infty \bar{F}_X(t)dt$, where $f(.)$ is the density, m is the mean, and $\bar{F}_X(.)$ is the survival function. Further, $\mathbb{M}(n) = 2\int_{nm}^\infty \bar{F}(x)dx$. Assume $\mu = \frac{1}{2}\sigma^2$, or $X = \exp\left(\sigma Z - \frac{\sigma^2}{2}\right)$ where Z is a standard normal variate. Let S_n be the sum $X_1 + \ldots + X_n$; we get $\mathbb{M}(n) = 2\int_n^\infty \mathbb{P}(S_n > t)dt$. Using the property of subexponentiality ([193]), $\mathbb{P}(S_n > t) \geq \mathbb{P}(\max_{0<i\leq n}(X_i) > t) \geq n\mathbb{P}(X_1 > t) - \binom{n}{2}\mathbb{P}\left(X_1 > t\right)^2$. Now $\mathbb{P}\left(X_1 > t\right) \overset{\sigma\to\infty}{\to} 1$ and the second term to 0 (using Hölder's inequality).

Skipping steps, we get $\liminf_{\sigma\to\infty} \frac{\mathbb{M}(n)}{\mathbb{M}(1)} \geq n$, while at the same time we need to satisfy the bound $\frac{\mathbb{M}(n)}{\mathbb{M}(1)} \leq n$. So for $\sigma \to \infty$, $\frac{\mathbb{M}(n)}{\mathbb{M}(1)} = n$, hence $\kappa_{1,n} \overset{\sigma\to\infty}{\to} 1$.

Pearson Family Approach for Computation For computational purposes, for the σ parameter not too large (below $\approx .3$, we can use the Pearson family for computational convenience –although the lognormal does not belong to the Pearson class (the normal does, but we are close enough for computation). Intuitively, at low σ, the first four moments can be sufficient because of the absence of large deviations; not at higher σ for which conserving the lognormal would be the right method.

The use of Pearson class is practiced in some fields such as information/communication theory, where there is a rich literature: for summation of lognormal variates see Nie and Chen, [178], and for Pearson IV, [41], [64].

The Pearson family is defined for an appropriately scaled density f satisfying the following differential equation.

$$f'(x) = -\frac{(a_0 + a_1 x)}{b_0 + b_1 x + b_2 x^2} f(x) \tag{8.13}$$

We note that our parametrization of a_0, b_2, etc. determine the distribution within the Pearson class –which appears to be the Pearson IV. Finally we get an expression of mean deviation as a function of n, σ, and μ.

Let m be the mean. Diaconis et al [66] from an old trick by De Moivre, Suzuki [218] show that we can get explicit mean absolute deviation. Using, again, the identity $\mathbb{E}(|X-m|) = 2\int_m^\infty (x-m)f(x)\mathrm{d}x$ and integrating by parts,

$$\mathbb{E}(|X-m|) = \frac{2\left(b_0 + b_1 m + b_2 m^2\right)}{a_1 - 2b_2} f(m) \tag{8.14}$$

6 Review of the paper version; Loulakis proposed a formal proof in place of the heuristic derivation.

We use cumulants of the n-summed lognormal to match the parameters. Setting $a_1 = 1$, and $m = \frac{b_1 - a_0}{1 - 2b_2}$, we get

$$\begin{cases} a_0 = \frac{e^{\mu+\frac{\sigma^2}{2}}\left(-12n^2+(3-10n)e^{4\sigma^2}+6(n-1)e^{\sigma^2}+12(n-1)e^{2\sigma^2}-(8n+1)e^{3\sigma^2}+3e^{5\sigma^2}+e^{6\sigma^2}+12\right)}{2\left(6(n-1)+e^{2\sigma^2}\left(e^{\sigma^2}\left(5e^{\sigma^2}+4\right)-3\right)\right)} \\ b_2 = \frac{e^{2\sigma^2}\left(e^{\sigma^2}-1\right)\left(2e^{\sigma^2}+3\right)}{2\left(6(n-1)+e^{2\sigma^2}\left(e^{\sigma^2}\left(5e^{\sigma^2}+4\right)-3\right)\right)} \\ b_1 = \frac{\left(e^{\sigma^2}-1\right)e^{\mu+\frac{\sigma^2}{2}}\left(e^{\sigma^2}\left(e^{\sigma^2}\left(e^{\sigma^2}\left(-4n+e^{\sigma^2}\left(e^{\sigma^2}+4\right)+7\right)-6n+6\right)+6(n-1)\right)+12(n-1)\right)}{2\left(6(n-1)+e^{2\sigma^2}\left(e^{\sigma^2}\left(5e^{\sigma^2}+4\right)-3\right)\right)} \\ b_0 = -\frac{n\left(e^{\sigma^2}-1\right)e^{2\left(\mu+\sigma^2\right)}\left(e^{\sigma^2}\left(-2(n-1)e^{\sigma^2}-3n+e^{3\sigma^2}+3\right)+6(n-1)\right)}{2\left(6(n-1)+e^{2\sigma^2}\left(e^{\sigma^2}\left(5e^{\sigma^2}+4\right)-3\right)\right)} \end{cases}$$

Polynomial Expansions Other methods, such as Gram-Charlier expansions, such as Schleher [207], Beaulieu,[14], proved less helpful to obtain κ_n. At high values of σ, the approximations become unstable as we include higher order Lhermite polynomials. See review in Dufresne [69] and [70].

8.6.3 Exponential

The exponential is the "entry level" fat tails, just at the border.

$$f(x) = \quad \lambda e^{-\lambda x}, \quad x \geq 0.$$

By convolution the sum $Z = X_1, X_2, \ldots X_n$ we get, by recursion, since $f(y) = \int_0^y f(x)f(y-x)\,dx = \lambda^2 y e^{-\lambda y}$:

$$f_n(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!} \tag{8.15}$$

which is the gamma distribution; we get the mean deviation for n summands:

$$\mathbb{M}(n) = \frac{2e^{-n}n^n}{\lambda\Gamma(n)}, \tag{8.16}$$

hence:

$$\kappa_{1,n} = 2 - \frac{\log(n)}{n\log(n) - n - \log(\Gamma(n)) + 1} \tag{8.17}$$

We can see the asymptotic behavior is equally slow (similar to the student) although the exponential distribution is sitting at the cusp of subexponentiality:

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

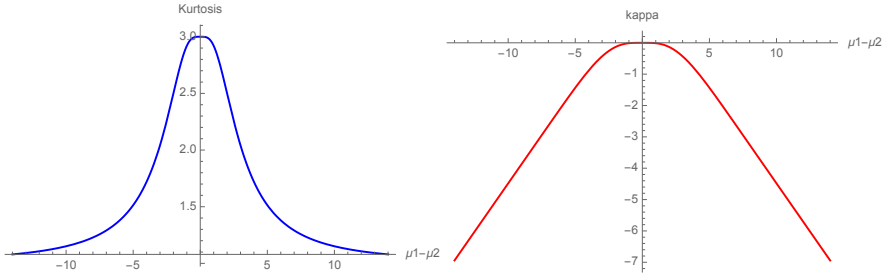


Figure 8.5: *Negative kurtosis from A.3 and corresponding kappa.*

8.6.4 Negative Kappa, Negative Kurtosis

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

These situations with thinner tails than the Gaussian are encountered with bimodal situations where μ_1 and μ_2 are separated; the effect becomes acute when they are separated by several standard deviations. Let d= $\mu_1 - \mu_2$ and $\sigma = \sigma_1 = \sigma_2$ (to achieve minimum kurtosis),

$$\kappa_1 = \frac{\log(4)}{\log(\pi) - 2\log\left(\frac{\sqrt{\pi} d e^{\frac{d^2}{4\sigma^2}} \operatorname{erf}\left(\frac{d}{2\sigma}\right) + 2\sqrt{\sigma^2} e^{\frac{d^2}{4\sigma^2}} + 2\sigma}{d e^{\frac{d^2}{4\sigma^2}} \operatorname{erf}\left(\frac{d}{2\sqrt{2}\sigma}\right) + 2\sqrt{\frac{2}{\pi}}\sigma e^{\frac{d^2}{8\sigma^2}}}\right)} + 2 \tag{8.18}$$

which we see is negative for wide values of $\mu_1 - \mu_2$.

NEXT

Next we consider some simple diagnostics for power laws with application to the SP500 . We show the differences between naive methods and those based on ML estimators that allow extrapolation into the tails.

9 EXTREME VALUES AND HIDDEN RISK *,†

WHEN the data is thick tailed, there is a hidden part of the distribution, not shown in past samples. Past extrema (maximum or minimum) is not a good predictor of future extrema – visibly records take place and past higher water mark is a naive estimation, what is referred to in Chapter 3 as the Lucretius fallacy , which as we saw can be paraphrased as: *the fool believes that the tallest river and tallest mountain there is equals the tallest ones he has personally seen.*

This chapter, after a brief introduction to extreme value theory, focuses on its application to thick tails. When the data is power law distributed, the maximum of n observations follows a distribution easy to build from scratch. We show practically how the Fréchet distribution is, asymptotically, the maximum domain of attraction MDA of power law distributed variables.

More generally extreme value theory allows a rigorous approach to deal with extremes and the extrapolation past the sample maximum. We present some results on the "hidden mean", as it relates to a variety of fallacies in the risk management literature.

9.1 PRELIMINARY INTRODUCTION TO EVT

Let $X_1, \ldots X_n$ be independent and distributed Pareto random variables with CDF $F(.)$

Exposition chapter with somme research.

Lucretius in *De Rerum Natura*:
Scilicet et fluvius qui visus maximus ei,
Qui non ante aliquem majorem vidit; et ingens
Arbor, homoque videtur, et omnia de genere omni
Maxima quae vidit quisque, haec ingentia fingit.

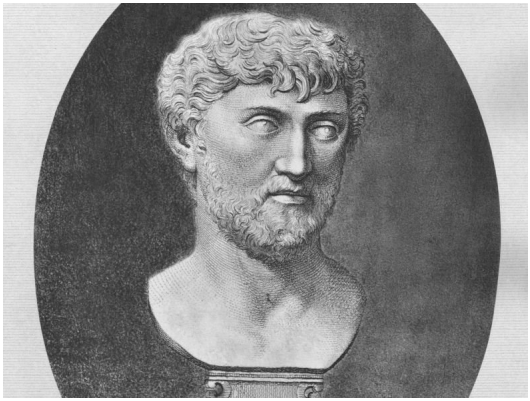


Figure 9.1: *The Roman philosophical poet Lucretius.*

We can get an exact distribution of the max (or minimum). The CDF of the maximum of the n variables will be

$$\begin{aligned}\mathbb{P}\left(X_{max} \leq x\right) =& \mathbb{P}\left(X_1 \leq x, \ldots, X_n \leq x\right) = \mathbb{P}\left(X_1 \leq x\right) \\ & \cdots \mathbb{P}\left(X_n \leq x\right) = F(x)^n\end{aligned} \tag{9.1}$$

that is, the probability that all values of x falling at or below X_{max}. The PDF is the first derivative : $\psi(x) = \frac{\partial F(x)^n}{\partial x}$.

The extreme value distribution concerns that of the maximum r.v., when $x \to x^*$, where $x^* = \sup\{x : F(x) < 1\}$ (the right "endpoint" of the distribution) is in the maximum domain of attraction, MDA [115]. In other words,

$$\max(X_1, \ldots X_n) \xrightarrow{P} x^*,$$

where $\xrightarrow{P}$ denotes convergence in probability. The central question becomes: what is the distribution of x^*? We said that we have the exact distribution, so as engineers we could be satisfied with the PDF from Eq. 9.1. As a matter of fact, we could get all test statistics from there, provided we have patience, computer power, and the will to investigate –it is the only way to deal with preasymptotics, that is "what happens when n is small enough so x is not quite x^*.

But it is quite useful for general statistical work to understand the general asymptotic structure.

The Fisher-Tippett-Gnedenko theorem (Embrech et al. [81], de Haan and Ferreira [115]) states the following. If there exist sequences of "norming" constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \underset{n\to\infty}{\to} G(x), \tag{9.2}$$

then

$$G(x) \propto \exp\left(-(1+\xi x)^{-1/\xi}\right)$$

where ξ is the extreme value index, and governs the tail behavior of the distribution. G is called the (generalized) extreme value distribution, GED . The subfamilies defined by $\xi = 0$, $\xi > 0$ and $\xi < 0$ correspond, respectively, to the Gumbel, Fréchet and Weibull families:

Gumbel distribution (Type 1) Here $\xi = 0$; rather $\lim_{\xi \to 0} \exp\left(-(\xi x + 1)^{-\frac{1}{\xi}}\right)$:

$$G(x) = \exp\left(-\exp\left(-\left(\frac{x - b_n}{a_n}\right)\right)\right) \text{ for } x \in \mathbb{R}.$$

when the distribution of M_n has an exponential tail.

Fréchet distribution(Type 2) Here $\xi = \frac{1}{\alpha}$:

$$G(x) = \begin{cases} 0 & x \leq b_n \\ \exp\left(-\left(\frac{x-b_n}{a_n}\right)^{-\alpha}\right) & x > b_n. \end{cases}$$

when the distribution of M_n has power law right tail, as we saw earlier. Note that $\alpha > 0$.

Weibull distribution (Type 3) Here $\xi = -\frac{1}{\alpha}$

$$G(x) = \begin{cases} \exp\left(-\left(-\left(\frac{x-b_n}{a_n}\right)\right)^{\alpha}\right) & x < b_n \\ 1 & x \geq b \end{cases}$$

when the distribution of M_n has a finite support on the right (i.e., bounded maximum). Note here again that $\alpha > 0$.

9.1.1 How Any Power Law Tail Leads to Fréchet

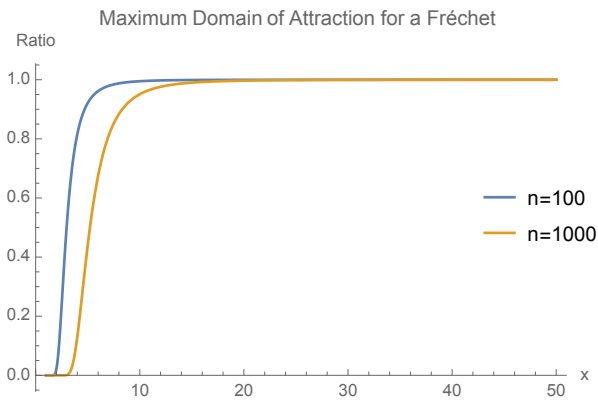


Figure 9.2: *Shows the ration of distributions of the CDF of the exact distribution over that of a Fréchet. We can visualize the acceptable level of approximation and see how x reaches the Maximum Domain of Attraction, MDA. Here $\alpha = 2$, $L = 1$. We note that the ratio for the PDF shows the same picture, unlike the Gaussian, as we will see further down.*

Let us proceed now like engineers rather than mathematicians, and consider two existing distributions, the Pareto and the Fréchet, and see how one can me made to converge to the other, in other words rederive the Fréchet from the asymptotic properties of power laws.

The reasoning we will follow next can be generalized to any Pareto-tailed variable considered above the point where slowly varying function satisfactorily approximates a constant –the "Karamata point".

The CDF of the Pareto with minimum value (and scale) L and tail exponent α:

$$F(x) = 1 - \left(\frac{L}{x}\right)^{\alpha},$$

so the PDF of the maximum of n observations:

$$\psi(x) = \frac{\alpha n \left(\frac{L}{x}\right)^{\alpha} \left(1 - \left(\frac{L}{x}\right)^{\alpha}\right)^{n-1}}{x}. \tag{9.3}$$

The PDF of the Frechét:

$$\varphi(x) = \alpha \beta^{\alpha} x^{-\alpha-1} e^{\beta^{\alpha}\left(-x^{-\alpha}\right)}. \tag{9.4}$$

Let us now look for x "very large" where the two functions equate, or $\psi(x^*) \to \varphi(x^*)$.

$$\lim_{x\to\infty} \frac{\psi(x)}{\varphi(x)} = n \left(\frac{1}{\beta}\right)^{\alpha} L^{\alpha}. \tag{9.5}$$

Accordingly, for x deemed "large", we can use $\beta = Ln^{1/\alpha}$. Equation 9.5 shows us how the tail α conserves across transformations of distribution:

Property 4

The tail exponent of the maximum of i.i.d random variables is the same as that of the random variables themselves.

Now, in practice, "where" do we approximate is shown in figure 9.2.

Property 5

We get an exact asymptotic fitting for power law extrema.

9.1.2 Gaussian Case

The Fréchet case is quite simple –power laws are usually simpler analytically, and we can get limiting parametrizations. For the Gaussian and other distributions, more involved derivations and approximations are required to fit the norming constants a_n and b_n, usually entailing quantile functions. The seminal paper by

Fisher and Tippet [93] warns us that "from the normal distribution, the limiting distribution is approached with extreme slowness" (cited by Gasull et al. [100]).

In what follows we look for norming constants for a Gaussian, based on [119] and later developments.

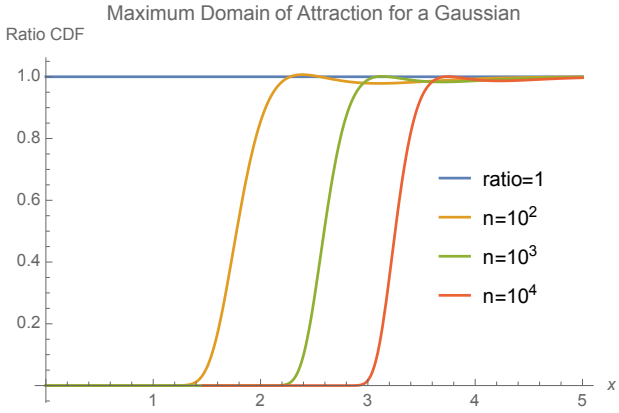


Figure 9.3: *The behavior of the Gaussian; it is hard to get a good parametrization, unlike with power laws. The y axis shows the ratio for the CDF of thee exact maximum distribution for n variables over that of the parametrized EVT.*

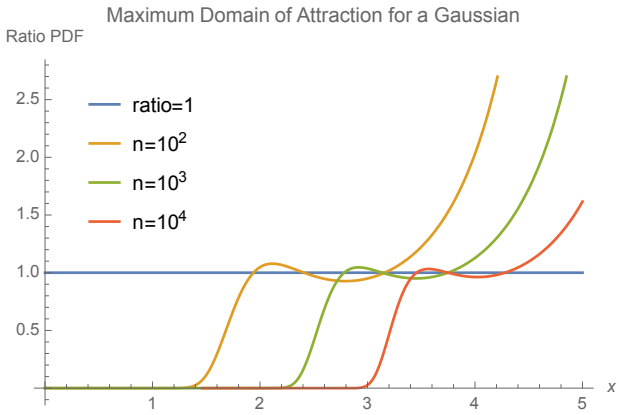


Figure 9.4: *The same as figure 9.3 but using PDF. It is not possible to obtain a good approximation in the tails.*

Consider $M_n = a_n x + b_n$ in Eq. 9.2. We assume then that M_n follows the Extreme Value Distribution EVT (the CDF is e^{-e^x}, the mirror distribution of the Gumbel for minima, obtained by transforming the distribution of $-M_n$ where $\frac{M_n - b_n}{a_n}$ follows a Gumbel with CDF $1 - e^{-e^x}$.) [3] The parametrized CDF for M_n is $e^{-e^{-\frac{x-b_n}{a_n}}}$.

An easy shortcut comes from the following approximation[4]: $a_n = \frac{b_n}{b_n^2+1}$ and

3 The convention we follow considers the Gumbel for minima only, with the properly parametrized EVT for the maxima.

4 Embrechts et al [81] proposes $a_n = \frac{1}{\sqrt{2\log(n)}}$, $b_n = \sqrt{2\log(n)} - \frac{\log(\log(n))+\log(4\pi)}{2\sqrt{2\log(n)}}$, the second term for b_n only needed for large values of n. The approximation is of order $\sqrt{\log(n)}$.

$b_n = -\sqrt{2}\text{erfc}^{-1}\left(2\left(1-\frac{1}{n}\right)\right)$, where erfc^{-1} is the inverse complementary error function.

Figure 9.5: *The high watermark: the level of flooding in Paris in 1910 as a maxima. Clearly one has to consider that such record will be topped some day in the future and proper risk management consists in "how much" more than such a level one should seek protection. We have been repeating the Lucretius fallacy forever.*

Property 6

For tail risk and properties, it is vastly preferable to work with the exact distribution for the Gaussian, namely for n variables, we have the exact distribution of the maximum from the CDF of the Standard Gaussian $F^{(g)}$:

$$\frac{\partial F^{(g)}(K)}{\partial K} = \frac{e^{-\frac{K^2}{2}} 2^{\frac{1}{2}-n} n\, erfc\left(-\frac{K}{\sqrt{2}}\right)^{n-1}}{\sqrt{\pi}}, \tag{9.6}$$

where erfc is the complementary error function.

9.1.3 The Picklands-Balkema-de Haan Theorem

The conditional excess distribution function is the equivalent in density to the "Lindy" conditional expectation of excess deviation [115, 187], –we will make use of it in Chapter 16.

Consider an unknown distribution function F of a random variable X; we are interested in estimating the conditional distribution function F_u of the variable X above a certain threshold u, defined as

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)} \tag{9.7}$$

for $0 \leq y \leq x^* - u$, where x^* is the finite or infinite right endpoint of the underlying distribution F. Then there exists a measurable function $\sigma(u)$ such that

$$\lim_{u \to x^*} \sup_{0 \leq x < x^* - u} \left| F_u(x) - G_{\xi,\sigma(u)}(x) \right| = 0 \tag{9.8}$$

and vice versa where $G_{\xi,\sigma(u)}(x)$ is the generalized Pareto distribution (GPD) :

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - (1 + \xi x/\sigma)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-x/\sigma) & \text{if } \xi = 0 \end{cases} \tag{9.9}$$

If $\xi > 0$, $G_{.,.}$ is a Pareto distribution. If $\xi = 0$, $G_{.,.}$ (as we saw above) is a an exponential distribution. If $\xi = -1$, $G_{.,.}$ is uniform.

The theorem allows us to do some data inference by isolating exceedances. More on it in our discussion of wars and trends of violence in Chapter 16.

9.2 THE INVISIBLE TAIL FOR A POWER LAW

Consider K_n the maximum of a sample of n independent identically distributed variables in the power law class; $K_n = \max(X_1, X_2, \ldots, X_n)$. Let $\phi(.)$ be the density of the underlying distribution. We can decompose the moments in two parts, with the "hidden" moment above K_0, as shown in Fig 9.6:

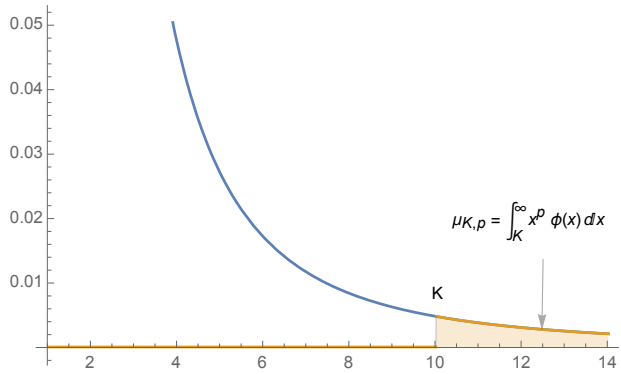


Figure 9.6: *The p^{th} moment above K*

$$\mathbb{E}(X^p) = \underbrace{\int_L^{K_n} x^p \phi(x)\, dx}_{\mu_{0,p}} + \underbrace{\int_{K_n}^{\infty} x^p \phi(x)\, dx}_{\mu_{K,p}}$$

where μ_0 is the visible part of the distribution and μ_n the hidden one.

We can also consider using ϕ_e as the empirical distribution by normalizing. Since:

$$\underbrace{\left(\int_L^{K_n} \phi_e(x) dx - \int_{K_n}^{\infty} \phi(x)\, dx \right)}_{\text{Corrected}} + \int_{K_n}^{\infty} \phi(x)\, dx = 1, \tag{9.10}$$

we can use the Radon-Nikodym derivative

$$\mathbb{E}(X^p) = \int_L^{K_n} x^p \frac{\partial \mu(x)}{\partial \mu_e(x)} \phi_e(x) dx + \int_{K_n}^{\infty} x^p \phi(x)\, dx. \tag{9.11}$$

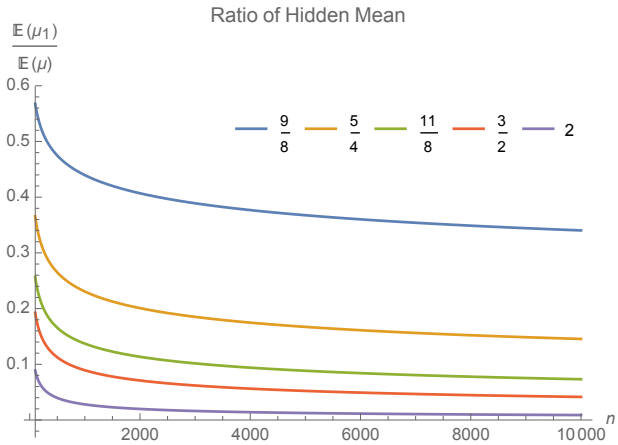


Figure 9.7: *Proportion of the hidden mean in relation to the total mean, for different parametrizations of the tail exponent α.*

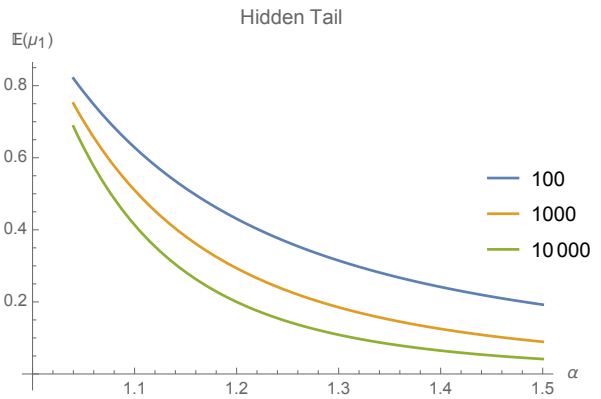


Figure 9.8: *Proportion of the hidden mean in relation to the total mean, for different sample sizes n.*

Proposition 9.1
Let K^* be point where the survival function of the random variable X can be satisfactorily approximated by a constant, that is $\mathbb{P}(X > x) \approx L^{-\alpha} x^{-\alpha}$.

Under the assumptions that $K > K^*$, the distribution for the hidden moment, $\mu_{K,p}$, for n observation has for density $g_{(.,.,.)}(.)$:

$$g_{n,p,\alpha}(z) = nL^{\frac{\alpha p}{p-\alpha}} \left(z - \frac{pz}{\alpha}\right)^{\frac{p}{\alpha-p}} \exp\left(n\left(-L^{\frac{\alpha p}{p-\alpha}}\right)\left(z - \frac{pz}{\alpha}\right)^{-\frac{\alpha}{p-\alpha}}\right) \tag{9.12}$$

for $z \geq 0$, $p > \alpha$, and $L > 0$.

The expectation of the p^{th} moment above K, with $K > L > 0$ can be derived as

$$\mathbb{E}(\mu_{K,p}) = \frac{\alpha \left(L^p - L^\alpha K^{p-\alpha}\right)}{\alpha - p}. \tag{9.13}$$

We note that the distribution of the sample survival function (that is, $p = 0$) is an exponential distribution with pdf:

$$g_{n,0,\alpha}(z) = ne^{-nz} \tag{9.14}$$

which we can see depends only on n. Exceedance probability for an empirical distribution does not depend on the fatness of the tails.

To get the mean, we just need to get the integral with a stochastic lower bound $K > K_{min}$:

$$\int_{K_{min}}^{\infty} \left(\underbrace{\int_{K_n}^{\infty} x^p \phi(x)\, dx}_{\mu_{K,p}} \right) f_K(K) dK.$$

For the full distribution $g_{n,p,\alpha}(z)$, let us decompose the mean of a Pareto with scale L, so $K_{min} = L$.

By standard transformation, a change of variable, $K \sim \mathcal{F}(\alpha, Ln^{\frac{1}{\alpha}})$ a Fréchet distribution with PDF: $f_K(K) = \alpha n K^{-\alpha-1} L^\alpha e^{n\left(-\left(\frac{L}{K}\right)^\alpha\right)}$, we get the required result.

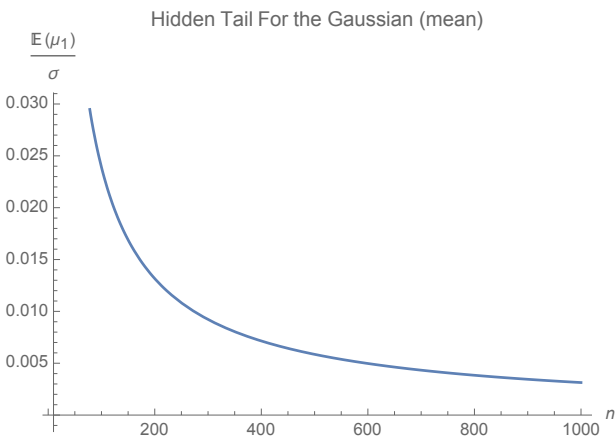


Figure 9.9: *Proportion of the hidden mean in relation to the standard deviation, for different values of n.*

9.2.1 Comparison with the Normal Distribution

For a Gaussian with PDF $\phi^{(g)}(.)$ indexed by (g), $\mu_K^{(g)} = \int_K^\infty \phi^{(g)}(x)dx = \frac{2^{\frac{p}{2}-1}\Gamma\left(\frac{p+1}{2}, \frac{K^2}{2}\right)}{\sqrt{\pi}}$. As we saw earlier, without going through the Gumbel (rather EVT or "mirror-Gumbel"), it is preferable to the exact distribution of the maximum from the CDF of the Standard Gaussian $F^{(g)}$:

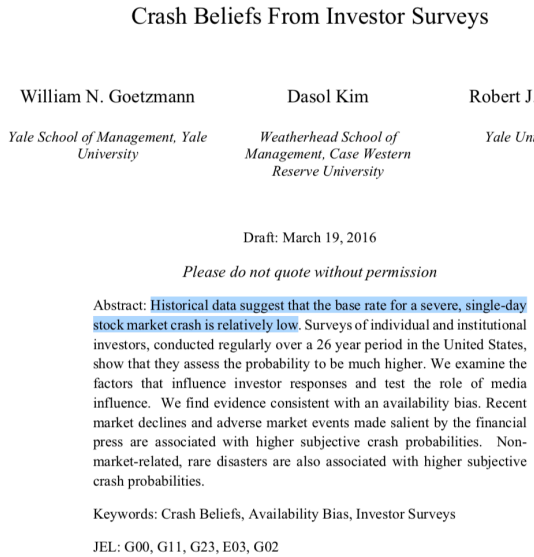
$$\frac{\partial F^{(g)}(K)}{\partial K} = \frac{e^{-\frac{K^2}{2}} 2^{\frac{1}{2}-n} n \operatorname{erfc}\left(-\frac{K}{\sqrt{2}}\right)^{n-1}}{\sqrt{\pi}},$$

where ertc is the complementary error function

For $p = 0$, the expectation of the "invisible tail" $\approx \frac{1}{n}$.

$$\int_0^\infty \frac{e^{-\frac{K^2}{2}} 2^{-n-\frac{1}{2}} n \Gamma\left(\frac{1}{2}, \frac{K^2}{2}\right)\left(\operatorname{erf}\left(\frac{K}{\sqrt{2}}\right)+1\right)^{n-1}}{\pi} dK = \frac{1-2^{-n}}{n+1}.$$

9.3 APPENDIX: THE EMPIRICAL DISTRIBUTION IS NOT EMPIRICAL



Crash Beliefs From Investor Surveys

William N. Goetzmann — *Yale School of Management, Yale University*

Dasol Kim — *Weatherhead School of Management, Case Western Reserve University*

Robert J. Shiller — *Yale University*

Draft: March 19, 2016

Please do not quote without permission

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

Keywords: Crash Beliefs, Availability Bias, Investor Surveys

JEL: G00, G11, G23, E03, G02

Figure 9.10: *The base rate fallacy, revisited —or, rather in the other direction. The "base rate" is an empirical evaluation that bases itself on the worst past observations, an error identified in [224] as the fallacy identified by the Roman poet Lucrecius in De rerum natura of thinking the tallest future mountain equals the tallest on has previously seen. Quoted without permission after warning the author.*

There is a prevalent confusion about the nonparametric empirical distribution based on the following powerful property: as n grows, the errors around the empirical histogram for cumulative frequencies are Gaussian *regardless of the base distribution*, even if the true distribution is fat-tailed (assuming infinite support). For the CDF (or survival functions) are both uniform on $[0, 1]$, and, further, by the

Donsker theorem, the sequence $\sqrt{n}\,(F_n(x) - F(x))$ (F_n is the observed CDF or survival function for n summands, F the true CDF or survival function) converges in distribution to a Normal Distribution with mean 0 and variance $F(x)\,(1 - F(x))$ (one may find even stronger forms of convergence via the Glivenko– Cantelli theorem).

Owing to this remarkable property, one may mistakenly assume that the effect of tails of the distribution converge in the same manner independently of the distribution. Further, and what contributes to the confusion, the variance, $F(x)\,(1 - F(x))$ for both empirical CDF and survival function, drops at the extremes –though not its corresponding payoff.

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

For the U.S. stock market indices, while the first method is deemed to be ludicrous, using the second method leads to an underestimation of the payoff in the tails of between 5 and 70 times, as can be shown in Figure 9.11. The topic is revisited again in Chapter 11 with our discussion of the difference between binary and continuous payoffs, and the conflation between probability and real world payoffs when said payoffs are from a fat tailed distribution.

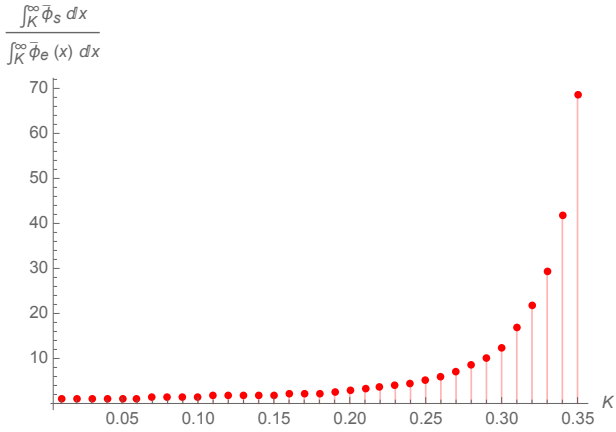


Figure 9.11: *This figure shows the relative value of tail CVar-style measure compared to that from the (smoothed) empirical distribution. The deep tail is underestimated up to 70 times by current methods, even those deemed "empirical".*

B | THE LARGE DEVIATION PRINCIPLE, IN BRIEF

LET US RETURN to the Cramer bound with a rapid exposition of the surrounding literature. The idea behind the tall vs. rich outliers in 3.1 is that under some conditions, their tail probabilities decay exponentially. Such a property that is central in risk management –as we mentioned earlier, the catastrophe principle explains that for diversification to be effective, such exponential decay is necessary.

The large deviation principle helps us understand such a tail behavior.It also helps us figure out why things do not blow-up under thin-tailedness –but, more significantly, why they could under fat tails, or where the Cramèr condition is not satisfied [117].

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

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

where $I(.)$ is the Cramer (or rate) function (Varadhan [255], Denbo and Zeitouni [58]). If we know the distribution of X, then, by Legendre transformation, $I(x) = \sup_{\theta>0}(\theta x - \lambda(\theta))$, where $\lambda(\theta) = \log \mathbb{E}\left(\mathrm{e}^{\theta(X)}\right)$ is the cumulant generating function.

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

A special case for Bernoulli variables is the Chernoff Bound, which provides tight bounds for that class of discrete variables.

SIMPLE CASE: CHERNOFF BOUND

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

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

and for $0 < \delta \leq 1$

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

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

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

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

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

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

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

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

Substituting in B.1, we get:

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

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

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

An extension of Chernoff bounds was made by Hoeffding [128] who broadened it to bounded independent random variables, but not necessarily Bernouilli..

C | CALIBRATING UNDER PARETIANITY

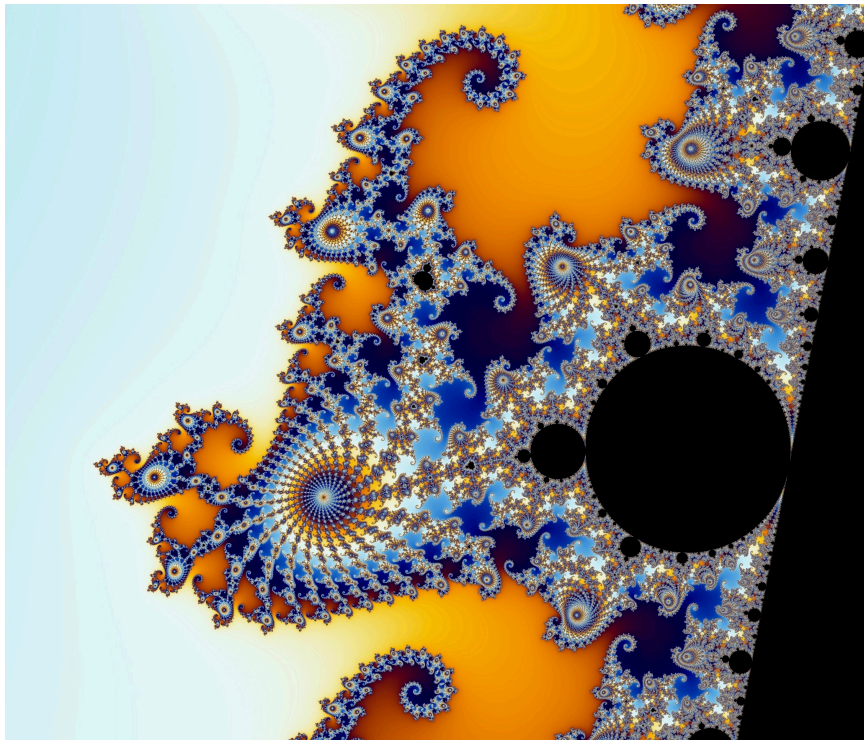


Figure C.1: *The great Benoit Mandelbrot linked fractal geometry to statistical distributions via self-affinity at all scales. When asked to explain his work, he said: "rugosité", meaning"roughness" –it took him fifty years to realize that was his specialty. (Seahorse Created by Wolfgang Beyer, Wikipedia Commons.)*

We start with a refresher:

Definition C.1 (Power Law Class $\mathfrak{P}$)
The r.v. $X \in \mathbb{R}$ belongs to $\mathfrak{P}$, the class of slowly varying functions (a.k.a. Paretian tail or

power law-tailed) if its survival function (for the variable taken in absolute value) decays asymptotically at a fixed exponent α, or α', that is

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

(right tail) or

$$\mathbb{P}(-X > x) \sim L(x)\, x^{-\alpha'} \tag{C.2}$$

(left tail)
where $\alpha, \alpha' > 0$ *and* $L : (0, \infty) \to (0, \infty)$ *is a slowly varying function, defined as* $\lim_{x\to\infty} \frac{L(kx)}{L(x)} = 1$ *for all* $k > 0$.

The happy result is that the parameter α obeys an inverse gamma distribution that converges rapidly to a Gaussian and does not require a large n to get a good estimate. This is illustrated in Figure C.2, where we can see the difference in fit.

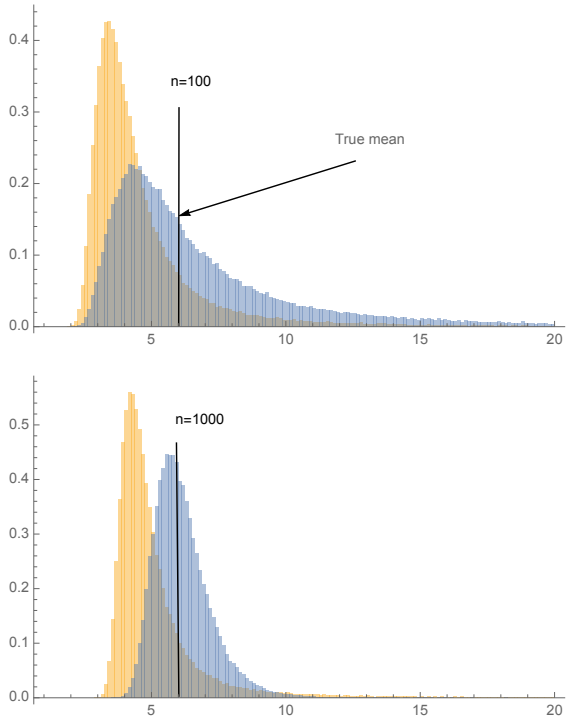


Figure C.2: *Monte Carlo Simulation* (10^5) *of a comparison of sample mean (Methods 1 and 2) vs maximum likelihood mean estimations (Method 3) for a Pareto Distribution with* $\alpha = 1.2$ *(yellow and blue respectively), for* $n = 100, 1000$. *We can see how the MLE tracks the distribution more reliably. We can also observe the bias as Methods 1 and 2 underestimate the sample mean in the presence of skewness in the data. We need* 10^7 *more data in order to get the same error rate.*

As we saw, there is a problem with the so-called finite variance power laws: finiteness of variance doesn't help as we saw in Chapter 8.

C.1 DISTRIBUTION OF THE SAMPLE TAIL EXPONENT

Consider the standard Pareto distribution for a random variable X with pdf:

$$\phi_X(x) = \alpha L^\alpha x^{-\alpha-1}, x > L \tag{C.3}$$

Assume $L = 1$ by scaling.

The likelihood function is $\mathcal{L} = \prod_{i=1}^{n} \alpha x_i^{-\alpha-1}$. Maximizing the Log of the likelihood function (assuming we set the minimum value) $\log(\mathcal{L}) = n(\log(\alpha) + \alpha \log(L)) - (\alpha + 1)\sum_{i=1}^{n} \log(x_i)$ yields: $\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log(x_i)}$. Now consider $l = -\frac{\sum_{i=1}^{n} \log X_i}{n}$. Using the characteristic function to get the distribution of the average logarithm yield:

$$\psi(t)^n = \left(\int_1^\infty f(x) \exp\left(\frac{it \log(x)}{n} \right) dx \right)^n = \left(\frac{\alpha n}{\alpha n - it} \right)^n$$

which is the characteristic function of the gamma distribution $(n, \frac{1}{\alpha n})$. A standard result is that $\hat{\alpha}' \triangleq \frac{1}{l}$ will follow the inverse gamma distribution with density:

$$\phi_{\hat{\alpha}}(a) = \frac{e^{-\frac{\alpha n}{\hat{\alpha}}} \left(\frac{\alpha n}{\hat{\alpha}}\right)^n}{\hat{\alpha}\Gamma(n)}, a > 0$$

.

Debiasing Since $\mathbb{E}(\hat{\alpha}) = \frac{n}{n-1}\alpha$ we elect another –unbiased– random variable $\hat{\alpha'} = \frac{n-1}{n}\hat{\alpha}$ which, after scaling, will have for distribution $\phi_{\hat{\alpha'}}(a) = \frac{e^{\frac{\alpha-\alpha n}{a}} \left(\frac{\alpha(n-1)}{a}\right)^{n+1}}{\alpha\Gamma(n+1)}$.

Truncating for $\alpha > 1$ Given that values of $\alpha \leq 1$ lead to absence of mean we restrict the distribution to values greater than $1 + \epsilon$, $\epsilon > 0$. Our sampling now applies to lower-truncated values of the estimator, those strictly greater than 1, with a cut point $\epsilon > 0$, that is, $\sum \frac{n-1}{\log(x_i)} > 1 + \epsilon$, or $\mathbb{E}(\hat{\alpha}|_{\hat{\alpha}>1+\epsilon})$: $\phi_{\hat{\alpha''}}(a) = \frac{\phi_{\hat{\alpha'}}(a)}{\int_{1+\epsilon}^{\infty} \phi_{\hat{\alpha'}}(a)\,da}$, hence the distribution of the values of the exponent conditional of it being greater than 1 becomes:

$$\phi_{\hat{\alpha''}}(a) = \frac{e^{\frac{\alpha n^2}{a-an}} \left(\frac{\alpha n^2}{a(n-1)}\right)^n}{a\left(\Gamma(n) - \Gamma\left(n, \frac{n^2\alpha}{(n-1)(\epsilon+1)}\right)\right)}, a \geq 1 + \epsilon \tag{C.4}$$

So as we can see in Figure C.2, the "plug-in" mean via the tail α might be a good approach under one-tailed Paretianity.

10 "IT IS WHAT IT IS": DIAGNOSING THE SP500†

THIS IS A DIAGNOSTICS tour of the properties of the SP500 index in its history. We engage in a battery of tests and check what statistical picture emerges. Clearly, its returns are power law distributed (with some added complications, such as an asymmetry between upside and downside) which, again, invalidates common methods of analysis. We look, among other things to:

- The behavior of Kurtosis under aggregation (as we lengthen the observation window)
- The behavior of the conditional expectation $\mathbb{E}(X|_{X>K})$ for various values of K.
- The maximum-to-sum plot(MS Plot).
- Drawdowns (that is, maximum excursions over a time window)
- Extremes and records to see if extremes are independent.

These diagnostics allow us to confirm that an entire class of analyses in $L2$ such as modern portfolio theory, factor analysis, GARCH, conditional variance, or stochastic volatility are methodologically (and practically) invalid.

10.1 PARETIANITY AND MOMENTS

The problem As we said in the Prologue, hanging from thin-tailed to fat-tailed is not just *changing the color of the dress*. The finance and economic rent seekers hold the message "we know it is fat tailed" but then fail to grasp the consequences on many things such as the slowness of the law of large numbers and the failure of

† This is largely a graphical chapter made to be read from the figures more than from the text as the arguments largely repose on the absence of convergence in the graphs.

sample means or higher moments to be sufficient statistic (as well as the ergodicity effect, among others).

Paretianity is clearly defined by the absence of some higher moment, exhibited by lack of convergence under LLN.

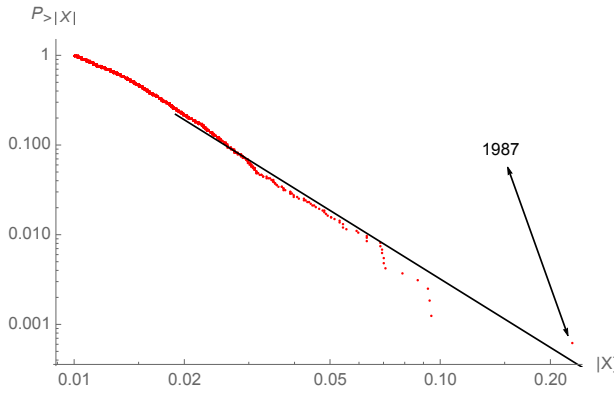


Figure 10.1: *Visual Identification of Paretianity on a standard log-log plot with (absolute) returns on the horizontal axis and the survival function on the vertical one. If one removes the data point corresponding to the crash of 1987, a lognormal would perhaps work, or some fat tailed mixed distribution outside the power law class. For we can see the survival function becoming vertical, indicative of an infinite asymptotic tail exponent. But as the saying goes, all one needs is a single event...*

Remark 7

Given that:

1) the regularly varying class has no higher moments than α, more precisely,

- *if $p > \alpha$, $\mathbb{E}(X^p) = \infty$ if p is even or the distribution has one-tailed support and*
- *$\mathbb{E}(X^p)$ is undefined if p is odd and the distribution has two-tailed support,*

and

2) distributions outside the regularly varying class have all moments $\forall p \in \mathbb{N}^+, \mathbb{E}(X^p) < \infty$.

$\exists p \in \mathbb{N}^+$ s.t. $\mathbb{E}(X^p)$ is either undefined or infinite $\Leftrightarrow X \in \mathfrak{P}$.

Next we examine ways to detect "infinite" moments. Much confusion attends the notion of infinite moments and its identification since by definition sample moments are finite and measurable under the counting measure. We will rely on the nonconvergence of moments. Let $\|\mathbf{X}\|_p$ be the weighted p-norm

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

we have the property of power laws:

$$\mathbb{E}(X^p) \not< \infty \Leftrightarrow \|\mathbf{x}\|_p \text{ is not convergent.}$$

Question How does belonging to the class of Power Law tails (with $\alpha \leq 4$) cancel much of the methods in *L2*?

Section 5.8 shows the distribution of the mean deviation of the second moment for a finite variance power law. Simply, even if the fourth moment does not exist, under infinite higher moments, the second moment of the variance has itself infinite variance, and we fall in the sampling problems seen before: just as with a power law of α close to 1 (though slightly above it), the mean exists but will never be observed, in a situation of infinite third moment, the observed second moment will fail to be informative as it will almost never converge to its value.

10.2 CONVERGENCE TESTS

Convergence laws can help us *exclude* some classes of probability distributions.

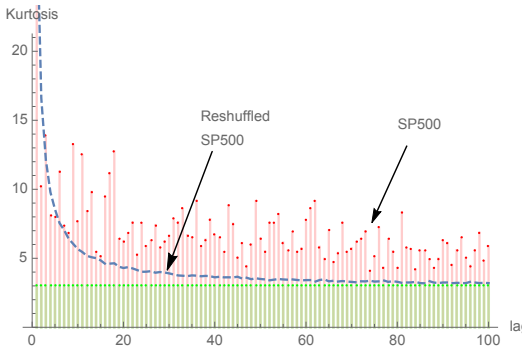


Figure 10.2: *Visual convergence diagnostics for the kurtosis of the SP500 over the past 17000 observations. We compute the kurtosis at different lags for the raw SP500 and reshuffled data. While the 4^{th} norm is not convergent for raw data, it is clearly so for the reshuffled series. We can thus assume that the "fat tailedness" is attributable to the temporal structure of the data, particularly the clustering of its volatility. See Table 7.1 for the expected drop at speed $1/n$ for thin-tailed distributions.*

10.2.1 Test 1: Kurtosis under Aggregation

If Kurtosis existed, it would end up converging to that of a Gaussian as one lengthens the time window. So we test for the computations of returns over longer and longer lags, as we can see in Fig 10.2.

Result The verdict as shown in Figure 10.2 is that the one-month kurtosis is not lower than the daily kurtosis and, as we add data, no drop in kurtosis is observed. Further we would expect a drop $\sim n^{-1}$. This allows us to safely eliminate numerous classes, which includes stochastic volatility in its simple formulations such as gamma variance. Next we will get into the technicals of the point and the strength of the evidence.

A typical misunderstanding is as follows. In a note "What can Taleb learn from Markowitz" [250], Jack L. Treynor, one of the founders of portfolio theory, defended the field with the argument that the data may be fat tailed "short term" but in something called the "long term" things become Gaussian. Sorry, it is not so.

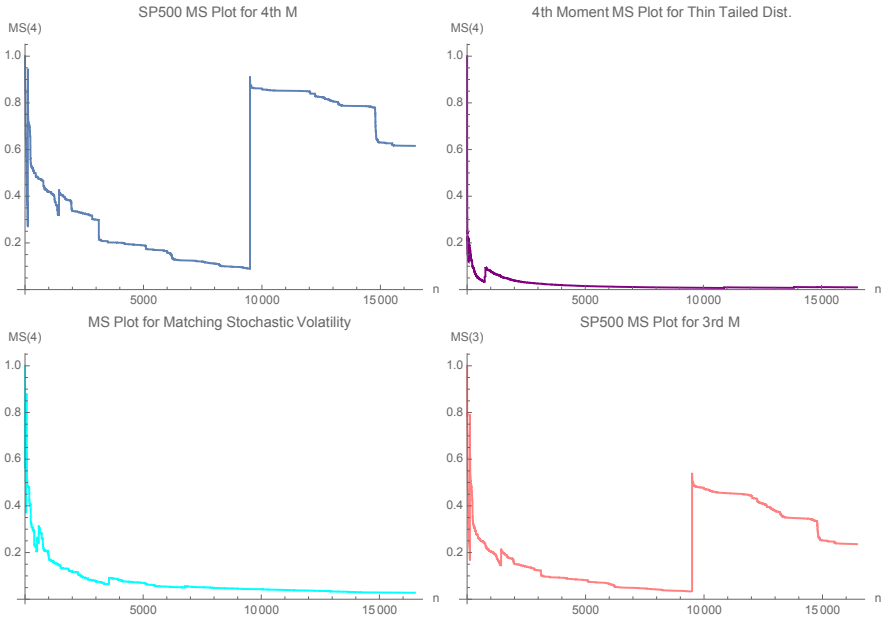


Figure 10.3: *MS Plot (or "law of large numbers for p moments") for $p = 4$ for the SP500 compared to $p = 4$ for a Gaussian and stochastic volatility for a matching Kurtosis (30) over the entire period. Convergence, if any, does not take place in any reasonable time. MS Plot for moment $p = 3$ for the SP500 compared to $p = 4$ for a Gaussian. We can safely say that the 4^{th} moment is infinite and the 3^{rd} one is indeterminate*

(We add the ergodic problem that blurs, if not eliminate, the distinction between long term and short term).

The reason is that, simply we cannot possibly talk about "Gaussian" if kurtosis is infinite, even when lower moments exist. Further, for $\alpha \approx 3$, Central limit operates very slowly, requires n of the order of 10^6 to become acceptable, not what we have in the history of markets. [27]

10.2.2 Maximum Drawdowns

For a time series for asset S taken over $(t_0, t_0 + \Delta t, t_0 + n\Delta t)$, we are interested in the behavior of

$$\delta\left(t_0, t, \Delta t\right) = \text{Min}\left(S_{i\Delta t+t_0} - \left(\text{Min} S_{j\Delta t+t_0}\right)_{j=i+1}^{n}\right)_{i=0}^{n} \tag{10.1}$$

We can consider the relative drawdown by using the log of that minimum, as we do with returns. The window for the drawdown can be $n = 5,100,252$ days. As seen in Figure **??**, drawdowns are Paretian.

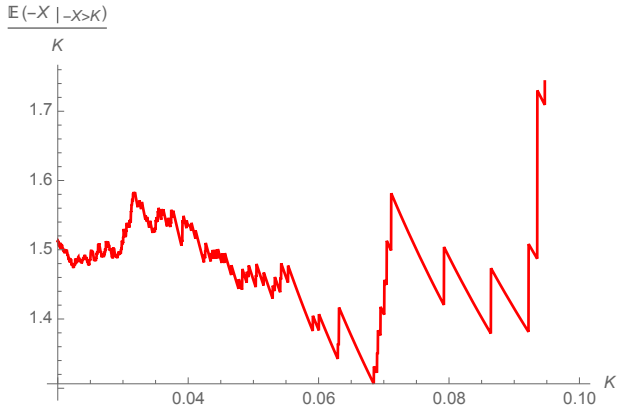


Figure 10.4: *The "Lindy test" or Condexp, using the conditional expectation below K as K varies as test of scalability. As we move K, the measure should drop.*

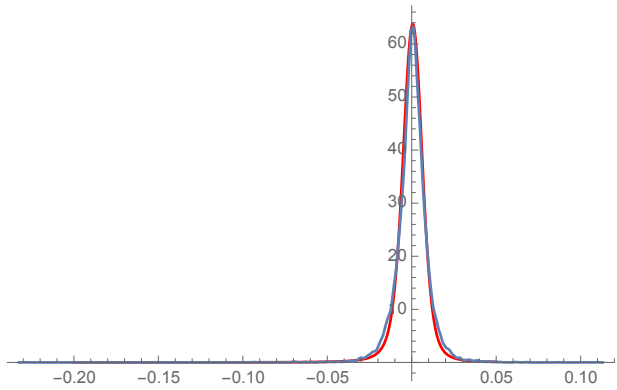


Figure 10.5: *The empirical distribution could conceivably fit a Lévy stable distribution with $\alpha_l = 1.62$.*

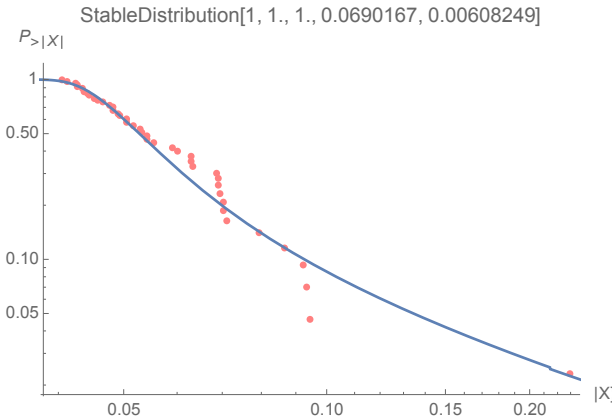


Figure 10.6: *The tails can even possibly fit an infinite mean stable distribution with $\alpha_l = 1$.*

10.2.3 Empirical Kappa

From our kappa equation in Chapter 8:

$$\kappa(n_0, n) = 2 - \frac{\log(n) - \log(n_0)}{\log\left(\frac{\mathbb{M}(n)}{\mathbb{M}(n_0)}\right)}. \tag{10.2}$$

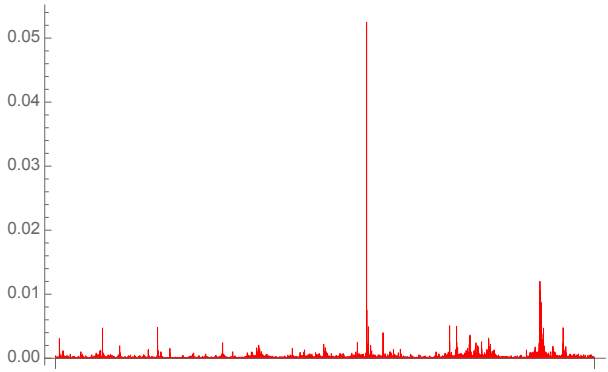


Figure 10.7: *SP500 squared returns for 16500 observations. No GARCH(1,1) can produce such jaggedness or what the great Benoit Mandelbrot called "rugosité".*

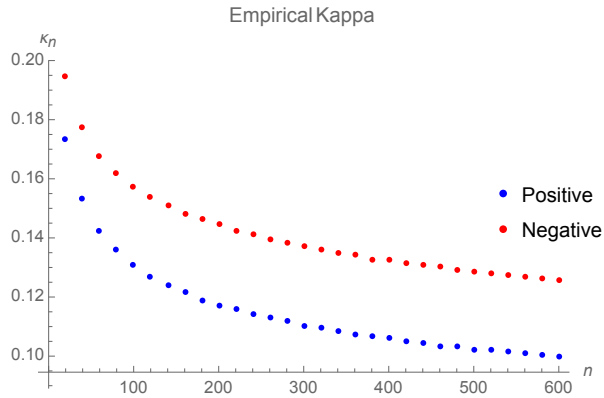


Figure 10.8: *kappa-n estimated empirically.*

with shortcut $\kappa_n = \kappa(1, n)$. We estimate empirically via bootstrapping and can effectively see how it maps to that of a power law – with $\alpha < 3$ for the negative returns.

10.2.4 Test 2: Excess Conditional Expectation

Result: The verdict from this test is that, as we can see in Figure 10.4, that the conditional expectation of X (and $-X$), conditional on X is greater than some arbitrary value K, remains proportional to K.

Definition 10.1
Let K be in $\mathbb{R}^+$*, the relative excess conditional expectation:*

$$\varphi_K^+ \triangleq \frac{\mathbb{E}(X)|_{X>K}}{K}$$

$$\varphi_K^- \triangleq \frac{\mathbb{E}(-X)|_{X>K}}{K}$$

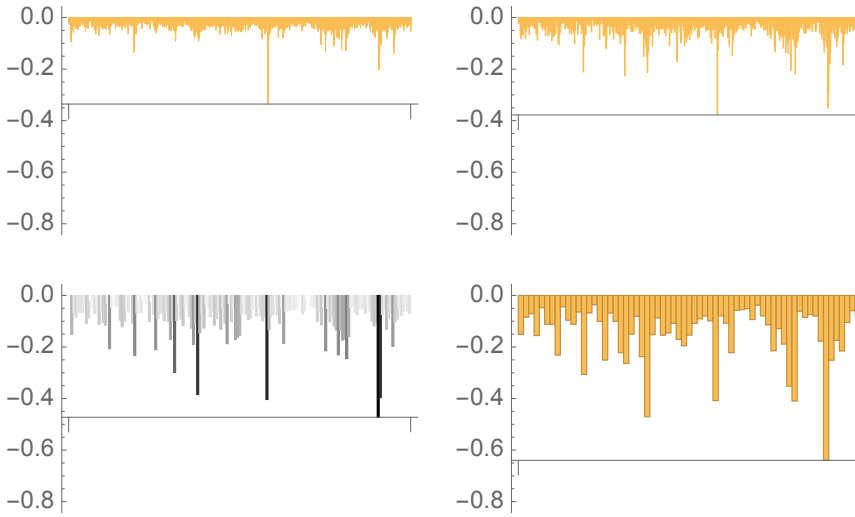


Figure 10.9: *Drawdowns for windows n = 5, 30, 100, and 252 days, respectively. Maximum drawdowns are excursions mapped in Eq. 10.1. We use here the log of the minimum of S over a window of n days following a given S.*

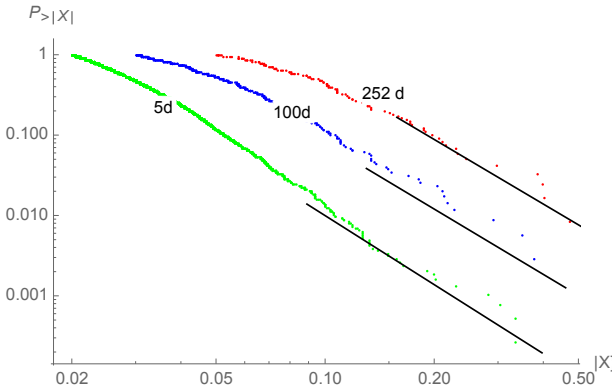


Figure 10.10: *Paretianity of Drawdowns and Scale*

We have

$$\lim_{K\to\infty} \varphi_K = 0$$

for distributions outside the power law basin, and

$$\lim_{K\to\infty} \varphi_K/K = \frac{\alpha}{1-\alpha}$$

for distribution satisfying Definition 1. Note the van der Wijk's law [44],[225].

Figure 10.4 shows the following: the conditional expectation does not drop for large values, which is incompatible with non-Paretian distributions.

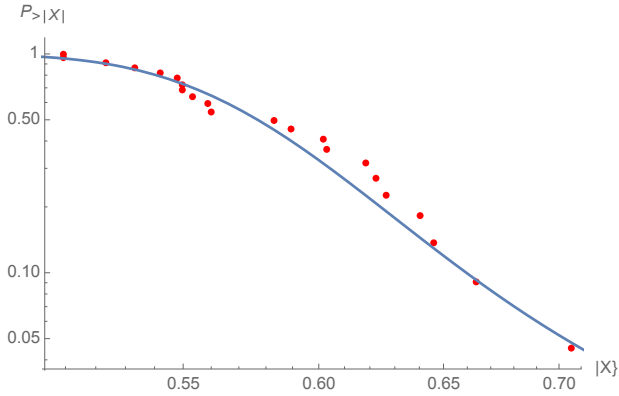


Figure 10.11: *Fitting a Stable Distribution to drawdowns*

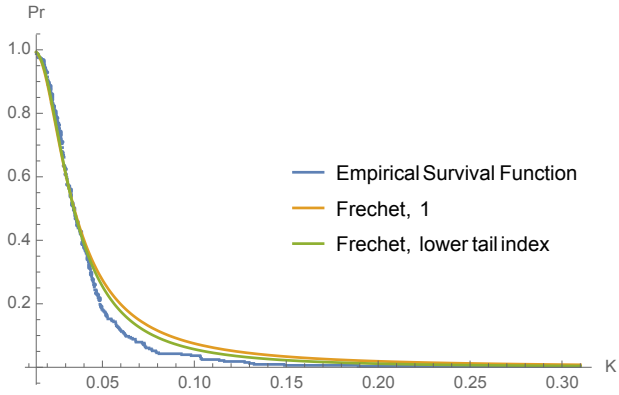


Figure 10.12: *Correcting the empirical distribution function with a Frechet for the SP500*

10.2.5 Test 3- Instability of 4^{th} moment

A main argument in [225] is that in 50 years of SP500 observations, a single one represents >80 % of the Kurtosis. Similar effect are seen with other socioeconomic variables, such as gold, oil, silver other stock markets, soft commodities. Such sample dependence of the kurtosis means that the fourth moment does not have the stability, that is, does not exist.

10.2.6 Test 4: MS Plot

An additional approach to detect if $\mathbb{E}(X^p)$ exists consists in examining convergence according to the law of large numbers (or, rather, absence of), by looking the behavior of higher moments in a given sample. One convenient approach is the Maximum-to-Sum plot, or MS plot as shown in Figure 10.3. The MS Plot relies on a consequence of the law of large numbers [181] when it comes to the maximum

of a variable. For a sequence $X_1, X_2, ..., X_n$ of nonnegative i.i.d. random variables, if for $p = 1, 2, 3, \ldots$, $\mathbb{E}[X^p] < \infty$, then

$$R_n^p = M_n^p / S_n^p \rightarrow^{a.s.} 0$$

as $n \rightarrow \infty$, where $S_n^p = \sum_{i=1}^{n} X_i^p$ is the partial sum, and $M_n^p = \max(X_1^p, ..., X_n^p)$ the partial maximum. (Note that we can have X the absolute value of the random variable in case the r.v. can be negative to allow the approach to apply to odd moments.)

We show by comparison the MS plot for a Gaussian and that for a Student T with a tail exponent of 3. We observe that the SP500 show the typical characteristics of a steep power law, as in 16,000 observations (50 years) it does not appear to drop to the point of allowing the functioning of the law of large numbers.

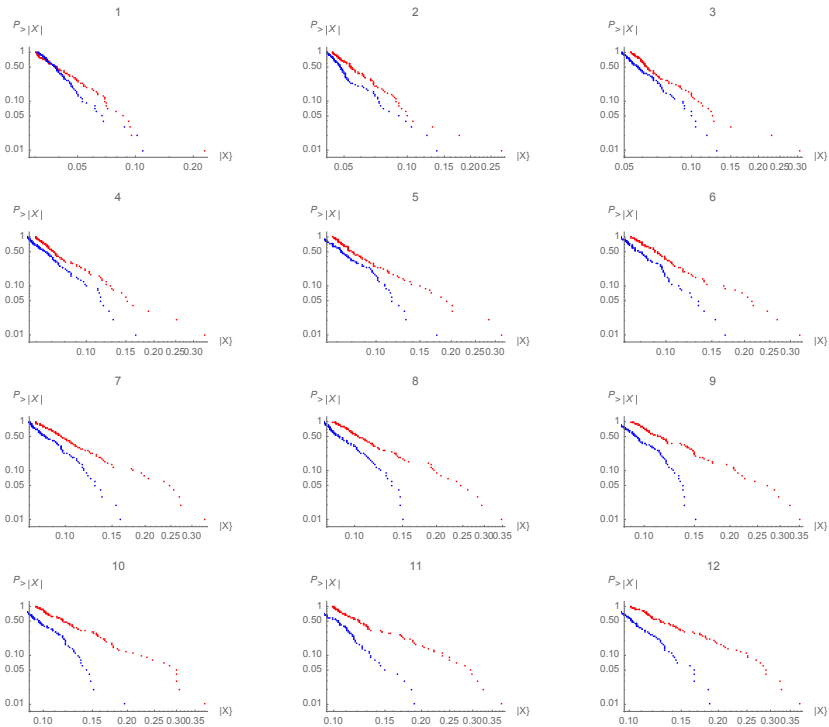


Figure 10.13: *We separate positive and negative logarithmic returns and use overlapping cumulative returns from 1 up to 15. Clearly the negative returns appear to follow a Power Law while the Paretianity of the right one is more questionable.*

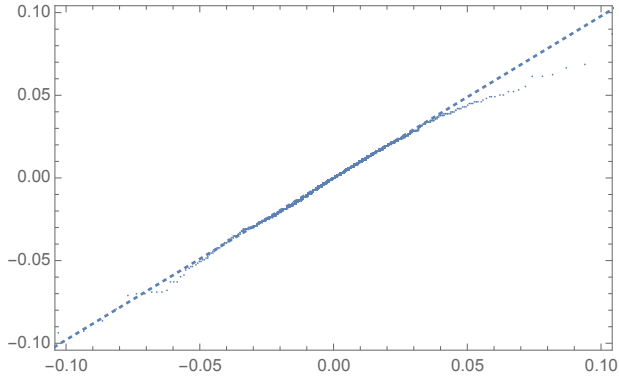


Figure 10.14: *QQ Plot comparing the Student T to the empirical distribution of the SP500: the left tail fits, not the right tail.*

10.2.7 Records and Extrema

The Gumbel record methods is as follows (Embrechts et al [81]). Let $X_1, X_2, \ldots$ be a discrete time series, with a maximum at period $t \geq 2$, $M_t = \max(X_1, X_2, \ldots, X_t)$, we have the record counter $N_{1,t}$ for n data points.

$$N_{1,t} = 1 + \sum_{k=2}^{t} \mathbb{1}_{X_t > M_{t-1}} \tag{10.3}$$

Regardless of the underlying distribution, the expectation $\mathbb{E}(N_t)$ is the Harmonic Number H_t, and the variance $H_t - H_t^2$, where $H_t = \sum_{i=1}^{t} \frac{1}{i^r}$. We note that the harmonic number is concave and very slow in growth, logarithmic, as it can be approximated with $log(n) + \gamma$, where γ is the Euler Mascheroni constant. The approximation is such that $\frac{1}{2(t+1)} \leq H_t - log(t) - \gamma \leq \frac{1}{2t}$ (*Wolfram Mathworld* [258]).

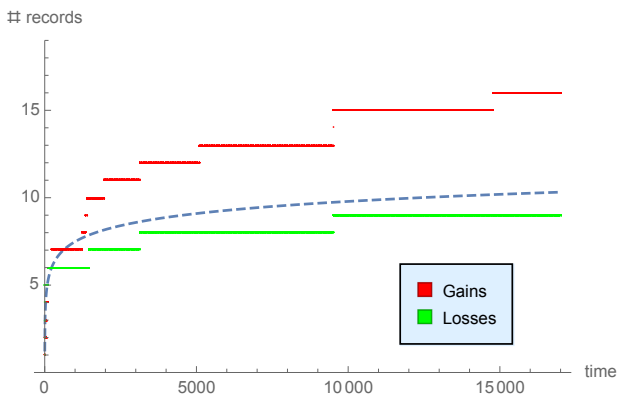


Figure 10.15: *The record test shows independence for extremes of negative returns, dependence for positive ones. The number of records for independent observations grows over time at the harmonic number $H(t)$ (dashed line), $\approx$ logarithmic but here appears to grow $>$ 2.5 standard deviations faster for positive returns, hence we cannot assume independence for extremal gains. The test does not make claims about dependence outside extremes.*

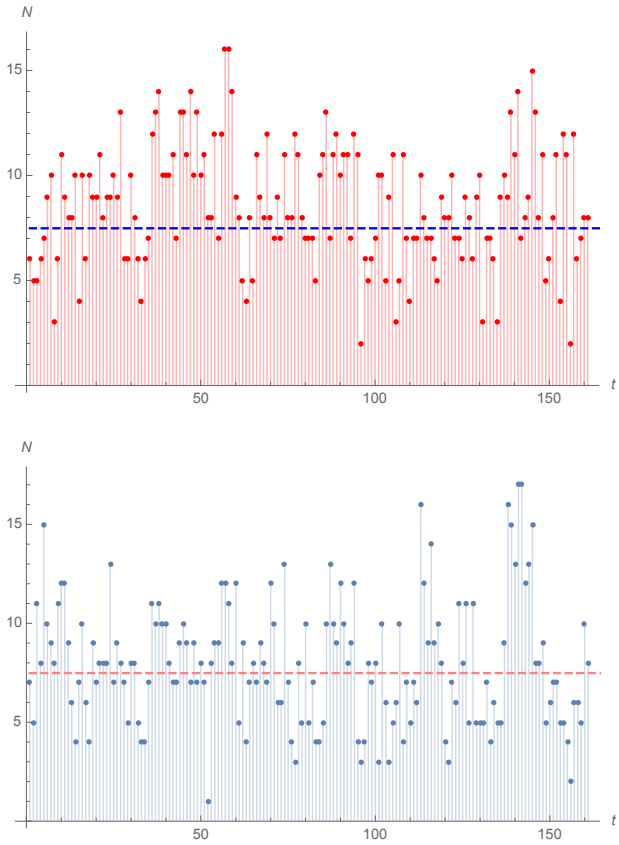


Figure 10.16: *Running shorter period, $t = 1000$ days of overlapping observations for the records of maxima(top) and minima (bottom), compared to the expected Harmonic number $H(1000)$.*

Remark 8

The Gumbel test of independence above is sufficient condition for the convergence of extreme negative values of the log-returns of the SP500 to the Maximum Domain of Attraction (MDA) of the extreme value distribution.

Entire series We reshuffled the SP500 (i.e. bootstrapped without replacement, using a sample size equal to the original ≈ 17000 points, with 10^3 repeats) and ran records across all of them. As shown in Fig. 10.18 and 10.17, the mean was 10.4 (approximated by the harmonic number, with a corresponding standard deviation.) The survival function S(.) of $N_{1.7\times10^4} = 16$, $S(16) = \frac{1}{40}$ which allows us to consider the independence of positive extrema implausible.

On the other hand the negative extrema (9 counts) show realizations close to what is expected (10.3), diverting by $\frac{1}{2}$ a s.t.d. from expected, enough to justify a failure to reject independence.

Subrecords If instead of taking the data as one block over the entire period, we broke the period into sub-periods, we get (because of the concavity of the measure and Jensen's inequality), $N_{t_1+\delta,t_1+\Delta+\delta}$, we obtain T/δ observations. We took $\Delta = 10^3$ and $\delta = 10^2$, thus getting 170 subperiods for the $T \approx 17 \times 10^3$ days. The picture as shown in Fig. 10.16 cannot reject independence for both positive and reject observations.

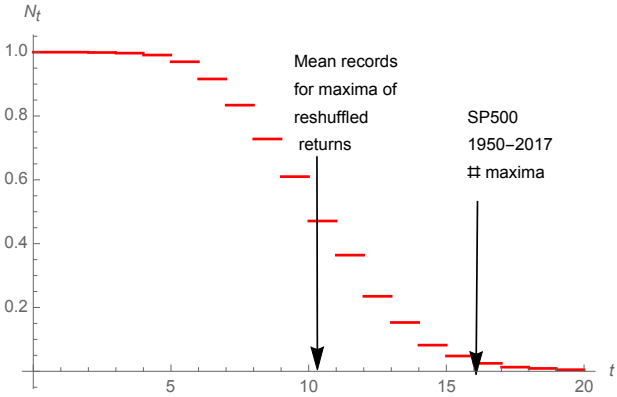


Figure 10.17: *The survival function of the records of positive maxima for the resampled SP500 (10^3 times) by keeping all returns but reshuffling them, thus removing the temporal structure. The mass above 16 (observed number of maxima records for SP500 over the period) is $\frac{1}{40}$.*

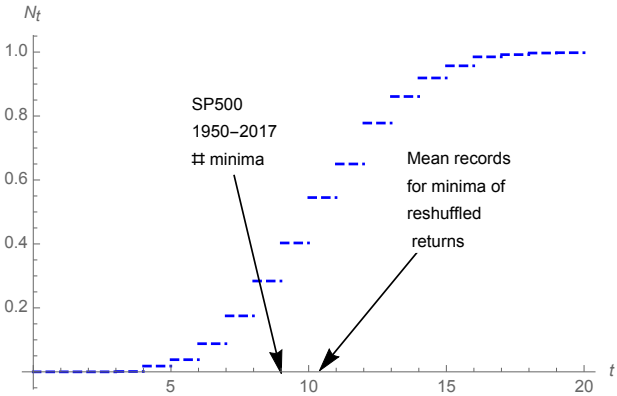


Figure 10.18: *The CDF of the records of negative extrema for the resampled SP500 (10^3 times) reshuffled as above. The mass above 9 (observed number of minima records for SP500 over the period) is $\frac{2}{5}$.*

Conclusion for subrecords We can at least apply EVT methods for negative observations.

10.2.8 Asymmetry right-left tail

We note an asymmetry as seen in Figure 10.13, with the left tail considerably thicker than the right one. It may be a nightmare for modelers looking for some precise process, but not necessarily to people interested in risk, and option trading.

10.3 CONCLUSION: IT IS WHAT IT IS

This chapter allowed us to explore a simple topic: returns on the SP500 index (which represents the bulk of the U.S. stock market capitalization) are, simply power law distributed –by Wittgenstein's ruler, it is irresponsible to model them in any other manner. Standard methods such as Modern Portfolio Theory (MPT) or "base rate crash" verbalisms (claims that people overestimate the probabilities of tail events) are totally bogus –we are talking of $> 70,000$ papers and entire cohorts of research, not counting about 10^6 papers in general economics with results depending on "variance" and "correlation". You need to live with the fact that these metrics are bogus. As the ancients used to say, *dura lex sed lex*, or in more modern mafia terms:

It is what it is.

D | THE PROBLEM WITH ECONOMETRICS

THERE IS SOMETHING WRONG with econometrics, as almost all papers don't replicate in the real world. Two reliability tests in Chapter 10, one about parametric methods the other about robust statistics, show that there must be something rotten in econometric methods, fundamentally wrong, and that the methods are not dependable enough to be of use in anything remotely related to risky decisions. Practitioners keep spinning inconsistent *ad hoc* statements to explain failures. This is a brief nontechnical exposition from the results in [225].

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

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

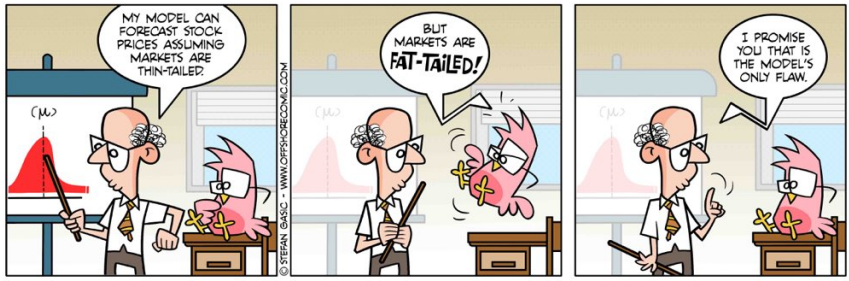


Figure D.1: *Credit: Stefan Gasic*

D.1 PERFORMANCE OF STANDARD PARAMETRIC RISK ESTIMATORS

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

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

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

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

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

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

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

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

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

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).

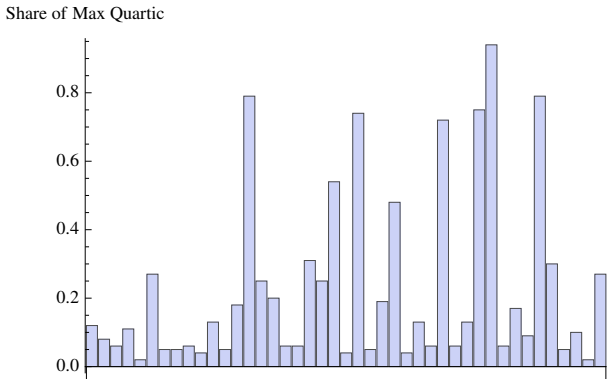


Figure D.2: *Max quartic across securities in Table D.1.*

D.2 PERFORMANCE OF STANDARD NONPARAMETRIC RISK ESTIMATORS

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

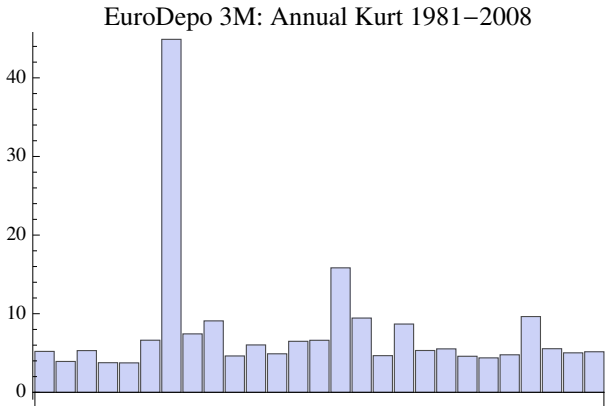


Figure D.3: *Kurtosis across nonoverlapping periods for Eurodeposits.*

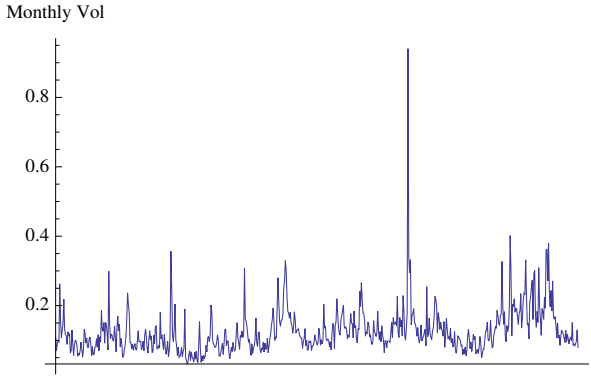


Figure D.4: *Monthly delivered volatility in the SP500 (as measured by standard deviations). The only structure it seems to have comes from the fact that it is bounded at 0. This is standard.*

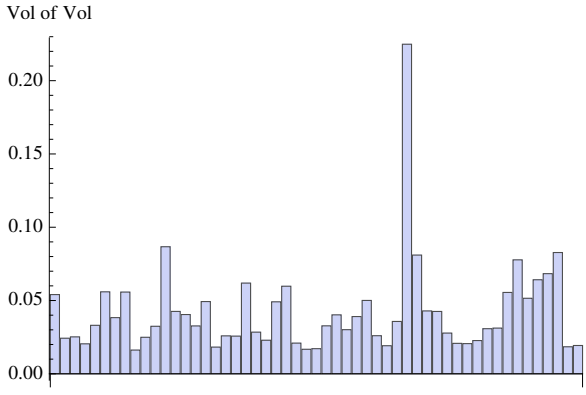


Figure D.5: *Montly volatility of volatility from the same dataset in Table D.1, predictably unstable.*

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 empirically observed moments and conditional moments

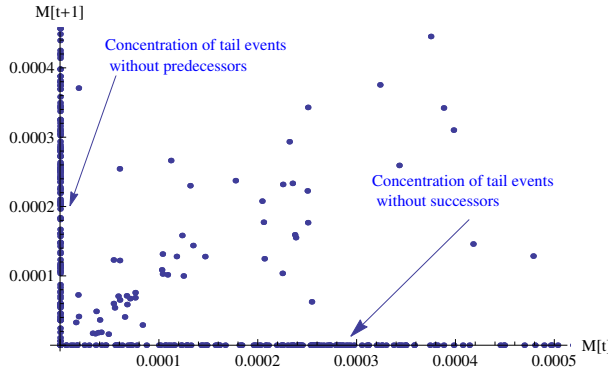


Figure D.6: *Comparing one absolute deviation M[t] and the subsequent one M[t+1] over a certain threshold (here 4% in stocks); illustrated how large deviations have no (or few) predecessors, and no (or few) successors– over the past 50 years of data.*

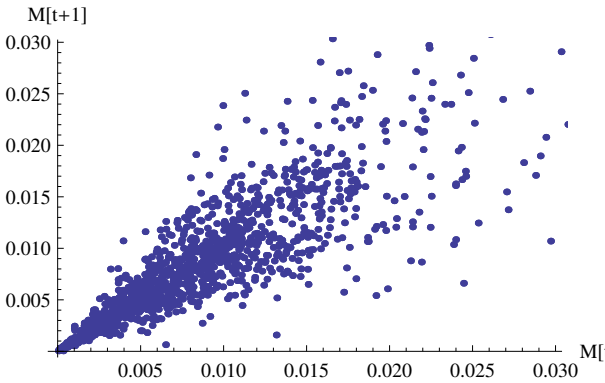


Figure D.7: *The "regular" is predictive of the regular, that is mean deviation. Comparing one absolute deviation M[t] and the subsequent one M[t+1] for macroeconomic data.*

extremely unstable. Tail events for a vector are vastly more difficult to calibrate, and increase in dimensions.

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

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

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

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

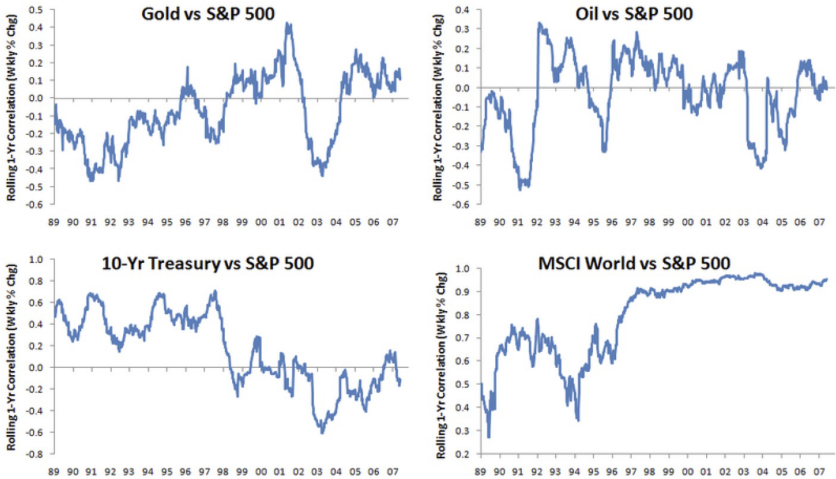


Figure D.8: *Correlations are also problematic, which flows from the instability of single variances and the effect of multiplication of the values of random variables. Under such stochasticity of correlations it makes no sense, no sense whatsoever, to use covariance-based methods such as portfolio theory.*

since much of the analytics and policies that can be destroyed by tail events are inapplicable.

Peso problem confusion about the Black Swan problem :

> "(...) "Black Swans" (Taleb, 2007). These cultural icons refer to disasters that occur so infrequently that they are virtually impossible to analyze using standard statistical inference. However, we find this perspective less than helpful because it suggests a state of hopeless ignorance in which we resign ourselves to being buffeted and battered by the unknowable."
>
> Andrew Lo, who obviously did not bother to read the book he was citing.

Lack of skin in the game. Indeed one wonders why econometric methods keep being used while being wrong, so shockingly wrong, how "University" researchers (adults) can partake of such acts of artistry. Basically these capture the ordinary and mask higher order effects. Since blowups are not frequent, these events do not show in data and the researcher looks smart most of the time while being fundamentally wrong. At the source, researchers, "quant" risk manager, and academic economist do not have skin in the game so they are not hurt by wrong risk measures: other people are hurt by them. And the artistry should continue perpetually so long as people are allowed to harm others with impunity. (More in Taleb and Sandis [242], Taleb [233]).

E MACHINE LEARNING CONSIDERATIONS

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

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

a This appears to be an independent discovery by traders of the universal approximation theorem, initially for sigmoid functions, which are discussed further down (Cybenko [51]).

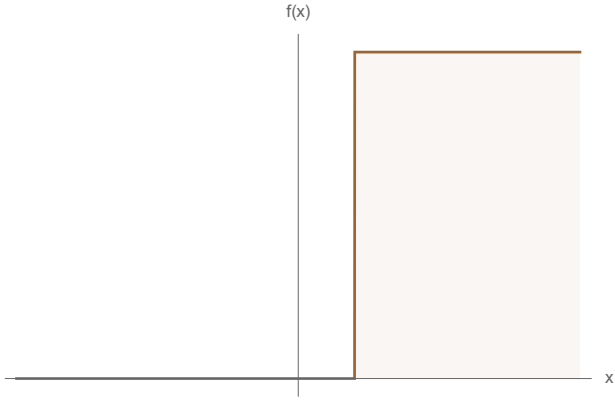


Figure E.1: *The heaviside θ function: note that it is the payoff of the "binary option" and can be decomposed as* $\lim_{\Delta K \to 0} \frac{C(K) - C(K + \Delta K)}{\delta K}$.

So this discussion is about ...fattailedness and how the different building blocks can accommodate them. Statistical machine learning switched to "ReLu" or "ramp"

functions that act exactly like call options rather than an aggregation of "S" curves. Researchers then discovered that it allows better handling of out of sample tail events (since there are by definition no unexpected tail events in sample) owing to the latter's extrapolation properties.

What is a sigmoid? Consider a payoff function as shown in E.7 that can be expressed with formula $S : (-\infty, \infty) \to (0, 1)$, $S(x) = \frac{1}{2}\tanh\left(\frac{\kappa x}{\pi}\right) + \frac{1}{2}\Big)$, or, more precisely, a three parameter function $S_i : (-\infty, \infty) \to (0, a_1)$ $S_i(x) = \frac{a_i}{e^{(c_i - b_i x)} + 1}$. It can also be the cumulative normal distribution, $\mathcal{N}(\mu, \sigma)$ where σ controls the smoothness (it then becomes the Heaviside of Fig. E.7 at the limit of $\sigma \to 0$). The (bounded) sigmoid is the smoothing using parameters of the Heaviside function.

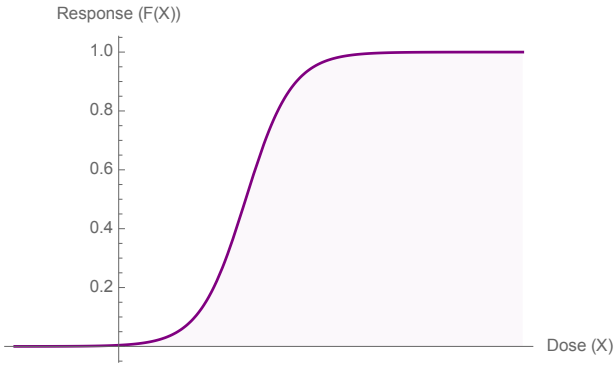


Figure E.2: *The sigmoid function ; note that it is bounded to both the left and right sides owing to saturation: it looks like a smoothed Heaviside θ.*

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

Remark 9

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

We need an infinity of summands for an unbounded function. So wherever the "empirical distribution" will be maxed, the last observation will match the flat part of the sig. For the definition of an empirical distribution see 3.3.

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

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

We can sum the function as $\sum_i = 1^n \rho(x, K_i, p_i)$ to fit a nonlinear function, which in fact replicates what we did with call options –the parameters p_i allow to smooth time value.

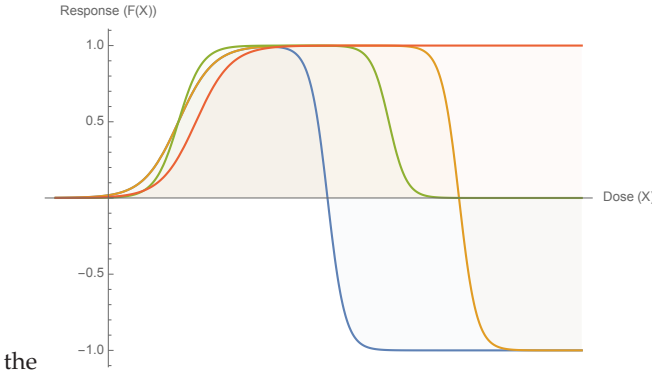


Figure E.3: *A sum of sigmoids will always be bounded, so one needs an infinite sum to replicate an "open" payoff, one that is not subjected to saturation.*

the

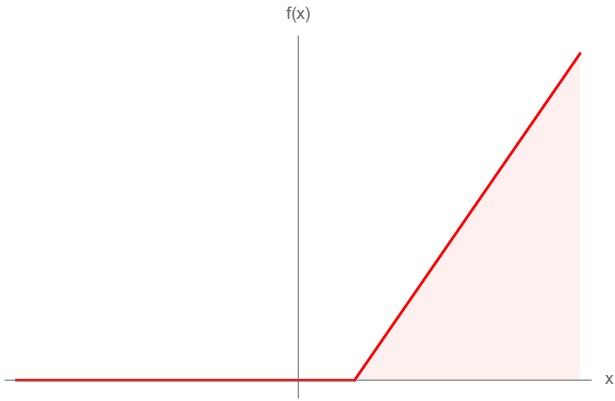


Figure E.4: *An option payoff at expiration, open on the right.*

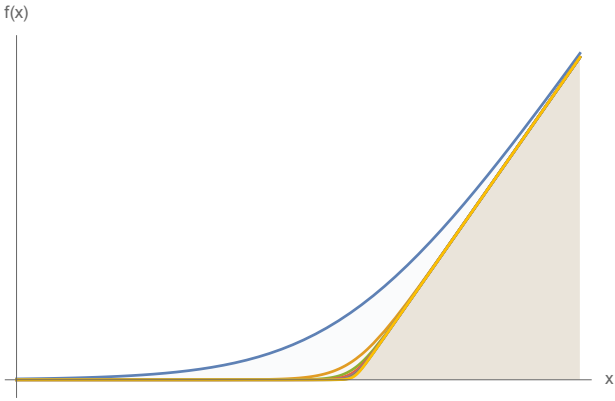


Figure E.5: *ρ function, from Eq. 11.18 , with k = 0. We calibrate and smooth the payoff with different values of p.*

E.0.1 Calibration via Angles

From figure E.6 we can see that, in the equation, $S = \sum_i^n \omega_i \, C(K_i, t_i)$, the ω_i corresponds to the arc tangent of the angle made –if positive (as illustrated in figure E.7), or the negative of the arctan of the supplementary angle.

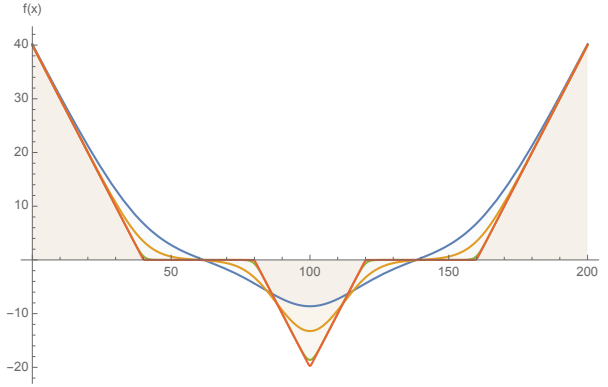


Figure E.6: *A butterfly (built via a sum of options/ReLu, not sigmoids), with open tails on both sides and flipping first and second derivatives. This example is particularly potent as it has no verbalistic correspondence but can be understood by option traders and machine learning.*

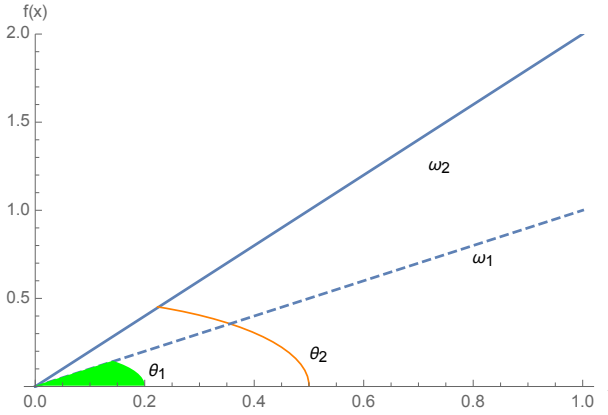


Figure E.7: *How* $\omega = \arctan\theta$. *By fitting angles we can translate a nonlinear function into its option summation.*

Summary

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

Part III

PREDICTIONS, FORECASTING, AND UNCERTAINTY

11 PROBABILITY CALIBRATION CALIBRATION UNDER FAT TAILS ‡

WHAT DO BINARY (or probabilistic) forecasting abilities have to do with performance? We map the difference between (univariate) binary predictions or "beliefs" (expressed as a specific "event" will happen/will not happen) and real-world continuous payoffs (numerical benefits or harm from an event) and show the effect of their conflation and mischaracterization in the decision-science literature.

The effects are:

A) Spuriousness of psychological research particularly those documenting that humans overestimate tail probabilities and rare events, or that they overreact to fears of market crashes, ecological calamities, etc. Many perceived "biases" are just mischaracterizations by psychologists. There is also a misuse of Hayekian arguments in promoting prediction markets.

B) Being a "good forecaster" in binary space doesn't lead to having a good performance, and vice versa, especially under nonlinearities. A binary forecasting record is likely to be a reverse indicator under some classes of distributions. Deeper uncertainty or more complicated and realistic probability distribution worsen the conflation .

C) Machine Learning: Some nonlinear payoff functions, while not lending themselves to verbalistic expressions and "forecasts", are well captured by ML or expressed in option contracts.

D) M Competitions Methods: The score for the M4-M5 competitions appear to be closer to real world variables than the Brier score.

The appendix shows the mathematical properties and exact distribution of the various payoffs, along with an exact distribution for the Brier score helpful for significance testing and sample sufficiency.

---

‡ Research chapter.

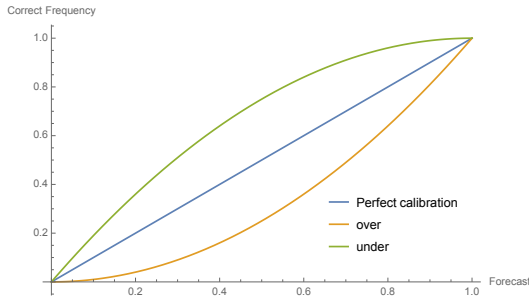


Figure 11.1: *Probabilistic calibration as seen in the psychology literature. The x axis shows the estimated probability produced by the forecaster, the y axis the actual realizations, so if a weather forecaster predicts* 30% *chance of rain, and rain occurs* 30% *of the time, they are deemed "calibrated". We hold that calibration in frequency (probability) space is an academic exercise (in the bad sense of the word) that mistracks real life outcomes outside narrow binary bets. It is particularly fallacious under fat tails.*

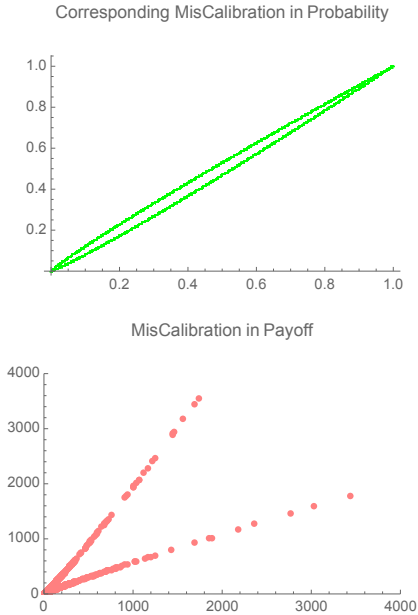


Figure 11.2: *How miscalibration in probability corresponds to miscalibration in payoff under power laws. The distribution under consideration is Pareto with tail index* $\alpha = 1.15$

11.1 CONTINUOUS VS. DISCRETE PAYOFFS: DEFINITIONS AND COMMENTS

Example 11.1 ("One does not eat beliefs and (binary) forecasts")
In the first volume of the Incerto (Fooled by Randomness, 2001 [223]), the narrator, a trader, is asked by the manager "do you predict that the market is going up or down?" "Up", he answered, with confidence. Then the boss got angry when, looking at the firm's exposures, he discovered that the narrator was short the market, i.e., would benefit from the market going down.

The trader had difficulties conveying the idea that there was no contradiction, as someone could hold the (binary) belief that the market had a higher probability of going up than down, but that, should it go down, there is a very small probability that it could go down

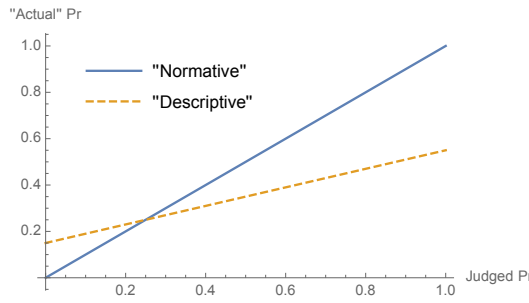


Figure 11.3: *"Typical patterns" as stated and described in [13], a representative claim in psychology of decision making that people overestimate small probability events. The central findings are in 1977 and 1978 [150] and [151]. We note that to the left, in the estimation part, 1) events such as floods, tornados, botulism, mostly patently thick tailed variables, matters of severe consequences that agents might have incorporated in the probability, 2) these probabilities are subjected to estimation error that, when endogenised, increase the estimation.*

considerably, hence a short position had a positive expected return and the rational response was to engage in a short exposure. "You do not eat forecasts, but P/L" (or "one does not monetize forecasts") goes the saying among traders.

If exposures and beliefs do not go in the same direction, it is because beliefs are verbalistic reductions that contract a higher dimensional object into a single dimension. To express the manager's error in terms of decision-making research, there can be a conflation in something as elementary as the notion of a binary event (related to the zeroth moment) or the *probability* of an event and *expected payoff* from it (related to the first moment and, when nonlinear, to all higher moments) as the payoff functions of the two can be similar in some circumstances and different in others.

Commentary 11.1
In short, probabilistic calibration requires estimations of the zeroth moment while the real world requires all moments (outside of gambling bets or artificial environments such as psychological experiments where payoffs are necessarily truncated), and it is a central property of thick tails that higher moments are explosive (even "infinite") and count more and more.

11.1.1 Away from the Verbalistic

While the trader story is mathematically trivial (though the mistake is committed a bit too often), more serious gaps are present in decision making and risk management, particularly when the payoff function is more complicated, or nonlinear (and related to higher moments). So once we map the contracts or exposures mathematically, rather than focus on words and verbal descriptions, some serious distributional issues arise.

Definition 11.1 (Event)
A (real-valued) random variable $X: \Omega \to \mathbb{R}$ *defined on the probability space* $(\Omega, \mathcal{F}, P)$ *is a function* $X(\omega)$ *of the outcome* $\omega \in \Omega$. *An event is a measurable subset (countable or not) of* Ω, *measurable meaning that it can be defined through the value(s) of one of several random variable(s).*

Definition 11.2 (Binary forecast/payoff)
A binary forecast (belief, or payoff) is a random variable taking two values

$$X : \Omega \to \{X_1, X_2\},$$

with realizations $X_1, X_2 \in \mathbb{R}$.

In other words, it lives in the binary set (say $\{0,1\}$, $\{-1,1\}$, etc.), i.e., the specified event will or will not take place and, if there is a payoff, such payoff will be mapped into two finite numbers (a fixed sum if the event happened, another one if it didn't). Unless otherwise specified, in this discussion we default to the $\{0,1\}$ set.

Example of situations in the real world where the payoff is binary:

- Casino gambling, lotteries , coin flips, "ludic" environments, or binary options paying a fixed sum if, say, the stock market falls below a certain point and nothing otherwise –deemed a form of gambling[2].
- Elections where the outcome is binary (e.g., referenda, U.S. Presidential Elections), though not the economic effect of the result of the election.[3]
- Medical prognoses for a single patient entailing survival or cure over a specified duration, though not the duration itself as variable, or disease-specific survival expressed in time, or conditional life expectancy. Also exclude anything related to epidemiology.
- Whether a given person who has an online profile will buy or not a unit or more of a specific product at a given time (not the quantity or units).

Commentary 11.2 (A binary belief is equivalent to a payoff)
A binary "belief" should map to an economic payoff (under some scaling or normalization necessarily to constitute a probability), an insight owed to De Finetti [56] who held that a "belief" and a "prediction" (when they are concerned with two distinct outcomes) map into the equivalent of the expectation of a binary random variable and bets with a payoff in $\{0,1\}$. *An "opinion" becomes a choice price for a gamble, and one at which one is equally willing to buy or sell. Inconsistent opinions therefore would lead to a violation of arbitrage rules, such as the "Dutch book", where a combination of mispriced bets can guarantee a future loss.*

Definition 11.3 (Real world open continuous payoff)

$$X : \Omega \to [a, \infty) \vee (-\infty, b] \vee (-\infty, \infty)$$

A continuous payoff "lives" in an interval, not a finite set. It corresponds to an unbounded random variable either doubly unbounded or semi-bounded, with the bound on one side (one tailed variable).

2 Retail binary options are typically used for gambling and have been banned in many jurisdictions, such as, for instance, by the European Securities and Markets Authority (ESMA), www.esma.europa.eu, as well as the United States where it is considered another form of internet gambling, triggering a complaint by a collection of decision scientists, see Arrow et al. [3]. We consider such banning as justified since bets have practically no economic value, compared to financial markets that are widely open to the public, where natural exposures can be properly offset.

3 Note the absence of spontaneously forming gambling markets with binary payoffs for continuous variables. The exception might have been binary options but these did not remain in fashion for very long, from the experiences of the author, for a period between 1993 and 1998, largely motivated by tax gimmicks.

Caveat We are limiting for the purposes of our study the consideration to binary vs. continuous and open-ended (i.e., no compact support). Many discrete payoffs are subsumed into the continuous class using standard arguments of approximation. We are also omitting triplets, that is, payoffs in, say $\{-1, 0, 3\}$, as these obey the properties of binaries (and can be constructed using a sum of binaries). Further, many variable with a floor and a remote ceiling (hence, formally with compact support), such as the number of victims or a catastrophe, are analytically and practically treated as if they were open-ended [46].

Example of situations in the real world where the payoff is continuous:

- Wars casualties, calamities due to earthquake, medical bills, etc.
- Magnitude of a market crash, severity of a recession, rate of inflation
- Income from a strategy
- Sales and profitability of a new product
- In general, anything covered by an insurance contract

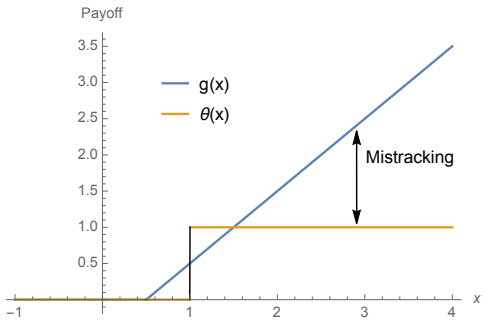


Figure 11.4: *Comparing the payoff of a binary bet (The Heaviside $\theta(.)$) to a continuous open-ended exposure $g(x)$. Visibly there is no way to match the (mathematical) derivatives for any form of hedging.*

Most natural and socio-economic variables are continuous and their statistical distribution does not have a compact support in the sense that we do not have a handle of an exact upper bound.

Example 11.2
Predictive analytics in binary space $\{0, 1\}$ can be successful in forecasting if, from his online activity, online consumer Iannis Papadopoulos will purchase a certain item, say a wedding ring, based solely on computation of the probability. But the probability of the "success" for a potential new product might be –as with the trader's story– misleading. Given that company sales are typically thick tailed, a very low probability of success might still be satisfactory to make a decision. Consider venture capital or option trading –an out of the money option can often be attractive yet may have less than 1 in 1000 probability of ever paying off.

More significantly, the tracking error for probability guesses will not map to that of the performance. $\lambda^{(M_4)}$ would.

This difference is well known by option traders as there are financial derivative contracts called "binaries" that pay in the binary set $\{0, 1\}$ (say if the underlying asset S, say, exceeds a strike price K), while others called "vanilla" that pay in $[0, \infty)$, i.e. $\max(S - K, 0)$ (or, worse, in $(-\infty, 0)$ for the seller can now be exposed

to bankruptcy owing to the unbounded exposure). The considerable mathematical and economic difference between the two has been discussed and is the subject of *Dynamic Hedging: Managing Vanilla and Exotic Options* [222]. Given that the former are bets paying a fixed amount and the latter have full payoff, one cannot be properly replicated (or hedged) using another, especially under fat tails and parametric uncertainty –meaning performance in one does not translate to performance into the other. While this knowledge is well known in mathematical finance it doesn't seem to have been passed on to the decision-theory literature.

Commentary 11.3 (Derivatives theory)
Our approach here is inspired from derivatives (or option) theory and practice where there are different types of derivative contracts, 1) those with binary payoffs (that pay a fixed sum if an event happens) and 2) "vanilla" ones (standard options with continuous payoffs). It is practically impossible to hedge one with another [222]. Furthermore a bet with a strike price K and a call option with same strike K, with K in the tails of the distribution, almost always have their valuations react in opposite ways when one increases the kurtosis of the distribution, (while preserving the first three moments) or, in an example further down in the lognormal environment, when one increases uncertainty via the scale of the distribution.

Commentary 11.4 (Term sheets)
Note that, thanks to "term sheets" that are necessary both legally and mathematically, financial derivatives practive provides precise legalistic mapping of payoffs in a way to make their mathematical, statistical, and economic differences salient.

There has been a tension between prediction markets and real financial markets. As we can show here, prediction markets may be useful for gamblers, but they cannot hedge economic exposures.

The mathematics of the difference and the impossibility of hedging can be shown in the following. Let X be a random variable in $\mathbb{R}$, we have the payoff of the bet or the prediction $\theta_K : \mathbb{R} \to \{0, 1\}$,

$$\theta_K(x) = \begin{cases} 1, & x \geq K \\ 0 & \text{otherwise,} \end{cases} \tag{11.1}$$

and $g : \mathbb{R} \to \mathbb{R}$ that of the natural exposure. Since $\frac{\partial}{\partial x}\theta_K(x)$ is a Dirac delta function at K, $\delta(K)$ and $\frac{\partial)}{\partial x}g_k(x)$ is at least once differentiable for $x \geq K$ (or constant in case the exposure is globally linear or, like an option, piecewise linear above K), matching derivatives for the purposes of offsetting variations is not a possible strategy.[4] The point is illustrated in Fig 11.4.

11.1.2 There is no defined "collapse", "disaster", or "success" under fat tails

The fact that an "event" has some uncertainty around its magnitude carries some mathematical consequences. Some verbalistic papers in 2019 still commit the fallacy of binarizing an event in $[0, \infty)$: A recent paper on calibration of beliefs says

4 To replicate an open-ended continuous payoff with binaries, one needs an infinite series of bets, which cancels the entire idea of a prediction market by transforming it into a financial market. Distributions with compact support always have finite moments, not the case of those on the real line.

"...if a person claims that the United States is on the verge of an economic collapse or that a climate disaster is imminent..." An economic "collapse" or a climate "disaster" must not be expressed as an event in $\{0, 1\}$ when in the real world it can take many values. For that, a characteristic scale is required. In fact under fat tails, there is no "typical" collapse or disaster, owing to the absence of characteristic scale, hence verbal binary predictions or beliefs cannot be used as gauges.

We present the difference between thin tailed and fat tailed domains as follows.

Definition 11.4 (Characteristic scale)
Let X be a random variable that lives in either $(0, \infty)$ or $(-\infty, \infty)$ and $\mathbb{E}$ the expectation operator under "real world" (physical) distribution. By classical results [81]:

$$\lim_{K \to \infty} \frac{1}{K} \mathbb{E}(X|_{X>K}) = \lambda, \tag{11.2}$$

- *If $\lambda = 1$, X is said to be in the thin tailed class $\mathcal{D}_1$ and has a characteristic scale*
- *If $\lambda > 1$, X is said to be in the fat tailed regular variation class $\mathcal{D}_2$ and has no characteristic scale*
- *If*

  $$\lim_{K \to \infty} \mathbb{E}(X|_{X>K}) - K = \mu$$

  where $\mu > 0$, then X is in the borderline exponential class

The point can be made clear as follows. One cannot have a binary contract that adequately hedges someone against a "collapse", given that one cannot know in advance the size of the collapse or how much the face value or such contract needs to be. On the other hand, an insurance contract or option with continuous payoff would provide a satisfactory hedge. Another way to view it: reducing these events to verbalistic "collapse", "disaster" is equivalent to a health insurance payout of a lump sum if one is "very ill" –regardless of the nature and gravity of the illness – and 0 otherwise.

And it is highly flawed to separate payoff and probability in the integral of expected payoff.[5] Some experiments of the type shown in Figure 11 ask agents what is their estimates of deaths from botulism or some such disease: agents are blamed for misunderstanding the probability. This is rather a problem with the experiment: people do not necessarily separate probabilities from payoffs.

5 Practically all economic and informational variables have been shown since the 1960s to belong to the $\mathcal{D}_2$ class, or at least the intermediate subexponential class (which includes the lognormal), [98, 160, 161, 162, 223], along with social variables such as size of cities, words in languages, connections in networks, size of firms, incomes for firms, macroeconomic data, monetary data, victims from interstate conflicts and civil wars[46, 196], operational risk, damage from earthquakes, tsunamis, hurricanes and other natural calamities, income inequality [40], etc. Which leaves us with the more rational question: where are Gaussian variables? These appear to be at best one order of magnitude fewer in decisions entailing formal predictions.

11.2 SPURIOUS OVERESTIMATION OF TAIL PROBABILITY IN PSYCHOLOGY

Definition 11.5 (Substitution of integral)
Let $K \in \mathbb{R}^+$ be a threshold, $f(.)$ a density function and $p_K \in [0,1]$ the probability of exceeding it, and $g(x)$ an impact function. Let I_1 be the expected payoff above K:

$$I_1 = \int_K^\infty g(x) f(x)\, \mathrm{d}x,$$

and Let I_2 be the impact at K multiplied by the probability of exceeding K:

$$I_2 = g(K) \int_K^\infty f(x)\, dx = g(K) p_K.$$

The substitution comes from conflating I_1 and I_2, which becomes an identity if and only if $g(.)$ is constant above K (say $g(x) = \theta_K(x)$, the Heaviside theta function). For $g(.)$ a variable function with positive first derivative, I_1 can be close to I_2 only under thin-tailed distributions, not under the fat tailed ones.

For the discussions and examples in this section assume $g(x) = x$ as we will consider the more advanced nonlinear case in Section 11.5.

Theorem 1: Convergence of $\frac{I_1}{I_2}$

If X is in the thin tailed class $\mathcal{D}_1$ as described in 11.2,

$$\lim_{K \to \infty} \frac{I_1}{I_2} = 1 \tag{11.3}$$

If X is in the regular variation class $\mathcal{D}_2$,

$$\lim_{K \to \infty} \frac{I_1}{I_2} = \lambda > 1. \tag{11.4}$$

Proof. From Eq. 11.2. Further comments:

11.2.1 Thin tails

By our very definition of a thin tailed distribution (more generally any distribution outside the subexponential class, indexed by (g)), where $f^{(g)}(.)$ is the PDF:

$$\lim_{K \to \infty} \frac{\int_K^\infty x f^{(g)}(x)\, dx}{K \int_K^\infty f^{(g)}(x)\, dx} = \frac{I_1}{I_2} = 1. \tag{11.5}$$

Special case of a Gaussian: Let $g(.)$ be the PDF of predominantly used Gaussian distribution (centered and normalized),

$$\int_K^\infty x g(x)\, dx = \frac{e^{-\frac{K^2}{2}}}{\sqrt{2\pi}} \tag{11.6}$$

and $K_p = \frac{1}{2}\text{erfc}\left(\frac{K}{\sqrt{2}}\right)$, where erfc is the complementary error function, and K_p is the threshold corresponding to the probability p.

We note that $K_p \frac{I_1}{I_2}$ corresponds to the inverse Mills ratio used in insurance.

11.2.2 Fat tails

For all distributions in the regular variation class, defined by their tail survival function: for K large,

$$\mathbb{P}(X > K) \approx LK^{-\alpha},\ \alpha > 1,$$

where $L > 0$ and $f^{(p)}$ is the PDF of a member of that class:

$$\lim_{K_p \to \infty} \frac{\int_K^\infty x f^{(p)}(x)\, dx}{K \int_{K_p}^\infty f^{(p)}(x)\, dx} = \frac{\alpha}{\alpha - 1} > 1 \tag{11.7}$$

□

11.2.3 Conflations

Conflation of I_1 and I_2 In numerous experiments, which include the prospect theory paper by Kahneman and Tversky (1978) [137], it has been repeatedly established that agents overestimate small probabilities in experiments where the odds are shown to them, and when the outcome corresponds to a single payoff. The well known Kahneman-Tversky result proved robust, but interpretations make erroneous claims from it. Practically all the subsequent literature, relies on I_2 and conflates it with I_1, what this author has called *the ludic fallacy* in *The Black Swan* [223], as games are necessarily truncating a dimension from reality. The psychological results might be robust, in the sense that they replicate when repeated in the exact similar conditions, but all the claims outside these conditions and extensions to real risks will be an exceedingly dubious generalization –given that our exposures in the real world rarely map to I_1. Furthermore, one can overestimate the probability yet underestimate the expected payoff.

Stickiness of the conflation The misinterpretation is still made four decades after Kahneman-Tversky (1979). In a review of behavioral economics, with emphasis on miscaculation of probability, Barberis (2003) [12] treats $I_1 = I_2$. And Arrow et al. [3], a long list of decision scientists pleading for deregulation of the betting markets also misrepresented the fitness of these binary forecasts to the real world (particularly in the presence of real financial markets).

Another stringent –and dangerous –example is the "default VaR" (Value at risk) which is explicitly given as I_2 , i.e. default probability $x(1 -$ expected recovery rate), which can be quite different from the actual loss expectation in case of default. Fi-

nance presents erroneous approximations of CVaR[6], and the approximation is the risk-management flaw that may have caused the crisis of 2008 [240].

The fallacious argument is that they compute the recovery rate as the expected value of collateral, without conditioning by the default event. The expected value of the collateral conditionally to a default is often far less then its unconditional expectation. In 2007, after a massive series of foreclosures, the value of most collaterals dropped to about 1/3 of its expected value!

Misunderstanding of Hayek's knowledge arguments "Hayekian" arguments for the consolidation of beliefs via prices does not lead to prediction markets as discussed in such pieces as [30], or Sunstein's [217]: prices exist in financial and commercial markets; prices are not binary bets. For Hayek [125] consolidation of knowledge is done via prices and *arbitrageurs* (his words)–and arbitrageurs trade products, services, and financial securities, not binary bets.

Definition 11.6 (Corrected probability in binarized experiments)
Let p^* be the equivalent probability to make $I_1 = I_2$ and eliminate the effect of the error, so

$$p^* = \{p : I_1 = I_2 = K\}$$

Now let' s solve for K_p "in the tails", working with a probability p. For the Gaussian, $K_p = \sqrt{2}\text{erfc}^{-1}(2p)$; for the Paretian tailed distribution, $K_p = p^{-1/\alpha}$.

Hence, for a Paretian distribution, the ratio of real continuous probability to the binary one

$$\frac{p^*}{p} = \frac{\alpha}{1-\alpha}$$

which can allow in absurd cases p^* to exceed 1 when the distribution is grossly misspecified.

Tables 11.1 and 11.2 show, for a probability level p, the corresponding tail level K_p, such as

$$K_p = \{\inf K : \mathbb{P}(X > K) > p\},$$

and the corresponding adjusted probability p^* that de-binarize the event [7][8]– probabilities here need to be in the bottom half, i.e., $p < .5$. Note that we are operating under the mild case of known probability distributions, as it gets worse under parametric uncertainty.[9]

6 The mathematical expression of the Value at Risk, VaR, for a random variable X with distribution function F and threshold $\alpha \in [0, 1]$

$$\text{VaR}_\alpha(X) = -\inf\{x \in \mathbb{R} : F_X(x) > \alpha\},$$

and the corresponding CVar

$$\text{ES}_\alpha(X) = \mathbb{E}\left(-X \mid_{X \leq -\text{VaR}_\alpha(X)}\right)$$

7 The analysis is invariant to whether we use the right or left tail .By convention, finance uses negative value for losses, whereas other areas of risk management express the negative of the random variance, hence focus on the right tail.

8 K_p is equivalent to the Value at Risk VaR_p in finance, where p is the probability of loss.

9 Note the van der Wijk's law, see Cirillo [44]: $\frac{I_1}{I_2}$ is related to what is called in finance the expected shortfall for K_p.

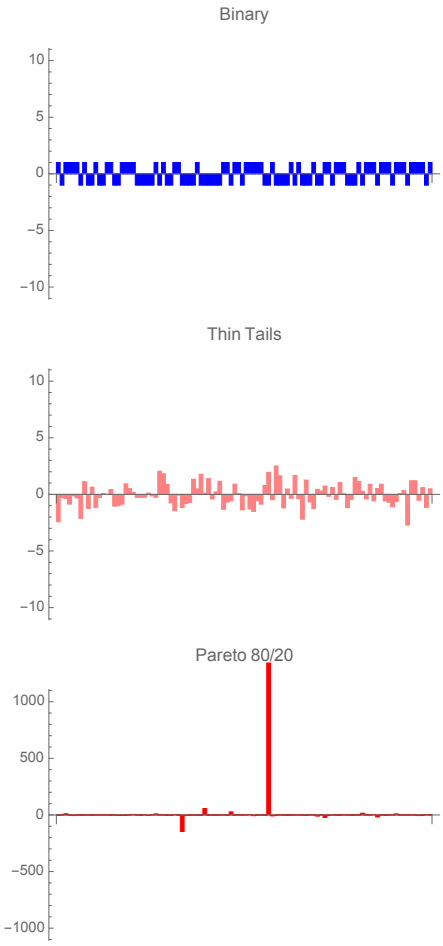


Figure 11.5: *Comparing the three payoffs under two distributions –the binary has the same profile regardless of whether the distribution is thin or fat tailed. The first two subfigures are to scale, the third (representing the Pareto 80/20 with $\alpha = 1.16$ requires multiplying the scale by two orders of magnitude.*

The most commonly known distribution among the public, the "Pareto 80/20" (based on Pareto discovering that 20 percent of the people in Italy owned 80 percent of the land), maps to a tail index $\alpha = 1.16$, so the adjusted probability is > 7 times the naive one.

Example of probability and expected payoff reacting in opposite direction under increase in uncertainty An example showing how, under a skewed distribution, the binary and the expectation reacting in opposite directions is as follows. Consider the risk-neutral lognormal distribution $\mathcal{L}(X_0 - \frac{1}{\sigma^2}, \sigma)$ with pdf $f_L(.)$, mean X_0 and variance $\left(e^{\sigma^2} - 1\right) X_0^2$. We can increase its uncertainty with the parameter σ. We have the expectation of a contract above X_0, $\mathbb{E}_{>X_0}$:

$$\mathbb{E}_{>X_0} = \int_{X_0}^{\infty} x f_L(x)\, \mathrm{d}x = \frac{1}{2} X_0 \left(1 + \mathrm{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)\right)$$

Table 11.1: *Gaussian pseudo-overestimation*

p	K_p	$\int_{K_p}^{\infty} xf(x)dx$	$K_p \int_{K_p}^{\infty} f(x)dx$	p^*	$\frac{p^*}{p}$
$\frac{1}{10}$	1.28	1.75×10^{-1}	1.28×10^{-1}	1.36×10^{-1}	1.36
$\frac{1}{100}$	2.32	2.66×10^{-2}	2.32×10^{-2}	1.14×10^{-2}	1.14
$\frac{1}{1000}$	3.09	3.36×10^{-3}	3.09×10^{-3}	1.08×10^{-3}	1.08
$\frac{1}{10000}$	3.71	3.95×10^{-4}	3.71×10^{-4}	1.06×10^{-4}	1.06

Table 11.2: *Paretian pseudo-overestimation*

p	K_p	$\int_{K_p}^{\infty} xf(x)dx$	$K_p \int_{K_p}^{\infty} f(x)dx$	p^*	$\frac{p^*}{p}$
$\frac{1}{10}$	8.1	8.92	0.811	1.1 (sic)	11.
$\frac{1}{100}$	65.7	7.23	0.65	0.11	11.
$\frac{1}{1000}$	533	5.87	0.53	0.011	11.
$\frac{1}{10000}$	4328	4.76	0.43	0.0011	11.

and the probability of exceeding X_0,

$$\mathbb{P}(X > X_0) = \frac{1}{2}\left(1 - \text{erf}\left(\frac{\sigma}{2\sqrt{2}}\right)\right),$$

where erf is the error function. As σ rises $\text{erf}\left(\frac{\sigma}{2\sqrt{2}}\right) \to 1$, with $\mathbb{E}_{>X_0} \to X_0$ and $\mathbb{P}(X > X_0) \to 0$. This example is well known by option traders (see *Dynamic Hedging* [222]) as the binary option struck at X_0 goes to 0 while the standard call of the same strike rises considerably to reach the level of the asset –regardless of strike. This is typically the case with venture capital: the riskier the project, the less likely it is to succeed but the more rewarding in case of success. So, the expectation can go to $+\infty$ while to probability of success goes to 0.

11.2.4 Distributional Uncertainty

Remark 10: Distributional uncertainty

Owing to Jensen's inequality, the discrepancy $(I_1 - I_2)$ increases under parameter uncertainty, expressed in higher kurtosis, via stochasticity of σ the scale of the thin-tailed distribution, or that of α the tail index of the Paretian one.

Proof. First, the Gaussian world. We consider the effect of $I_1 - I_2 = \int_K^{\infty} xf^{(g)}(x) - \int_K^{\infty} f^{(g)}(x)$ under stochastic volatility, i.e. the parameter from increase of volatility. Let σ be the scale of the Gaussian, with K constant:

$$\frac{\partial^2(\int_K^{\infty} xf^{(g)}(x)dx)}{\partial\sigma^2} - \frac{\partial^2(\int_K^{\infty} f^{(g)}(x)dx)}{\partial\sigma^2} = \frac{e^{-\frac{K^2}{2\sigma^2}}\left((K-1)K^3 - (K-2)K\sigma^2\right)}{\sqrt{2\pi}\sigma^5}, \quad (11.8)$$

which is positive for all values of $K > 0$ (given that $K^4 - K^3 - K^2 + 2K > 0$ for K positive).

Second, consider the sensitivity of the ratio $\frac{I_1}{I_2}$ to parameter uncertainty for α in the Paretian case (for which we can get a streamlined expression compared to the difference). For $\alpha > 1$ (the condition for a finite mean):

$$\frac{\partial^2 \left(\int_K^\infty x f^{(p)}(x) dx \Big/ \int_K^\infty f^{(p)}(x) dx \right)}{\partial \alpha^2} = \frac{2K}{(\alpha - 1)^3} \tag{11.9}$$

which is positive and increases markedly at lower values of α, meaning the fatter the tails, the worse the uncertainty about the expected payoff and the larger the difference between I_1 and I_2.

□

11.3 CALIBRATION AND MISCALIBRATION

The psychology literature also examines the "calibration" of probabilistic assessment –an evaluation of how close someone providing odds of events turns out to be on average (under some operation of the law of large number deemed satisfactory) [150], [141], see Fig. 11.2. The methods, for the reasons we have shown here, are highly flawed except in narrow circumstances of purely binary payoffs (such as those entailing a "win/lose" outcome) –and generalizing from these payoffs is either not possible or produces misleading results. Accordingly, Fig. 11 makes little sense empirically.

At the core, calibration metrics such as the Brier score are always thin-tailed, when the variable under measurement is fat-tailed, which worsens the tractability.

To use again the saying "You do not eat forecasts", most businesses have severely skewed payoffs, so being calibrated in probability is meaningless.

Remark 11: Distributional differences

Binary forecasts and calibration metrics via the Brier score belong to the thin-tailed class.

We will show proofs next.

11.4 SCORING METRICS

This section, summarized in Table 11.3, compares the probability distributions of the various metrics used to measure performance, either by explicit formulation or by linking it to a certain probability class. Clearly one may be mismeasuring performance if the random variable is in the wrong probability class. Different underlying distributions will require a different number of sample sizes owing to the differences in the way the law of numbers operates across distributions. A series of binary forecasts will converge very rapidly to a thin-tailed Gaussian

Table 11.3: *Scoring Metrics for Performance Evaluation*

Metric	Name	Fitness to reality
$P^{(r)}(T)$	Cumulative P/L	Adapted to real world distributions, particularly under a survival filter
$P^{(p)}(n)$	Tally of Bets	Misrepresents the performance under fat tails, works only for binary bets and/or thin tailed domains.
$\lambda(n)$	Brier Score	Misrepresents performance precision under fat tails, ignores higher moments.
$\lambda_n^{(M4)}$	M4 Score	Represents precision not exactly real world performance but maps to real distribution of underlying variables.
$\lambda_n^{(M5)}$	Proposed M5 Score	Represents both precision and survival conditions by predicting extrema of time series.
$g(.)$	Machine learning nonlinear payoff function (not a metric)	Expresses exposures without verbalism and reflects true economic or other P/L. Resembles financial derivatives term sheets.

even if the underlying distribution is fat-tailed, but an economic P/L tracking performance for someone with a real exposure will require a considerably larger sample size if, say, the underlying is Pareto distributed [232].

We start by precise expressions for the four possible ones:

1. Real world performance under conditions of survival, or, in other words, P/L or a quantitative cumulative score.
2. A tally of bets, the naive sum of how often a person's binary prediction is correct
3. De Finetti's Brier score $\lambda(B)_n$
4. The M4 score $\lambda_n^{M_4}$ for n observations used in the M4 competition, and its prosed sequel M5.

P/L in Payoff Space (under survival condition) The "P/L" is short for the natural profit and loss index, that is, a cumulative account of performance. Let X_i be realizations of an unidimensional generic random variable X with support in $\mathbb{R}$ and $t = 1, 2, \dots n$. Real world payoffs $P_r(.)$ are expressed in a simplified way as

$$P_r(n) = P(0) + \sum_{k \leq N} g(x_t), \tag{11.10}$$

where $g_t : \mathbb{R} \to \mathbb{R}$ is a measurable function representing the payoff; g may be path dependent (to accommodate a survival condition), that is, it is a function of

the preceding period $\tau < t$ or on the cumulative sum $\sum_{\tau \leq t} g(x_\tau)$ to introduce an absorbing barrier, say, bankruptcy avoidance, in which case we write:

$$P^{(r)}(T) = P^{(r)}(0) + \sum_{t \leq n} \mathbb{1}_{(\sum_{\tau<t} g(x_\tau)>b)} \; g(x_t), \tag{11.11}$$

where b is is any arbitrary number in $\mathbb{R}$ that we call the survival mark and $\mathbb{1}_{(.)}$ an indicator function $\in \{0, 1\}$.

The last condition from the indicator function in Eq. 11.11 is meant to handle ergodicity or lack of it [223].

Commentary 11.5
P/L tautologically corresponds to the real world distribution, with an absorbing barrier at the survival condition.

Frequency Space, The standard psychology literature has two approaches.

A-When tallying forecasts as a counter

$$P^{(p)}(n) = \frac{1}{n} \sum_{i \leq n} \mathbb{1}_{X_t \in \chi}, \tag{11.12}$$

where $\mathbb{1}_{X_t \in \chi} \in \{0, 1\}$ is an indicator that the random variable $x \in \chi_t$ in in the "forecast range", and T the total number of such forecasting events. where $f_t \in [0, 1]$ is the probability announced by the forecaster for event t

B-When dealing with a score (calibration method) in the absence of a visible net performance, researchers produce some more advanced metric or score to measure calibration. We select below the gold standard", De Finetti's Brier score(DeFinetti, [57]). It is favored since it doesn't allow arbitrage and requires perfect probabilistic calibration: someone betting than an event has a probability 1 of occurring will get a perfect score only if the event occurs all the time.

$$\lambda_n^{(B)} = \frac{1}{n} \sum_{t \leq n} (f_t - \mathbb{1}_{X_t \in \chi})^2. \tag{11.13}$$

which needs to be minimized for a perfect probability assessor.

Applications: M4 and M5 Competitions The M series (Makridakis [157]) evaluate forecasters using various methods to predict a point estimate (along with a range of possible values). The last competition in 2018, M4, largely relied on a series of scores, λ^{M4_j}, which works well in situations where one has to forecast the first moment of the distribution and the dispersion around it.

Definition 11.7 (The M4 first moment forecasting scores)
The M4 competition precision score (Makridakis et al. [157]) judges competitors on the following metrics indexed by $j = 1, 2$

$$\lambda_n^{(M4)_j} = \frac{1}{n}\sum_i^n \frac{\left|X_{f_i} - X_{r_i}\right|}{s_j} \tag{11.14}$$

where $s_1 = \frac{1}{2}\left(|X_{f_i}|+|X_{r_i}|\right)$ *and* s_2 *is (usually) the raw mean absolute deviation for the observations available up to period i (i.e., the mean absolute error from either "naive" forecasting or that from in sample tests),* X_{f_i} *is the forecast for variable i as a point estimate,* X_{r_i} *is the realized variable and n the number of experiments under scrutiny.*

In other word, it is an application of the Mean Absolute Scaled Error (MASE) and the symmetric Mean Absolute Percentage Error (sMAPE) [131].

The suggested M5 score (expected for 2020) adds the forecasts of extrema of the variables under considerations and repeats the same tests as the one for raw variables in Definition 11.7.

11.4.1 Deriving Distributions

Distribution of $P^{(p)}(n)$

Remark 12

The tally of binary forecast $P^{(p)}(n)$ *is asymptotically normal with mean p and standard deviation* $\sqrt{\frac{1}{n}(p-p^2)}$ *regardless of the distribution class of the random variable X.*

The results are quite standard, but see appendix for the re-derivations.

Distribution of the Brier Score λ_n

Theorem 2

Regardless of the distribution of the random variable X, without even assuming independence of $(f_1 - \mathbb{1}_{A1}), \ldots, (f_n - \mathbb{1}_{An})$*, for* $n < +\infty$*, the score* λ_n *has all moments of order q,* $\mathbb{E}(\lambda_n^q) < +\infty$*.*

Proof. For all i, $(f_i - \mathbb{1}_{Ai})^2 \leq 1$. □

We can get actually closer to a full distribution of the score across independent betting policies. Assume binary predictions f_i are independent and follow a beta distribution $\mathcal{B}(a, b)$ (which approximates or includes all unimodal distributions in

$[0,1]$ (plus a Bernoulli via two Dirac functions), and let p be the rate of success $p = \mathbb{E}\left(\mathbb{1}_{Ai}\right)$, the characteristic function of λ_n for n evaluations of the Brier score is

$$\begin{aligned}\varphi_n(t) = \pi^{n/2} \Bigg(& 2^{-a-b+1}\Gamma(a+b) \\ & \left(p\, {}_2\tilde{F}_2\left(\frac{b+1}{2}, \frac{b}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); \frac{it}{n}\right)\right. \\ & \left. - (p-1)\, {}_2\tilde{F}_2\left(\frac{a+1}{2}, \frac{a}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); \frac{it}{n}\right)\right)\Bigg)\end{aligned} \tag{11.15}$$

Here ${}_2\tilde{F}_2$ is the generalized hypergeometric function regularized ${}_2\tilde{F}_2(.,.;.,.;.) = \frac{{}_2F_2(a;b;z)}{\left(\Gamma(b_1)...\Gamma(b_q)\right)}$ and ${}_pF_q(a;b;z)$ has series expansion $\sum_{k=0}^{\infty} \frac{(a_1)_k...(a_p)_k}{(b_1)_k...(b_p)_k} z^k/k!$, were $(a)_{(.)}$ is the Pochhammer symbol.

Hence we can prove the following: under the conditions of independence of the summands stated above,

$$\lambda_n \xrightarrow{D} \mathcal{N}\left(\mu, \sigma_n\right) \tag{11.16}$$

where $\mathcal{N}$ denotes the Gaussian distribution with for first argument the mean and for second argument the standard deviation.

The proof and parametrization of μ and σ_n is in the appendix.

Distribution of the economic P/L or quantitative measure P_r

Remark 13

Conditional on survival to time T, the distribution of the quantitative measure $P^{(r)}(T)$ will follow the distribution of the underlying variable $g(x)$.

The discussion is straightforward if there is no absorbing barrier (i.e., no survival condition).

Distribution of the M4 score The distribution of an absolute deviation is in the same probability class as the variable itself. Thee Brier score is in the norm L2 and is based on the second moment (which always exists) as De Finetti has shown that it is more efficient to just a probability in square deviations. However for nonbinaries, it is vastly more efficient under fat tails to rely on absolute deviations, even when the second moment exists [236].

11.5 NON-VERBALISTIC PAYOFF FUNCTIONS AND THE GOOD NEWS FROM MACHINE LEARNING

Earlier examples focused on simple payoff functions, with some cases where the conflation I_1 and I_2 can be benign (under the condition of being in a thin tailed environment). However

Inseparability of probability under nonlinear payoff function Now when we introduce a payoff function $g(.)$ that is nonlinear, that is that the economic or other quantifiable response to the random variable X varies with the levels of X, the discrepancy becomes greater and the conflation worse.

Commentary 11.6 (Probability as an integration kernel)
Probability is just a kernel inside an integral or a summation, not a real thing on its own. The economic world is about quantitative payoffs.

Remark 14: Inseparability of probability

Let $F : \mathcal{A} \to [0,1]$ be a probability distribution (with derivative f) and $g : \mathbb{R} \to \mathbb{R}$ a measurable function, the "payoff"". Clearly, for $\mathcal{A}'$ a subset of $\mathcal{A}$:

$$\begin{aligned}\int_{\mathcal{A}'} g(x)\mathrm{d}F(x) &= \int_{\mathcal{A}'} f(x)g(x)\mathrm{d}x \\ &\neq \int_{\mathcal{A}'} f(x)\mathrm{d}x \; g\left(\int_{\mathcal{A}'} \mathrm{d}x\right)\end{aligned}$$

In discrete terms, with $\pi(.)$ a probability mass function:

$$\begin{aligned}\sum_{x\in\mathcal{A}'} \pi(x)g(x) &\neq \sum_{x\in\mathcal{A}'} \pi(x)\; g\left(\frac{1}{n}\sum_{x\in\mathcal{A}'} x\right) \\ &= \textit{probability of event } \times \textit{payoff of average event}\end{aligned} \tag{11.17}$$

Proof. Immediate by Jensen's inequality. □

In other words, the probability of an event is an expected payoff only when, as we saw earlier, $g(x)$ is a Heaviside theta function.

Next we focus on functions tractable mathematically or legally but not reliable verbalistically via "beliefs" or "predictions".

Misunderstanding g Figure **??** showing the mishedging story of Morgan Stanley is illustrative of verbalistic notions such as "collapse" mis-expressed in nonlinear exposures. In 2007 the Wall Street firm Morgan Stanley decided to "hedge" against a real estate "collapse", before the market in real estate started declining. The problem is that they didn't realize that "collapse" could take many values, some worse than they expected, and set themselves up to benefit if there were a mild decline, but lose much if there is a larger one. They ended up right in predicting the crisis, but lose $10 billion from the "hedge".

Figure **??** shows a more complicated payoff, dubbed a "butterfly"

The function g and machine learning We note that g maps to various machine learning functions that produce exhaustive nonlinearities via the universal universal approximation theorem (Cybenko [51]), or the generalized option payoff decompositions (see *Dynamic Hedging* [222]).

Consider the function $\rho : (-\infty, \infty) \to [K, \infty)$, with K, the r.v. $X \in \mathbb{R}$:

$$\rho_{K,p}(x) = k + \frac{\log\left(e^{p(x-K)} + 1\right)}{p} \tag{11.18}$$

We can express all nonlinear payoff functions g as, with the weighting $\omega_i \in \mathbb{R}$:

$$g(x) = \sum_i \omega_i \, \rho_{K_i,p}(x) \tag{11.19}$$

by some similarity, $\rho_{K,p}(x)$ maps to the value a call price with strike K and time t to expiration normalized to 1, all rates set at 0, with sole other parameter σ the standard deviation of the underlying.

We note that the expectation of $g(.)$ is the sum of expectation of the ReLu functions:

$$\mathbb{E}\left(g(x)\right) = \sum_i \omega_i \, \mathbb{E}\left(\rho_{K_i,p}(x)\right) \tag{11.20}$$

The variance and other higher order statistical measurements are harder to obtain in closed or simple form.

Commentary 11.7
Risk management is about changing the payoff function $g(.)$ rather than making "good forecasts".

We note than λ is not a metric but a target to which one can apply various metrics.

Survival

Decision making is sequential. Accordingly, miscalibration may be a good idea if it reduces the odds of being absorbed. See the appendix of *Skin in the Game* [223], which shows the difference between ensemble probability and time probability. The expectation of the sum of n gamblers over a given day is different from that of a single gambler over n days, owing to the conditioning.

In that sense, measuring the performance of an agent who will eventually go bust (with probability one) is meaningless.[10]

11.6 CONCLUSION:

Finally, that in the real world, it is the net performance (economic or other) that counts, and making "calibration" mistakes where it doesn't matter or can be helpful should be encouraged, not penalized. The bias variance argument is well known in machine learning [122] as means to increase performance, in discussions of rationality (see *Skin in the Game* [223]) as a necessary mechanism for survival, and a very helpful psychological adaptation (Brighton and Gigerenzer [33] show a potent

10 The M5 competition is expected to correct for that by making "predictors" predict the minimum (or maximum) in a time series.

argument that if it is a bias, it is a pretty useful one.) If a mistake doesn't cost you anything –or helps you survive or improve your outcomes– it is clearly not a mistake. And if it costs you something, and has been present in society for a long time, consider that there may be hidden evolutionary advantages to these types of mistakes –of the following sort: **mistaking a bear for a stone** is worse than **mistaking a stone for a bear**.

We have shown that, in risk management, one should never operate in probability space.

11.7 APPENDIX: PROOFS AND DERIVATIONS

11.7.1 Distribution of Binary Tally $P^{(p)}(n)$

We are dealing with an average of Bernoulli random variables, with well known results but worth redoing. The characteristic function of a Bernoulli distribution with parameter p is $\psi(t) = 1 - p + E^{(}It)p$. We are concerned with the N-summed cumulant generating function $\psi'(\omega) = \log \psi(\frac{\omega}{N})^N$. We have $\kappa(p)$ the cumulant of order p:

$$\kappa(p) = -i^p \frac{\partial^p \psi'}{\partial t^p}\bigg|_{t\to 0}$$

So: $\kappa(1) = p$, $\kappa(2) = \frac{(1-p)p}{N}$, $\kappa(3) = \frac{(p-1)p(2p-1)}{N^2}$, $\kappa(4) = \frac{(1-p)p(6(p-1)p+1)}{N^3}$, which proves that $P^{(p)}(N)$ converges by the law of large numbers at speed $\sqrt{N}$, and by the central limit theorem arrives to the Gaussian at a rate of $\frac{1}{N}$, (since from the cumulants above, its kurtosis $= 3 - \frac{6(p-1)p+1}{n(p-1)p}$).

11.7.2 Distribution of the Brier Score

Base probability f First, we consider the distribution of f the base probability. We use a beta distribution that covers both the conditional and unconditional case (it is a matter of parametrization of a and b in Eq. 11.15).

Distribution of the probability Let us refresh a standard result behind nonparametric discussions and tests, dating from Kolmogorov [144] to show the rationale behind the claim that the probability distribution of probability (sic) is robust –in other words the distribution of the probability of X doesn't depend on the distribution of X, ([67] [141]).

The probability integral transform is as follows. Let X have a continuous distribution for which the cumulative distribution function (CDF) is F_X. Then –in the absence of additional information –the random variable U defined as $U = F_X(X)$ is uniform between 0 and 1.The proof is as follows: For $t \in [0, 1]$,

$$\mathbb{P}(Y \leq u) = P(F_X(X) \leq u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u \quad (11.21)$$

which is the cumulative distribution function of the uniform. This is the case regardless of the probability distribution of X.

Clearly we are dealing with 1) f beta distributed (either as a special case the uniform distribution when purely random, as derived above, or a beta distribution when one has some accuracy, for which the uniform is a special case), and 2) $\mathbb{1}_{At}$ a Bernoulli variable with probability p.

Let us consider the general case. Let $g_{a,b}$ be the PDF of the Beta:

$$g_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},\ 0 < x < 1$$

The results, a bit unwieldy but controllable:

$$\mu = \frac{\left(a^2(-(p-1)) - ap + a + b(b+1)p\right)\Gamma(a+b)}{\Gamma(a+b+2)}$$

$$\sigma_n^2 = -\frac{1}{n(a+b)^2(a+b+1)^2}\left(a^2(p-1)+a(p-1)-b(b+1)p\right)^2 + \frac{1}{(a+b+2)(a+b+3)}(a+b)(a+b+1)(p(a-b)(a+b+3)(a(a+3)+(b+1)(b+2)) - a(a+1)(a+2)(a+3))$$

We can further verify that the Brier score has thinner tails than the Gaussian as its kurtosis is lower than 3.

Proof. We start with $y_j = (f - \mathbb{1}_{Aj})$, the difference between a continuous Beta distributed random variable and a discrete Bernoulli one, both indexed by j. The characteristic function of y_j, $\Psi_f^{(y)} = \left(1 + p\left(-1 + e^{-it}\right)\right)\ {}_1F_1(a; a+b; it)$ where ${}_1F_1(.;.;.)$ is the Kummer confluent hypergeometric function ${}_1F_1(a;b;z) = \sum_{k=0}^{\infty} \frac{a_k \frac{z^k}{k}!}{b_k}$.

From here we get the characteristic function for $y_j^2 = (f_j - \mathbb{1}_{Aj})^2$

$$\Psi^{(y^2)}(t) = \sqrt{\pi}2^{-a-b+1}\Gamma(a+b)\left(p\ {}_2\tilde{F}_2\left(\frac{b+1}{2}, \frac{b}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); it\right) - (p-1)\ {}_2\tilde{F}_2\left(\frac{a+1}{2}, \frac{a}{2}; \frac{a+b}{2}, \frac{1}{2}(a+b+1); it\right)\right) \tag{11.22}$$

where ${}_2\tilde{F}_2$ is the generalized hypergeometric function regularized ${}_2\tilde{F}_2(.,.;.,.;.) = \frac{{}_2F_2(a;b;z)}{\left(\Gamma(b_1)...\Gamma(b_q)\right)}$ and ${}_pF_q(a;b;z)$ has series expansion $\sum_{k=0}^{\infty} \frac{(a_1)_k...(a_p)_k}{(b_1)_k...(b_p)_k} z^k/k!$, were $(a)_{(.)}$ is the Pochhammer symbol.

We can proceed to prove directly from there the convergence in distribution for the average $\frac{1}{n}\sum_i^n y_i^2$:

$$\lim_{n\to\infty} \Psi_{y^2}(t/n)^n = \exp\left(-\frac{it(p(a-b)(a+b+1) - a(a+1))}{(a+b)(a+b+1)}\right) \tag{11.23}$$

which is that of a degenerate Gaussian (Dirac) with location parameter $\frac{p(b-a)+\frac{a(a+1)}{a+b+1}}{a+b}$.

We can finally assess the speed of convergence, the rate at which higher moments map to those of a Gaussian distribution: consider the behavior of the 4^{th} cumulant $\kappa_4 = -i\frac{\partial^4 \log \Psi_{(.)}}{\partial t^4}|_{t\to 0}$:

1) in the maximum entropy case of $a = b = 1$:

$$\kappa_4\big|_{a=1,b=1} = -\frac{6}{7n}$$

regardless of p.

2) In the maximum variance case, using l'Hôpital:

$$\lim_{\substack{a\to 0\\ b\to 0}} \kappa_4 = -\frac{6(p-1)p+1}{n(p-1)p}$$

Se we have $\frac{\kappa_4}{\kappa_2^2} \underset{n\to\infty}{\to} 0$ at rate n^{-1}. □

Further, we can extract its probability density function of the Brier score for $N = 1$: for $0 < z < 1$,

$$p(z) = \frac{\Gamma(a+b)\left((p-1)z^{a/2}\left(1-\sqrt{z}\right)^b - p\left(1-\sqrt{z}\right)^a z^{b/2}\right)}{2\left(\sqrt{z}-1\right)z\Gamma(a)\Gamma(b)}. \tag{11.24}$$

12 ELECTION PREDICTIONS AS MARTINGALES: AN ARBITRAGE APPROACH‡

We examine the effect of uncertainty on binary outcomes, with application to elections. A standard result in quantitative finance is that when the volatility of the underlying security increases, arbitrage pressures push the corresponding binary option to trade closer to 50%, and become less variable over the remaining time to expiration. Counterintuitively, the higher the uncertainty of the underlying security, the lower the volatility of the binary option. This effect should hold in all domains where a binary price is produced – yet we observe severe violations of these principles in many areas where binary forecasts are made, in particular those concerning the U.S. presidential election of 2016. We observe stark errors among political scientists and forecasters, for instance with 1) assessors giving the candidate D. Trump between 0.1% and 3% chances of success , 2) jumps in the revisions of forecasts from 48% to 15%, both made while invoking uncertainty.

Conventionally, the quality of election forecasting has been assessed statically by De Finetti's method, which consists in minimizing the Brier score , a metric of divergence from the final outcome (the standard for tracking the accuracy of probability assessors across domains, from elections to weather). No intertemporal evaluations of changes in estimates appear to have been imposed outside the

Research chapter.

The author thanks Dhruv Madeka and Raphael Douady for detailed and extensive discussions of the paper as well as thorough auditing of the proofs across the various iterations, and, worse, the numerous changes of notation. Peter Carr helped with discussions on the properties of a bounded martingale and the transformations. I thank David Shimko,Andrew Lesniewski, and Andrew Papanicolaou for comments. I thankArthur Breitman for guidance with the literature for numerical approximations of the various logistic-normal integrals. I thank participants of the Tandon School of Engineering and Bloomberg Quantitative Finance Seminars. I also thank Bruno Dupire, MikeLawler, the Editors-In-Chief of Quantitative Finance, and various friendly people on social media. DhruvMadeka, then at Bloomberg, while working on a similar problem, independently came up with the same relationships between the volatility of an estimate and its bounds and the same arbitrage bounds. All errors are mine.

quantitative finance practice and literature. Yet De Finetti's own principle is that a probability should be treated like a two-way "choice" price, which is thus violated by conventional practice.

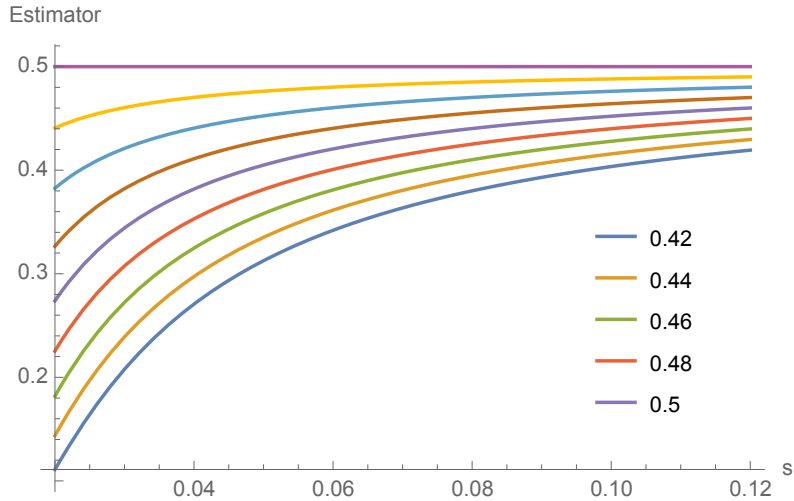


Figure 12.1: *Election arbitrage "estimation" (i.e., valuation) at different expected proportional votes* $Y \in [0,1]$*, with* s *the expected volatility of* Y *between present and election results. We can see that under higher uncertainty, the estimation of the result gets closer to 0.5, and becomes insensitive to estimated electoral margin.*

In this chapter we take a dynamic, continuous-time approach based on the principles of quantitative finance and argue that a probabilistic estimate of an election outcome by a given "assessor" needs be treated like a tradable price, that is, as a binary option value subjected to arbitrage boundaries (particularly since binary options are actually used in betting markets). Future revised estimates need to be compatible with martingale pricing, otherwise intertemporal arbitrage is created, by "buying" and "selling" from the assessor.

A mathematical complication arises as we move to continuous time and apply the standard martingale approach: namely that as a probability forecast, the underlying security lives in $[0,1]$. Our approach is to create a dual (or "shadow") martingale process Y, in an interval $[L,H]$ from an arithmetic Brownian motion, X in $(-\infty,\infty)$ and price elections accordingly. The dual process Y can for example represent the numerical votes needed for success. A complication is that, because of the transformation from X to Y, if Y is a martingale, X cannot be a martingale (and vice-versa).

The process for Y allows us to build an arbitrage relationship between the volatility of a probability estimate and that of the underlying variable, e.g. the vote number. Thus we are able to show that when there is a high uncertainty about the final outcome, 1) indeed, the arbitrage value of the forecast (as a binary option)

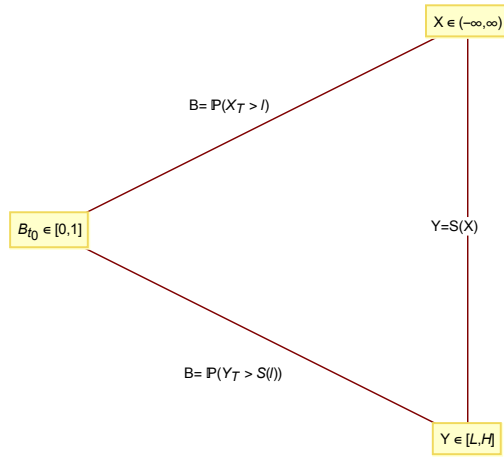


Figure 12.2: *X is an open non observable random variable (a shadow variable of sorts) on* $\mathbb{R}$*, Y, its mapping into "votes" or "electoral votes" via a sigmoidal function S(.), which maps one-to-one, and the binary as the expected value of either using the proper corresponding distribution.*

gets closer to 50% and 2) the estimate should not undergo large changes even if polls or other bases show significant variations.[3]

The pricing links are between 1) the binary option value (that is, the forecast probability), 2) the estimation of Y and 3) the volatility of the estimation of Y over the remaining time to expiration (see Figures 12.1 and 12.2).

12.0.1 Main results

For convenience, we start with our notation.

Notation

3 A central property of our model is that it prevents $B(.)$ from varying more than the estimated Y: in a two candidate contest, it will be capped (floored) at Y if lower (higher) than .5. In practice, we can observe probabilities of winning of 98% vs. 02% from a narrower spread of estimated votes of 47% vs. 53%; our approach prevents, under high uncertainty, the probabilities from diverging away from the estimated votes. But it remains conservative enough to not give a higher proportion.

Y_0 the observed estimated proportion of votes expressed in $[0,1]$ at time t_0. These can be either popular or electoral votes, so long as one treats them with consistency.

T period when the irrevocable final election outcome Y_T is revealed, or expiration.

t_0 present evaluation period, hence $T - t_0$ is the time until the final election, expressed in years.

s annualized volatility of Y, or uncertainty attending outcomes for Y in the remaining time until expiration. We assume s is constant without any loss of generality –but it could be time dependent.

$B(.)$ "forecast probability", or estimated continuous-time arbitrage evaluation of the election results, establishing arbitrage bounds between $B(.)$, Y_0 and the volatility s.

Main results

$$B(Y_0, \sigma, t_0, T) = \frac{1}{2}\text{erfc}\left(\frac{l - \text{erf}^{-1}(2Y_0 - 1)e^{\sigma^2(T-t_0)}}{\sqrt{e^{2\sigma^2(T-t_0)} - 1}}\right), \tag{12.1}$$

where

$$\sigma \approx \frac{\sqrt{\log\left(2\pi s^2 e^{2\text{erf}^{-1}(2Y_0-1)^2} + 1\right)}}{\sqrt{2}\sqrt{T - t_0}}, \tag{12.2}$$

l is the threshold needed (defaults to .5), and erfc(.) is the standard complementary error function, 1-erf(.), with $\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$. □

We find it appropriate here to answer the usual comment by statisticians and people operating outside of mathematical finance: "why not simply use a Beta-style distribution for Y?". The answer is that 1) the main purpose of the paper is establishing (arbitrage-free) time consistency in binary forecasts, and 2) we are not aware of a continuous time stochastic process that accommodates a beta distribution or a similarly bounded conventional one.

12.0.2 Organization

The remaining parts of the paper are organized as follows. First, we show the process for Y and the needed transformations from a specific Brownian motion. Second, we derive the arbitrage relationship used to obtain equation (12.1). Finally, we discuss De Finetti's approach and show how a martingale valuation relates to minimizing the conventional standard in the forecasting industry, namely the Brier Score.

A comment on absence of closed form solutions for σ We note that for Y we lack a closed form solution for the integral reflecting the total variation:

$\int_{t_0}^{T} \frac{\sigma}{\sqrt{\pi}} e^{-\text{erf}^{-1}(2y_s-1)^2} ds$, though the corresponding one for X is computable. Accordingly, we have relied on propagation of uncertainty methods to obtain a closed form solution for the probability density of Y, though not explicitly its moments as the logistic normal integral does not lend itself to simple expansions [192].

Time slice distributions for X and Y The time slice distribution is the probability density function of Y from time t, that is the one-period representation, starting at t with $y_0 = \frac{1}{2} + \frac{1}{2}\text{erf}(x_0)$. Inversely, for X given y_0, the corresponding x_0, X may be found to be normally distributed for the period $T - t_0$ with

$$\mathbb{E}(X, T) = X_0 e^{\sigma^2 (T-t_0)},$$
$$\mathbb{V}(X, T) = \frac{e^{2\sigma^2 (T-t_0)} - 1}{2}$$

and a kurtosis of 3. By probability transformation we obtain φ, the corresponding distribution of Y with initial value y_0 is given by

$$\varphi(y; y_0, T) = \frac{1}{\sqrt{e^{2\sigma^2(t-t_0)} - 1}} \exp\left\{ \text{erf}^{-1}(2y-1)^2 - \frac{1}{2}\left(\coth\left(\sigma^2 t\right) - 1\right)\left(\text{erf}^{-1}(2y-1) - \text{erf}^{-1}(2y_0-1) e^{\sigma^2(t-t_0)}\right)^2 \right\} \tag{12.3}$$

and we have $\mathbb{E}(Y_t) = Y_0$.

As to the variance, $\mathbb{E}(Y^2)$, as mentioned above, does not lend itself to a closed-form solution derived from $\varphi(.)$, nor from the stochastic integral; but it can be easily estimated from the closed form distribution of X using methods of propagation of uncertainty for the first two moments (the delta method).

Since the variance of a function f of a finite moment random variable X can be approximated as $V\left(f(X)\right) = f'\left(\mathbb{E}(X)\right)^2 V(X)$:

$$\left.\frac{\partial S^{-1}(y)}{\partial y}\right|_{y=Y_0} s^2 \approx \frac{e^{2\sigma^2(T-t_0)} - 1}{2}$$

$$s \approx \sqrt{\frac{e^{-2\text{erf}^{-1}(2Y_0-1)^2}\left(e^{2\sigma^2(T-t_0)} - 1\right)}{2\pi}}. \tag{12.4}$$

Likewise for calculations in the opposite direction, we find

$$\sigma \approx \frac{\sqrt{\log\left(2\pi s^2 e^{2\text{erf}^{-1}(2Y_0-1)^2} + 1\right)}}{\sqrt{2}\sqrt{T-t_0}},$$

which is (12.2) in the presentation of the main result.

Note that expansions including higher moments do not bring a material increase in precision – although s is highly nonlinear around the center, the range of values

for the volatility of the total or, say, the electoral college is too low to affect higher order terms in a significant way, in addition to the boundedness of the sigmoid-style transformations.

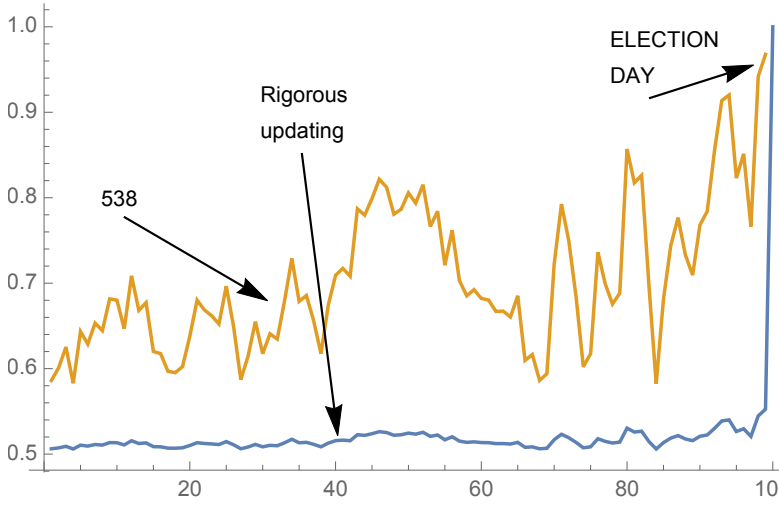


Figure 12.3: *Theoretical approach (top) vs practice (bottom). Shows how the estimation process cannot be in sync with the volatility of the estimation of (electoral or other) votes as it violates arbitrage boundaries*

12.0.3 A Discussion on Risk Neutrality

We apply risk neutral valuation, for lack of conviction regarding another way, as a default option. Although Y may not necessarily be tradable, adding a risk premium for the process involved in determining the arbitrage valuation would necessarily imply a negative one for the other candidate(s), which is hard to justify. Further, option values or binary bets, need to satisfy a no Dutch Book argument (the De Finetti form of no-arbitrage) (see [96]), i.e. properly priced binary options interpreted as probability forecasts give no betting "edge" in all outcomes without loss. Finally, any departure from risk neutrality would degrade the Brier score (about which, below) as it would represent a diversion from the final forecast.

Also note the absence of the assumptions of financing rate usually present in financial discussions.

12.1 THE BACHELIER–STYLE VALUATION

Let $F(.)$ be a function of a variable X satisfying

$$dX_t = \sigma^2 X_t dt + \sigma \, dW_t. \tag{12.5}$$

We wish to show that X has a simple Bachelier option price $B(.)$. The idea of no arbitrage is that a continuously made forecast must itself be a martingale.

Applying Itô's Lemma to $F \triangleq B$ for X satisfying (12.5) yields

$$dF = \left[\sigma^2 X \frac{\partial F}{\partial X} + \frac{1}{2}\sigma^2 \frac{\partial^2 F}{\partial X^2} + \frac{\partial F}{\partial t}\right] dt + \sigma \frac{F}{X} dW$$

so that, since $\frac{\partial F}{\partial t} \triangleq 0$, F must satisfy the partial differential equation

$$\frac{1}{2}\sigma^2 \frac{\partial^2 F}{\partial X^2} + \sigma^2 X \frac{\partial F}{\partial X} + \frac{\partial F}{\partial t} = 0, \tag{12.6}$$

which is the driftless condition that makes B a martingale.

For a binary (call) option, we have for terminal conditions $B(X,t) \triangleq F, F_T = \theta(x-l)$, where $\theta(.)$ is the Heaviside theta function and l is the threshold:

$$\theta(x) := \begin{cases} 1, & x \geq l \\ 0, & x < l \end{cases}$$

with initial condition x_0 at time t_0 and terminal condition at T given by:

$$\frac{1}{2}\text{erfc}\left(\frac{x_0 e^{\sigma^2 t} - l}{\sqrt{e^{2\sigma^2 t} - 1}}\right)$$

which is, simply, the survival function of the Normal distribution parametrized under the process for X.

Likewise we note from the earlier argument of one-to one (one can use Borel set arguments) that

$$\theta(y) := \begin{cases} 1, & y \geq S(l) \\ 0, & y < S(l), \end{cases}$$

so we can price the alternative process $B(Y,t) = \mathbb{P}(Y > \frac{1}{2})$ (or any other similarly obtained threshold l, by pricing

$$B(Y_0, t_0) = \mathbb{P}(x > S^{-1}(l)).$$

The pricing from the proportion of votes is given by:

$$B(Y_0, \sigma, t_0, T) = \frac{1}{2}\text{erfc}\left(\frac{l - \text{erf}^{-1}(2Y_0 - 1)e^{\sigma^2(T-t_0)}}{\sqrt{e^{2\sigma^2(T-t_0)} - 1}}\right),$$

the main equation (12.1), which can also be expressed less conveniently as

$$B(y_0, \sigma, t_0, T) = \frac{1}{\sqrt{e^{2\sigma^2 t} - 1}} \int_l^1 \exp\left(\text{erf}^{-1}(2y-1)^2 - \frac{1}{2}\left(\coth\left(\sigma^2 t\right) - 1\right)\left(\text{erf}^{-1}(2y-1) - \text{erf}^{-1}(2y_0 - 1)e^{\sigma^2 t}\right)^2\right) dy$$

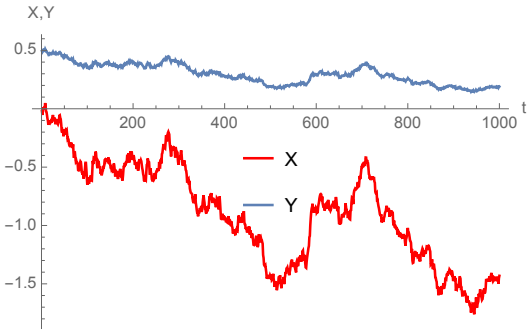


Figure 12.4: *Process and Dual Process*

12.2 BOUNDED DUAL MARTINGALE PROCESS

Y_T is the terminal value of a process on election day. It lives in $[0,1]$ but can be generalized to the broader $[L,H]$, $L,H \in [0,\infty)$. The threshold for a given candidate to win is fixed at l. Y can correspond to raw votes, electoral votes, or any other metric. We assume that Y_t is an intermediate realization of the process at t, either produced synthetically from polls (corrected estimates) or other such systems.

Next, we create, for an unbounded arithmetic stochastic process, a bounded "dual" stochastic process using a sigmoidal transformation. It can be helpful to map processes such as a bounded electoral process to a Brownian motion, or to map a bounded payoff to an unbounded one, see Figure 12.2.

Proposition 12.1
Under sigmoidal style transformations $S : x \mapsto y, \mathbb{R} \to [0,1]$ *of the form a)* $\frac{1}{2} + \frac{1}{2}erf(x)$, *or b)* $\frac{1}{1+\exp(-x)}$, *if* X *is a martingale,* Y *is only a martingale for* $Y_0 = \frac{1}{2}$, *and if* Y *is a martingale,* X *is only a martingale for* $X_0 = 0$.

Proof. The proof is sketched as follows. From Itô's lemma, the drift term for dX_t becomes 1) $\sigma^2 X(t)$, or 2) $\frac{1}{2}\sigma^2 \text{Tanh}\left(\frac{X(t)}{2}\right)$, where σ denotes the volatility, respectively with transformations of the forms a) of X_t and b) of X_t under a martingale for Y. The drift for dY_t becomes: 1) $\frac{\sigma^2 e^{-\text{erf}^{-1}(2Y-1)^2}\text{erf}^{-1}(2Y-1)}{\sqrt{\pi}}$ or 2) $\frac{1}{2}\sigma^2 Y(Y-1)(2Y-1)$ under a martingale for X. □

We therefore select the case of Y being a martingale and present the details of the transformation a). The properties of the process have been developed by Carr [35]. Let X be the arithmetic Brownian motion (12.5), with X-dependent drift and constant scale σ:

$$dX_t = \sigma^2 X_t dt + \sigma dW_t, \quad 0 < t < T < +\infty.$$

We note that this has similarities with the Ornstein-Uhlenbeck process normally written $dX_t = \theta(\mu - X_t)dt + \sigma dW$, except that we have $\mu = 0$ and violate the rules by using a negative mean reversion coefficient, rather more adequately described as "mean repelling", $\theta = -\sigma^2$.

We map from $X \in (-\infty, \infty)$ to its dual process Y as follows. With $S : \mathbb{R} \to [0,1]$, $Y = S(x)$,

$$S(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(x)$$

the dual process (by unique transformation since S is one to one, becomes, for $y \triangleq S(x)$, using Itô's lemma (since $S(.)$ is twice differentiable and $\partial S/\partial t = 0$):

$$dS = \left(\frac{1}{2}\sigma^2\frac{\partial^2 S}{\partial x^2} + X\sigma^2\frac{\partial S}{\partial x}\right)\mathrm{d}t + \sigma\frac{\partial S}{\partial x}dW$$

which with zero drift can be written as a process

$$dY_t = s(Y)dW_t,$$

for all $t > \tau, \mathbb{E}(Y_t|Y_\tau) = Y_\tau$. and scale

$$s(Y) = \frac{\sigma}{\sqrt{\pi}}e^{-\text{erf}^{-1}(2y-1)^2}$$

which as we can see in Figure 12.5, $s(y)$ can be approximated by the quadratic function $y(1-y)$ times a constant.

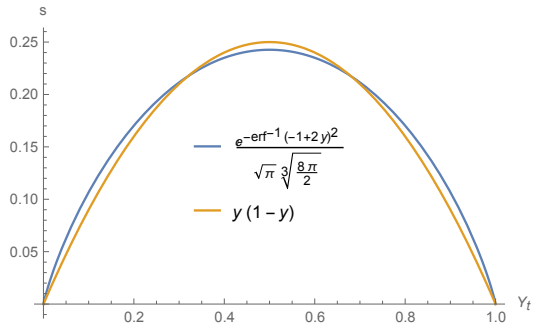


Figure 12.5: *The instantaneous volatility of Y as a function of the level of Y for two different methods of transformations of X, which appear to not be substantially different. We compare to the quadratic form $y - y^2$ scaled by a constant $\frac{1}{\sqrt[3]{\frac{8\pi}{2}}}$. The volatility declines as we move away from $\frac{1}{2}$ and collapses at the edges, thus maintaining Y in $(0,1)$. For simplicity we assumed $\sigma = t = 1$.*

We can recover equation (12.5) by inverting, namely $S^{-1}(y) = \text{erf}^{-1}(2y-1)$, and again applying Itô's Lemma. As a consequence of gauge invariance option prices are identical whether priced on X or Y, even if one process has a drift while the other is a martingale . In other words, one may apply one's estimation to the electoral threshold, or to the more complicated X with the same results. And, to summarize our method, pricing an option on X is familiar, as it is exactly a Bachelier-style option price.

12.3 RELATION TO DE FINETTI'S PROBABILITY ASSESSOR

This section provides a brief background for the conventional approach to probability assessment. The great De Finetti [57] has shown that the "assessment" of the "probability" of the realization of a random variable in $\{0,1\}$ requires a nonlinear

Figure 12.6: *Bruno de Finetti (1906-1985). A probabilist, philosopher, and insurance mathematician, he formulated the Brier score for probabilistic assessment which we show is compatible dynamically with a martingale. Source: DeFinetti.org*

loss function – which makes his definition of *probabilistic assessment* differ from that of the P/L of a trader engaging in binary bets.

Assume that a betting agent in an n-repeated two period model, t_0 and t_1, produces a strategy $\mathfrak{S}$ of bets $b_{0,i} \in [0,1]$ indexed by $i = 1, 2, \ldots, n$, with the realization of the binary r.v. $\mathbb{1}_{t_1,i}$. If we take the absolute variation of his P/L over n bets, it will be

$$L_1(\mathfrak{S}) = \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{1}_{t_1,i} - b_{t_0,i} \right| .$$

For example, assume that $\mathbb{E}(\mathbb{1}_{t_1}) = \frac{1}{2}$. Betting on the probability, here $\frac{1}{2}$, produces a loss of $\frac{1}{2}$ in expectation, which is the same as betting either 0 or 1 – hence not favoring the agent to bet on the exact probability.

If we work with the same random variable and non-time-varying probabilities, the L^1 metric would be appropriate:

$$L_1(\mathfrak{S}) = \frac{1}{n} \left| \mathbb{1}_{t_1,i} - \sum_{i=1}^{n} b_{t_0,i} \right| .$$

De Finetti proposed a "Brier score" type function, a quadratic loss function in $\mathcal{L}^2$:

$$L_2(\mathfrak{S}) = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}_{t_1,i} - b_{t_0,i})^2,$$

the minimum of which is reached for $b_{t_0,i} = \mathbb{E}(\mathbb{1}_{t_1})$.

In our world of continuous time derivative valuation, where, in place of a two period lattice model, we are interested, for the same final outcome at t_1, in the stochastic process b_t, $t_0 \geq t \geq t_1$, the arbitrage "value" of a bet on a binary outcome needs to match the expectation, hence, again, we map to the Brier score – *by an arbitrage argument*. Although there is no quadratic loss function involved, the fact that the bet is a function of a martingale, which is required to be itself a martingale, i.e. that the conditional expectation remains invariant to time, does not allow an arbitrage to take place. A "high" price can be "shorted" by the arbitrageur, a "low" price can be "bought", and so on repeatedly. The consistency between bets at period t and other periods $t + \Delta t$ enforces the probabilistic discipline. In other words, someone can "buy" from the forecaster then "sell" back to him, generating a positive expected "return" if the forecaster is out of line with martingale valuation.

As to the current practice by forecasters, although some election forecasters appear to be aware of the need to minimize their Brier score , the idea that the revisions of estimates should also be subjected to martingale valuation is not well established.

12.4 CONCLUSION AND COMMENTS

As can be seen in Figure 12.1, a binary option reveals more about uncertainty than about the true estimation, a result well known to traders, see [222].

In the presence of more than 2 candidates, the process can be generalized with the following heuristic approximation. Establish the stochastic process for $Y_{1,t}$, and just as $Y_{1,t}$ is a process in $[0, 1]$, $Y_{2,t}$ is a process $\in (Y_{1,t}, 1]$, with $Y_{3,t}$ the residual $1 - Y_{2,t} - Y_{1,t}$, and more generally $Y_{n-1,t} \in (Y_{n_2,t}, 1]$ and $Y_{n,t}$ is the residual $Y_n = 1 - \sum_{i=1}^{n-1} Y_{i,t}$. For n candidates, the n^{th} is the residual.

ADDENDUM: ALL ROADS LEAD TO QUANTITATIVE FINANCE

Background *Aubrey Clayton [ref] sent a letter to the editor complaining about the previous piece on grounds of "errors" in the above methodology. The author answered, with Dhruv Madeka, not quite to Clayton, rather to express the usefulness of quantitative finance methods in life.*

We are happy to respond to Clayton's (non-reviewed) letter, in spite of its confusions, as it will give us the opportunity to address more fundamental misunderstandings of the role of quantitative finance in general, and arbitrage pricing in particular, and proudly show how "all roads lead to quantitative finance", that is, that arbitrage approaches are universal and applicable to all manner of binary forecasting. It also allows the second author to comment from his paper, Madeka (2017)[156], which independently and simultaneously obtained similar results to Taleb (2018)[231].

Incorrect claims

Taleb's criticism of popular forecast probabilities, specifically the election forecasts of FiveThirtyEight..." and "*He [Taleb] claims this means the FiveThirtyEight forecasts must have "violate[d] arbitrage boundaries*" are factually incorrect.

There is no mention of FiveThirtyEight in Taleb (2018), and Clayton must be confusing scientific papers with Twitter debates. The paper is an attempt at addressing elections in a rigorous manner, not journalistic discussion, and only mentions the 2016 election in one illustrative sentence.[4]

Let us however continue probing Clayton's other assertions, in spite of his confusion and the nature of the letter.

Incorrect arbitrage valuation

Clayton's claims either an error ("First, one of the "standard results" of quantitative finance that his election forecast assessments rely on is false", he initially writes), or, as he confusingly retracts, something "only partially true". Again, let us set aside that Taleb(2018)[231] makes no "assessment" of FiveThirtyEight's record and outline his reasoning.

Clayton considers three periods, $t_0 = 0$, an intermediate period t and a terminal one T, with $t_0 \leq t < T$. Clayton shows a special case of the distribution of the forward probability, seen at t_0, for time starting at $t = \frac{T}{2}$ and ending at T. It is a uniform distribution for that specific time period. In fact under his construction, using the probability integral transform, one can show that the probabilities follow what resembles a symmetric beta distribution with parameters a and b, and with $a = b$. When $t = \frac{T}{2}$, we have $a = b = 1$ (hence the uniform distribution). Before $T/2$ it has a $\cap$ shape, with Dirac at $t = t_0$. Beyond $T/2$ it has a $\cup$ shape, ending with two Dirac sticks at 0 and 1 (like a Bernoulli) when t is close to T (and close to an arcsine distribution with $a = b = \frac{1}{2}$ somewhere in between).

Clayton's construction is indeed misleading, since he analyzes the distribution of the price at time t with the filtration at time t_0, particularly when discussing arbitrage pricing and arbitrage pressures. Agents value options between t and T at time t (not period t_0), with an underlying price: under such constraint, the binary option automatically converges towards $\frac{1}{2}$ as $\sigma \to \infty$, and that for *any* value of the underlying price, no matter how far away from the strike price (or threshold). The σ here is never past realized, only future unrealized volatility. This can be seen within the framework presented in Taleb (2018) [231] but also by taking any binary option pricing model. A price is not a probability (less even a probability distribution), but an expectation. Simply, as arbitrage operators, we look at *future* volatility given information about the underlying when pricing a binary option, not the distribution of probability itself in the unconditional abstract.

At infinite σ, it becomes all noise, and such a level of noise drowns all signals.

4 Incidentally, the problem with FiveThirtyEight isn't changing probabilities from .55 to .85 within a 5 months period, it is performing abrupt changes within a much shorter timespan –and *that* was discussed in Madeka (2017)[156].

Another way to view the pull of uncertainty towards $\frac{1}{2}$ is in using information theory and the notion of maximum entropy under deep uncertainty : the entropy (I) of a Bernoulli distribution with probabilities p and $(1-p)$, $I = -((1-p)\log(1-p) + p\log(p))$ is maximal at $\frac{1}{2}$.

To beat a $\frac{1}{2}$ pricing one needs to have enough information to beat the noise. As we will see in the next section, it is not easy.

Arbitrage counts

Another result from quantitative finance that puts bounds on the volatility of forecasting is as follows. Since election forecasts can be interpreted as a European binary option, we can exploit the fact that the price process of this option is bounded between 0 and 1 to make claims about the volatility of the price itself.

Essentially, if the price of the binary option varies too much, a simple trading strategy of buying low and selling high is guaranteed to produce a profit[5]. The argument can be summed up by noting that if we consider an arithmetic brownian motion that's bounded between $[L, H]$:

$$dB_t = \sigma dW_t \tag{12.7}$$

The stochastic integral $2\int^T (B_0 - B_t)dB_t = \sigma^2 T - (B_T - B_0)^2$ can be replicated at zero cost, indicating that the value of B_T is bounded by the maximum value of the square difference on the right hand side of the equation. That is, a forecaster who produces excessively volatile probabilities – if he or she is willing to trade on such a forecast (i.e. they have skin in the game) – can be arbitraged by following a strategy that sells (proportionally) when the forecast is too high and buys (proportionally) when the forecast is too low.

To conclude, any numerical probabilistic forecasting should be treated like a choice price —De Finetti's intuition is that forecasts should have skin in the game. Under these conditions, binary forecasting belongs to the rules of arbitrage and derivative pricing, well mapped in quantitative finance. Using a quantitative finance approach to produce binary forecasts does not prevent Bayesian methods (Taleb(2018) does not say probabilities should be $\frac{1}{2}$, only that there is a headwind towards that level owing to arbitrage pressures and constraints on how variable a forecast can be). It is just that there is one price that counts at the end, 1 or 0, which puts a structure on the updating.[6]

5 We take this result from Bruno Dupire's notes for his continuous time finance class at NYU's Courant Institute, particularly his final exam for the Spring of 2019.

6 Another way to see it, from outside our quantitative finance models: consider a standard probabilistic score. Let $X_1, \ldots, X_n$ be random variables in $[0, 1$ and a B_T a constant $B_T \in \{0, 1\}$, we have the λ score

$$\lambda_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - B_T)^2\,,$$

which needs to be minimized (on a single outcome B_T). For any given B_T and an average forecast $\bar{x} = \sum_{i=1}^{n} x_i$, the minimum value of λ_n is reached for $x_1 = \ldots = x_n$. To beat a Dirac forecast $x_1 = \ldots = x_n = \frac{1}{2}$ for which $\lambda = \frac{1}{4}$ with a high variance strategy, one needs to have 75% accuracy. (Note that a uniform forecast has a score of $\frac{1}{3}$.) This shows us the trade-off between volatility and signal.

The reason Clayton might have trouble with quantitative finance could be that probabilities and underlying polls may not be martingales in real life; traded probabilities (hence real forecasts) must be martingales. Which is why in Taleb (2018)[231] the process for the polls (which can be vague and nontradable) needs to be transformed into a process for probability in $[0, 1]$.

ACKNOWLEDGMENTS

Raphael Douady, Students at NYU Tandon School of Engineering, participants at the Bloomberg Quantitative Finance Seminar in New York.

Part IV

INEQUALITY ESTIMATORS UNDER FAT TAILS

13 GINI ESTIMATION UNDER INFINITE VARIANCE ‡

footnotetext(With A. Fontanari and P. Cirillo), coauthors

THIS CHAPTER is about the problems related to the estimation of the Gini index in presence of a fat-tailed data generating process, i.e. one in the stable distribution class with finite mean but infinite variance (i.e. with tail index $\alpha \in (1, 2)$). We show that, in such a case, the Gini coefficient cannot be reliably estimated using conventional nonparametric methods, because of a downward bias that emerges under fat tails. This has important implications for the ongoing discussion about economic inequality.

We start by discussing how the nonparametric estimator of the Gini index undergoes a phase transition in the symmetry structure of its asymptotic distribution, as the data distribution shifts from the domain of attraction of a light-tailed distribution to that of a fat-tailed one, especially in the case of infinite variance. We also show how the nonparametric Gini bias increases with lower values of α. We then prove that maximum likelihood estimation outperforms nonparametric methods, requiring a much smaller sample size to reach efficiency.

Finally, for fat-tailed data, we provide a simple correction mechanism to the small sample bias of the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution.

13.1 INTRODUCTION

Wealth inequality studies represent a field of economics, statistics and econophysics exposed to fat-tailed data generating processes, often with infinite variance [39, 142]. This is not at all surprising if we recall that the prototype of fat-tailed distributions, the Pareto, has been proposed for the first time to model household in-

‡ Research chapter.

comes [182]. However, the fat-tailedness of data can be problematic in the context of wealth studies, as the property of efficiency (and, partially, consistency) does not necessarily hold for many estimators of inequality and concentration [81, 142].

The scope of this work is to show how fat tails affect the estimation of one of the most celebrated measures of economic inequality, the Gini index [77, 109, 142], often used (and abused) in the econophysics and economics literature as the main tool for describing the distribution and the concentration of wealth around the world [39, 188**?**].

The literature concerning the estimation of the Gini index is wide and comprehensive (e.g. [77, 219] for a review), however, strangely enough, almost no attention has been paid to its behavior in presence of fat tails, and this is curious if we consider that: 1) fat tails are ubiquitous in the empirical distributions of income and wealth [142, 188], and 2) the Gini index itself can be seen as a measure of variability and fat-tailedness [75, 78, 79, 94].

The standard method for the estimation of the Gini index is nonparametric: one computes the index from the empirical distribution of the available data using Equation (13.5) below. But, as we show in this paper, this estimator suffers from a downward bias when we deal with fat-tailed observations. Therefore our goal is to close this gap by deriving the limiting distribution of the nonparametric Gini estimator in presence of fat tails, and propose possible strategies to reduce the bias. We show how the maximum likelihood approach, despite the risk of model misspecification, needs much fewer observations to reach efficiency when compared to a nonparametric one.[2]

Our results are relevant to the discussion about wealth inequality, recently rekindled by Thomas Piketty in [188], as the estimation of the Gini index under fat tails and infinite variance may cause several economic analyses to be unreliable, if not markedly wrong. Why should one trust a biased estimator?

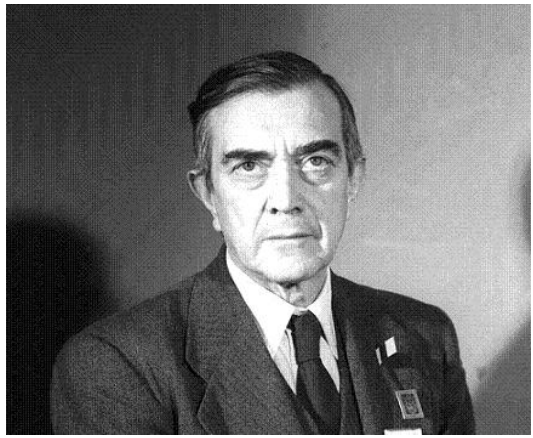


Figure 13.1: *The Italian statistician Corrado Gini, 1884-1965. source: Bocconi.*

2 A similar bias also affects the nonparametric measurement of quantile contributions, i.e. those of the type "the top 1% owns x% of the total wealth" [238]. This paper extends the problem to the more widespread Gini coefficient, and goes deeper by making links with the limit theorems.

By fat-tailed data we indicate those data generated by a positive random variable X with cumulative distribution function (c.d.f.) $F(x)$, which is regularly-varying of order α [134], that is, for $\bar{F}(x) := 1 - F(x)$, one has

$$\lim_{x\to\infty} x^{\alpha}\bar{F}(x) = L(x), \tag{13.1}$$

where $L(x)$ is a slowly-varying function such that $\lim_{x\to\infty} \frac{L(cx)}{L(x)} = 1$ with $c > 0$, and where $\alpha > 0$ is called the tail exponent .

Regularly-varying distributions define a large class of random variables whose properties have been extensively studied in the context of extreme value theory [81, 115], when dealing with the probabilistic behavior of maxima and minima. As pointed out in [44], regularly-varying and fat-tailed are indeed synonyms. It is known that, if $X_1, ..., X_n$ are i.i.d. observations with a c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (13.1), then their data generating process falls into the maximum domain of attraction of a Fréchet distribution with parameter ρ, in symbols $X \in MDA(\Phi(\rho))$[115]. This means that, for the partial maximum $M_n = \max(X_1, ..., X_n)$, one has

$$P\left(a_n^{-1}\left(M_n - b_n\right) \le x\right) \xrightarrow{d} \Phi(\rho) = e^{-x^{-\rho}}, \qquad \rho > 0, \tag{13.2}$$

with $a_n > 0$ and $b_n \in \mathbb{R}$ two normalizing constants. Clearly, the connection between the regularly-varying coefficient α and the Fréchet distribution parameter ρ is given by: $\alpha = \frac{1}{\rho}$ [81].
The Fréchet distribution is one of the limiting distributions for maxima in extreme value theory, together with the Gumbel and the Weibull; it represents the fat-tailed and unbounded limiting case [115]. The relationship between regularly-varying random variables and the Fréchet class thus allows us to deal with a very large family of random variables (and empirical data), and allows us to show how the Gini index is highly influenced by maxima, i.e. extreme wealth, as clearly suggested by intuition [94, 142], especially under infinite variance. Again, this recommends some caution when discussing economic inequality under fat tails.

It is worth remembering that the existence (finiteness) of the moments for a fat-tailed random variable X depends on the tail exponent α, in fact

$$\begin{aligned} E(X^{\delta}) &< \infty \text{ if } \delta \le \alpha, \\ E(X^{\delta}) &= \infty \text{ if } \delta > \alpha. \end{aligned} \tag{13.3}$$

In this work, we restrict our focus on data generating processes with finite mean and infinite variance, therefore, according to Equation (13.3), on the class of regularly-varying distributions with tail index $\alpha \in (1, 2)$.

Table 13.1 and Figure 13.2 present numerically and graphically our story, already suggesting its conclusion, on the basis of artificial observations sampled from a Pareto distribution (Equation (13.13) below) with tail parameter α equal to 1.1.

Table 13.1 compares the nonparametric Gini index of Equation (13.5) with the maximum likelihood (ML) tail-based one of Section 13.3. For the different sample sizes in Table 13.1, we have generated 10^8 samples, averaging the estimators via

Monte Carlo. As the first column shows, the convergence of the nonparametric estimator to the true Gini value ($g = 0.8333$) is extremely slow and monotonically increasing; this suggests an issue not only in the tail structure of the distribution of the nonparametric estimator but also in its symmetry.

Figure 13.2 provides some numerical evidence that the limiting distribution of the nonparametric Gini index loses its properties of normality and symmetry [90], shifting towards a skewed and fatter-tailed limit, when data are characterized by an infinite variance. As we prove in Section 13.2, when the data generating process is in the domain of attraction of a fat-tailed distribution, the asymptotic distribution of the Gini index becomes a skewed-to-the-right α-stable law. This change of behavior is responsible of the downward bias of the nonparametric Gini under fat tails. However, the knowledge of the new limit allows us to propose a correction for the nonparametric estimator, improving its quality, and thus reducing the risk of badly estimating wealth inequality, with all the possible consequences in terms of economic and social policies [142, 188].

Table 13.1: *Comparison of the Nonparametric (NonPar) and the Maximum Likelihood (ML) Gini estimators, using Paretian data with tail $\alpha = 1.1$ (finite mean, infinite variance) and different sample sizes. Number of Monte Carlo simulations:* 10^8.

n	Nonpar		ML		Error Ratio[3]
(number of obs.)	*Mean*	*Bias*	*Mean*	*Bias*	
10^3	*0.711*	*-0.122*	*0.8333*	*0*	1.4
10^4	*0.750*	*-0.083*	*0.8333*	*0*	3
10^5	*0.775*	*-0.058*	*0.8333*	*0*	6.6
10^6	0.790	-0.043	0.8333	0	156
10^7	0.802	-0.031	0.8333	0	10^5+

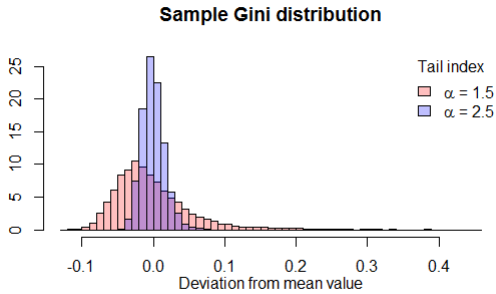


Figure 13.2: *Histograms for the Gini nonparametric estimators for two Paretian (type I) distributions with different tail indices, with finite and infinite variance (plots have been centered to ease comparison). Sample size:* 10^3. *Number of samples:* 10^2 *for each distribution.*

The rest of the paper is organized as follows. In Section 13.2 we derive the asymptotic distribution of the sample Gini index when data possess an infinite variance. In Section 13.3 we deal with the maximum likelihood estimator; in Section 13.4 we provide an illustration with Paretian observations; in Section 13.5 we propose a simple correction based on the mode-mean distance of the asymptotic distribution of the nonparametric estimator, to take care of its small-sample bias. Section 13.6

closes the paper. A technical Appendix contains the longer proofs of the main results in the work.

13.2 ASYMPTOTICS OF THE NONPARAMETRIC ESTIMATOR UNDER INFINITE VARIANCE

We now derive the asymptotic distribution for the nonparametric estimator of the Gini index when the data generating process is fat-tailed with finite mean but infinite variance.

The so-called stochastic representation of the Gini g is

$$g = \frac{1}{2}\frac{\mathbb{E}\left(|X' - X''|\right)}{\mu} \in [0, 1], \tag{13.4}$$

where X' and X'' are i.i.d. copies of a random variable X with c.d.f. $F(x) \in [c, \infty)$, $c > 0$, and with finite mean $\mathbb{E}(X) = \mu$. The quantity $\mathbb{E}\left(|X' - X''|\right)$ is known as the "Gini Mean Difference" (GMD) [219]. For later convenience we also define $g = \frac{\theta}{\mu}$ with $\theta = \frac{\mathbb{E}(|X'-X''|)}{2}$.

The Gini index of a random variable X is thus the mean expected deviation between any two independent realizations of X, scaled by twice the mean [80].

The most common nonparametric estimator of the Gini index for a sample $X_1, ..., X_n$ is defined as

$$G^{NP}(X_n) = \frac{\sum_{1 \le i < j \le n} |X_i - X_j|}{(n-1)\sum_{i=1}^{n} X_i}, \tag{13.5}$$

which can also be expressed as

$$G^{NP}(X_n) = \frac{\sum_{i=1}^{n}(2(\frac{i-1}{n-1}) - 1)X_{(i)}}{\sum_{i=1}^{n} X_{(i)}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)}}{\frac{1}{n}\sum_{i=1}^{n} X_i}, \tag{13.6}$$

where $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are the ordered statistics of $X_1, ..., X_n$, such that: $X_{(1)} < X_{(2)} < ... < X_{(n)}$ and $Z_{(i)} = 2\left(\frac{i-1}{n-1} - 1\right) X_{(i)}$. The asymptotic normality of the estimator in Equation (13.6) under the hypothesis of finite variance for the data generating process is known [142, 219]. The result directly follows from the properties of the U-statistics and the L-estimators involved in Equation (13.6)

A standard methodology to prove the limiting distribution of the estimator in Equation (13.6), and more in general of a linear combination of order statistics, is to show that, in the limit for $n \to \infty$, the sequence of order statistics can be approximated by a sequence of i.i.d random variables [55, 149]. However, this usually requires some sort of L^2 integrability of the data generating process, something we are not assuming here.

Lemma 13.1 (proved in the Appendix) shows how to deal with the case of sequences of order statistics generated by fat-tailed L^1-only integrable random variables.

Lemma 13.1
Consider the following sequence $R_n = \frac{1}{n}\sum_{i=1}^{n}(\frac{i}{n} - U_{(i)})F^{-1}(U_{(i)})$ *where* $U_{(i)}$ *are the order statistics of a uniformly distributed i.i.d random sample. Assume that* $F^{-1}(U) \in L^1$. *Then the following results hold:*

$$R_n \xrightarrow{L^1} 0, \tag{13.7}$$

and

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \xrightarrow{L^1} 0, \tag{13.8}$$

with $\alpha \in (1,2)$ *and* $L_0(n)$ *a slowly-varying function.*

13.2.1 A Quick Recap on α-Stable Random Variables

We here introduce some notation for α-stable distributions, as we need them to study the asymptotic limit of the Gini index.

A random variable X follows an α-stable distribution, in symbols $X \sim S(\alpha, \beta, \gamma, \delta)$, if its characteristic functionis

$$E(e^{itX}) = \begin{cases} e^{-\gamma^\alpha |t|^\alpha (1 - i\beta \operatorname{sign}(t)) \tan(\frac{\pi\alpha}{2}) + i\delta t} & \alpha \neq 1 \\ e^{-\gamma|t|(1 + i\beta \frac{2}{\pi} \operatorname{sign}(t)) \ln|t| + i\delta t} & \alpha = 1 \end{cases},$$

where $\alpha \in (0,2)$ governs the tail, $\beta \in [-1,1]$ is the skewness, $\gamma \in \mathbb{R}^+$ is the scale parameter, and $\delta \in \mathbb{R}$ is the location one. This is known as the $S1$ parametrization of α-stable distributions [179, 206].

Interestingly, there is a correspondence between the α parameter of an α-stable random variable, and the α of a regularly-varying random variable as per Equation (13.1): as shown in [90, 179], a regularly-varying random variable of order α is α-stable, with the same tail coefficient. This is why we do not make any distinction in the use of the α here. Since we aim at dealing with distributions characterized by finite mean but infinite variance, we restrict our focus to $\alpha \in (1,2)$, as the two α's coincide.

Recall that, for $\alpha \in (1,2]$, the expected value of an α-stable random variable X is equal to the location parameter δ, i.e. $\mathbb{E}(X) = \delta$. For more details, we refer to [179, 206].

The standardized α-stable random variable is expressed as

$$S_{\alpha,\beta} \sim S(\alpha, \beta, 1, 0). \tag{13.9}$$

We note that α-stable distributions are a subclass of infinitely divisible distributions. Thanks to their closure under convolution, they can be used to describe the limiting behavior of (rescaled) partials sums, $S_n = \sum_{i=1}^{n} X_i$, in the General central limit theorem (GCLT) setting [90]. For $\alpha = 2$ we obtain the normal distribution as a special case, which is the limit distribution for the classical CLTs, under the hypothesis of finite variance.

In what follows we indicate that a random variable is in the domain of attraction of an α-stable distribution, by writing $X \in DA(S_\alpha)$. Just observe that this condition for the limit of partial sums is equivalent to the one given in Equation (13.2) for the limit of partial maxima [81, 90].

13.2.2 The α-Stable Asymptotic Limit of the Gini Index

Consider a sample $X_1, ..., X_n$ of i.i.d. observations with a continuous c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (13.1), with tail index $\alpha \in (1, 2)$. The data generating process for the sample is in the domain of attraction of a Fréchet distribution with $\rho \in (\frac{1}{2}, 1)$, given that $\rho = \frac{1}{\alpha}$.

For the asymptotic distribution of the Gini index estimator, as presented in Equation (13.6), when the data generating process is characterized by an infinite variance, we can make use of the following two theorems: Theorem 1 deals with the limiting distribution of the Gini Mean Difference (the numerator in Equation (13.6)), while Theorem 2 extends the result to the complete Gini index. Proofs for both theorems are in the Appendix.

Theorem 1
Consider a sequence $(X_i)_{1 \le i \le n}$ of i.i.d random variables from a distribution X on $[c, +\infty)$ with $c > 0$, such that X is in the domain of attraction of an α-stable random variable, $X \in DA(S_\alpha)$, with $\alpha \in (1, 2)$. Then the sample Gini mean deviation (GMD) $\frac{\sum_{i=1}^n Z_{(i)}}{n}$ satisfies the following limit in distribution:

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} - \theta \right) \xrightarrow{d} S_{\alpha,1}, \tag{13.10}$$

where $Z_i = (2F(X_i) - 1)X_i$, $\mathbb{E}(Z_i) = \theta$, $L_0(n)$ is a slowly-varying function such that Equation (13.37) holds (see the Appendix), and $S_{\alpha,1}$ is a right-skewed standardized α-stable random variable defined as in Equation (13.9).

Moreover the statistic $\frac{1}{n} \sum_{i=1}^n Z_{(i)}$ is an asymptotically consistent estimator for the GMD, i.e. $\frac{1}{n} \sum_{i=1}^n Z_{(i)} \xrightarrow{P} \theta$.

Note that Theorem 1 could be restated in terms of the maximum domain of attraction $MDA(\Phi(\rho))$ as defined in Equation (13.2).

Theorem 2
Given the same assumptions of Theorem 1, the estimated Gini index $G^{NP}(X_n) = \frac{\sum_{i=1}^n Z_{(i)}}{\sum_{i=1}^n X_i}$ satisfies the following limit in distribution

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(G^{NP}(X_n) - \frac{\theta}{\mu} \right) \xrightarrow{d} Q, \tag{13.11}$$

where $\mathbb{E}(Z_i) = \theta$, $\mathbb{E}(X_i) = \mu$, $L_0(n)$ is the same slowly-varying function defined in Theorem 1 and Q is a right-skewed α-stable random variable $S(\alpha, 1, \frac{1}{\mu}, 0)$.

Furthermore the statistic $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$ *is an asymptotically consistent estimator for the Gini index, i.e.* $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i} \xrightarrow{P} \frac{\theta}{\mu} = g$.

In the case of fat tails with $\alpha \in (1, 2)$, Theorem 2 tells us that the asymptotic distribution of the Gini estimator is always right-skewed notwithstanding the distribution of the underlying data generating process. Therefore heavily fat-tailed data not only induce a fatter-tailed limit for the Gini estimator, but they also change the shape of the limit law, which definitely moves away from the usual symmetric Gaussian. As a consequence, the Gini estimator, whose asymptotic consistency is still guaranteed [149], will approach its true value more slowly, and from below. Some evidence of this was already given in Table 13.1.

13.3 THE MAXIMUM LIKELIHOOD ESTIMATOR

Theorem 2 indicates that the usual nonparametric estimator for the Gini index is not the best option when dealing with infinite-variance distributions, due to the skewness and the fatness of its asymptotic limit. The aim is to find estimators that still preserve their asymptotic normality under fat tails, which is not possible with nonparametric methods, as they all fall into the α-stable Central Limit Theorem case [81, 90]. Hence the solution is to use parametric techniques.

Theorem 3 shows how, once a parametric family for the data generating process has been identified, it is possible to estimate the Gini index via MLE. The resulting estimator is not just asymptotically normal, but also asymptotically efficient.

In Theorem 3 we deal with random variables X whose distribution belongs to the large and flexible exponential family [208], i.e. whose density can be represented as

$$f_\theta(x) = h(x)e^{(\eta(\theta)T(x) - A(\theta))},$$

with $\theta \in \mathbb{R}$, and where $T(x)$, $\eta(\theta)$, $h(x)$, $A(\theta)$ are known functions.

Theorem 3
Let $X \sim F_\theta$ *such that* F_θ *is a distribution belonging to the exponential family. Then the Gini index obtained by plugging-in the maximum likelihood estimator of* θ, $G^{ML}(X_n)_\theta$, *is asymptotically normal and efficient. Namely:*

$$\sqrt{n}\left(G^{ML}(X_n)_\theta - g_\theta\right) \xrightarrow{D} \mathcal{N}\left(0, g_\theta'^2 I^{-1}(\theta)\right), \tag{13.12}$$

where $g'_\theta = \frac{dg_\theta}{d\theta}$ *and* $I(\theta)$ *is the Fisher Information.*

$$\sqrt{n}\left(G^{ML}(X_n)_\theta - g_\theta\right) \xrightarrow{D} \mathcal{N}\left(0, g_\theta'^2 I^{-1}(\theta)\right),$$

Proof. The result follows easily from the asymptotic efficiency of the maximum likelihood estimators of the exponential family, and the invariance principle of MLE. In particular, the validity of the invariance principle for the Gini index is granted

by the continuity and the monotonicity of g_θ with respect to θ. The asymptotic variance is then obtained by application of the delta-method [208]. □

13.4 A PARETIAN ILLUSTRATION

We provide an illustration of the obtained results using some artificial fat-tailed data. We choose a Pareto I [182], with density

$$f(x) = \alpha c^\alpha x^{-\alpha-1}, x \geq c. \tag{13.13}$$

It is easy to verify that the corresponding survival function $\bar{F}(x)$ belongs to the regularly-varying class with tail parameter α and slowly-varying function $L(x) = c^\alpha$. We can therefore apply the results of Section 13.2 to obtain the following corollaries.

Corollary 13.1
Let $X_1, ..., X_n$ be a sequence of i.i.d. observations with Pareto distribution with tail parameter $\alpha \in (1, 2)$. The nonparametric Gini estimator is characterized by the following limit:

$$D_n^{NP} = G^{NP}(X_n) - g \sim S\left(\alpha, 1, \frac{C_\alpha^{-\frac{1}{\alpha}}}{n^{\frac{\alpha-1}{\alpha}}} \frac{(\alpha-1)}{\alpha}, 0\right). \tag{13.14}$$

Proof. Without loss of generality we can assume $c = 1$ in Equation (13.13). The results is a mere application of Theorem 2, remembering that a Pareto distribution is in the domain of attraction of α-stable random variables with slowly-varying function $L(x) = 1$. The sequence c_n to satisfy Equation (13.37) becomes $c_n = n^{\frac{1}{\alpha}} C_\alpha^{-\frac{1}{\alpha}}$, therefore we have $L_0(n) = C_\alpha^{-\frac{1}{\alpha}}$, which is independent of n. Additionally the mean of the distribution is also a function of α, that is $\mu = \frac{\alpha}{\alpha-1}$. □

Corollary 13.2
Let the sample $X_1, ..., X_n$ be distributed as in Corollary 13.1, let G_θ^{ML} be the maximum likelihood estimator for the Gini index as defined in Theorem 3. Then the MLE Gini estimator, rescaled by its true mean g, has the following limit:

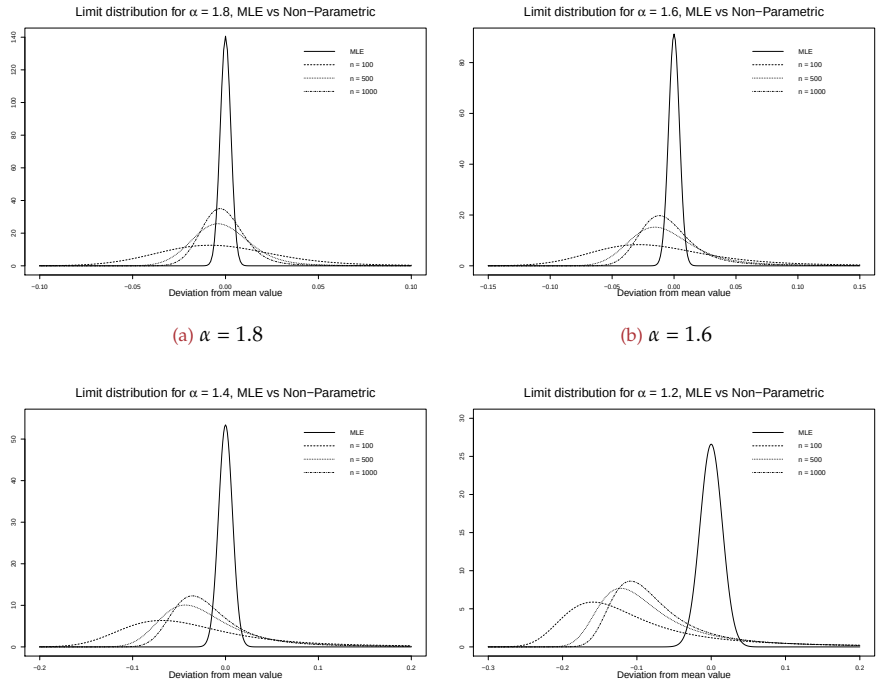
$$D_n^{ML} = G_\alpha^{ML}(X_n) - g \sim N\left(0, \frac{4\alpha^2}{n(2\alpha-1)^4}\right), \tag{13.15}$$

where N indicates a Gaussian.

Proof. The functional form of the maximum likelihood estimator for the Gini index is known to be $G_\theta^{ML} = \frac{1}{2\alpha^{ML}-1}$ [142]. The result then follows from the fact that the Pareto distribution (with known minimum value x_m) belongs to an exponential family and therefore satisfies the regularity conditions necessary for the asymptotic normality and efficiency of the maximum likelihood estimator. Also notice that the Fisher information for a Pareto distribution is $\frac{1}{\alpha^2}$. □

Now that we have worked out both asymptotic distributions, we can compare the quality of the convergence for both the MLE and the nonparametric case when dealing with Paretian data, which we use as the prototype for the more general class of fat-tailed observations.

In particular, we can approximate the distribution of the deviations of the estimator from the true value g of the Gini index for finite sample sizes, by using Equations (13.14) and (13.15).



(a) $\alpha = 1.8$ (b) $\alpha = 1.6$

(c) $\alpha = 1.4$ (d) $\alpha = 1.2$

Figure 13.3: *Comparisons between the maximum likelihood and the nonparametric asymptotic distributions for different values of the tail index α. The number of observations for MLE is fixed to $n = 100$. Note that, even if all distributions have mean zero, the mode of the distributions of the nonparametric estimator is different from zero, because of the skewness.*

Figure 13.3 shows how the deviations around the mean of the two different types of estimators are distributed and how these distributions change as the number of observations increases. In particular, to facilitate the comparison between the maximum likelihood and the nonparametric estimators, we fixed the number of observation in the MLE case, while letting them vary in the nonparametric one. We perform this study for different types of tail indices to show how large the impact is on the consistency of the estimator. It is worth noticing that, as the tail index decreases towards 1 (the threshold value for a infinite mean), the mode of the distribution of the nonparametric estimator moves farther away from the

mean of the distribution (centered on 0 by definition, given that we are dealing with deviations from the mean). This effect is responsible for the small sample bias observed in applications. Such a phenomenon is not present in the MLE case, thanks to the the normality of the limit for every value of the tail parameter.

We can make our argument more rigorous by assessing the number of observations $\tilde{n}$ needed for the nonparametric estimator to be as good as the MLE one, under different tail scenarios. Let's consider the likelihood-ratio-type function

$$r(c,n) = \frac{P_S(|D_n^{NP}|> c)}{P_N(|D_{100}^{ML}|> c)}, \tag{13.16}$$

where $P_S(|D_n^{NP}|> c)$ and $P_N(|D_{100}^{ML}|> c)$ are the probabilities (α-stable and Gaussian respectively) of the centered estimators in the nonparametric, and in the MLE cases, of exceeding the thresholds $\pm c$, as per Equations (13.15) and (13.14). In the nonparametric case the number of observations n is allowed to change, while in the MLE case it is fixed to 100. We then look for the value $\tilde{n}$ such that $r(c,\tilde{n}) = 1$ for fixed c.

Table 13.2 displays the results for different thresholds c and tail parameters α. In particular, we can see how the MLE estimator outperforms the nonparametric one, which requires a much larger number of observations to obtain the same tail probability of the MLE with n fixed to 100. For example, we need at least 80×10^6 observations for the nonparametric estimator to obtain the same probability of exceeding the ± 0.02 threshold of the MLE one, when $\alpha = 1.2$.

Table 13.2: *The number of observations $\tilde{n}$ needed for the nonparametric estimator to match the tail probabilities, for different threshold values c and different values of the tail index α, of the maximum likelihood estimator with fixed $n = 100$.*

	Threshold c as per Equation (13.16):			
α	0.005	0.01	0.015	0.02
1.8	27×10^3	12×10^5	12×10^6	63×10^5
1.5	21×10^4	21×10^4	46×10^5	81×10^7
1.2	33×10^8	67×10^7	20×10^7	80×10^6

Interestingly, the number of observations needed to match the tail probabilities in Equation (13.16) does not vary uniformly with the threshold. This is expected, since as the threshold goes to infinity or to zero, the tail probabilities remain the same for every value of n. Therefore, given the unimodality of the limit distributions, we expect that there will be a threshold maximizing the number of observations needed to match the tail probabilities, while for all the other levels the number of observations will be smaller.

We conclude that, when in presence of fat-tailed data with infinite variance, a plug-in MLE based estimator should be preferred over the nonparametric one.

13.5 SMALL SAMPLE CORRECTION

Theorem 2 can be also used to provide a correction for the bias of the nonparametric estimator for small sample sizes. The key idea is to recognize that, for unimodal distributions, most observations come from around the mode. In symmetric distributions the mode and the mean coincide, thus most observations will be close to the mean value as well, not so for skewed distributions: for right-skewed continuous unimodal distributions the mode is lower than the mean. Therefore, given that the asymptotic distribution of the nonparametric Gini index is right-skewed, we expect that the observed value of the Gini index will be usually lower than the true one (placed at the mean level). We can quantify this difference (i.e. the bias) by looking at the distance between the mode and the mean, and once this distance is known, we can correct our Gini estimate by adding it back[4].

Formally, we aim to derive a corrected nonparametric estimator $G^C(X_n)$ such that

$$G^C(X_n) = G^{NP}(X_n) + ||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||, \tag{13.17}$$

where $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ is the distance between the mode m and the mean of the distribution of the nonparametric Gini estimator $G^{NP}(X_n)$.

Performing the type of correction described in Equation (13.17) is equivalent to shifting the distribution of $G^{NP}(X_n)$ in order to place its mode on the true value of the Gini index.

Ideally, we would like to measure this mode-mean distance $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ on the exact distribution of the Gini index to get the most accurate correction. However, the finite distribution is not always easily derivable as it requires assumptions on the parametric structure of the data generating process (which, in most cases, is unknown for fat-tailed data [142]). We therefore propose to use the limiting distribution for the nonparametric Gini obtained in Section 13.2 to approximate the finite sample distribution, and to estimate the mode-mean distance with it. This procedure allows for more freedom in the modeling assumptions and potentially decreases the number of parameters to be estimated, given that the limiting distribution only depends on the tail index and the mean of the data, which can be usually assumed to be a function of the tail index itself, as in the Paretian case where $\mu = \frac{\alpha}{\alpha-1}$.

By exploiting the location-scale property of α-stable distributions and Equation (13.11), we approximate the distribution of $G^{NP}(X_n)$ for finite samples by

$$G^{NP}(X_n) \sim S\left(\alpha, 1, \gamma(n), g\right), \tag{13.18}$$

where $\gamma(n) = \frac{1}{n^{\frac{\alpha-1}{\alpha}}} \frac{L_0(n)}{\mu}$ is the scale parameter of the limiting distribution.

As a consequence, thanks to the linearity of the mode for α-stable distributions, we have

$$||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))|| \approx ||m(\alpha, \gamma(n)) + g - g|| = ||m(\alpha, \gamma(n))||,$$

4 Another idea, which we have tested in writing the paper, is to use the distance between the median and the mean; the performances are comparable.

where $m(\alpha, \gamma(n))$ is the mode function of an α-stable distribution with zero mean.

The implication is that, in order to obtain the correction term, knowledge of the true Gini index is not necessary, given that $m(\alpha, \gamma(n))$ does not depend on g. We then estimate the correction term as

$$\hat{m}(\alpha, \gamma(n)) = \arg\max_x s(x), \tag{13.19}$$

where $s(x)$ is the numerical density of the associated α-stable distribution in Equation (13.18), but centered on 0. This comes from the fact that, for α-stable distributions, the mode is not available in closed form, but it can be easily computed numerically [179], using the unimodality of the law.

The corrected nonparametric estimator is thus

$$G^C(X_n) = G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)), \tag{13.20}$$

whose asymptotic distribution is

$$G^C(X_n) \sim S\left(\alpha, 1, \gamma(n), g + \hat{m}(\alpha, \gamma(n))\right). \tag{13.21}$$

Note that the correction term $\hat{m}\left(\alpha, \gamma(n)\right)$ is a function of the tail index α and is connected to the sample size n by the scale parameter $\gamma(n)$ of the associated limiting distribution. It is important to point out that $\hat{m}(\alpha, \gamma(n))$ is decreasing in n, and that $\lim_{n\to\infty} \hat{m}(\alpha, \gamma(n)) \to 0$. This happens because, as n increases, the distribution described in Equation (13.18) becomes more and more centered around its mean value, shrinking to zero the distance between the mode and the mean. This ensures the asymptotic equivalence of the corrected estimator and the nonparametric one. Just observe that

$$\begin{aligned} \lim_{n\to\infty} |G(X_n)^C - G^{NP}(X_n)| &= \lim_{n\to\infty} |G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)) - G^{NP}(X_n)| \\ &= \lim_{n\to\infty} |\hat{m}(\alpha, \gamma(n))| \to 0. \end{aligned}$$

Naturally, thanks to the correction, $G^C(X_n)$ will always behave better in small samples. Consider also that, from Equation (13.21), the distribution of the corrected estimator has now for mean $g + \hat{m}(\alpha, \gamma(n))$, which converges to the true Gini g as $n \to \infty$.

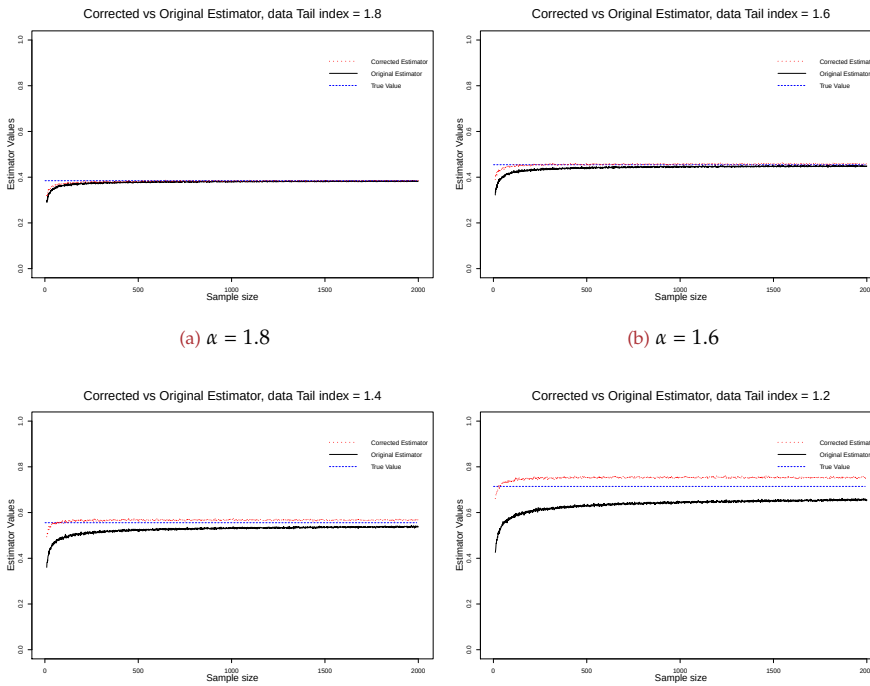
From a theoretical point of view, the quality of this correction depends on the distance between the exact distribution of $G^{NP}(X_n)$ and its α-stable limit; the closer the two are to each other, the better the approximation. However, given that, in most cases, the exact distribution of $G^{NP}(X_n)$ is unknown, it is not possible to give more details.

From what we have written so far, it is clear that the correction term depends on the tail index of the data, and possibly also on their mean. These parameters, if not assumed to be known a priori, must be estimated. Therefore the additional uncertainty due to the estimation will reflect also on the quality of the correction.

We conclude this Section with the discussion of the effect of the correction procedure with a simple example. In a Monte Carlo experiment, we simulate 1000

Paretian samples of increasing size, from $n = 10$ to $n = 2000$, and for each sample size we compute both the original nonparametric estimator $G^{NP}(X_n)$ and the corrected $G^{C}(X_n)$. We repeat the experiment for different α's. Figure 13.4 presents the results.

It is clear that the corrected estimators always perform better than the uncorrected ones in terms of absolute deviation from the true Gini value. In particular, our numerical experiment shows that for small sample sizes with $n \leq 1000$ the gain is quite remarkable for all the different values of $\alpha \in (1, 2)$. However, as expected, the difference between the estimators decreases with the sample size, as the correction term decreases both in n and in the tail index α. Notice that, when the tail index equals 2, we obtain the symmetric Gaussian distribution and the two estimators coincide, given that, thanks to the finiteness of the variance, the nonparametric estimator is no longer biased.



(a) $\alpha = 1.8$ (b) $\alpha = 1.6$

(c) $\alpha = 1.4$ (d) $\alpha = 1.2$

Figure 13.4: *Comparisons between the corrected nonparametric estimator (in red, the one on top) and the usual nonparametric estimator (in black, the one below). For small sample sizes the corrected one clearly improves the quality of the estimation.*

13.6 CONCLUSIONS

In this chapter we address the issue of the asymptotic behavior of the nonparametric estimator of the Gini index in presence of a distribution with infinite variance, an issue that has been curiously ignored by the literature. The central mistake in the nonparametric methods largely used is to believe that asymptotic consistency translates into equivalent pre-asymptotic properties.

We show that a parametric approach provides better asymptotic results thanks to the properties of maximum likelihood estimation. Hence we strongly suggest that, if the collected data are suspected to be fat-tailed, parametric methods should be preferred.

In situations where a fully parametric approach cannot be used, we propose a simple correction mechanism for the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution. Even if the correction works nicely, we suggest caution in its use owing to additional uncertainty from the estimation of the correction term.

TECHNICAL APPENDIX

Proof of Lemma 13.1

Let $U = F(X)$ be the standard uniformly distributed integral probability transform of the random variable X. For the order statistics, we then have [**?**]: $X_{(i)} \overset{a.s.}{=} F^{-1}(U_{(i)})$. Hence

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(i/n - U_{(i)})F^{-1}(U_{(i)}). \tag{13.22}$$

Now by definition of empirical c.d.f it follows that

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)}), \tag{13.23}$$

where $F_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{U_i \leq u}$ is the empirical c.d.f of uniformly distributed random variables.

To show that $R_n \xrightarrow{L^1} 0$, we are going to impose an upper bound that goes to zero. First we notice that

$$\mathbb{E}|R_n| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)})|. \tag{13.24}$$

To build a bound for the right-hand side (r.h.s) of (13.24), we can exploit the fact that, while $F^{-1}(U_{(i)})$ might be just L^1-integrable, $F_n(U_{(i)}) - U_{(i)}$ is L^∞ integrable, therefore we can use Hölder's inequality with $q = \infty$ and $p = 1$. It follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)})| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|\mathbb{E}|F^{-1}(U_{(i)})|. \tag{13.25}$$

Then, thanks to the Cauchy-Schwarz inequality, we get

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|\mathbb{E}|F^{-1}(U_{(i)})| \\ &\leq \left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|)^2\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}(F^{-1}(U_{(i)})))^2\right)^{\frac{1}{2}}.\end{aligned} \tag{13.26}$$

Now, first recall that $\sum_{i=1}^{n}F^{-1}(U_{(i)}) \overset{a.s.}{=} \sum_{i=1}^{n}F^{-1}(U_i)$ with U_i, $i = 1, \ldots, n$, being an i.i.d sequence, then notice that $\mathbb{E}(F^{-1}(U_i)) = \mu$, so that the second term of Equation (13.26) becomes

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|)^2\right)^{\frac{1}{2}}. \tag{13.27}$$

The final step is to show that Equation (13.27) goes to zero as $n \to \infty$.

We know that F_n is the empirical c.d.f of uniform random variables. Using the triangular inequality the inner term of Equation (13.27) can be bounded as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - F(U_{(i)}))|)^2 + \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F(U_{(i)}) - U_{(i)})|)^2.\end{aligned} \tag{13.28}$$

Since we are dealing with uniforms, we known that $F(U) = u$, and the second term in the r.h.s of (13.28) vanishes.

We can then bound $\mathbb{E}(\sup_{U_{(i)}}|(F_n(U_{(i)}) - F(U_{(i)})|)$ using the so called Vapnik-Chervonenkis (VC) inequality, a uniform bound for empirical processes [29, 54, 253], getting

$$\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - F(U_{(i)})| \leq \sqrt{\frac{\log(n+1) + \log(2)}{n}}. \tag{13.29}$$

Combining Equation (13.29) with Equation (13.27) we obtain

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|)^2\right)^{\frac{1}{2}} \leq \mu\sqrt{\frac{\log(n+1) + \log(2)}{n}}, \tag{13.30}$$

which goes to zero as $n \to \infty$, thus proving the first claim.

For the second claim, it is sufficient to observe that the r.h.s of (13.30) still goes to zero when multiplied by $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}$ if $\alpha \in (1,2)$.

Proof of Theorem 1

The first part of the proof consists in showing that we can rewrite Equation (13.10) as a function of i.i.d random variables in place of order statistics, to be able to apply a Central Limit Theorem (CLT) argument.

Let's start by considering the sequence

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(2\frac{i-1}{n-1} - 1\right) F^{-1}(U_{(i)}). \tag{13.31}$$

Using the integral probability transform $X \overset{d}{=} F^{-1}(U)$ with U standard uniform, and adding and removing $\frac{1}{n}\sum_{i=1}^{n}\left(2U_{(i)} - 1\right) F^{-1}(U_{(i)})$, the r.h.s. in Equation (13.31) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \frac{1}{n}\sum_{i=1}^{n}(2U_{(i)} - 1)F^{-1}(U_{(i)}) + \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1} - U_{(i)}\right) F^{-1}(U_{(i)}). \tag{13.32}$$

Then, by using the properties of order statistics [55] we obtain the following almost sure equivalence

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n}(2U_i - 1)F^{-1}(U_i) + \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1} - U_{(i)}\right) F^{-1}(U_{(i)}). \tag{13.33}$$

Note that the first term in the r.h.s of (13.33) is a function of i.i.d random variables as desired, while the second term is just a reminder, therefore

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n,$$

with $Z_i = (2U_i - 1)F^{-1}(U_i)$ and $R_n = \frac{1}{n}\sum_{i=1}^{n}(2(\frac{i-1}{n-1} - U_{(i)}))F^{-1}(U_{(i)})$.

Given Equation (13.10) and exploiting the decomposition given in (13.33) we can rewrite our claim as

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} - \theta\right) = \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta\right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n. \tag{13.34}$$

From the second claim of the Lemma 13.1 and Slutsky Theorem, the convergence in Equation (13.10) can be proven by looking at the behavior of the sequence

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \theta\right), \tag{13.35}$$

where $Z_i = (2U_i - 1)F^{-1}(U_i) = (2F(X_i) - 1)X_i$. This reduces to proving that Z_i is in the fat tails domain of attraction.

Recall that by assumption $X \in DA(S_\alpha)$ with $\alpha \in (1, 2)$. This assumption enables us to use a particular type of CLT argument for the convergence of the sum of fat-tailed random variables. However, we first need to prove that $Z \in DA(S_\alpha)$ as well, that is $P(|Z| > z) \sim L(z)z^{-\alpha}$, with $\alpha \in (1, 2)$ and $L(z)$ slowly-varying.

Notice that

$$P(|\tilde{Z}| > z) \leq P(|Z| > z) \leq P(2X > z),$$

where $\tilde{Z} = (2U - 1)X$ and $U \perp X$. The first bound holds because of the positive dependence between X and $F(X)$ and it can be proven rigorously by noting that $2UX \leq 2F(X)X$ by the so-called re-arrangement inequality [120]. The upper bound conversely is trivial.

Using the properties of slowly-varying functions, we have $P(2X > z) \sim 2^\alpha L(z)z^{-\alpha}$. To show that $\tilde{Z} \in DA(S_\alpha)$, we use the Breiman's Theorem, which ensure the stability of the α-stable class under product, as long as the second random variable is not too fat-tailed [262].

To apply the Theorem we re-write $P(|\tilde{Z}| > z)$ as

$$P(|\tilde{Z}| > z) = P(\tilde{Z} > z) + P(-\tilde{Z} > z) = P(\tilde{U}X > z) + P(-\tilde{U}X > z),$$

where $\tilde{U}$ is a standard uniform with $\tilde{U} \perp X$.

We focus on $P(\tilde{U}X > z)$ since the procedure is the same for $P(-\tilde{U}X > z)$. We have

$$P(\tilde{U}X > z) = P(\tilde{U}X > z|\tilde{U} > 0)P(\tilde{U} > 0) + P(\tilde{U}X > z|\tilde{U} \leq 0)P(\tilde{U} \leq 0),$$

for $z \to +\infty$.

Now, we have that $P(\tilde{U}X > z|\tilde{U} \leq 0) \to 0$, while, by applying Breiman's Theorem, $P(\tilde{U}X > z|\tilde{U} > 0)$ becomes

$$P(\tilde{U}X > z|\tilde{U} > 0) \to E(\tilde{U}^\alpha|U > 0)P(X > z)P(U > 0).$$

Therefore

$$P(|\tilde{Z}| > z) \to \frac{1}{2}E(\tilde{U}^\alpha|U > 0)P(X > z) + \frac{1}{2}E((-\tilde{U})^\alpha|U \leq 0)P(X > z).$$

From this

$$\begin{aligned} P(|\tilde{Z}| > z) &\to \frac{1}{2}P(X > z)[E(\tilde{U})^\alpha|U > 0) + E((-\tilde{U}^\alpha|U \leq 0)] \\ &= \frac{2^\alpha}{1-\alpha}P(X > z) \sim \frac{2^\alpha}{1-\alpha}L(z)z^{-\alpha}. \end{aligned}$$

We can then conclude that, by the squeezing Theorem [90],

$$P(|Z| > z) \sim L(z)z^{-\alpha},$$

as $z \to \infty$. Therefore $Z \in DA(S_\alpha)$.

We are now ready to invoke the Generalized Central Limit Theorem (GCLT) [81] for the sequence Z_i, i.e.

$$nc_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_i)\right) \xrightarrow{d} S_{\alpha,\beta}. \tag{13.36}$$

with $\mathbb{E}(Z_i) = \theta$, $S_{\alpha,\beta}$ a standardized α-stable random variable, and where c_n is a sequence which must satisfy

$$\lim_{n\to\infty} \frac{nL(c_n)}{c_n^\alpha} = \frac{\Gamma(2-\alpha)|\cos(\frac{\pi\alpha}{2})|}{\alpha - 1} = C_\alpha. \tag{13.37}$$

Notice that c_n can be represented as $c_n = n^{\frac{1}{\alpha}} L_0(n)$, where $L_0(n)$ is another slowly-varying function possibly different from $L(n)$.

The skewness parameter β is such that

$$\frac{P(Z > z)}{P(|Z| > z)} \to \frac{1+\beta}{2}.$$

Recalling that, by construction, $Z \in [-c, +\infty)$, the above expression reduces to

$$\frac{P(Z > z)}{P(Z > z) + P(-Z > z)} \to \frac{P(Z > z)}{P(Z > z)} = 1 \to \frac{1+\beta}{2}, \tag{13.38}$$

therefore $\beta = 1$. This, combined with Equation (13.34), the result for the reminder R_n of Lemma 13.1 and Slutsky Theorem, allows us to conclude that the same weak limits holds for the ordered sequence of $Z_{(i)}$ in Equation (13.10) as well.

Proof of Theorem 2

The first step of the proof is to show that the ordered sequence $\frac{\sum_{i=1}^n Z_{(i)}}{\sum_{i=1}^n X_i}$, characterizing the Gini index, is equivalent in distribution to the i.i.d sequence $\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i}$. In order to prove this, it is sufficient to apply the factorization in Equation (13.33) to Equation (13.11), getting

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu}\right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \frac{n}{\sum_{i=1}^n X_i}. \tag{13.39}$$

By Lemma 13.1 and the application of the continuous mapping and Slutsky Theorems, the second term in Equation (13.39) goes to zero at least in probability. Therefore to prove the claim it is sufficient to derive a weak limit for the following sequence

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)}\left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu}\right). \tag{13.40}$$

Expanding Equation (13.40) and recalling that $Z_i = (2F(X_i) - 1)X_i$, we get

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{n}{\sum_{i=1}^{n} X_i} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right) \right). \tag{13.41}$$

The term $\frac{n}{\sum_{i=1}^{n} X_i}$ in Equation (13.41) converges in probability to $\frac{1}{\mu}$ by an application of the continuous mapping Theorem, and the fact that we are dealing with positive random variables X. Hence it will contribute to the final limit via Slutsky Theorem.

We first start by focusing on the study of the limit law of the term

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^{n} X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right). \tag{13.42}$$

Set $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ and note that $\mathbb{E}(\hat{Z}_i) = 0$, since $\mathbb{E}(Z_i) = \theta$ and $\mathbb{E}(X_i) = \mu$.

In order to apply a GCLT argument to characterize the limit distribution of the sequence $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i$ we need to prove that $\hat{Z} \in DA(S_\alpha)$. If so then we can apply GCLT to

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{\sum_{i=1}^{n} \hat{Z}_i}{n} - \mathbb{E}(\hat{Z}_i) \right). \tag{13.43}$$

Note that, since $\mathbb{E}(\hat{Z}_i) = 0$, Equation (13.43) equals Equation (13.42).

To prove that $\hat{Z} \in DA(S_\alpha)$, remember that $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ is just $Z_i = X_i(2F(X_i) - 1)$ shifted by $\frac{\theta}{\mu}$. Therefore the same argument used in Theorem 1 for Z applies here to show that $\hat{Z} \in DA(S_\alpha)$. In particular we can point out that $\hat{Z}$ and Z (therefore also X) share the same α and slowly-varying function $L(n)$.

Notice that by assumption $X \in [c, \infty)$ with $c > 0$ and we are dealing with continuous distributions, therefore $\hat{Z} \in [-c(1 + \frac{\theta}{\mu}), \infty)$. As a consequence the left tail of $\hat{Z}$ does not contribute to changing the limit skewness parameter β, which remains equal to 1 (as for Z) by an application of Equation (13.38).

Therefore, by applying the GCLT we finally get

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)} \left(\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i} - \frac{\theta}{\mu} \right) \xrightarrow{d} \frac{1}{\mu} S(\alpha, 1, 1, 0). \tag{13.44}$$

We conclude the proof by noting that, as proven in Equation (13.39), the weak limit of the Gini index is characterized by the i.i.d sequence of $\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i}$ rather than the ordered one, and that an α-stable random variable is closed under scaling by a constant [206].

14 | ON THE SUPER-ADDITIVITY AND ESTIMATION BIASES OF QUANTILE CONTRIBUTIONS ‡

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

a With R. Douady

14.1 INTRODUCTION

Vilfredo Pareto noticed that 80% of the land in Italy belonged to 20% of the population, and vice-versa, thus both giving birth to the power law class of distributions and the popular saying 80/20. The self-similarity at the core of the property of power laws [160] and [161] allows us to recurse and reapply the 80/20 to the remaining 20%, and so forth until one obtains the result that the top percent of the population will own about 53% of the total wealth.

It looks like such a measure of concentration can be seriously biased, depending on how it is measured, so it is very likely that the true ratio of concentration of

‡ Research chapter.

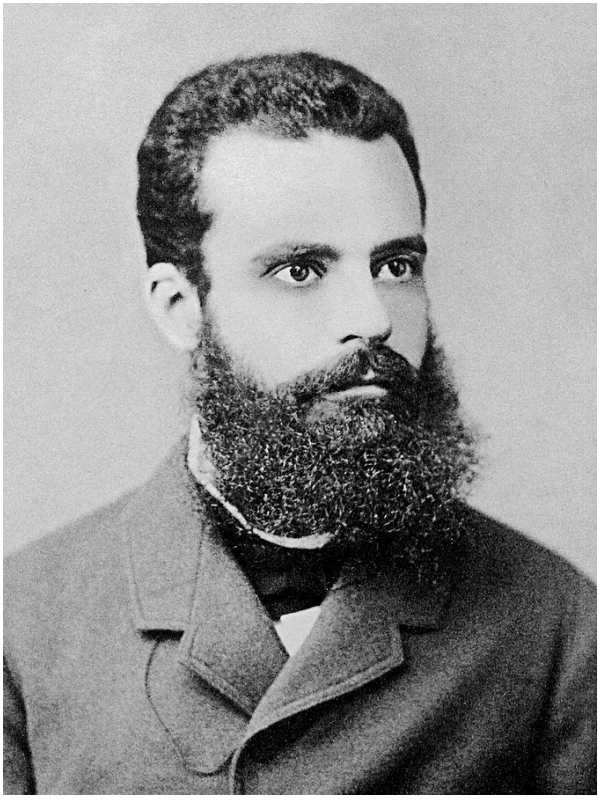


Figure 14.1: *The young Vilfredo Pareto, before he discovered power laws.*

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

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

Let us define the *quantile contribution*

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

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

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

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

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

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

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

14.2 ESTIMATION FOR UNMIXED PARETO-TAILED DISTRIBUTIONS

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

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{14.1}$$

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

There is little difference for small exceedance quantiles (<50%) between the various possible distributions such as Student's t, Lévy α-stable, Dagum,[52],[53] Singh-Maddala distribution [210], or straight Pareto.

For exponents $1 \leq \alpha \leq 2$, as observed in [232] (Chapter 8 in this book), the law of large numbers operates, though *extremely* slowly. The problem is acute for α around, but strictly above 1 and severe, as it diverges, for $\alpha = 1$.

14.2.1 Bias and Convergence

Simple Pareto Distribution Let us first consider $\phi_\alpha(x)$ the density of a α-Pareto distribution bounded from below by $x_{\min} > 0$, in other words: $\phi_\alpha(x) = \alpha x_{\min}^{\alpha} x^{-\alpha-1} \mathbb{1}_{x\geq x_{\min}}$, and $\mathbb{P}(X > x) = \left(\frac{x_{\min}}{x}\right)^{\alpha}$. Under these assumptions, the cutpoint of exceedance is $h(q) = x_{\min}\, q^{-1/\alpha}$ and we have:

$$\kappa_q = \frac{\int_{h(q)}^{\infty} x\,\phi(x)dx}{\int_{x_{min}}^{\infty} x\,\phi(x)dx} = \left(\frac{h(q)}{x_{\min}}\right)^{1-\alpha} = q^{\frac{\alpha-1}{\alpha}} \tag{14.2}$$

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

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

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

Table 14.1 shows the bias of $\widehat{\kappa}_q$ as an estimator of κ_q in the case of an α-Pareto distribution for $\alpha = 1.1$, a value chosen to be compatible with practical economic measures, such as the wealth distribution in the world or in a particular country, including developped ones.[2] In such a case, the estimator is extemely sensitive to *"small"* samples, *"small"* meaning in practice 10^8. We ran up to a trillion simulations across varieties of sample sizes. While $\kappa_{0.01} \approx 0.657933$, even a sample size of 100 million remains severely biased as seen in the table.

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

Table 14.1: *Biases of Estimator of* $\kappa = 0.657933$ *From* 10^{12} *Monte Carlo Realizations*

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

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

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

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

2 This value, which is lower than the estimated exponents one can find in the literature – around 2 – is, following [85], a lower estimate which cannot be excluded from the observations.

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

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

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

Proposition 14.1
Let $\boldsymbol{X} = (X)_{i=1}^{n}$ a random sample of size $n > \frac{1}{q}$, $Y = X_{n+1}$ an extra single random observation, and define: $\widehat{\kappa}_h(\boldsymbol{X} \sqcup Y) = \dfrac{\sum_{i=1}^{n} \mathbb{1}_{X_i>h} X_i + \mathbb{1}_{Y>h} Y}{\sum_{i=1}^{n} X_i + Y}$. We remark that, whenever $Y > h$, one has:

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

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

We face a different situation from the common small sample effect resulting from high impact from the rare observation in the tails that are less likely to show up in small samples, a bias which goes away by repetition of sample runs. The concavity of the estimator constitutes a upper bound for the measurement in finite n, clipping large deviations, which leads to problems of aggregation as we will state below in Theorem 1.

In practice, even in very large sample, the contribution of very large rare events to κ_q slows down the convergence of the sample estimator to the true value. For a better, unbiased estimate, one would need to use a different path: first estimating the distribution parameters $\left(\hat{\alpha}, \hat{\lambda}\right)$ and only then, estimating the theoretical tail contribution $\kappa_q(\hat{\alpha}, \hat{\lambda})$. Falk [85] observes that, even with a proper estimator of α and λ, the convergence is extremely slow, namely of the order of $n^{-\delta}/\ln n$, where

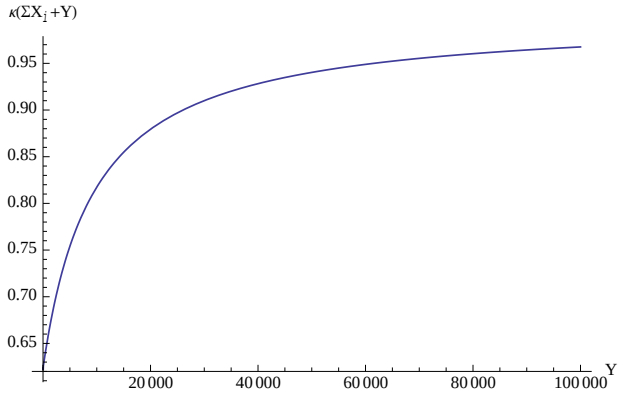


Figure 14.2: *Effect of additional observations on κ*

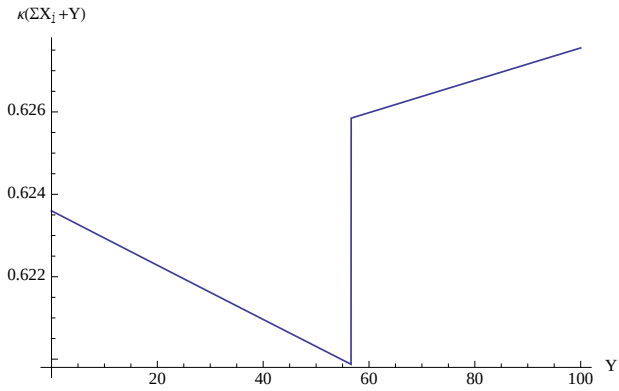


Figure 14.3: *Effect of additional observations on κ, we can see convexity on both sides of h except for values of no effect to the left of h, an area of order* $1/n$

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

14.3 AN INEQUALITY ABOUT AGGREGATING INEQUALITY

For the estimation of the mean of a fat-tailed r.v. $(X)_i^j$, in m sub-samples of size n_i each for a total of $n = \sum_{i=1}^{m} n_i$, the allocation of the total number of observations n between i and j does not matter so long as the total n is unchanged. Here the allocation of n samples between m sub-samples does matter because of the concavity of κ.[3] Next we prove that global concentration as measured by $\widehat{\kappa}_q$ on a broad set of data will appear higher than local concentration, so aggregating European data, for instance, would give a $\widehat{\kappa}_q$ higher than the average measure of concentration across countries – an *"inequality about inequality"*. In other words, we claim that the estimation bias when using $\widehat{\kappa}_q(n)$ is even increased when dividing

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

the sample into sub-samples and taking the weighted average of the measured values $\widehat{\kappa}_q(n_i)$.

Theorem 3

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

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

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

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

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

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

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

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

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

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

We observe that:

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

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

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

Let us now show that:

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

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

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

hence we will have:

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

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

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

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

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

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

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

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

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

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

This ends the proof of the theorem. □

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

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

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

Theorem 4

For any $n \in \mathbb{N}$, we have:

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

and

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

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

14.4 MIXED DISTRIBUTIONS FOR THE TAIL EXPONENT

Consider now a random variable X, the distribution of which $p(dx)$ is a mixture of parametric distributions with different values of the parameter: $p(dx) =$

Figure 14.4: *Pierre Simon, Marquis de Laplace. He got his name on a distribution and a few results, but was behind both the Cauchy and the Gaussian distributions (see the Stigler law of eponymy [215]). Posthumous portrait by Jean-Baptiste Paulin Guérin, 1838.*

$\sum_{i=1}^{m} \omega_i p_{\alpha_i}(dx)$. A typical n-sample of X can be made of $n_i = \omega_i n$ samples of X_{α_i} with distribution p_{α_i}. The above theorem shows that, in this case, we have:

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

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

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

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

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

and

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

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

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

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

Suppose now that X is a positive random variable with unknown distribution, except that its tail decays like a power low with unknown exponent. An unbiased estimation of the exponent, with necessarily some amount of uncertainty (i.e., a distribution of possible true values around some average), would lead to a downward biased estimate of κ_q.

Because the concentration measure only depends on the tail of the distribution, this inequality also applies in the case of a mixture of distributions with a power decay, as in Equation 23.1:

$$\mathbb{P}(X > x) \sim \sum_{j=1}^{N} \omega_i L_i(x) x^{-\alpha_j} \tag{14.3}$$

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

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

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

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

14.5 A LARGER TOTAL SUM IS ACCOMPANIED BY INCREASES IN $\widehat{\kappa}_q$

There is a large dependence between the estimator $\widehat{\kappa}_q$ and the sum $S = \sum_{j=1}^{n} X_j$: conditional on an increase in $\widehat{\kappa}_q$ the expected sum is larger. Indeed, as shown in theorem 3, $\widehat{\kappa}_q$ and S are positively correlated.

For the case in which the random variables under concern are wealth, we observe as in Figure 14.5 such conditional increase; in other words, since the distribution is of the class of thick tails under consideration, the maximum is of the same order as the sum, additional wealth means more measured inequality. Under such dynamics, is quite absurd to assume that additional wealth will arise from the bottom or even the middle. (The same argument can be applied to wars, epidemics, size or companies, etc.)

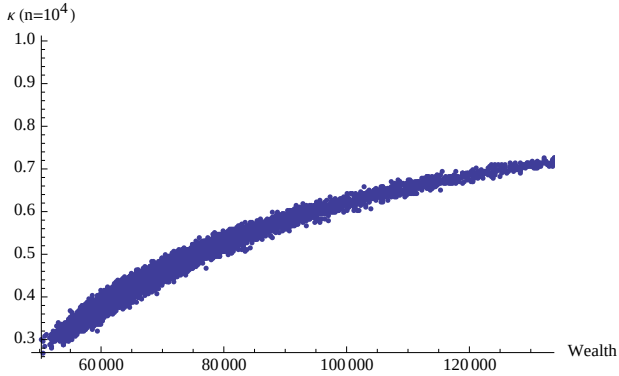


Figure 14.5: *Effect of additional wealth on* $\hat{\kappa}$

14.6 CONCLUSION AND PROPER ESTIMATION OF CONCENTRATION

Concentration can be high at the level of the generator, but in small units or subsections we will observe a lower κ_q. So examining times series, we can easily get a historical illusion of rise in, say, wealth concentration when it has been there all along at the level of the process; and an expansion in the size of the unit measured can be part of the explanation.[4]

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

One must not perform analyses of year-on-year changes in $\widehat{\kappa}_q$ without adjustment. It did not escape our attention that some theories are built based on claims of such "increase" in inequality, as in [188], without taking into account the true nature of

4 Accumulated wealth is typically thicker tailed than income, see [98].

κ_q, and promulgating theories about the "variation" of inequality without reference to the stochasticity of the estimation – and the lack of consistency of κ_q across time and sub-units. What is worse, rejection of such theories also ignored the size effect, by countering with data of a different sample size, effectively making the dialogue on inequality uninformational statistically.[5]

The mistake appears to be commonly made in common inference about fat-tailed data in the literature. The very methodology of using concentration and changes in concentration is highly questionable. For instance, in the thesis by Steven Pinker [191] that the world is becoming less violent, we note a fallacious inference about the concentration of damage from wars from a $\widehat{\kappa}_q$ with minutely small population in relation to the fat-tailedness.[6] Owing to the fat-tailedness of war casualties and consequences of violent conflicts, an adjustment would rapidly invalidate such claims that violence from war has statistically experienced a decline.

14.6.1 Robust methods and use of exhaustive data

We often face argument of the type "the method of measuring concentration from quantile contributions $\hat{\kappa}$ is robust and based on a complete set of data". Robust methods, alas, tend to fail with fat-tailed data, see Chapter 8. But, in addition, the problem here is worse: even if such "robust" methods were deemed unbiased, a method of direct centile estimation is still linked to a static and specific population and does not aggregegage. Accordingly, such techniques do not allow us to make statistical claims or scientific statements about the true properties which should necessarily carry out of sample.

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

14.6.2 How Should We Measure Concentration?

Practitioners of risk managers now tend to compute CVaR and other metrics, methods that are extrapolative and nonconcave, such as the information from the α exponent, taking the one closer to the lower bound of the range of exponents, as we

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

6 Using Richardson's data, [191]: "(Wars) followed an 80:2 rule: almost eighty percent of the deaths were caused by *two percent* (his emph.) of the wars". So it appears that both Pinker and the literature cited for the quantitative properties of violent conflicts are using a flawed methodology, one that produces a severe bias, as the centile estimation has extremely large biases with fat-tailed wars. Furthermore claims about the mean become spurious at low exponents.

saw in our extension to Theorem 2 and rederiving the corresponding κ, or, more rigorously, integrating the functions of α across the various possible states. Such methods of adjustment are less biased and do not get mixed up with problems of aggregation –they are similar to the "stochastic volatility" methods in mathematical finance that consist in adjustments to option prices by adding a "smile" to the standard deviation, in proportion to the variability of the parameter representing volatility and the errors in its measurement. Here it would be "stochastic alpha" or "stochastic tail exponent "[7] By extrapolative, we mean the built-in extension of the tail in the measurement by taking into account realizations outside the sample path that are in excess of the extrema observed.[8] [9]

ACKNOWLEDGMENT

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

7 Also note that, in addition to the centile estimation problem, some authors such as [189] when dealing with censored data, use Pareto interpolation for unsufficient information about the tails (based on tail parameter), filling-in the bracket with conditional average bracket contribution, which is not the same thing as using full power law extension; such a method retains a significant bias.

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

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

Part V

SHADOW MOMENTS PAPERS

15 ON THE SHADOW MOMENTS OF APPARENTLY INFINITE-MEAN PHENOMENA (WITH P. CIRILLO)‡

This Chapter proposes an approach to compute the conditional moments of fat-tailed phenomena that, only looking at data, could be mistakenly considered as having infinite mean. This type of problems manifests itself when a random variable Y has a heavy-tailed distribution with an extremely wide yet bounded support.

We introduce the concept of dual distribution, by means of a log-transformation that smoothly removes the upper bound. The tail of the dual distribution can then be studied using extreme value theory, without making excessive parametric assumptions, and the estimates one obtains can be used to study the original distribution and compute its moments by reverting the transformation.

The central difference between our approach and a simple truncation is in the smoothness of the transformation between the original and the dual distribution, allowing use of extreme value theory.

War casualties, operational risk, environment blight, complex networks and many other econophysics phenomena are possible fields of application.

15.1 INTRODUCTION

Consider a heavy-tailed random variable Y with finite support $[L, H]$. W.l.o.g. set $L >> 0$ for the lower bound, while for upper one H, assume that its value is remarkably large, yet finite. It is so large that the probability of observing values in its vicinity is extremely small, so that in data we tend to find observations only below a certain $M << H < \infty$.

‡ Research chapter.

Figure 15.1 gives a graphical representation of the problem. For our random variable Y with remote upper bound H the real tail is represented by the continuous line. However, if we only observe values up to $M << H$, and - willing or not - we ignore the existence of H, which is unlikely to be seen, we could be inclined to believe the the tail is the dotted one, the apparent one. The two tails are indeed essentially indistinguishable for most cases, as the divergence is only evident when we approach H.

Now assume we want to study the tail of Y and, since it is fat-tailed and despite $H < \infty$, we take it to belong to the so-called Fréchet class[2]. In extreme value theory [181], a distribution F of a random variable Y is said to be in the Fréchet class if $\bar{F}(y) = 1 - F(y) = y^{-\alpha}L(y)$, where $L(y)$ is a slowly varying function . In other terms, the Fréchet class is the class of all distributions whose right tail behaves as a power law.

Looking at the data, we could be led to believe that the right tail is the dotted line in Figure 15.1, and our estimation of α shows it be smaller than 1. Given the properties of power laws, this means that $E[Y]$ is not finite (as all the other higher moments). This also implies that the sample mean is essentially useless for making inference, in addition to any considerations about robustness [166]. But if H is finite, this cannot be true: all the moments of a random variable with bounded support are finite.

A solution to this situation could be to fit a parametric model, which allows for fat tails and bounded support, such as for example a truncated Pareto [1]. But what happens if Y only shows a Paretian behavior in the upper tail, and not for the whole distribution? Should we fit a mixture model?

In the next section we propose a simple general solution, which does not rely on strong parametric assumptions.

15.2 THE DUAL DISTRIBUTION

Instead of altering the tails of the distribution we find it more convenient to transform the data and rely on distributions with well known properties. In Figure 15.1, the real and the apparent tails are indistinguishable to a great extent. We can use this fact to our advantage, by transforming Y to remove its upper bound H, so that the new random variable Z - the dual random variable - has the same tail as the apparent tail. We can then estimate the shape parameter α of the tail of Z and come back to Y to compute its moments or, to be more exact, to compute its excess moments, the conditional moments above a given threshold, view that we will just extract the information from the tail of Z.

Take Y with support $[L, H]$, and define the function

$$\varphi(Y) = L - H \log\left(\frac{H - Y}{H - L}\right). \tag{15.1}$$

2 Note that treating Y as belonging to the Fréchet class is a mistake. If a random variable has a finite upper bound, it cannot belong to the Fréchet class, but rather to the Weibull class [115].

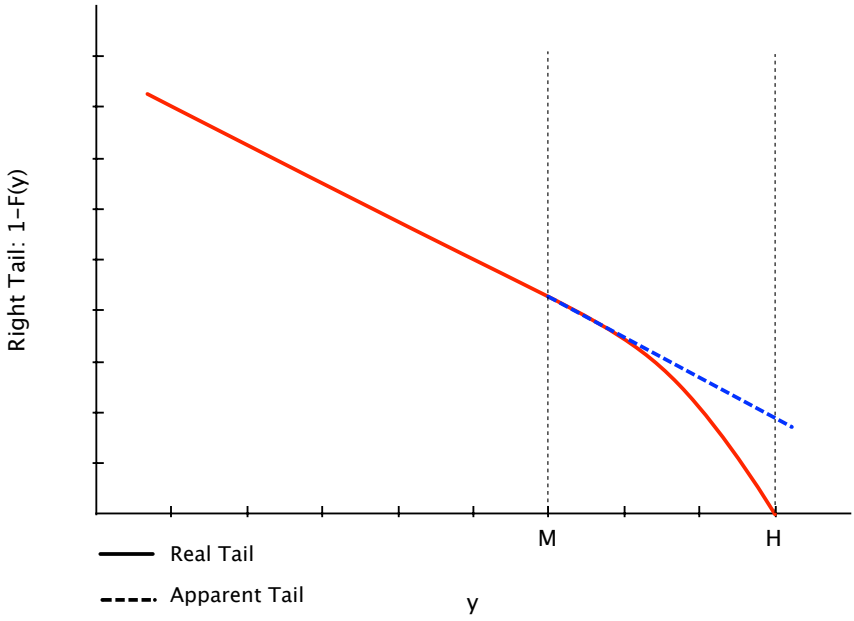


Figure 15.1: *Graphical representation of what may happen if one ignores the existence of the finite upper bound H, since only M is observed.*

We can verify that φ is "smooth": $\varphi \in C^{\infty}$, $\varphi^{-1}(\infty) = H$, and $\varphi^{-1}(L) = \varphi(L) = L$. Then $Z = \varphi(Y)$ defines a new random variable with lower bound L and an infinite upper bound. Notice that the transformation induced by $\varphi(\cdot)$ does not depend on any of the parameters of the distribution of Y.

By construction, $z = \varphi(y) \approx y$ for very large values of H. This means that for a very large upper bound, unlikely to be touched, the results we get for the tail of Y and $Z = \varphi(Y)$ are essentially the same, until we do not reach H. But while Y is bounded, Z is not. Therefore we can safely model the unbounded dual distribution of Z as belonging to the Fréchet class, study its tail, and then come back to Y and its moments, which under the dual distribution of Z could not exist.[3]

The tail of Z can be studied in different ways, see for instance [181] and [86]. Our suggestions is to rely on the so-called de Pickands, Balkema and de Haan's Theorem [115]. This theorem allows us to focus on the right tail of a distribution, without caring too much about what happens below a given threshold threshold u. In our case $u \geq L$.

3 Note that the use of logarithmic transformation is quite natural in the context of utility.

Consider a random variable Z with distribution function G, and call G_u the conditional df of Z above a given threshold u. We can then define the r.v. W, representing the rescaled excesses of Z over the threshold u, so that

$$G_u(w) = P(Z - u \leq w | Z > u) = \frac{G(u+w) - G(u)}{1 - G(u)},$$

for $0 \leq w \leq z_G - u$, where z_G is the right endpoint of G.

Pickands, Balkema and de Haan have showed that for a large class of distribution functions G, and a large u, G_u can be approximated by a Generalized Pareto distribution, i.e. $G_u(w) \to GPD(w; \xi, \sigma)$, as $u \to \infty$ where

$$GPD(w; \xi, \sigma) = \begin{cases} 1 - (1 + \xi \frac{w}{\sigma})^{-1/\xi} & if\ \xi \neq 0 \\ 1 - e^{-\frac{w}{\sigma}} & if\ \xi = 0 \end{cases}, \quad w \geq 0. \tag{15.2}$$

The parameter ξ, known as the shape parameter, and corresponding to $1/\alpha$, governs the fatness of the tails, and thus the existence of moments. The moment of order p of a Generalized Pareto distributed random variable only exists if and only if $\xi < 1/p$, or $\alpha > p$ [181]. Both ξ and σ can be estimated using MLE or the method of moments [115].[4]

15.3 BACK TO y: THE SHADOW MEAN (OR POPULATION MEAN)

With f and g, we indicate the densities of Y and Z.

We know that $Z = \varphi(Y)$, so that $Y = \varphi^{-1}(Z) = (L - H)e^{\frac{L-Z}{H}} + H$.

Now, let's assume we found $u = L^* \geq L$, such that $G_u(w) \approx \text{GPD}(w; \xi, \sigma)$. This implies that the tail of Y, above the same value L^* that we find for Z, can be obtained from the tail of Z, i.e. G_u.

First we have

$$\int_{L^*}^{\infty} g(z)\,dz = \int_{L^*}^{\varphi^{-1}(\infty)} f(y)\,dy. \tag{15.3}$$

And we know that

$$g(z; \xi, \sigma) = \frac{1}{\sigma}\left(1 + \frac{\xi z}{\sigma}\right)^{-\frac{1}{\xi}-1}, \qquad z \in [L^*, \infty). \tag{15.4}$$

Setting $\alpha = \xi^{-1}$, we get

$$f(y; \alpha, \sigma) = \frac{H\left(1 + \frac{H(\log(H-L) - \log(H-y))}{\alpha\sigma}\right)^{-\alpha-1}}{\sigma(H-y)}, \quad y \in [L^*, H], \tag{15.5}$$

4 There are alternative methods to face finite (or concave) upper bounds, i.e., the use of tempered power laws (with exponential dampening)[194] or stretched exponentials [147]; while being of the same nature as our exercise, these methods do not allow for immediate applications of extreme value theory or similar methods for parametrization.

Figure 15.2: *C.F. Gauss, painted by Christian Albrecht Jensen. Gauss has his name on the distribution, generally attributed to Laplace.*

or, in terms of distribution function,

$$F(y;\alpha,\sigma) = 1 - \left(1 + \frac{H(\log(H-L) - \log(H-y))}{\alpha\sigma}\right)^{-\alpha}. \tag{15.6}$$

Clearly, given that φ is a one-to-one transformation, the parameters of f and g obtained by maximum likelihood methods will be the same —the likelihood functions of f and g differ by a scaling constant.

We can derive the shadow mean[5] of Y, conditionally on $Y > L^*$, as

$$E[Y|Y > L^*] = \int_{L^*}^{H} y\, f(y;\alpha,\sigma)\, \mathrm{d}y, \tag{15.7}$$

5 We call the population average –as opposed to the sample one – "shadow", as it is not immediately visible from the data.

obtaining

$$E[Y|Z > L^*] = (H - L^*)e^{\frac{\alpha\sigma}{H}} \left(\frac{\alpha\sigma}{H}\right)^{\alpha} \Gamma\left(1 - \alpha, \frac{\alpha\sigma}{H}\right) + L^*. \tag{15.8}$$

The conditional mean of Y above $L^* \geq L$ can then be estimated by simply plugging in the estimates $\hat{\alpha}$ and $\hat{\sigma}$, as resulting from the GPD approximation of the tail of Z. It is worth noticing that if $L^* = L$, then $E[Y|Y > L^*] = E[Y]$, i.e. the conditional mean of Y above Y is exactly the mean of Y.

Naturally, in a similar way, we can obtain the other moments, even if we may need numerical methods to compute them.

Our method can be used in general, but it is particularly useful when, from data, the tail of Y appears so fat that no single moment is finite, as it is often the case when dealing with operational risk losses, the degree distribution of large complex networks, or other econophysical phenomena.

For example, assume that for Z we have $\xi > 1$. Then both $E[Z|Z > L^*]$ and $E[Z]$ are not finite[6]. Figure 15.1 tells us that we might be inclined to assume that also $E[Y]$ is infinite - and this is what the data are likely to tell us if we estimate $\hat{\xi}$ from the tail[7] of Y. But this cannot be true because $H < \infty$, and even for $\xi > 1$ we can compute the expected value $E[Y|Z > L^*]$ using equation (15.8).

Value-at-Risk and Expected Shortfall

Thanks to equation (15.6), we can compute by inversion the quantile function of Y when $Y \geq L^*$, that is

$$Q(p; \alpha, \sigma, H, L) = e^{-\gamma(p)} \left(L^* e^{\frac{\alpha\sigma}{H}} + He^{\gamma(p)} - He^{\frac{\alpha\sigma}{H}}\right), \tag{15.9}$$

where $\gamma(p) = \frac{\alpha\sigma(1-p)^{-1/\alpha}}{H}$ and $p \in [0, 1]$. Again, this quantile function is conditional on Y being larger than L^*.

From equation (15.9), we can easily compute the Value-at-Risk (VaR) of $Y|Y \geq L^*$ for whatever confidence level. For example, the 95% VaR of Y, if Y represents operational losses over a 1-year time horizon, is simply $VaR^Y_{0.95} = Q(0.95; \alpha, \sigma, H, L)$.

Another quantity we might be interested in when dealing with the tail risk of Y is the so-called expected shortfall (ES), that is $E[Y|Y > u \geq L^*]$. This is nothing more than a generalization of equation (15.8).

We can obtain the expected shortfall by first computing the mean excess function of $Y|Y \geq L^*$, defined as

$$e_u(Y) = E[Y - u|Y > u] = \frac{\int_u^{\infty}(u - y)f(y; \alpha, \sigma)\mathrm{d}y}{1 - F(u)},$$

6 Remember that for a GPD random variable Z, $E[Z^p] < \infty$ iff $\xi < 1/p$.

7 Because of the similarities between $1 - F(y)$ and $1 - G(z)$, at least up until M, the GPD approximation will give two statistically undistinguishable estimates of ξ for both tails [181].

for $y \geq u \geq L^*$. Using equation (15.5), we get

$$e_u(Y) = (H-L)e^{\frac{\alpha\sigma}{H}} \left(\frac{\alpha\sigma}{H}\right)^{\alpha} \left(\frac{H \log\left(\frac{H-L}{H-u}\right)}{\alpha\sigma} + 1\right)^{\alpha} \times \Gamma\left(1-\alpha, \frac{\alpha\sigma}{H} + \log\left(\frac{H-L}{H-u}\right)\right). \tag{15.10}$$

The Expected Shortfall is then simply computed as

$$E[Y|Y > u \geq L^*] = e_u(Y) + u.$$

As in finance and risk management, ES and VaR can be combined. For example we could be interested in computing the 95% ES of Y when $Y \geq L^*$. This is simply given by $VaR^Y_{0.95} + e_{VaR^Y_{0.95}}(Y)$.

15.4 COMPARISON TO OTHER METHODS

There are three ways to go about explicitly cutting a Paretian distribution in the tails (not counting methods to stretch or "temper" the distribution).

1) The first one consists in hard truncation, i.e. in setting a single endpoint for the distribution and normalizing. For instance the distribution would be normalized between L and H, distributing the excess mass across all points.

2) The second one would assume that H is an absorbing barrier, that all the realizations of the random variable in excess of H would be compressed into a Dirac delta function at H –as practiced in derivative models. In that case the distribution would have the same density as a regular Pareto except at point H.

3) The third is the one presented here.

The same problem has cropped up in quantitative finance over the use of truncated normal (to correct for Bachelier's use of a straight Gaussian) vs. logarithmic transformation (Sprenkle, 1961 [213]), with the standard model opting for logarithmic transformation and the associated one-tailed lognormal distribution. Aside from the additivity of log-returns and other such benefits, the models do not produce a "cliff", that is an abrupt change in density below or above, with the instability associated with risk measurements on non-smooth function.

As to the use of extreme value theory, Breilant et al. (2014)[**?**] go on to truncate the distribution by having an excess in the tails with the transformation $Y^{-\alpha} \rightarrow (Y^{-\alpha} - H^{-\alpha})$ and apply EVT to the result. Given that the transformation includes the estimated parameter, a new MLE for the parameter α is required. We find issues with such a non-smooth transformation. The same problem occurs as with financial asset models, particularly the presence an abrupt "cliff" below which there is a density, and above which there is none. The effect is that the expectation obtained in such a way will be higher than ours, particularly at values of $\alpha < 1$, as seen in Figure **??**.

We can demonstrate the last point as follows. Assume we observe distribution is Pareto that is in fact truncated but treat it as a Pareto. The density is $f(x) = \frac{1}{\sigma}\left(\frac{x-L}{\alpha\sigma}+1\right)^{-\alpha-1}$, $x \in [L,\infty)$. The truncation gives $g(x) = \frac{\left(\frac{x-L}{\alpha\sigma}+1\right)^{-\alpha-1}}{\sigma(1-\alpha^{\alpha}\sigma^{\alpha}(\alpha\sigma+H-L)^{-\alpha})}$, $x \in [L,H]$.

Moments of order p of the truncated Pareto (i.e. what is seen from realizations of the process), $M(p)$ are:

$$M(p) = \alpha e^{-i\pi p}(\alpha\sigma)^{\alpha}(\alpha\sigma - L)^{p-\alpha} \frac{\left(B_{\frac{H}{L-\alpha\sigma}}(p+1,-\alpha) - B_{\frac{L}{L-\alpha\sigma}}(p+1,-\alpha)\right)}{\left(\frac{\alpha\sigma}{\alpha\sigma+H-L}\right)^{\alpha}-1} \tag{15.11}$$

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

We end up with $r(H,\alpha)$, the ratio of the mean of the soft truncated to that of the truncated Pareto.

$$r(H,\alpha) = e^{-\frac{\alpha}{H}}\left(\frac{\alpha}{H}\right)^{\alpha}\left(\frac{\alpha}{\alpha+H}\right)^{-\alpha}\left(\frac{\alpha+H}{\alpha}\right)^{-\alpha} \frac{\left(-\left(\frac{\alpha+H}{\alpha}\right)^{\alpha}+H+1\right)}{(\alpha-1)\left(\left(\frac{\alpha}{H}\right)^{\alpha}-\left(\frac{\alpha+H}{H}\right)^{\alpha}\right)E_{\alpha}\left(\frac{\alpha}{H}\right)} \tag{15.12}$$

where $E_{\alpha}\left(\frac{\alpha}{H}\right)$ is the exponential integral $e_{\alpha}z = \int_1^{\infty}\frac{e^{t(-\alpha)}}{t^n}\,dt$.

15.5 APPLICATIONS

Operational risk The losses for a firm are bounded by the capitalization, with well-known maximum losses.

Capped Reinsurance Contracts Reinsurance contracts almost always have caps (i.e., a maximum claim); but a reinsurer can have many such contracts on the same source of risk and the addition of the contract pushes the upper bound in such a way as to cause larger potential cumulative harm.

Violence While wars are extremely fat-tailed, the maximum effect from any such event cannot exceed the world's population.

Credit risk A loan has a finite maximum loss, in a way similar to reinsurance contracts.

City size While cities have been shown to be Zipf distributed, the size of a given city cannot exceed that of the world's population.

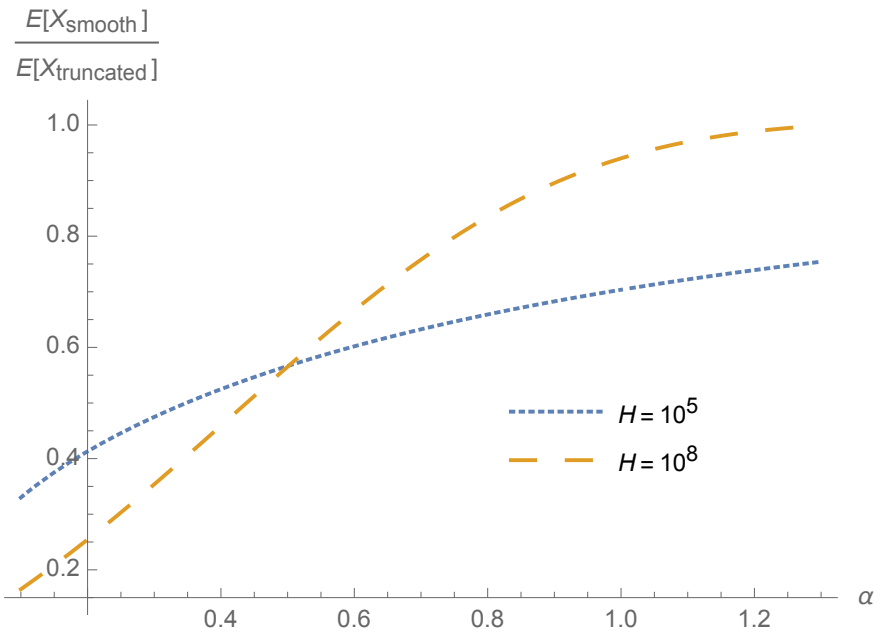


Figure 15.3: *Ratio of the expectation of smooth transformation to truncated.*

Environmental harm While these variables are exceedingly fat-tailed, the risk is confined by the size of the planet (or the continent on which they take place) as a firm upper bound.

Complex networks The number of connections is finite.

Company size The sales of a company is bound by the GDP.

Earthquakes The maximum harm from an earthquake is bound by the energy.

Hydrology The maximum level of a flood can be determined.

16 | ON THE TAIL RISK OF VIOLENT CONFLICT AND ITS UNDERESTIMATION (WITH P. CIRILLO)‡

We examine all possible statistical pictures of violent conflicts over common era history with a focus on dealing with incompleteness and unreliability of data. We apply methods from extreme value theory on log-transformed data to remove compact support, then, owing to the boundedness of maximum casualties, retransform the data and derive expected means. We find the estimated mean likely to be at least three times larger than the sample mean, meaning severe underestimation of the severity of conflicts from naive observation. We check for robustness by sampling between high and low estimates and jackknifing the data. We study inter-arrival times between tail events and find (first-order) memorylessless of events. The statistical pictures obtained are at variance with the claims about "long peace".

16.1 INTRODUCTION/SUMMARY

This study is as much about new statistical methodologies with thick tailed (and unreliable data), as well as bounded random variables with local Power Law behavior, as it is about the properties of violence.[2]

Violence is much more severe than it seems from conventional analyses and the prevailing "long peace" theory which claims that violence has declined. Adapting methods from extreme value theory, and adjusting for errors in reporting of conflicts and historical estimates of casualties, we look at the various statistical pictures of violent conflicts, with focus for the parametrization on those with more than 50k

‡ Research chapter.

2 Acknowledgments: Captain Mark Weisenborn engaged in the thankless and gruesome task of compiling the data, checking across sources and linking each conflict to a narrative on Wikipedia (see Appendix 1). We also benefited from generous help on social networks where we put data for scrutiny, as well as advice from historians thanked in the same appendix. We also thank the late Benoit Mandelbrot for insights on the tail properties of wars and conflicts, as well as Yaneer Bar-Yam, Raphael Douady...

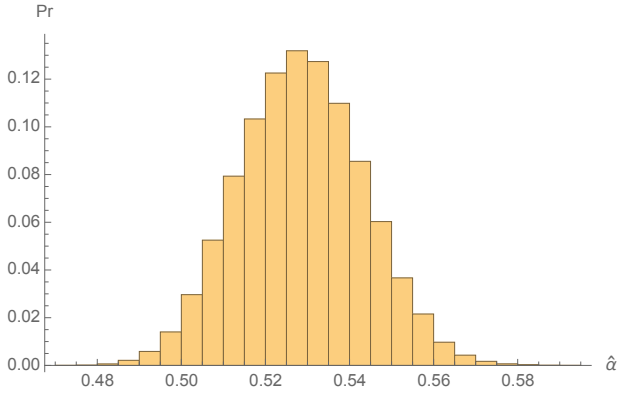


Figure 16.1: *Values of the tail exponent α from Hill estimator obtained across 100,000 different rescaled casualty numbers uniformly selected between low and high estimates of conflict. The exponent is slightly (but not meaningfully) different from the Maximum Likelihood for all data as we focus on top 100 deviations.*

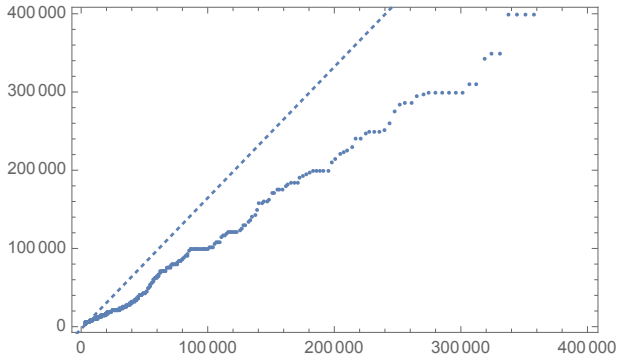


Figure 16.2: *Q-Q plot of the rescaled data in the near-tail plotted against a Pareto II-Lomax Style distribution.*

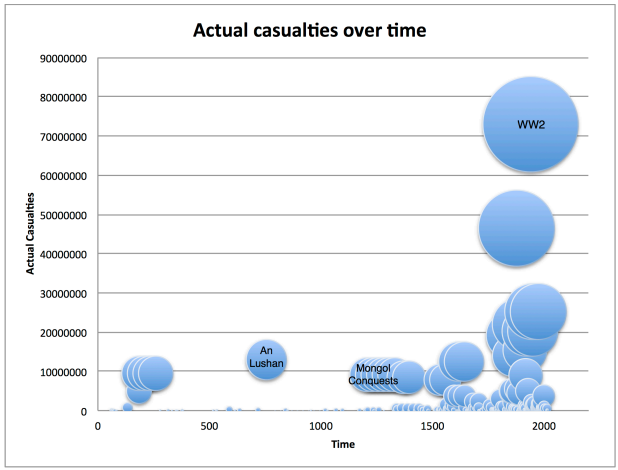


Figure 16.3: *Death toll from "named conflicts" over time. Conflicts lasting more than 25 years are disaggregated into two or more conflicts, each one lasting 25 years.*

victims (in equivalent ratio of today's population, which would correspond to ≈ 5k in the 18^{th} C.). Contrary to current discussions, *all* statistical pictures thus obtained show that 1) the risk of violent conflict has not been decreasing, but is rather

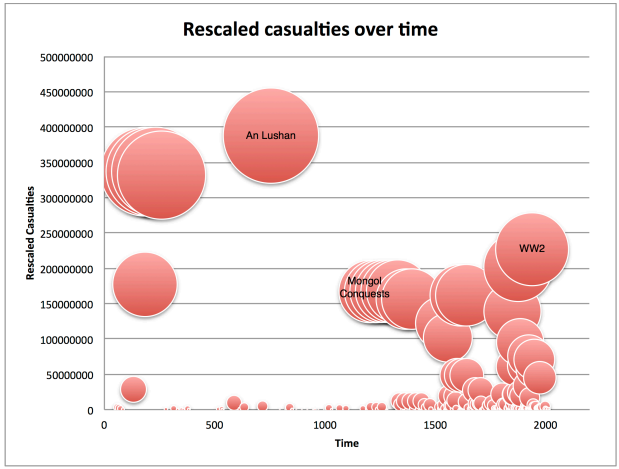


Figure 16.4: *Rescaled death toll of armed conflict and regimes over time. Data are rescaled w.r.t. today's world population. Conflicts lasting more than 25 years are disaggregated into two or more conflicts, each one lasting 25 years.*

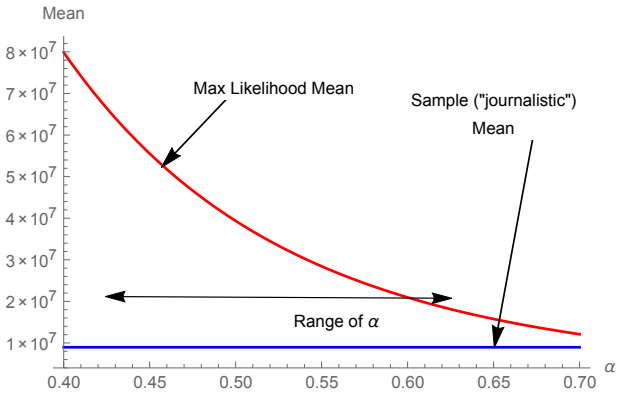


Figure 16.5: *Observed "journalistic" mean compared to MLE mean (derived from rescaling back the data to compact support) for different values of α (hence for permutations of the pair (σ_α, α)). The "range of α is the one we get from possible variations of the data from bootstrap and reliability simulations.*

underestimated by techniques relying on naive year-on-year changes in the mean, or using sample mean as an estimator of the true mean of an extremely fat-tailed phenomenon; 2) armed conflicts have memoryless inter-arrival times, thus incompatible with the idea of a time trend. Our analysis uses 1) raw data, as recorded and estimated by historians; 2) a naive transformation, used by certain historians and sociologists, which rescales past conflicts and casualties with respect to the actual population; 3) more importantly, a log transformation to account for the fact that the number of casualties in a conflict cannot be larger than the world population. (This is similar to the transformation of data into log-returns in mathematical finance in order to use distributions with support on the real line.)

All in all, among the different classes of data (raw and rescaled), we observe that 1) casualties are Power Law distributed.[3] In the case of log-rescaled data we observe $.4 \leq \alpha \leq .7$, thus indicating an extremely fat-tailed phenomenon with an

3 Many earlier studies have found Paretianity in data, [**?**],[38]. Our study, aside from the use of extreme value techniques, reliability bootstraps, and compact support transformations, varies in both calibrations and interpretation.

undefined mean (a result that is robustly obtained); 2) the inter-arrival times of conflicts above the 50k threshold follow a homogeneous Poisson process, indicating no particular trend, and therefore contradicting a popular narrative about the decline of violence; 3) the true mean to be expected in the future, and the most compatible with the data, though highly stochastic, is $\approx 3\times$ higher than past mean.

Further, we explain: 1) how the mean (in terms of expected casualties) is severely underestimated by conventional data analyses as the observed mean is not an estimator of true mean (unlike the tail exponent that provides a picture with smaller noise); 2) how misconceptions arise from the deceiving lengthy (and volatile) inter-arrival times between large conflicts.

To remedy the inaccuracies of historical numerical assessments, we provide a standard bootstrap analysis of our estimates, in addition to Monte Carlo checks for unreliability of wars and absence of events from currently recorded history.

16.2 SUMMARY STATISTICAL DISCUSSION

16.2.1 Results

Paretian tails Peak-Over-Threshold methods show (both raw and rescaled variables) exhibit strong Paretian tail behavior, with survival probability $\mathbb{P}(X > x) \sim \lambda(x)x^{-\alpha}$, where $\lambda : [L, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x\to+\infty} \frac{\lambda(kx)}{\lambda(x)} = 1$ for any $k > 0$.

We parametrize G(.), a Generalized Pareto Distribution (GPD) , see Table 16.4, $G(x) = 1 - (1 + \xi y/\beta)^{-1/\xi}$, with $\xi \approx 1.88, \pm.14$ for rescaled data which corresponds to a tail $\alpha = \frac{1}{\xi} = .53, \pm.04$.

Memorylessness of onset of conflicts Tables 16.2 and 16.3 show inter-arrival times, meaning one can wait more than a hundred years for an event such as WWII without changing one's expectation. There is no visible autocorrelation, no statistically detectable temporal structure (i.e. we cannot see the imprint of a self-exciting process), see Figure 16.8.

Full distribution(s) Rescaled data fits a Lomax-Style distribution with same tail as obtained by POT, with strong goodness of fit. For events with casualties $> L = 10K, 25K, 50K, etc.$ we fit different Pareto II (Lomax) distributions with corresponding tail α (fit from GPD), with scale $\sigma = 84,360$, i.e., with density $\frac{\alpha\left(\frac{-L+\sigma+x}{\sigma}\right)^{-\alpha-1}}{\sigma}$, $x \geq L$.

We also consider a wider array of statistical "pictures" from pairs α, σ_α across the data from potential alternative values of α, with recalibration of maximum likelihood σ, see Figure 16.5.

Difference between sample mean and maximum likelihood mean : Table 16.1 shows the true mean using the parametrization of the Pareto distribution above and inverting the transformation back to compact support. "True" or maximum likelihood, or "statistical" mean is between 3 and 4 times observed mean.

This means the "journalistic" observation of the mean, aside from the conceptual mistake of relying on sample mean, underestimates the true mean by at least 3 times and higher future observations would not allow the conlusion that violence has "risen".

Table 16.1: *Sample means and estimated maximum likelihood mean across minimum values L – Rescaled data.*

L	Sample Mean	ML Mean	Ratio
10K	9.079×10^6	3.11×10^7	3.43
25K	9.82×10^6	3.62×10^7	3.69
50K	1.12×10^7	4.11×10^7	3.67
100K	1.34×10^7	4.74×10^7	3.53
200K	1.66×10^7	6.31×10^7	3.79
500K	2.48×10^7	8.26×10^7	3.31

16.2.2 Conclusion

History as seen from tail analysis is far more risky, and conflicts far more violent than acknowledged by naive observation of behavior of averages in historical time series.

Table 16.2: *Average inter-arrival times and their mean absolute deviation for events with more than 1, 2, 5 and 10 million casualties, using actual estimates.*

Threshold	Average	MAD
1	26.71	31.66
2	42.19	47.31
5	57.74	68.60
10	101.58	144.47

16.3 METHODOLOGICAL DISCUSSION

16.3.1 Rescaling method

We remove the compact support to be able to use power laws as follows (see earlier chapters). Using X_t as the r.v. for number of incidences from conflict at times t,

Table 16.3: *Average inter-arrival times and their mean absolute deviation for events with more than 1, 2, 5, 10, 20, and 50 million casualties, using rescaled amounts.*

Threshold	Average	MAD
1	11.27	12.59
2	16.84	18.13
5	26.31	27.29
10	37.39	41.30
20	48.47	52.14
50	67.88	78.57

Table 16.4: *Estimates (and standard errors) of the Generalized Pareto Distribution parameters for casualties over a 50k threshold. For both actual and rescaled casualties, we also provide the number of events lying above the threshold (the total number of events in our data is 99).*

Data	Nr. Excesses	ξ	β
Raw Data	307	1.5886	3.6254
		(0.1467)	(0.8191)
Naive Rescaling	524	1.8718	14.3254
		(0.1259)	(2.1111)
Log-rescaling	524	1.8717	14.3261
		(0.1277)	(2.1422)

consider first a naive rescaling of $X'_t = \frac{X_t}{H_t}$, where H_t is the total human population at period t. See appendix for methods of estimation of H_t.

Next, with today's maximum population H and L the naively rescaled minimum for our definition of conflict, we introduce a smooth rescaling function $\varphi : [L, H] \to [L, \infty)$ satisfying:

i φ is "smooth": $\varphi \in C^\infty$,

ii $\varphi^{-1}(\infty) = H$,

iii $\varphi^{-1}(L) = \varphi(L) = L$.

In particular, we choose:

$$\varphi(x) = L - H \log\left(\frac{H - x}{H - L}\right). \tag{16.1}$$

We can perform appropriate analytics on $x_r = \varphi(x)$ given that it is unbounded, and properly fit Power Law exponents. Then we can rescale back for the properties of X. Also notice that the $\varphi(x) \approx x$ for very large values of H. This means that for a very large upper bound, the results we will get for x and $\varphi(x)$ will be essentially the same. The big difference is only from a philosophical/methodological point of view, in the sense that we remove the upper bound (unlikely to be reached).

In what follows we will use the naively rescaled casualties as input for the $\varphi(\cdot)$ function.

We pick $H = P_{t_0}$ for the exercise.

The distribution of x can be rederived as follows from the distribution of x_r:

$$\int_L^{\infty} f(x_r)\,\mathrm{d}x_r = \int_L^{\varphi^{-1}(\infty)} g(x)\,\mathrm{d}x, \tag{16.2}$$

where $\varphi^{-1}(u) = (L-H)e^{\frac{L-u}{H}} + H$

In this case, from the Pareto-Lomax selected:

$$f(x_r) = \frac{\alpha\left(\frac{-L+\sigma+x_r}{\sigma}\right)^{-\alpha-1}}{\sigma}, x_r \in [L,\infty) \tag{16.3}$$

$$g(x) = \frac{\alpha H\left(\frac{\sigma - H\log\left(\frac{H-x}{H-L}\right)}{\sigma}\right)^{-\alpha-1}}{\sigma(H-x)}, x \in [L,H],$$

which verifies $\int_L^H x\,g(x)\,\mathrm{d}x = 1$. Hence the expectation

$$\mathbb{E}_g(x;L,H,\sigma,\alpha) = \int_L^H x\,g(x)\,\mathrm{d}x, \tag{16.4}$$

$$\mathbb{E}_g(X;L,H,\sigma,\alpha) = \alpha H\left(\frac{1}{\alpha} - \frac{(H-L)e^{\sigma/H}E_{\alpha+1}\left(\frac{\sigma}{H}\right)}{H}\right) \tag{16.5}$$

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

Note that we rely on the invariance property:

Remark 15

If $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ, then for an absolutely continuous function ϕ, $\phi(\hat{\theta})$ is the MLE estimator of $\phi(\theta)$.

For further details see [208].

16.3.2 Expectation by Conditioning (less rigorous)

We would be replacing a smooth function in C^{∞} by a Heaviside step function, that is the indicator function $\mathbb{1} : \mathbb{R} \to \{0,1\}$, written as $\mathbb{1}_{X\in[L,H]}$:

$$\mathbb{E}(\mathbb{1}_{X\in[L,H]}) = \frac{\int_L^H x\,f(x)\,\mathrm{d}x}{\int_L^H f(x)\,\mathrm{d}x}$$

which for the Pareto Lomax becomes:

$$\mathbb{E}(\mathbb{1}_{X\in[L,H]}) = \frac{\frac{\alpha\sigma^{\alpha}(H-L)}{\sigma^{\alpha}-(H-L+\sigma)^{\alpha}} + (\alpha-1)L + \sigma}{\alpha-1} \tag{16.6}$$

16.3.3 Reliability of data and effect on tail estimates

Data from violence is largely anecdotal, spreading via citations, often based on some vague estimate, without anyone's ability to verify the assessments using period sources. An event that took place in the seventh century, such as the an Lushan rebellion, is "estimated" to have killed 26 million people, with no precise or reliable methodology to allow us to trust the number. The independence war of Algeria has various estimates, some from France, others from the rebels, and nothing scientifically or professionally obtained.

As said earlier, in this chapter, we use different data: raw data, naively rescaled data w.r.t. the current world population, and log-rescaled data to avoid the theoretical problem of the upper bound.

For some observations, together with the estimated number of casualties, as resulting from historical sources, we also have a lower and upper bound available. Let X_t be the number of casualties in a given conflict at time t. In principle, we can define triplets like

- $\{X_t, X_t^l, X_t^u\}$ for the actual estimates (raw data), where X_t^l and X_t^u represent the lower and upper bound, if available.
- $\{Y_t = X_t \frac{P_{2015}}{P_t}, Y_t^l = X_t^l \frac{P_{2015}}{P_t}, Y_t^u = X_t^u \frac{P_{2015}}{P_t}\}$ for the naively rescaled data, where P_{2015} is the world population in 2015 and P_t is the population at time $t = 1, ..., 2014$.
- $\{Z_t = \varphi(Y_t), Z_t^l = \varphi(Y_t^l), Z_t^u = \varphi(Y_t^u)\}$ for the log-rescaled data.

To prevent possible criticism about the use of middle estimates, when bounds are present, we have decided to use the following Monte Carlo procedure (for more details [198]), obtaining no significant different in the estimates of all the quantities of interest (like the tail exponent $\alpha = 1/\xi$):

1. For each event X for which bounds are present, we have assumed casualties to be uniformly distributed between the lower and the upper bound, i.e. $X \sim U(X^l, X^u)$. The choice of the uniform distribution is to keep things simple. All other bounded distributions would in fact generate the same results in the limit, thanks to the central limit theorem.
2. We have then generated a large number of Monte Carlo replications, and in each replication we have assigned a random value to each event X according to $U(X^l, X^u)$.
3. For each replication we have computed the statistics of interest, typically the tail exponent, obtaining values that we have later averaged.

This procedure has shown that the precision of estimates does not affect the tail of the distribution of casualties, as the tail exponent is rather stable.

For those events for which no bound is given, the options were to use them as they are, or to perturb them by creating fictitious bounds around them (and then treat them as the other bounded ones in the Monte Carlo replications). We have chosen the second approach.

The above also applies to Y_t and Z_t.

Note that the tail α derived from an average is different from an average alpha across different estimates, which is the reason we perform the various analyses across estimates.

Technical comment These simulations are largely looking for a "stochastic alpha" bias from errors and unreliability of data (Chapter x). With a sample size of n, a parameter $\hat{\theta}_m$ will be the average parameter obtained across a large number of Monte Carlo runs. Let X_i be a given Monte Carlo simulated vector indexed by i and X_μ is the middle estimate between high and low bounds. Since, with $\frac{1}{m}\sum_{\leq m}\|X_j\|_1 = \|X_\mu\|_1$ across Monte Carlo runs but $\forall j, \|X_j\|_1 \neq \|X_\mu\|_1$, $\widehat{\theta}_m = \frac{1}{m}\sum_{\leq m}\widehat{\theta}(X_j) \neq \widehat{\theta}(X_\mu)$. For instance, consider the maximum likelihood estimation of a Paretian tail, $\widehat{\alpha}(X_i) \triangleq n\left(\sum_{1\leq i\leq n}\log\left(\frac{x_i}{L}\right)\right)^{-1}$. With $\Delta \geq x_m$, define

$$\widehat{\alpha}(X_i \sqcup \Delta) \triangleq \frac{1}{2}\left(\frac{n}{\sum_{i=1}^{n}\log\left(\frac{x_i}{L}\right) - \log\left(\frac{\Delta}{L}\right)} + \frac{n}{\sum_{i=1}^{n}\log\left(\frac{x_i}{L}\right) + \log\left(\frac{\Delta}{L}\right)}\right)$$

which, owing to the concavity of the logarithmic function, gives the inequality

$$\forall \Delta \geq x_m, \widehat{\alpha}(X_i \sqcup \Delta) \geq \widehat{\alpha}(X_i).$$

16.3.4 Definition of an "event"

"Named" conflicts are an arbitrary designation that, often, does not make sense statistically: a conflict can have two or more names; two or more conflicts can have the same name, and we found no satisfactory hierarchy between war and conflict. For uniformity, we treat events as the shorter of event or its disaggregation into units with a maximum duration of 25 years each. Accordingly, we treat Mongolian wars, which lasted more than a century and a quarter, as more than a single event. It makes little sense otherwise as it would be the equivalent of treating the period from the Franco-Prussian war to WW II as "German(ic) wars", rather than multiple events because these wars had individual names in contemporary sources. Effectively the main sources such as the *Encyclopedia of War* [186] list numerous conflicts in place of "Mongol Invasions" –the more sophisticated the historians in a given area, the more likely they are to break conflicts into different "named" events and, depending on historians, Mongolian wars range between 12 and 55 conflicts.

What controversy about the definition of a "name" can be, once again, solved by bootstrapping. Our conclusion, incidentally, is invariant to the bundling or unbundling of the Mongolian wars.

Further, in the absence of a clearly defined protocol in historical studies, it has been hard to disentangle direct death from wars and those from less direct effects on populations (say blocades, famine). For instance the First Jewish War has confused historians as an estimated 30K death came from the war, and a considerably higher (between 350K and the number 1M according to Josephus) from the famine or civilian casualties.

16.3.5 Missing events

We can assume that there are numerous wars that are not part of our sample, even if we doubt that such events are in the "tails" of the distribution, given that large conflicts are more likely to be reported by historians. Further, we also assume that their occurrence is random across the data (in the sense that they do not have an effect on clustering).

But we are aware of a bias from differential in both accuracy and reporting across time: events are more likely to be recorded in modern times than in the past. Raising the minimum value L the number of such "missed" events and their impact are likely to drop rapidly. Indeed, as a robustness check, raising the bar to a minimum $L = 500K$ does not change our analysis.

A simple jackknife procedure, performed by randomly removing a proportion of events from the sample and repeating analyses, shows us the dependence of our analysis on missing events, dependence that we have found to be insignificant, when focusing on the tail of the distribution of casualties. In other words, given that we are dealing with extremes, if removing 30% of events and checking the effects on parameters produce no divergence from initial results, then we do not need to worry of having missed 30% of events, as missing events are not likely to cause thinning of the tails.[4]

16.3.6 Survivorship Bias

We did not take into account of the survivorship biases in the analysis, assuming it to be negligible before 1960, as the probability of a conflict affecting all of mankind was negligible. Such probability (and risk) became considerably higher since, especially because of nuclear and other mass destruction weapons.

16.4 DATA ANALYSIS

Figures 16.3 and 16.4 graphically represent our data: the number of casualties over time. Figure 16.3 refers to the estimated actual number of victims, while Figure 16.4 shows the rescaled amounts, obtained by rescaling the past observation with respect to the world population in 2015 (around 7.2 billion people)[5]. Figures 16.3 might suggest an increase in the death toll of armed conflicts over time, thus supporting the idea that war violence has increased. Figure 16.4, conversely, seems to suggest a decrease in the (rescaled) number of victims, especially in the last hundred years, and possibly in violence as well. In what follows we show that both interpretations are surely naive, because they do not take into consideration the fact that we are dealing with extreme events.

4 The opposite is not true, which is at the core of the Black Swan asymmetry: such procedure does not remedy the missing of tail, "Black Swan" events from the record. A single "Black Swan" event can considerably fatten the tail. In this case the tail is fat enough and no missing information seems able to make it thinner.

5 Notice that, in equation (16.1), for $H = 7.2$ billion, $\varphi(x) \approx x$. Therefore Figure 16.4 is also representative for log-rescaled data.

16.4.1 Peaks over Threshold

Given the fat-tailed nature of the data, which can be easily observed with some basic graphical tools like histograms on the logs and QQplots (Figure 16.6 shows the QQplot of actual casualties against an exponential distribution: the clear concavity is a signal of fat-tailed distribution), it seems appropriate to use a well-known method of extreme value theory to model war casualties over time: the Peaks-over-Threshold or POT [181].

According to the POT method, excesses of an i.i.d. sequence over a high threshold u (that we have to identify) occur at the times of a homogeneous Poisson process, while the excesses themselves can be modeled with a Generalized Pareto Distribution (GPD). Arrival times and excesses are assumed to be independent of each other.

In our case, assuming the independence of the war events does not seem a strong assumption, given the time and space separation among them. Regarding the other assumptions, on the contrary, we have to check them.

We start by identifying the threshold u above which the GPD approximation may hold. Different heuristic tools can be used for this purpose, from Zipf plot to mean excess function plots, where one looks for the linearity which is typical of fat-tailed phenomena [44, 81]. Figure 16.7 shows the mean excess function plot for actual casualties[6]: an upward trend is clearly present, already starting with a threshold equal to 5k victims. For the goodness of fit, it might be appropriate to choose a slightly larger threshold, like $u = 50k$[7].

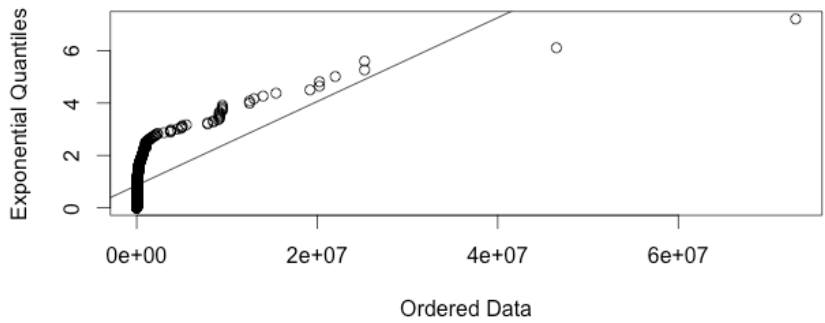


Figure 16.6: *QQplot of actual casualties against standard exponential quantile. The concave curvature of data points is a clear signal of heavy tails.*

6 Similar results hold for the rescaled amounts (naive and log). For the sake of brevity we always show plots for one of the two variables, unless a major difference is observed.

7 This idea has also been supported by subsequent goodness-of-fit tests.

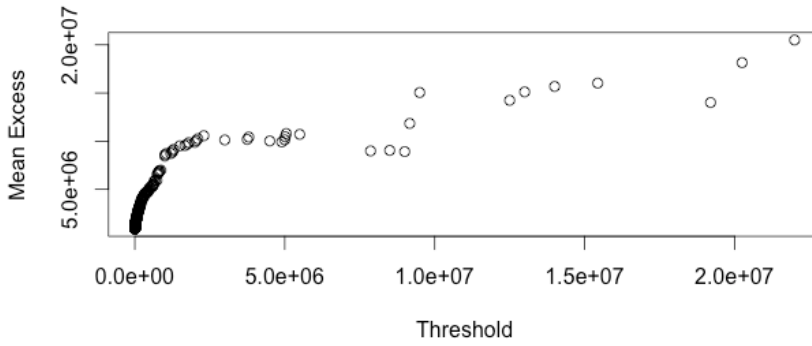


Figure 16.7: *Mean excess function plot (MEPLOT) for actual casualties. An upward trend - almost linear in the first part of the graph - is present, suggesting the presence of a fat right tail. The variability of the mean excess function for higher thresholds is due to the small number of observation exceeding those thresholds and should not be taken into consideration.*

16.4.2 Gaps in Series and Autocorrelation

To check whether events over time occur according to a homogeneous Poisson process, a basic assumption of the POT method, we can look at the distribution of the inter-arrival times or gaps, which should be exponential. Gaps should also show no autocorrelation.

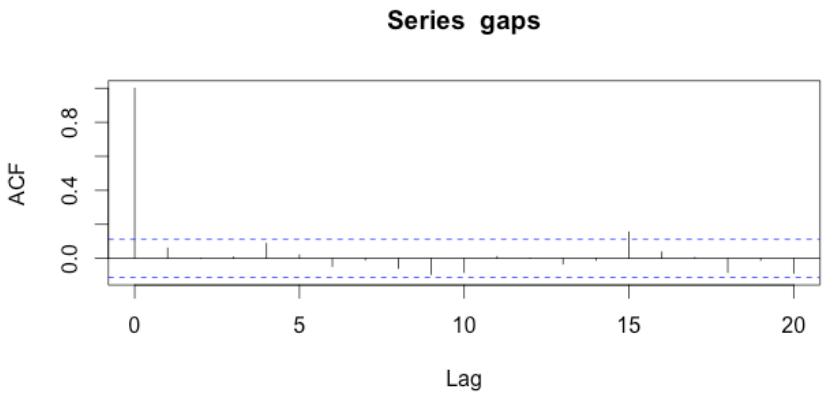


Figure 16.8: *ACF plot of gaps for actual casualties, no significant autocorrelation is visible.*

Figure 16.8 clearly shows the absence of autocorrelation. The plausibility of an exponential distribution for the inter-arrival times can be positively checked using both heuristic and analytic tools. Here we omit the positive results for brevity.

However, in order to provide some extra useful information, in Tables 16.2 and 16.3 we provide some basic statistics about the inter-arrival times for very catastrophic events in terms of casualties[8]. The simple evidence there contained should already be sufficient to underline how unreliable can be the statement that war violence has been decreasing over time. For an events with more than 10 million victims, if we refer to actual estimates, the average time delay is 101.58 years, with a mean absolute deviation of 144.47 years[9]. This means that it is totally plausible that in the last few years we have not observed such a large event. It could simply happen tomorrow or some time in the future. This also means that every trend extrapolation makes no great sense for this type of extreme events. Finally, we have to consider that an event as large as WW2 happened only once in 2014 years, if we deal with actual casualties (for rescaled casualties we can consider the An Lushan rebellion); in this case the possible waiting time is even longer.

16.4.3 Tail Analysis

Given that the POT assumptions about the Poisson process seem to be confirmed by data, it is finally the time to fit a Generalized Pareto Distribution to the exceedances.

Consider a random variable X with df F, and call F_u the conditional df of X above a given threshold u. We can then define a r.v. Y, representing the rescaled excesses of X over the threshold u, getting [181]

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}$$

for $0 \leq y \leq x_F - u$, where x_F is the right endpoint of the underlying distribution F. Pickands [187], Balkema and de Haan [8], [9] and [10] showed that for a large class of underlying distribution functions F (following in the so-called domain of attraction of the GEV distribution [181]), and a large u, F_u can be approximated by a Generalized Pareto distribution: $F_u(y) \to G(y)$, as $u \to \infty$ where

$$G(y) = \begin{cases} 1 - (1 + \xi y/\beta)^{-1/\xi} & if\ \xi \neq 0 \\ 1 - e^{-y/\beta} & if\ \xi = 0. \end{cases}. \tag{16.7}$$

It can be shown that the GPD distribution is a distribution interpolating between the exponential distribution (for $\xi = 0$) and a class of Pareto distributions. We refer to [181] for more details.

The parameters in (16.7) can be estimated using methods like maximum likelihood or probability weighted moments [181]. The goodness of fit can then be tested using bootstrap-based tests [257].

8 Table 16.2 does not show the average delay for events with 20M(50M) or more casualties. This is due to the limited amount of these observations in actual, non-rescaled data. In particular, all the events with more than 20 million victims have occurred during the last 150 years, and the average inter-arrival time is below 20 years. Are we really living in more peaceful world?

9 In case of rescaled amounts, inter-arrival times are shorter, but the interpretation is the same

Table 16.4 contains our mle estimates for actual and rescaled casualties above a 50k victims threshold. This threshold is in fact the one providing the best compromise between goodness of fit and a sufficient number of observation, so that standard errors are reliable. The actual and both the rescaled data show two different sets of estimates, but their interpretation is strongly consistent. For this reason we just focus on actual casualties for the discussion.

The parameter ξ is the most important for us: it is the parameter governing the fatness of the right tail. A ξ greater than 1 (we have 1.5886) signifies that no moment is defined for our Generalized Pareto: a very fat-tailed situation. Naturally, in the sample, we can compute all the moments we are interested in, but from a theoretical point of view they are completely unreliable and their interpretation is extremely flawed (a very common error though). According to our fitting, very catastrophic events are not at all improbable. It is worth noticing that the estimates is significant, given that its standard error is 0.1467.

Figures 16.9 and 16.10 compare our fittings to actual data. In both figures it is possible to see the goodness of the GPD fitting for most of the observations above the 50k victims threshold. Some problems arise for the very large events, like WW2 and the An Lushan rebellion [10]. In this case it appears that our fitting expects larger events to have happened. This is a well-known problem for extreme data [181]. The very large event could just be behind the corner.

Similarly, events with 5 to 10 million victims (not at all minor ones!) seem to be slightly more frequent than what is expected by our GPD fitting. This is another signal of the extreme character of war casualties, which does not allow for the extrapolation of simplistic trends.

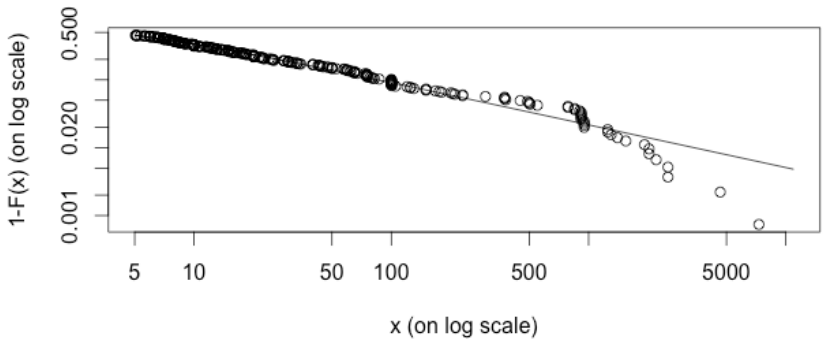


Figure 16.9: *GPD tail fitting to actual casualties' data (in 10k). Parameters as per Table 16.4, first line.*

10 If we remove the two largest events from the data, the GPD hypothesis cannot be rejected at the 5% significance level.

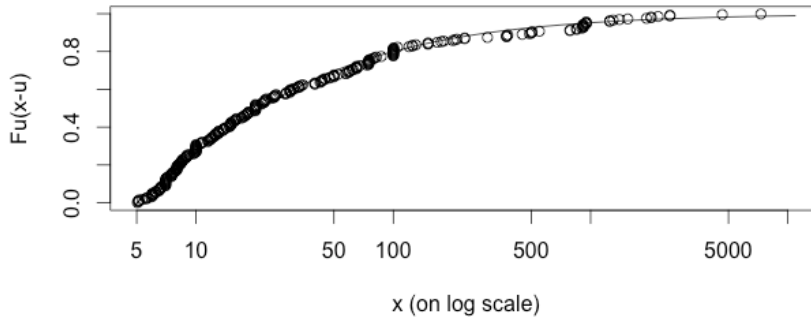


Figure 16.10: *GPD cumulative distribution fitting to actual casualties' data (in 10k). Parameters as per Table 16.4, first line.*

16.4.4 An Alternative View on Maxima

Another method is the block-maxima approach of extreme value theory. In this approach data are divided into blocks, and within each block only the maximum value is taken into consideration. The Fisher-Tippet theorem [181] then guarantees that the normalized maxima converge in distribution to a Generalized Extreme Value Distribution, or GEV.

$$GEV(x;\xi) = \begin{cases} \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right) & \xi \neq 0 \\ \exp\left(-\exp(-x)\right) & \xi = 0 \end{cases}, \quad 1+\xi x > 0$$

This distribution is naturally related to the GPD, and we refer to [181] for more details.

If we divide our data into 100-year blocks, we obtain 21 observation (the last block is the residual one from 2001 to 2014). Maximum likelihood estimations give a ξ larger than 2, indicating that we are in the so-called FrÃl'chet maximum domain of attraction, compatible with very heavy-tailed phenomena. A value of ξ greater than 2 under the GEV distribution further confirms the idea of the absence of moments, a clear signal of a very heavy right tail.

16.4.5 Full Data Analysis

Naturally, being aware of limitations, we can try to fit all our data, while for casualties in excess of 10000, we fit the Pareto Distribution from Equation 16.3 with $\alpha \approx 0.53$ throughout. The goodness of fit for the "near tail" (L=10K) can be see in Figure 16.2. Similar results to Figure 16.2 are seen for different values in table below, all with the same goodness of fit.

L	σ
$10K$	84,260
$25K$	899,953
$50K$	116,794
$100K$	172,733
$200K$	232,358
$500K$	598,292

The different possible values of the mean in Equation 16.4 can be calculated across different set values of α, with one single degree of freedom: the corresponding σ is a MLE estimate using such α as fixed: for a sample size n, and x_i the observations higher than L, $\sigma_\alpha = \left\{\sigma : \frac{\alpha n}{\sigma} - (\alpha+1)\sum_{i=1}^{n} \frac{1}{x_i - L + \sigma} = 0, \sigma > 0\right\}$.

The sample average for $L = 10K$ is 9.12×10^6, across 100K simulations, with the spread in values showed in Figure 16.15.

The "true" mean from Equation 16.4 yields $3.1 * 10^7$, and we repeated for L =10K, 20K, 50K, 100K, 200K, and 500K, finding ratios of true estimated mean to observed safely between 3 and 4., see Table 16.1. Notice that this value for the mean of ≈ 3.5 times the observed sample mean is only a general guideline, since, being stochastic, does not reveal any precise information other than prevent us from taking the naive mean estimation seriously.

For under fat tails, the mean derived from estimates of α is more rigorous and has a smaller error, since the estimate of α is asymptotically Gaussian while the average of a power law, even when it exists, is considerably more stochastic. See the discussion on "slowness of the law of large numbers" in 8 in connection with the point.

We get the mean by truncation for L=10K a bit lower, under equation 16.6; around 1.8835×10^7.

We finally note that, for values of L considered, 96 % of conflicts with more than 10,000 victims are below the mean: where m is the mean,

$$\mathbb{P}(X < m) = 1 - \left(1 - \frac{H \log\left(\alpha e^{\sigma/H} E_{\alpha+1}\left(\frac{\sigma}{H}\right)\right)}{\sigma}\right)^{-\alpha}.$$

16.5 ADDITIONAL ROBUSTNESS AND RELIABILITY TESTS

16.5.1 Bootstrap for the GPD

In order to check our sensitivity to the quality/precision of our data, we have decided to perform some bootstrap analysis. For both raw data and the rescaled ones we have generated 100K new samples by randomly selecting 90% of the observations, with replacement. Figures 16.11, 16.12 and 16.13 show the stability of our ξ estimates. In particular $\xi > 0$ in all samples, indicating the extreme fat-tailedness of the number of victims in armed conflicts. The ξ estimates in Table 16.4 appear

to be good approximations for our GPD real shape parameters, notwithstanding imprecisions and missing observations in the data.

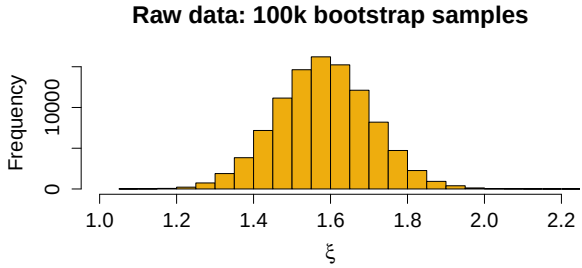


Figure 16.11: *ξ parameter's distribution over 100K bootstrap samples for actual data. Each sample is randomly selected with replacement using 90% of the original observations.*

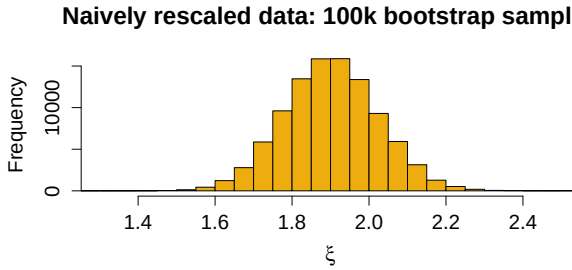


Figure 16.12: *ξ parameter's distribution over 100K bootstrap samples for naively rescaled data. Each sample is randomly selected with replacement using 90% of the original observations.*

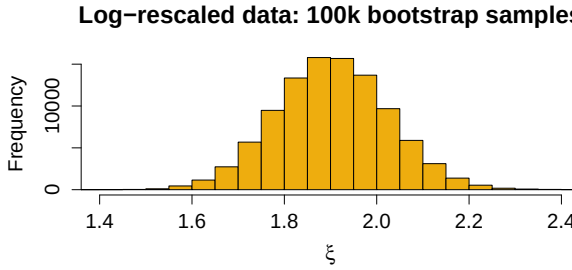


Figure 16.13: *ξ parameter's distribution over 100K bootstrap samples for log-rescaled data. Each sample is randomly selected with replacement using 90% of the original observations.*

16.5.2 Perturbation Across Bounds of Estimates

We performed analyses for the "near tail" using the Monte Carlo techniques discussed in section 16.3.3. We look at second order "p-values", that is the sensitivity of the p-values across different estimates in Figure 16.14 –practically all results meet the same statistical significance and goodness of fit.

In addition, we look at values of both the sample means and the alpha-derived MLE mean across permutations, see Figures 16.15 and 16.16.

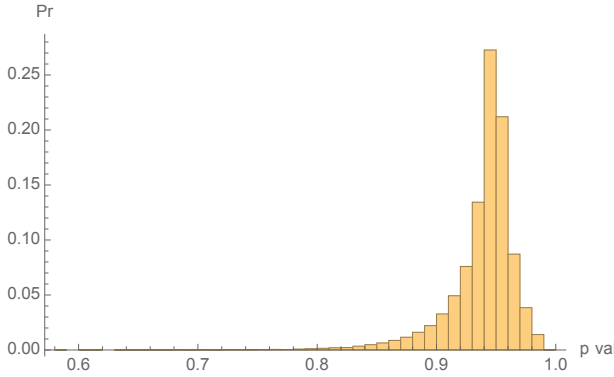


Figure 16.14: *P-Values of Pareto-Lomax across 100K combinations. This is not to ascertain the p-value, rather to check the robustness by looking at the variations across permutations of estimates.*

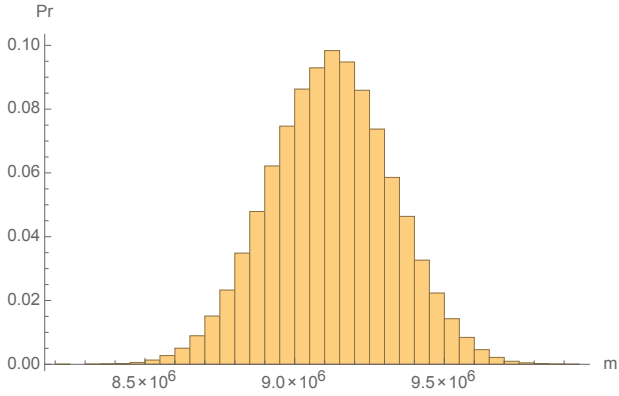


Figure 16.15: *Rescaled sample mean across 100K estimates between high-low.*

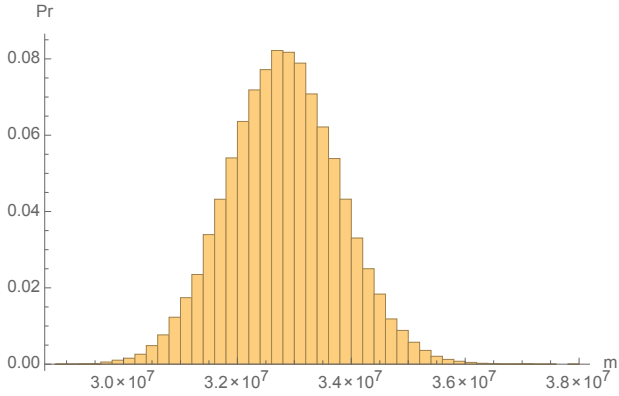


Figure 16.16: *Rescaled MLE mean across 100K estimates between high-low.*

16.6 CONCLUSION: IS THE WORLD MORE UNSAFE THAN IT SEEMS?

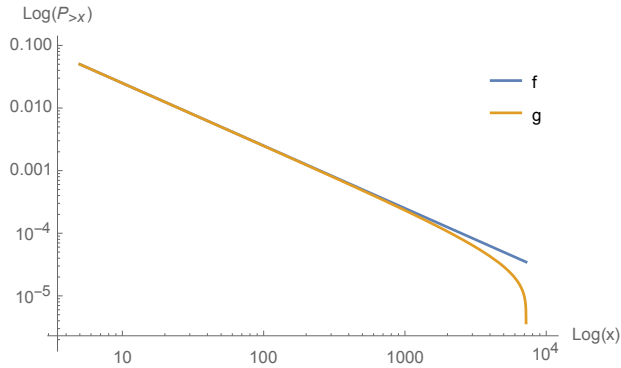


Figure 16.17: *Loglogplot comparison of f and g, showing a pasting-boundary style capping around H.*

To put our conclusion in the simplest of terms: the occurrence of events that would raise the average violence by a multiple of 3 would not cause us to rewrite this chapter, nor to change the parameters calibrated within.

- Indeed, from statistical analysis alone, the world is more unsafe than casually examined numbers. Violence is underestimated by journalistic nonstatistical looks at the mean and lack of understanding of the stochasticity of under inter-arrival times.
- The transformation into compact support allowed us to perform the analyses and gauge such underestimation which , if noisy, gives us an idea of the underestimation and its bounds.
- In other words, a large event and even a rise in observed mean violence would not be inconsistent with statistical properties, meaning it would justify a "nothing has changed" reaction.
- We avoided discussions of homicide since we limited L to values $> 10,000$, but its rate doesn't appear to have a particular bearing on the tails. It could be a drop in the bucket. It obeys different dynamics. We may have observed lower rate of homicide in societies but most risks of death come from violent conflict. (Casualties from homicide by rescaling from the rate 70 per $100k$, gets us 5.04×10^6 casualties per annum at today's population. A drop to minimum levels stays below the difference between errors on the mean of violence from conflicts with higher than 10,000 casualties.)
- We ignored survivorship bias in the data analysis (that is, the fact that *had the world been more violent, we wouldn't be here to talk about it*). Adding it would increase the risk. The presence of tail effects today makes further analysis require taking it into account. Since 1960, a single conflict –which almost happened– has the ability to reach the max casualties, something we did not have before.(We can rewrite the model with one of fragmentation of the world, constituted of "separate" isolated n independent random variables X_i, each with a maximum value H_i, with the total $\sum_n \omega_i H_i = H$, with all $w_i > 0$,

$\sum_n \omega_i = 1$. In that case the maximum (that is worst conflict) could require the joint probabilities that all $X_1, X_2, \cdots X_n$ are near their maximum value, which, under subexponentiality, is an event of much lower probability than having a single variable reach its maximum.)

16.7 ACKNOWLEDGMENTS

The data was compiled by Captain Mark Weisenborn. We thank Ben Kiernan for comments on East Asian conflicts.

WHAT ARE THE CHANCES OF A THIRD WORLD WAR?*,†

THIS IS FROM AN ARTICLE that is part of the debate with public intellectuals who claim that violence have dropped "from data", without realizing that science is hard; significance requires further data under fat tails and more careful examination. Our response (by the author and P. Cirillo) provides a way to summarize the main problem with naive empiricism under fat tails.

In a recent issue of *Significance* Mr. Peter McIntyre asked what the chances are that World War III will occur this century. Prof. Michael Spagat wrote that nobody knows, nobody can really answer–and we totally agree with him on this. Then he adds that "a really huge war is possible but, in my view, extremely unlikely." To support his statement, Prof. Spagat relies partly on the popular science work of Prof. Steven Pinker, expressed in The Better Angels of our Nature and journalistic venues. Prof. Pinker claims that the world has experienced a long-term decline in violence, suggesting a structural change in the level of belligerence of humanity.

It is unfortunate that Prof. Spagat, in his answer, refers to our paper (this volume, Chapter 16), which is part of a more ambitious project we are working on related to fat-tailed variables.

What characterizes fat tailed variables? They have their properties (such as the mean) dominated by extreme events, those "in the tails". The most popularly known version is the "Pareto 80/20".

We show that, simply, data do not support the idea of a structural change in human belligerence. So Prof. Spagat's first error is to misread our claim: we are making neither pessimistic nor optimistic declarations: we just believe that statisticians should abide by the foundations of statistical theory and avoid telling data what to say.

Let us go back to first principles.

Discussion chapter.

Figure F.1: *After Napoleon, there was a lull in Europe. Until nationalism came to change the story.*

Foundational Principles

Fundamentally, statistics is about ensuring people do not build scientific theories from hot air, that is without significant departure from random. Otherwise, it is patently "fooled by randomness".

Further, for fat tailed variables, the conventional mechanism of the law of large numbers is considerably slower and significance requires more data and longer periods. Ironically, there are claims that can be done on little data: inference is asymmetric under fat-tailed domains. ***We require more data to assert that there are no Black Swans than to assert that there are Black Swans*** hence we would need much more data to claim a drop in violence than to claim a rise in it.

Finally, statements that are not deemed statistically significant –and shown to be so –should never be used to construct scientific theories.

These foundational principles are often missed because, typically, social scientists' statistical training is limited to mechanistic tools from thin tailed domains [2]. In physics, one can often claim evidence from small data sets, bypassing standard statistical methodologies, simply because the variance for these variables is low. The higher the variance, the more data one needs to make statistical claims. For fat-tails, the variance is typically high and underestimated in past data.

The second –more serious –error Spagat and Pinker made is to believe that tail events and the mean are somehow different animals, not realizing that the mean includes these tail events.

> ***For fat-tailed variables, the mean is almost entirely determined by extremes. If you are uncertain about the tails, then you are uncertain about the mean.***

It is thus incoherent to say that violence has dropped but maybe not the risk of tail events; it would be like saying that someone is "extremely virtuous except during the school shooting episode when he killed 30 students".

Robustness

Our study tried to draw the most robust statistical picture of violence, relying on methods from extreme value theory and statistical methods adapted to fat tails. We also put robustness checks to deal with the imperfection of data collected some thousand years ago: our results need to hold even if a third (or more) of the data were wrong.

Inter-arrival times

We show that the inter-arrival times among major conflicts are extremely long, and consistent with a homogenous Poisson process: therefore no specific trend can be established: we as humans can not be deemed as less belligerent than usual. For a conflict generating at least 10 million casualties, an event less bloody than WW1 or WW2, the waiting time is on average 136 years, with a mean absolute deviation of 267 (or 52 years and 61 deviations for data rescaled to today's population). The seventy years of what is called the "Long Peace" are clearly not enough to state much about the possibility of WW3 in the near future.

Underestimation of the mean

We also found that the average violence observed in the past underestimates the true statistical average by at least half. Why? Consider that about 90-97% of the observations fall below the mean, which requires some corrections with the help of extreme value theory. (Under extreme fat tails, the statistical mean can be closer to the past maximum observation than sample average.)

A common mistake

Similar mistakes have been made in the past. In 1860, one H.T. Buckle[2] used the same unstatistical reasoning as Pinker and Spagat.

> That this barbarous pursuit is, in the progress of society, steadily declining, must be evident, even to the most hasty reader of European history. If we compare one country with another, we shall find that for a very long period wars have been becoming less frequent; and now so clearly is the movement marked, that, until the late commencement of hostilities, we had remained at peace for nearly forty years: a circumstance unparalleled (...) The question arises, as to what share our moral feelings have had in bringing about this great improvement.

Moral feelings or not, the century following Mr. Buckle's prose turned out to be the most murderous in human history.

2 Buckle, H.T. (1858) *History of Civilization in England, Vol. 1*, London: John W. Parker and Son.

We conclude by saying that we find it fitting –and are honored –to expose fundamental statistical mistakes in a journal called *Significance*, as the problem is precisely about significance and conveying notions of statistical rigor to the general public.

Part VI

METAPROBABILITY PAPERS

17 HOW THICK TAILS EMERGE FROM RECURSIVE EPISTEMIC UNCERTAINTY†

THE Opposite of Central Limit: With the Central Limit Theorem, we start with a specific distribution and end with a Gaussian. The opposite is more likely to be true. Recall how we fattened the tail of the Gaussian by stochasticizing the variance? Now let us use the same metaprobability method, putting additional layers of uncertainty.

The Regress Argument (Error about Error) The main problem behind *The Black Swan* is the limited understanding of model (or representation) error, and, for those who get it, a lack of understanding of second order errors (about the methods used to compute the errors) and by a regress argument, an inability to continuously reapplying the thinking all the way to its limit (***particularly when one provides no reason to stop***). Again, there is no problem with stopping the recursion, provided it is accepted as a declared *a priori* that escapes quantitative and statistical methods.

Epistemic not statistical re-derivation of power laws Note that previous derivations of power laws have been statistical (cumulative advantage, preferential attachment, winner-take-all effects, criticality), and the properties derived by Yule, Mandelbrot, Zipf, Simon, Bak, and others result from structural conditions or breaking the independence assumptions in the sums of random variables allowing for the application of the central limit theorem, [89] [209][99] [160] [159] . This work is entirely epistemic, based on standard philosophical doubts and regress arguments.

Discussion chapter.

A version of this chapter was presented at Benoit Mandelbrot's Scientific Memorial on April 29, 2011,in New Haven, CT.

17.1 METHODS AND DERIVATIONS

Figure 17.1: *A version of this chapter was presented at Benoit Mandelbrot's memorial.*

17.1.1 Layering Uncertainties

Take a standard probability distribution, say the Gaussian. The measure of dispersion, here σ, is estimated, and we need to attach some measure of dispersion around it. The uncertainty about the rate of uncertainty, so to speak, or higher order parameter, similar to what called the "volatility of volatility" in the lingo of option operators (see Taleb, 1997, Derman, 1994, Dupire, 1994, Hull and White, 1997) –here it would be "uncertainty rate about the uncertainty rate". And there is no reason to stop there: we can keep nesting these uncertainties into higher orders, with the uncertainty rate of the uncertainty rate of the uncertainty rate, and so forth. There is no reason to have certainty anywhere in the process.

17.1.2 Higher Order Integrals in the Standard Gaussian Case

We start with the case of a Gaussian and focus the uncertainty on the assumed standard deviation. Define $\phi(\mu,\sigma;x)$ as the Gaussian PDF for value x with mean μ and standard deviation σ.

A 2^{nd} order stochastic standard deviation is the integral of ϕ across values of $\sigma \in \mathbb{R}^+$, under the PDF $f(\bar{\sigma}, \sigma_1; \sigma)$, with σ_1 its scale parameter (our approach to trach the error of the error), not necessarily its standard deviation; the expected value of σ_1 is $\overline{\sigma_1}$.

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

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

$$f(x)_N = \int_0^\infty \dots \int_0^\infty \phi(\mu, \sigma, x) f(\bar{\sigma}, \sigma_1, \sigma)\, f(\overline{\sigma_1}, \sigma_2, \sigma_1) \dots f(\overline{\sigma_{N-1}}, \sigma_N, \sigma_{N-1})\, \mathrm{d}\sigma\, \mathrm{d}\sigma_1\, \mathrm{d}\sigma_2 \dots \mathrm{d}\sigma_N \quad (17.1)$$

The problem is that this approach is parameter-heavy and requires the specifications of the subordinated distributions (in finance, the lognormal has been traditionally used for σ^2 (or Gaussian for the ratio $\text{Log}[\frac{\sigma_t^2}{\sigma^2}]$ since the direct use of a Gaussian allows for negative values). We would need to specify a measure f for each layer of error rate. Instead this can be approximated by using the mean deviation for σ, as we will see next.

Discretization using nested series of two-states for σ- a simple multiplicative process

We saw in the last chapter a quite effective simplification to capture the convexity, the ratio of (or difference between) $\phi(\mu,\sigma,x)$ and $\int_0^\infty \phi(\mu, \sigma, x) f(\bar{\sigma}, \sigma_1, \sigma)\, \mathrm{d}\sigma$ (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. Unlike the earlier situation we are not preserving the variance, rather the STD. Thus the distribution using the first order stochastic standard deviation can be expressed as:

$$f(x)_1 = \frac{1}{2}\left(\phi(\mu, \sigma\,(1 + a(1)), x) + \phi(\mu, \sigma(1 - a(1)), x)\right) \quad (17.2)$$

Now assume uncertainty about the error rate a(1), expressed by a(2), in the same manner as before. Thus in place of a(1) we have $\frac{1}{2}$ a(1)(1$\pm$ a(2)).

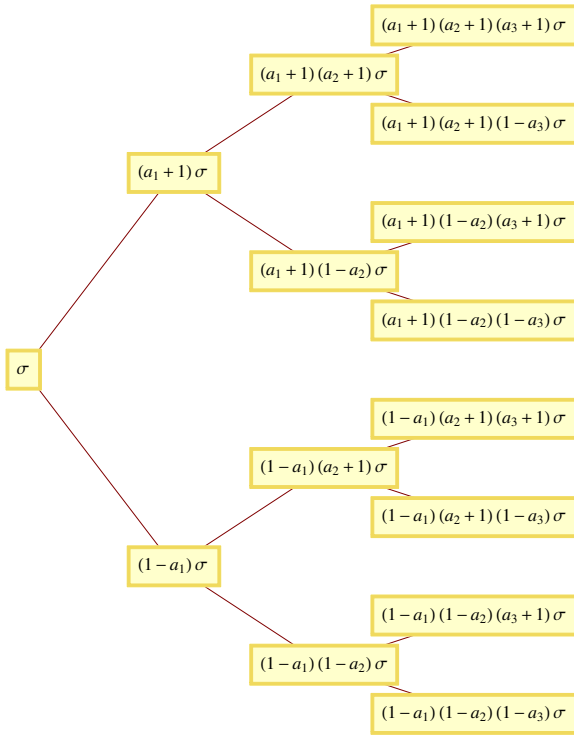


Figure 17.2: *Three levels of error rates for σ following a multiplicative process*

The second order stochastic standard deviation:

$$f(x)_2 = \frac{1}{4}\left(\phi\left(\mu, \sigma(1+a(1)(1+a(2))), x\right) + \phi\left(\mu, \sigma(1-a(1)(1+a(2))), x\right) + \phi\left(\mu, \sigma(1+a(1)(1-a(2)), x\right) + \phi\left(\mu, \sigma(1-a(1)(1-a(2))), x\right)\right) \tag{17.3}$$

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 $M^N\ (2^N \times 1)$

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

and $\mathbf{T}_{i,j}$ the element of i^{th}line and j^{th}column of the matrix of the exhaustive combination of n-Tuples of the set $\{-1, 1\}$,that is the sequences of n length $(1, 1, 1, ...)$ representing all combinations of 1 and -1.

for N=3,

$$T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{pmatrix}$$

and

$$M^3 = \begin{pmatrix} (1 - a(1))(1 - a(2))(1 - a(3)) \\ (1 - a(1))(1 - a(2))(a(3) + 1) \\ (1 - a(1))(a(2) + 1)(1 - a(3)) \\ (1 - a(1))(a(2) + 1)(a(3) + 1) \\ (a(1) + 1)(1 - a(2))(1 - a(3)) \\ (a(1) + 1)(1 - a(2))(a(3) + 1) \\ (a(1) + 1)(a(2) + 1)(1 - a(3)) \\ (a(1) + 1)(a(2) + 1)(a(3) + 1) \end{pmatrix}$$

So $M_1^3 = \{(1 - a(1))(1 - a(2))(1 - a(3))\}$, etc.

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.

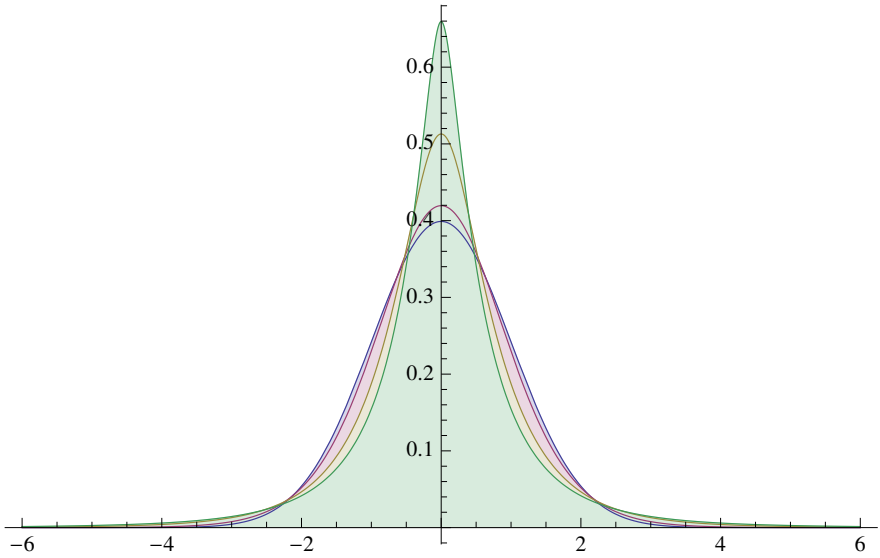


Figure 17.3: *Thicker tails (higher peaks) for higher values of N; here* $N = 0, 5, 10, 25, 50$, *all values of* $a=\frac{1}{10}$

Next let us consider the different regimes for higher order errors.

REGIME 1 (EXPLOSIVE): CASE OF A CONSTANT PARAMETER *a*

Special case of constant ***a***: Assume that a(1)=a(2)=...a(N)=a, i.e. the case of flat proportional error rate a. The Matrix M collapses into a conventional binomial tree for the dispersion at the level N.

$$f(x|\mu,\sigma,M,N) = 2^{-N}\sum_{j=0}^{N}\binom{N}{j}\phi\left(\mu,\sigma(a+1)^j(1-a)^{N-j},x\right) \tag{17.4}$$

Because of the linearity of the sums, when a is constant, we can use the binomial distribution as weights for the moments (note again the artificial effect of constraining the first moment μ in the analysis to a set, certain, and known *a priori*).

$$\begin{pmatrix} & \text{Moment} \\ 1 & \mu \\ 2 & \sigma^2\left(a^2+1\right)^N+\mu^2 \\ 3 & 3\mu\sigma^2\left(a^2+1\right)^N+\mu^3 \\ 4 & 6\mu^2\sigma^2\left(a^2+1\right)^N+\mu^4+3\left(a^4+6a^2+1\right)^N\sigma^4 \end{pmatrix}$$

Note again the oddity that in spite of the explosive nature of higher moments, the expectation of the absolute value of x is both independent of a and N, since the perturbations of σ do not affect the first absolute moment = $\sqrt{\frac{2}{\pi}}\sigma$ (that is, the initial assumed σ). The situation would be different under addition of x.

Every recursion multiplies the variance of the process by $(1 + a^2)$. The process is similar to a stochastic volatility model, with the standard deviation (not the variance) following a lognormal distribution, the volatility of which grows with M, hence will reach infinite variance at the limit.

Consequences

For a constant $a > 0$, and in the more general case with variable a where a(n) $\geq$ a(n-1), the moments explode.

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

B- In these conditions, we need to use power laws for epistemic reasons, or, at least, distributions outside the $\mathcal{L}^2$ norm, regardless of observations of past data.

Note that we need an *a priori* reason (in the philosophical sense) to cutoff the N somewhere, hence bound the expansion of the second moment.

Convergence to Properties Similar to Power Laws

We can see on the example next Log-Log plot (Figure 1) how, at higher orders of stochastic volatility, with equally proportional stochastic coefficient, (where a(1)=a(2)=...=a(N)= $\frac{1}{10}$) how the density approaches that of a Power Law (just like the Lognormal distribution at higher variance), as shown in flatter density on the LogLog plot. The probabilities keep rising in the tails as we add layers of uncertainty until they seem to reach the boundary of the power law, while ironically the first moment remains invariant.

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

17.1.3 Effect on Small Probabilities

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

Take the exceedant probability,that is, the probability of exceeding K, given N, for parameter a constant:

$$P > K|N = \sum_{j=0}^{N} 2^{-N-1} \binom{N}{j} \operatorname{erfc}\left(\frac{K}{\sqrt{2}\sigma(a+1)^j(1-a)^{N-j}}\right) \quad (17.5)$$

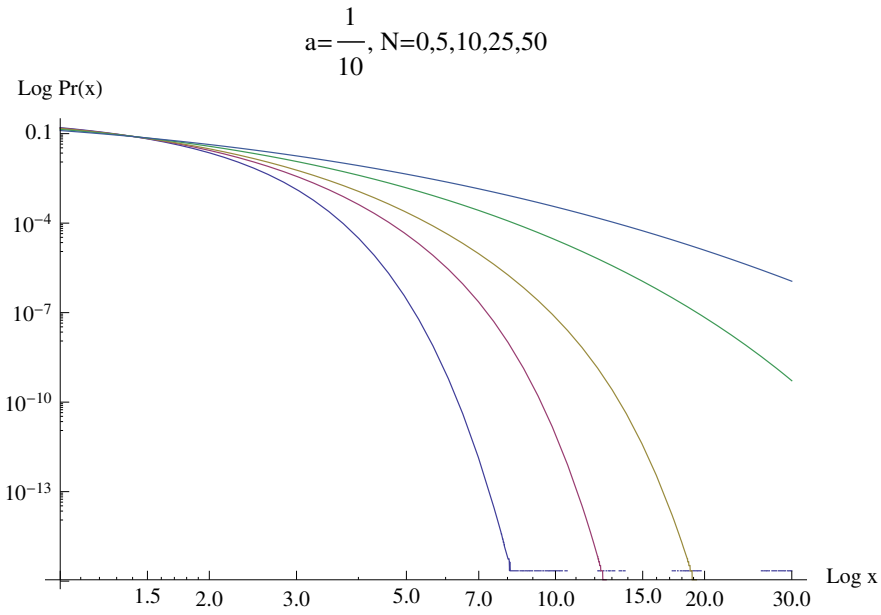


Figure 17.4: *LogLog Plot of the probability of exceeding x showing power law-style flattening as N rises. Here all values of a= 1/10*

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

Convexity effect The next Table shows the ratio of exceedant probability under different values of N divided by the probability in the case of a standard Gaussian.

Table 17.1: *Case of* $a = \frac{1}{10}$

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

Table 17.2: *Case of* $a = \frac{1}{100}$

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

17.2 REGIME 2: CASES OF DECAYING PARAMETERS $a(n)$

As we said, we may have (actually we need to have) *a priori* reasons to decrease the parameter a or stop N somewhere. When the higher order of a(i) decline, then the moments tend to be capped (the inherited tails will come from the lognormality of σ).

17.2.1 Regime 2-a;"Bleed" of Higher Order Error

Take a "bleed" of higher order errors at the rate λ, o$\leq \lambda <$ 1 , such as $a(N) = \lambda$ a(N-1), hence a(N) =λ^N a(1), with a(1) the conventional intensity of stochastic standard deviation. Assume μ=0.

With N=2 , the second moment becomes:

$$M_2(2) = \left(a(1)^2 + 1\right) \sigma^2 \left(a(1)^2 \lambda^2 + 1\right)$$

With N=3,

$$M_2(3) = \sigma^2 \left(1 + a(1)^2\right) \left(1 + \lambda^2 a(1)^2\right) \left(1 + \lambda^4 a(1)^2\right)$$

finally, for the general N:

$$M_3(N) = \left(a(1)^2 + 1\right) \sigma^2 \prod_{i=1}^{N-1} \left(a(1)^2 \lambda^{2i} + 1\right) \tag{17.6}$$

We can reexpress 17.6 using the Q-Pochhammer symbol $(a;q)_N = \prod_{i=1}^{N-1} \left(1 - aq^i\right)$

$$M_2(N) = \sigma^2 \left(-a(1)^2; \lambda^2\right)_N$$

Which allows us to get to the limit

$$\lim_{N \to \infty} M_2(N) = \sigma^2 \frac{\left(\lambda^2; \lambda^2\right)_2 \left(a(1)^2; \lambda^2\right)_\infty}{\left(\lambda^2 - 1\right)^2 \left(\lambda^2 + 1\right)}$$

As to the fourth moment:

By recursion:

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

$$M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2} - 3\right) a(1)^2; \lambda^2\right)_N \left(-\left(3 + 2\sqrt{2}\right) a(1)^2; \lambda^2\right)_N \tag{17.7}$$

$$\lim_{N\to\infty} M_4(N) = 3\sigma^4 \left(\left(2\sqrt{2}-3\right) a(1)^2;\lambda^2\right)_\infty \left(-\left(3+2\sqrt{2}\right) a(1)^2;\lambda^2\right)_\infty \qquad (17.8)$$

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

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

For N recursions,

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

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

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

$$A^N = \left(a^j\right)_{j=1,...N}$$

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.

17.3 LIMIT DISTRIBUTION

See Taleb and Cirillo **??** for the treatment of the limit distribution which will be a lognormal under the right conditions. In fact lognormal approximations work well when errors on errors are in constant proportion.

18 STOCHASTIC TAIL EXPONENT FOR ASYMMETRIC POWER LAWS†

We examine random variables in the power law/slowly varying class with stochastic tail exponent , the exponent α having its own distribution. We show the effect of stochasticity of α on the expectation and higher moments of the random variable. For instance, the moments of a right-tailed or right-asymmetric variable, when finite, increase with the variance of α; those of a left-asymmetric one decreases. The same applies to conditional shortfall (CVar), or mean-excess functions.

We prove the general case and examine the specific situation of lognormally distributed $\alpha \in [b, \infty), b > 1$.

The stochasticity of the exponent induces a significant bias in the estimation of the mean and higher moments in the presence of data uncertainty. This has consequences on sampling error as uncertainty about α translates into a higher expected mean.

The bias is conserved under summation, even upon large enough a number of summands to warrant convergence to the stable distribution. We establish inequalities related to the asymmetry.

We also consider the situation of capped power laws (i.e. with compact support), and apply it to the study of violence by Cirillo and Taleb (2016). We show that uncertainty concerning the historical data increases the true mean.

Research chapter.

Conference: Extremes and Risks in Higher Dimensions, Lorentz Center, Leiden, The Netherlands, September 2016.

18.1 BACKGROUND

stochastic volatility has been introduced heuristically in mathematical finance by traders looking for biases on option valuation, where a Gaussian distribution is considered to have several possible variances, either locally or at some specific future date. Options far from the money (i.e. concerning tail events) increase in value with uncertainty on the variance of the distribution, as they are convex to the standard deviation.

This led to a family of models of Brownian motion with stochastic variance (see review in Gatheral [101]) and proved useful in tracking the distributions of the underlying and the effect of the nonGaussian character of random processes on functions of the process (such as option prices).

Just as options are convex to the scale of the distribution, we find many situations where expectations are convex to the Power Law tail exponent . This note examines two cases:

- The standard power laws, one-tailed or asymmetric.
- The pseudo-power law, where a random variable appears to be a Power law but has compact support, as in the study of violence [46] where wars have the number of casualties capped at a maximum value.

18.2 ONE TAILED DISTRIBUTIONS WITH STOCHASTIC ALPHA

18.2.1 General Cases

Definition 18.1
Let X be a random variable belonging to the class of distributions with a "power law" right tail, that is support in $[x_0, +\infty), \in \mathbb{R}$*:*

Subclass $\mathfrak{P}_1$*:*

$$\{X : \mathbb{P}(X > x) = L(x)x^{-\alpha}, \frac{\partial^q L(x)}{\partial x^q} = 0 \text{ for } q \geq 1\} \tag{18.1}$$

We note that x_o can be negative by shifting, so long as $x_0 > -\infty$.

Class $\mathfrak{P}$*:*

$$\{X : \mathbb{P}(X > x) \sim L(x)\, x^{-\alpha}\} \tag{18.2}$$

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

We further assume that:

$$\lim_{x\to\infty} L'(x)\, x = 0 \tag{18.3}$$

$$\lim_{x\to\infty} L''(x)\, x = 0 \tag{18.4}$$

We have

$$\mathfrak{P}_1 \subset \mathfrak{P}$$

We note that the first class corresponds to the Pareto distributions (with proper shifting and scaling), where L is a constant and $\mathfrak{P}$ to the more general one-sided power laws.

18.2.2 Stochastic Alpha Inequality

Throughout the rest of the paper we use for notation X' for the stochastic alpha version of X, the constant α case.

Proposition 18.1
Let $p = 1, 2, ...$, X' be the same random variable as X above in $\mathfrak{P}_1$ (the one-tailed regular variation class), with $x_0 \geq 0$, except with stochastic α with all realizations $> p$ that preserve the mean $\bar{\alpha}$,

$$\mathbb{E}(X'^p) \geq \mathbb{E}(X^p).$$

Proposition 18.2
Let K be a threshold. With X in the $\mathfrak{P}$ class, we have the expected conditional shortfall (CVar):

$$\lim_{K\to\infty} \mathbb{E}(X'|_{X'>K}) \geq \lim_{K\to\infty} \mathbb{E}(X|_{X>K}).$$

The sketch of the proof is as follows.

We remark that $\mathbb{E}(X^p)$ is convex to α, in the following sense. Let X_{α_i} be the random variable distributed with constant tail exponent α_i, with $\alpha_i > p, \forall i$, and ω_i be the normalized positive weights: $\sum_i \omega_i = 1$, $0 \leq |\omega_i| \leq 1$, $\sum_i \omega_i \alpha_i = \bar{\alpha}$. By Jensen's inequality:

$$\omega_i \sum_i \mathbb{E}(X^p_{\alpha_i}) \geq \mathbb{E}(\sum_i (\omega_i X^p_{\alpha_i})).$$

As the classes are defined by their survival functions, we first need to solve for the corresponding density: $\varphi(x) = \alpha x^{-\alpha-1} L(x,\alpha) - x^{-\alpha} L^{(1,0)}(x,\alpha)$ and get the normalizing constant.

$$L(x_0,\alpha) = x_0^\alpha - \frac{2x_0 L^{(1,0)}(x_0,\alpha)}{\alpha - 1} - \frac{2x_0^2 L^{(2,0)}(x_0,\alpha)}{(\alpha-1)(\alpha-2)}, \tag{18.5}$$

$\alpha \neq 1, 2$ when the first and second derivative exist, respectively. The slot notation $L^{(p,0)}(x_0,\alpha)$ is short for $\frac{\partial^p L(x,\alpha)}{\partial x^p}|_{x=x_0}$.

By the Karamata representation theorem, [22],[243], a function L on $[x_0, +\infty)$ is slowly moving (Definition) if and only if it can be written in the form

$$L(x) = \exp\left(\int_{x_0}^{x} \frac{\epsilon(t)}{t}\, dt\right) + \eta(x)$$

where $\eta(.)$ is a bounded measurable function converging to a finite number as $x \to +\infty$, and $\epsilon(x)$ is a bounded measurable function converging to zero as $x \to +\infty$.

Accordingly, $L'(x)$ goes to 0 as $x \to \infty$. (We further assumed in 18.3 and 18.4 that $L'(x)$ goes to o faster than x and $L''(x)$ goes to o faster than x^2). Integrating by parts,

$$\mathbb{E}(X^p) = x_0^p + p \int_{x_0}^{\infty} x^{p-1}\, \mathrm{d}\bar{F}(x)$$

where $\bar{F}$ is the survival function in Eqs. 23.1 and 18.2. Integrating by parts three additional times and eliminating derivatives of $L(.)$ of higher order than 2:

$$\mathbb{E}(X^p) = \frac{x_0^{p-\alpha} L(x_0, \alpha)}{p-\alpha} - \frac{x_0^{p-\alpha+1} L^{(1,0)}(x_0, \alpha)}{(p-\alpha)(p-\alpha+1)} + \frac{x_0^{p-\alpha+2} L^{(2,0)}(x_0, \alpha)}{(p-\alpha)(p-\alpha+1)(p-\alpha+2)} \tag{18.6}$$

which, for the special case of X in $\mathfrak{P}_1$ reduces to:

$$\mathbb{E}(X^p) = x_0^p \frac{\alpha}{\alpha - p} \tag{18.7}$$

As to Proposition 2, we can approach the proof from the property that $\lim_{x\to\infty} L'(x) = 0$. This allows a proof of var der Mijk's law that Paretian inequality is invariant to the threshold in the tail, that is $\frac{\mathbb{E}(X|_{X>K})}{K}$ converges to a constant as $K \to +\infty$. Equation 18.6 presents the exact conditions on the functional form of $L(x)$ for the convexity to extend to sub-classes between $\mathfrak{P}_1$ and $\mathfrak{P}$.

Our results hold to distributions that are transformed by shifting and scaling, of the sort:

$x \mapsto x - \mu + x_0$ (Pareto II), or with further transformations to Pareto types II and IV.

We note that the representation $\mathfrak{P}_1$ uses the same parameter, x_0, for both scale and minimum value, as a simplification.

We can verify that the expectation from Eq. 18.7 is convex to α: $\frac{\partial \mathbb{E}(X^p)}{\partial \alpha^2} = x_0^p \frac{2}{(\alpha-1)^3}$.

18.2.3 Approximations for the Class $\mathfrak{P}$

For $\mathfrak{P} \setminus \mathfrak{P}_1$, our results hold when we can write an approximation the expectation of X as a constant multiplying the integral of $x^{-\alpha}$, namely

$$\mathbb{E}(X) \approx k \frac{\nu(\alpha)}{\alpha - 1} \tag{18.8}$$

where k is a positive constant that does not depend on α and $\nu(.)$ is approximated by a linear function of α (plus a threshold). The expectation will be convex to α.

Example: Student T Distribution For the Student T distribution with tail α, the "sophisticated" slowly varying function in common use for symmetric power laws in quantitative finance, the half-mean or the mean of the one-sided distribution (i.e. with support on $\mathbb{R}^+$ becomes

$$2\nu(\alpha) = 2\frac{\sqrt{\alpha}\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha}{2}\right)} \approx \alpha\frac{(1+\log(4))}{\pi},$$

where $\Gamma(.)$ is the gamma function.

18.3 SUMS OF POWER LAWS

As we are dealing from here on with convergence to the stable distribution, we consider situations of $1 < \alpha < 2$, hence $p = 1$ and will be concerned solely with the mean.

We observe that the convexity of the mean is invariant to summations of Power Law distributed variables as X above. The Stable distribution has a mean that in conventional parameterizations does not appear to depend on α –but in fact depends on it.

Let Y be distributed according to a Pareto distribution with density $f(y) \triangleq \alpha\lambda^\alpha y^{-\alpha-1}, y \geq \lambda > 0$ and with its tail exponent $1 < \alpha < 2$. Now, let $Y_1, Y_2, \dots Y_n$ be identical and independent copies of Y. Let $\chi(t)$ be the characteristic function for $f(y)$. We have $\chi(t) = \alpha(-it)^\alpha\Gamma(-\alpha, -it)$, where $\gamma(.,.)$ is the incomplete gamma function. We can get the mean from the characteristic function of the average of n summands $\frac{1}{n}(Y_1 + Y_2 + ...Y_n)$, namely $\chi(\frac{t}{n})^n$. Taking the first derivative:

$$-i\frac{\partial\chi(\frac{t}{n})^n}{\partial t} = (-i)^{\alpha(n-1)}n^{1-\alpha n}\alpha^n\lambda^{\alpha(n-1)}t^{\alpha(n-1)-1}\Gamma\left(-\alpha, -\frac{it\lambda}{n}\right)^{n-1}\left((-i)^\alpha\alpha\lambda^\alpha t^\alpha\Gamma\left(-\alpha, -\frac{it\lambda}{n}\right) - n^\alpha e^{\frac{i\lambda t}{n}}\right) \tag{18.9}$$

and

$$\lim_{n\to\infty} -i\frac{\partial\chi(\frac{t}{n})^n}{\partial t}\bigg|_{t=0} = \lambda\frac{\alpha}{\alpha-1} \tag{18.10}$$

Thus we can see how the converging asymptotic distribution for the average will have for mean the scale times $\frac{\alpha}{\alpha-1}$, which does not depends on n.

Let $\chi^S(t)$ be the characteristic function of the corresponding stable distribution $S_{\alpha,\beta,\mu,\sigma}$, from the distribution of an infinitely summed copies of Y. By the Lévy continuity theorem, we have

- $\frac{1}{n}\Sigma_{i\leq n} Y_i \xrightarrow{\mathcal{D}} S$, with distribution $S_{\alpha,\beta,\mu,\sigma}$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution

  and

- $\chi^S(t) = \lim_{n\to\infty} \chi(t/n)^n$

are equivalent.

So we are dealing with the standard result [267],[206], for exact Pareto sums [264], replacing the conventional μ with the mean from above:

$$\chi^S(t) = \exp\left(i\left(\lambda\frac{\alpha t}{\alpha-1} + |t|^\alpha\left(\beta\tan\left(\frac{\pi\alpha}{2}\right)\operatorname{sgn}(t)+i\right)\right)\right).$$

18.4 ASYMMETRIC STABLE DISTRIBUTIONS

We can verify by symmetry that, effectively, flipping the distribution in subclasses $\mathfrak{P}_1$ and $\mathfrak{P}_2$ around y_0 to make it negative yields a negative value of the mean d higher moments, hence degradation from stochastic α.

The central question becomes:

Remark 16: Preservation of Asymmetry

A normalized sum in $\mathfrak{P}_1$ one-tailed distribution with expectation that depends on α of the form in Eq. 18.8 will necessarily converge in distribution to an asymmetric stable distribution $S_{\alpha,\beta,\mu,1}$, with $\beta \neq 0$.

Remark 17

Let Y' be Y under mean-preserving stochastic α. The convexity effect becomes

$$sgn\left(\mathbb{E}(Y') - \mathbb{E}(Y)\right) = sgn(\beta).$$

The sketch of the proof is as follows. Consider two slowly varying functions as in 23.1, each on one side of the tails. We have $L(y) = \mathbb{1}_{y<y_\theta} L^-(y) + \mathbb{1}_{y\geq y_\theta} L^+(y)$:

$$\begin{cases} L^+(y), L:[y_\theta,+\infty], & \lim_{y\to\infty} L^+(y) = c \\ \\ L^-(y), L:[-\infty,y_\theta], & \lim_{y\to-\infty} L^-(y) = d. \end{cases}$$

From [206],

$$\text{if } \begin{cases} \mathbb{P}(X>x) \sim cx^{-\alpha}, x\to+\infty \\ \\ \mathbb{P}(X<x) \sim d|x|^{-\alpha}, x\to+\infty, \end{cases} \text{ then } Y \text{ converges in distribution to } S_{\alpha,\beta,\mu,1}$$

with the coefficient $\beta = \frac{c-d}{c+d}$.

We can show that the mean can be written as $(\lambda_+ - \lambda_-)\frac{\alpha}{\alpha-1}$ where:

$$\lambda_+ \geq \lambda_- \text{ if } \int_{y_\theta}^{\infty} L^+(y)\mathrm{d}y, \geq \int_{-\infty}^{y_\theta} L^-(y)\mathrm{d}y$$

18.5 PARETO DISTRIBUTION WITH LOGNORMALLY DISTRIBUTED α

Now assume α is following a shifted Lognormal distribution with mean α_0 and minimum value b, that is, $\alpha - b$ follows a Lognormal $\mathcal{L}\left(\log(\alpha_0) - \frac{\sigma^2}{2}, \sigma\right)$. The parameter b allows us to work with a lower bound on the tail exponent in order to satisfy finite expectation. We know that the tail exponent will eventually converge to b but the process may be quite slow.

Proposition 18.3
Assuming finite expectation for X' and for exponent the lognormally distributed shifted variable $\alpha - b$ with law $\mathcal{L}\left(\log(\alpha_0) - \frac{\sigma^2}{2}, \sigma\right)$, $b \geq 1$ mininum value for α, and scale λ:

$$\mathbb{E}(Y') = \mathbb{E}(Y) + \lambda \frac{(e^{\sigma^2} - b)}{\alpha_0 - b} \tag{18.11}$$

We need $b \geq 1$ to avoid problems of infinite expectation.

Let $\phi(y, \alpha)$ be the density with stochastic tail exponent. With $\alpha > 0, \alpha_0 > b, b \geq 1, \sigma > 0, Y \geq \lambda > 0$,

$$\begin{aligned}
\mathbb{E}(Y) &= \int_b^{\infty} \int_L^{\infty} y\phi(y;\alpha)\,\mathrm{d}y\,\mathrm{d}\alpha \\
&= \int_b^{\infty} \lambda \frac{\alpha}{\alpha - 1} \frac{1}{\sqrt{2\pi}\sigma(\alpha - b)} \\
&\quad \exp\left(- \frac{\left(\log(\alpha - b) - \log(\alpha_0 - b) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2} \right) \mathrm{d}\alpha \\
&= \frac{\lambda\left(\alpha_0 + e^{\sigma^2} - b\right)}{\alpha_0 - b}.
\end{aligned} \tag{18.12}$$

Approximation of the Density

With $b = 1$ (which is the lower bound for b),we get the density with stochastic α:

$$\phi(y;\alpha_0,\sigma) = \lim_{k\to\infty} \frac{1}{Y^2} \sum_{i=0}^{k} \frac{1}{i!} L(\alpha_0 - 1)^i e^{\frac{1}{2}i(i-1)\sigma^2} (\log(\lambda) - \log(y))^{i-1} (i + \log(\lambda) - \log(y)) \tag{18.13}$$

This result is obtained by expanding α around its lower bound b (which we simplified to $b = 1$) and integrating each summand.

18.6 PARETO DISTRIBUTION WITH GAMMA DISTRIBUTED ALPHA

Proposition 18.4
Assuming finite expectation for X' scale λ, and for exponent a gamma distributed shifted variable $\alpha - 1$ with law $\varphi(.)$, mean α_0 and variance s^2, all values for α greater than 1:

$$\mathbb{E}(X') = \mathbb{E}(X') + \frac{s^2}{(\alpha_0 - 1)(\alpha_0 - s - 1)(\alpha_0 + s - 1)} \tag{18.14}$$

Proof.

$$\varphi(\alpha) = \frac{e^{-\frac{(\alpha-1)(\alpha_0-1)}{s^2}} \left(\frac{s^2}{(\alpha-1)(\alpha_0-1)}\right)^{-\frac{(\alpha_0-1)^2}{s^2}}}{(\alpha-1)\Gamma\left(\frac{(\alpha_0-1)^2}{s^2}\right)}, \quad \alpha > 1 \tag{18.15}$$

$$\int_1^\infty \alpha \lambda^\alpha x^{-\alpha-1} \varphi(\alpha)\, d\alpha \tag{18.16}$$

$$= \int_1^\infty \frac{\alpha \left(e^{-\frac{(\alpha-1)(\alpha_0-1)}{s^2}} \left(\frac{s^2}{(\alpha-1)(\alpha_0-1)}\right)^{-\frac{(\alpha_0-1)^2}{s^2}}\right)}{(\alpha-1)\left((\alpha-1)\Gamma\left(\frac{(\alpha_0-1)^2}{s^2}\right)\right)} d\alpha$$

$$= \frac{1}{2}\left(\frac{1}{\alpha_0 + s - 1} + \frac{1}{\alpha_0 - s - 1} + 2\right)$$

□

18.7 THE BOUNDED POWER LAW IN CIRILLO AND TALEB (2016)

In [46] and [45], the studies make use of bounded power laws, applied to violence and operational risk, respectively. Although with $\alpha < 1$ the variable Z has finite expectations owing to the upper bound.

The methods offered were a smooth transformation of the variable as follows: we start with $z \in [L, H), L > 0$ and transform it into $x \in [L, \infty)$, the latter legitimately being Power Law distributed.

So the smooth logarithmic transformation):

$$x = \varphi(z) = L - H \log\left(\frac{H - z}{H - L}\right),$$

and

$$f(x) = \frac{\left(\frac{x-L}{\alpha\sigma} + 1\right)^{-\alpha-1}}{\sigma}.$$

We thus get the distribution of Z which will have a finite expectation for all positive values of α.

$$\frac{\partial^2 \mathbb{E}(Z)}{\partial \alpha^2} = \frac{1}{H^3}(H-L)\left(e^{\frac{\alpha\sigma}{H}}\left(2H^3 G^{4,0}_{3,4}\left(\frac{\alpha\sigma}{H}\middle| \begin{matrix}\alpha+1,\alpha+1,\alpha+1\\ 1,\alpha,\alpha,\alpha\end{matrix}\right)\right.\right.$$
$$-2H^2(H+\sigma)G^{3,0}_{2,3}\left(\frac{\alpha\sigma}{H}\middle| \begin{matrix}\alpha+1,\alpha+1\\ 1,\alpha,\alpha\end{matrix}\right) \quad (18.17)$$
$$\left.\left.+\sigma\left(\alpha\sigma^2+(\alpha+1)H^2+2\alpha H\sigma\right)E_\alpha\left(\frac{\alpha\sigma}{H}\right)\right)-H\sigma(H+\sigma)\right)$$

which appears to be positive in the range of numerical perturbations in [46].[3] At such a low level of α, around $\frac{1}{2}$, the expectation is extremely convex and the bias will be accordingly extremely pronounced.

This convexity has the following practical implication. Historical data on violence over the past two millennia, is fundamentally unreliable [46]. Hence an imprecision about the tail exponent , from errors embedded in the data, need to be present in the computations. The above shows that uncertainty about α, is more likely to make the "true" statistical mean (that is the mean of the process as opposed to sample mean) higher than lower, hence supports the statement that more uncertainty increases the estimation of violence.

18.8 ADDITIONAL COMMENTS

The bias in the estimation of the mean and shortfalls from uncertainty in the tail exponent can be added to analyses where data is insufficient, unreliable, or simply prone to forgeries.

In additional to statistical inference, these result can extend to processes, whether a compound Poisson process with power laws subordination [214] (i.e. a Poisson arrival time and a jump that is Power Law distributed) or a Lévy process. The latter can be analyzed by considering successive "slice distributions" or discretization of the process [49]. Since the expectation of a sum of jumps is the sum of expectation, the same convexity will appear as the one we got from Eq. 18.8.

18.9 ACKNOWLEDGMENTS

Marco Avellaneda, Robert Frey, Raphael Douady, Pasquale Cirillo.

3 $G^{4,0}_{3,4}\left(\frac{\alpha\sigma}{H}\middle| \begin{matrix}\alpha+1,\alpha+1,\alpha+1\\ 1,\alpha,\alpha,\alpha\end{matrix}\right)$ is the Meijer G function.

19 META-DISTRIBUTION OF P-VALUES AND P-HACKING‡

We present an exact probability distribution (meta-distribution) for p-values across ensembles of statistically identical phenomena, as well as the distribution of the minimum p-value among m independents tests. We derive the distribution for small samples $2 < n \leq n^* \approx 30$ as well as the limiting one as the sample size n becomes large. We also look at the properties of the "power" of a test through the distribution of its inverse for a given p-value and parametrization.

P-values are shown to be extremely skewed and volatile, regardless of the sample size n, and vary greatly across repetitions of exactly same protocols under identical stochastic copies of the phenomenon; such volatility makes the minimum p value diverge significantly from the "true" one. Setting the power is shown to offer little remedy unless sample size is increased markedly or the p-value is lowered by at least one order of magnitude.

The formulas allow the investigation of the stability of the reproduction of results and "p-hacking" and other aspects of meta-analysis –including a metadistribution of p-hacked results.

From a probabilistic standpoint, neither a p-value of .05 nor a "power" at .9 appear to make the slightest sense.

Assume that we know the "true" p-value, p_s, what would its realizations look like across various attempts on statistically identical copies of the phenomena? By true value p_s, we mean its expected value by the law of large numbers across an m ensemble of possible samples for the phenomenon under scrutiny, that is $\frac{1}{m}\sum_{\leq m} p_i \xrightarrow{P} p_s$ (where $\xrightarrow{P}$ denotes convergence in probability). A similar convergence argument can be also made for the corresponding "true median" p_M. The main result of the paper is that the the distribution of n small samples can be made explicit (albeit with special inverse functions), as well as its parsimonious limiting

Research chapter.

one for n large, with no other parameter than the median value p_M. We were unable to get an explicit form for p_s but we go around it with the use of the median. Finally, the distribution of the minimum p-value under can be made explicit, in a parsimonious formula allowing for the understanding of biases in scientific studies.

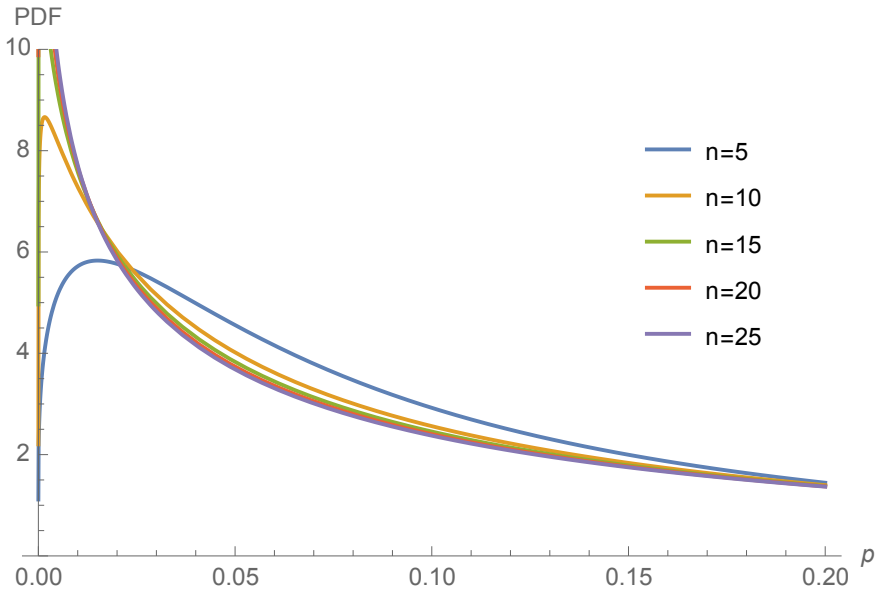


Figure 19.1: *The different values for Equ. 19.1 showing convergence to the limiting distribution.*

It turned out, as we can see in Fig. 19.2 the distribution is extremely asymmetric (right-skewed), to the point where 75% of the realizations of a "true" p-value of .05 will be <.05 (a borderline situation is 3× as likely to pass than fail a given protocol), and, what is worse, 60% of the true p-value of .12 will be below .05.

Although with compact support, the distribution exhibits the attributes of extreme fat-tailedness. For an observed p-value of, say, .02, the "true" p-value is likely to be >.1 (and very possibly close to .2), with a standard deviation >.2 (sic) and a mean deviation of around .35 (sic, sic). Because of the excessive skewness, measures of dispersion in $\mathcal{L}^1$ and $\mathcal{L}^2$ (and higher norms) vary hardly with p_s, so the standard deviation is not proportional, meaning an in-sample .01 p-value has a significant probability of having a true value $> .3$.

So clearly we don't know what we are talking about when we talk about p-values.

Earlier attempts for an explicit meta-distribution in the literature were found in [130] and [205], though for situations of Gaussian subordination and less parsimonious parametrization. The severity of the problem of *significance of the so-called "statistically significant"* has been discussed in [104] and offered a remedy via Bayesian

methods in [136], which in fact recommends the same tightening of standards to p-values $\approx .01$. But the gravity of the extreme skewness of the distribution of p-values is only apparent when one looks at the meta-distribution.

For notation, we use n for the sample size of a given study and m the number of trials leading to a p-value.

19.1 PROOFS AND DERIVATIONS

Proposition 19.1
Let P be a random variable $\in [0,1]$) corresponding to the sample-derived one-tailed p-value from the paired T-test statistic (unknown variance) with median value $\mathbb{M}(P) = p_M \in [0,1]$ derived from a sample of n size. The distribution across the ensemble of statistically identical copies of the sample has for PDF

$$\varphi(p; p_M) = \begin{cases} \varphi(p; p_M)_L & \text{for } p < \frac{1}{2} \\ \varphi(p; p_M)_H & \text{for } p > \frac{1}{2} \end{cases}$$

$$\varphi(p; p_M)_L = \lambda_p^{\frac{1}{2}(-n-1)} \sqrt{-\frac{\lambda_p\left(\lambda_{p_M}-1\right)}{\left(\lambda_p-1\right)\lambda_{p_M} - 2\sqrt{\left(1-\lambda_p\right)\lambda_p}\sqrt{\left(1-\lambda_{p_M}\right)\lambda_{p_M}}+1}} \left(\frac{1}{\frac{1}{\lambda_p} - \frac{2\sqrt{1-\lambda_p}\sqrt{\lambda_{p_M}}}{\sqrt{\lambda_p}\sqrt{1-\lambda_{p_M}}} + \frac{1}{1-\lambda_{p_M}} - 1}\right)^{n/2}$$

$$\varphi(p; p_M)_H = \left(1-\lambda'_p\right)^{\frac{1}{2}(-n-1)} \left(\frac{\left(\lambda'_p-1\right)\left(\lambda_{p_M}-1\right)}{\lambda'_p\left(-\lambda_{p_M}\right) + 2\sqrt{\left(1-\lambda'_p\right)\lambda'_p}\sqrt{\left(1-\lambda_{p_M}\right)\lambda_{p_M}}+1}\right)^{\frac{n+1}{2}} \tag{19.1}$$

where $\lambda_p = I^{-1}_{2p}\left(\frac{n}{2}, \frac{1}{2}\right)$, $\lambda_{p_M} = I^{-1}_{1-2p_M}\left(\frac{1}{2}, \frac{n}{2}\right)$, $\lambda'_p = I^{-1}_{2p-1}\left(\frac{1}{2}, \frac{n}{2}\right)$, and $I^{-1}_{(.)}(.,.)$ is the inverse beta regularized function.

Remark 18

For $p=\frac{1}{2}$ the distribution doesn't exist in theory, but does in practice and we can work around it with the sequence $p_{m_k} = \frac{1}{2} \pm \frac{1}{k}$, as in the graph showing a convergence to the Uniform distribution on $[0,1]$ in Figure 19.3. Also note that what is called the "null" hypothesis is effectively a set of measure 0.

Proof. Let Z be a random normalized variable with realizations ζ, from a vector $\vec{v}$ of n realizations, with sample mean m_v, and sample standard deviation s_v, $\zeta = \frac{m_v - m_h}{\frac{s_v}{\sqrt{n}}}$ (where m_h is the level it is tested against), hence assumed to $\sim$ Student T with n degrees of freedom, and, crucially, supposed to deliver a mean of $\bar{\zeta}$,

$$f(\zeta; \bar{\zeta}) = \frac{\left(\frac{n}{(\bar{\zeta}-\zeta)^2+n}\right)^{\frac{n+1}{2}}}{\sqrt{n}B\left(\frac{n}{2}, \frac{1}{2}\right)}$$

where B(.,.) is the standard beta function. Let $g(.)$ be the one-tailed survival function of the Student T distribution with zero mean and n degrees of freedom:

$$g(\zeta) = \mathbb{P}(Z > \zeta) = \begin{cases} \frac{1}{2} I_{\frac{n}{\zeta^2+n}}\left(\frac{n}{2}, \frac{1}{2}\right) & \zeta \geq 0 \\ \frac{1}{2}\left(I_{\frac{\zeta^2}{\zeta^2+n}}\left(\frac{1}{2}, \frac{n}{2}\right) + 1\right) & \zeta < 0 \end{cases}$$

where $I_{(.,.)}$ is the incomplete Beta function.

We now look for the distribution of $g \circ f(\zeta)$. Given that g(.) is a legit Borel function, and naming p the probability as a random variable, we have by a standard result for the transformation:

$$\varphi(p, \bar{\zeta}) = \frac{f\left(g^{(-1)}(p)\right)}{|g'\left(g^{(-1)}(p)\right)|}$$

We can convert $\bar{\zeta}$ into the corresponding median survival probability because of symmetry of Z. Since one half the observations fall on either side of $\bar{\zeta}$, we can ascertain that the transformation is median preserving: $g(\bar{\zeta}) = \frac{1}{2}$, hence $\varphi(p_M, .) = \frac{1}{2}$. Hence we end up having $\{\bar{\zeta} : \frac{1}{2} I_{\frac{n}{\bar{\zeta}^2+n}}\left(\frac{n}{2}, \frac{1}{2}\right) = p_M\}$ (positive case) and $\{\bar{\zeta} : \frac{1}{2}\left(I_{\frac{\zeta^2}{\bar{\zeta}^2+n}}\left(\frac{1}{2}, \frac{n}{2}\right) + 1\right) = p_M\}$ (negative case). Replacing we get Eq.19.1 and Proposition 19.1 is done.

□

We note that n does not increase significance, since p-values are computed from normalized variables (hence the universality of the meta-distribution); a high n corresponds to an increased convergence to the Gaussian. For large n, we can prove the following proposition:

Proposition 19.2
Under the same assumptions as above, the limiting distribution for $\varphi(.)$:

$$\lim_{n\to\infty} \varphi(p; p_M) = e^{-erfc^{-1}(2p_M)\left(erfc^{-1}(2p_M) - 2erfc^{-1}(2p)\right)} \tag{19.2}$$

where erfc(.) is the complementary error function and $erfc(.)^{-1}$ its inverse.

The limiting CDF $\Phi(.)$

$$\Phi(k; p_M) = \frac{1}{2} erfc\left(erf^{-1}(1-2k) - erf^{-1}(1-2p_M)\right) \tag{19.3}$$

Proof. For large n, the distribution of $Z = \frac{m_v}{\frac{s_v}{\sqrt{n}}}$ becomes that of a Gaussian, and the one-tailed survival function $g(.) = \frac{1}{2}\text{erfc}\left(\frac{\zeta}{\sqrt{2}}\right)$, $\zeta(p) \to \sqrt{2}\text{erfc}^{-1}(p)$. □

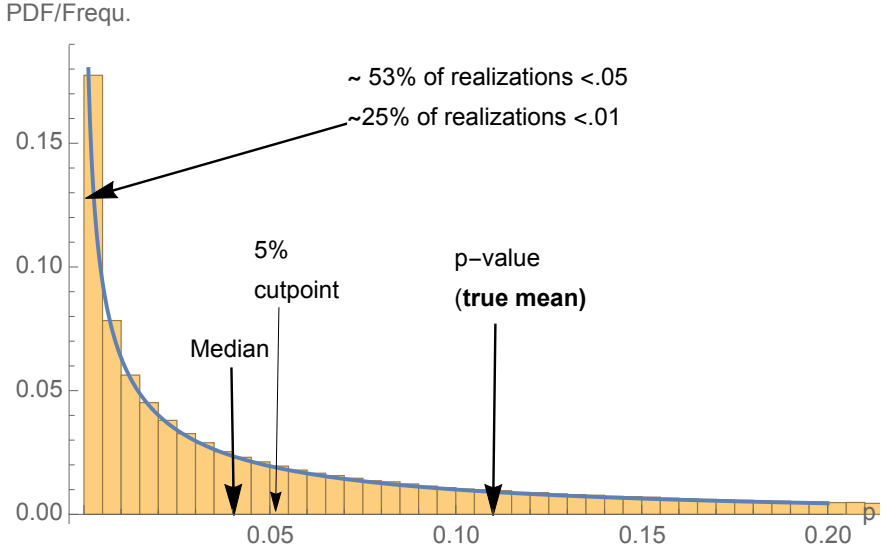


Figure 19.2: *The probability distribution of a one-tailed p-value with expected value .11 generated by Monte Carlo (histogram) as well as analytically with* $\varphi(.)$ *(the solid line). We draw all possible subsamples from an ensemble with given properties. The excessive skewness of the distribution makes the average value considerably higher than most observations, hence causing illusions of "statistical significance".*

This limiting distribution applies for paired tests with known or assumed sample variance since the test becomes a Gaussian variable, equivalent to the convergence of the T-test (Student T) to the Gaussian when n is large.

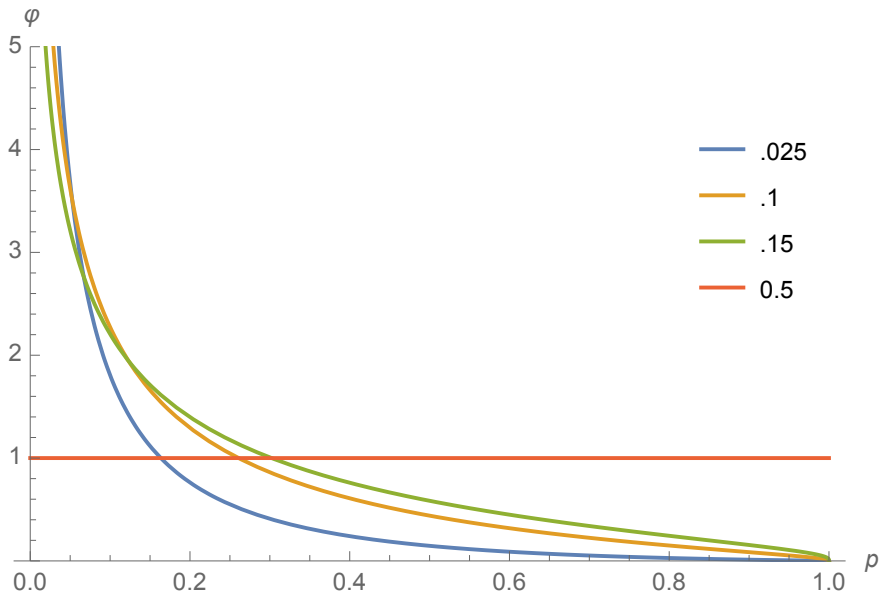


Figure 19.3: *The probability distribution of p at different values of p_M. We observe how $p_M = \frac{1}{2}$ leads to a uniform distribution.*

Remark 19

For values of p close to 0, φ in Equ. 19.2 can be usefully calculated as:

$$\varphi(p; p_M) = \sqrt{2\pi} p_M \sqrt{\log\left(\frac{1}{2\pi p_M^2}\right)} \, e^{\sqrt{-\log\left(2\pi \log\left(\frac{1}{2\pi p^2}\right)\right) - 2\log(p)} \sqrt{-\log\left(2\pi \log\left(\frac{1}{2\pi p_M^2}\right)\right) - 2\log(p_M)}} + O(p^2). \quad (19.4)$$

The approximation works more precisely for the band of relevant values $0 < p < \frac{1}{2\pi}$.

From this we can get numerical results for convolutions of φ using the Fourier Transform or similar methods.

We can and get the distribution of the minimum p-value per m trials across statistically identical situations thus get an idea of "p-hacking", defined as attempts by researchers to get the lowest p-values of many experiments, or try until one of the tests produces statistical significance.

Proposition 19.3
The distribution of the minimum of m observations of statistically identical p-values becomes (under the limiting distribution of proposition 19.2):

$$\varphi_m(p; p_M) = m\, e^{erfc^{-1}(2p_M)\left(2erfc^{-1}(2p) - erfc^{-1}(2p_M)\right)} \left(1 - \frac{1}{2} erfc\left(erfc^{-1}(2p) - erfc^{-1}(2p_M)\right)\right)^{m-1} \tag{19.5}$$

Proof. $P(p_1 > p, p_2 > p, \ldots, p_m > p) = \bigcap_{i=1}^{n} \Phi(p_i) = \bar{\Phi}(p)^m$. Taking the first derivative we get the result. □

Outside the limiting distribution: we integrate numerically for different values of m as shown in figure 19.4. So, more precisely, for m trials, the expectation is calculated as:

$$\mathbb{E}(p_{min}) = \int_0^1 -m\, \varphi(p; p_M) \left(\int_0^p \varphi(u, .)\, \mathrm{d}u \right)^{m-1} \mathrm{d}p$$

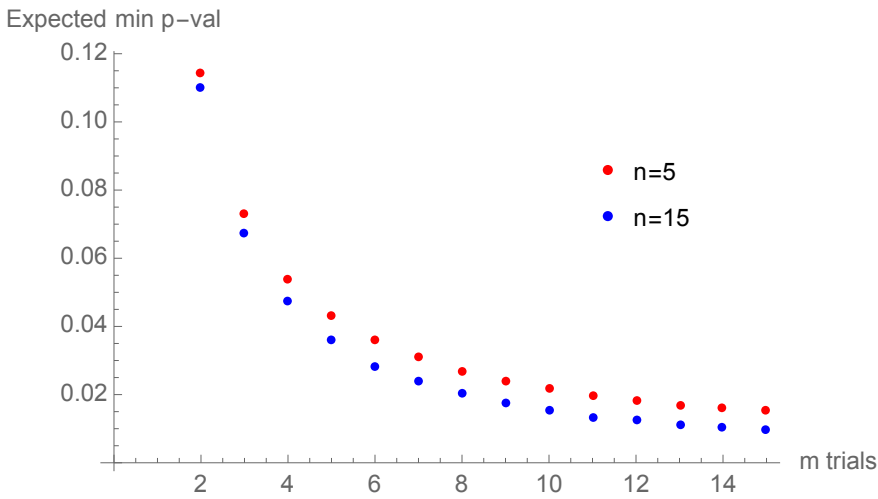
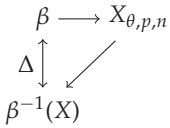


Figure 19.4: *The "p-hacking" value across m trials for $p_M = .15$ and $p_s = .22$.*

19.2 INVERSE POWER OF TEST

Let β be the power of a test for a given p-value p, for random draws X from unobserved parameter θ and a sample size of n. To gauge the reliability of β as a true measure of power, we perform an inverse problem:



Proposition 19.4
Let β_c be the projection of the power of the test from the realizations assumed to be student T distributed and evaluated under the parameter θ. We have

$$\Phi(\beta_c) = \begin{cases} \Phi(\beta_c)_L & \text{for } \beta_c < \frac{1}{2} \\ \Phi(\beta_c)_H & \text{for } \beta_c > \frac{1}{2} \end{cases}$$

where

$$\Phi(\beta_c)_L = \sqrt{1-\gamma_1}\gamma_1^{-\frac{n}{2}} \frac{\left(-\frac{\gamma_1}{2\sqrt{\frac{1}{\gamma_3}-1}\sqrt{-(\gamma_1-1)\gamma_1}-2\sqrt{-(\gamma_1-1)\gamma_1}+\gamma_1\left(2\sqrt{\frac{1}{\gamma_3}-1}-\frac{1}{\gamma_3}\right)-1}\right)^{\frac{n+1}{2}}}{\sqrt{-(\gamma_1-1)\gamma_1}} \tag{19.6}$$

$$\Phi(\beta_c)_H = \sqrt{\gamma_2}(1-\gamma_2)^{-\frac{n}{2}} B\left(\frac{1}{2},\frac{n}{2}\right) \frac{\left(\frac{1}{\frac{-2\left(\sqrt{-(\gamma_2-1)\gamma_2}+\gamma_2\right)\sqrt{\frac{1}{\gamma_3}-1}+2\sqrt{\frac{1}{\gamma_3}-1}+2\sqrt{-(\gamma_2-1)\gamma_2}-1}{\gamma_2-1}+\frac{1}{\gamma_3}}\right)^{\frac{n+1}{2}}}{\sqrt{-(\gamma_2-1)\gamma_2}B\left(\frac{n}{2},\frac{1}{2}\right)} \tag{19.7}$$

where $\gamma_1 = I^{-1}_{2\beta_c}\left(\frac{n}{2},\frac{1}{2}\right)$, $\gamma_2 = I^{-1}_{2\beta_c-1}\left(\frac{1}{2},\frac{n}{2}\right)$, *and* $\gamma_3 = I^{-1}_{(1,2p_s-1)}\left(\frac{n}{2},\frac{1}{2}\right)$.

19.3 APPLICATION AND CONCLUSION

- One can safely see that under such stochasticity for the realizations of p-values and the distribution of its minimum, to get what people mean by 5% confidence (and the inferences they get from it), they need a p-value of at least one order of magnitude smaller.
- Attempts at replicating papers, such as the open science project [48], should consider a margin of error in *its own* procedure and a pronounced bias towards favorable results (Type-I error). There should be no surprise that a previously deemed significant test fails during replication –in fact it is the replication of results deemed significant at a close margin that should be surprising.

- The "power" of a test has the same problem unless one either lowers p-values or sets the test at higher levels, such at .99.

ACKNOWLEDGMENT

Marco Avellaneda, Pasquale Cirillo, Yaneer Bar-Yam, friendly people on twitter ...

G | SOME CONFUSIONS IN BEHAVIORAL ECONOMICS

We saw earlier that the problem of "overestimation of the tails" is more attributable to the use of a wrong normative model. Here we use two similar cases, selected because one of them involves the author of "nudging", an invasive method by psychologists aiming at manipulating decisions by citizens.

G.1 CASE STUDY: HOW THE MYOPIC LOSS AVERSION IS MISSPECIFIED

The so-called "equity premium puzzle", originally detected by Mehra and Prescott [167], is called so because equities have historically yielded too high a return over fixed income investments; the puzzle is why is isn't arbitraged away.

We can easily figure out that the analysis misses the absence of ergodicity in such domain, as we saw in Chapter 3: agents do not really capture market returns unconditionally; it is foolish to use ensemble probabilities and the law of large numbers for individual investors who only have one life.

Benartzi and Thaler [17] claims that the Kahneman-Tversky prospect theory [137] explains such behavior owing to myopia. This might be true but such an analysis falls apart under thick tails.

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

Myopic loss aversion

Take the prospect theory valuation w function for x changes in wealth x, parametrized with λ and α.

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

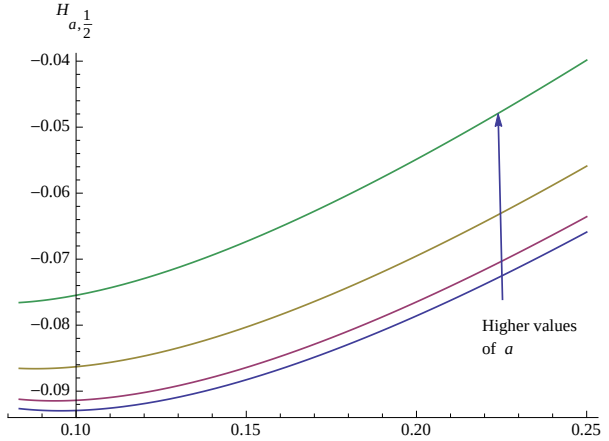


Figure G.1: *The effect of $H_{a,p}(t)$ "utility" or prospect theory of under second order effect on variance. Here $\sigma = 1, \mu = 1$ and t variable.*

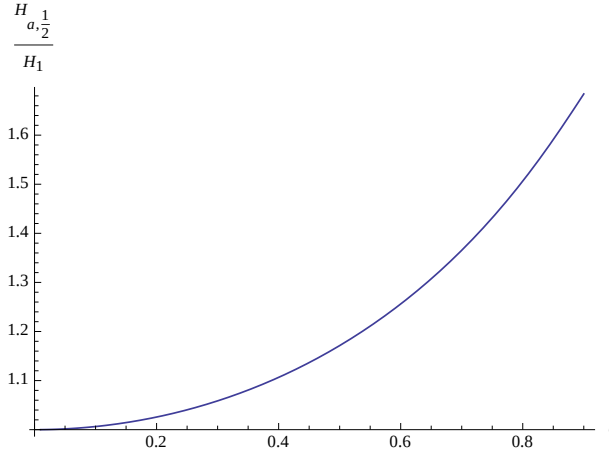


Figure G.2: *The ratio $\frac{H_{a,\frac{1}{2}}(t)}{H_0}$ or the degradation of "utility" under second order effects.*

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

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

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

$$= \frac{1}{\sqrt{\pi}} 2^{\frac{\alpha}{2}-2} \left(\frac{1}{\sigma^2 t}\right)^{-\frac{\alpha}{2}} \left(\Gamma\left(\frac{\alpha+1}{2}\right)\left(\sigma^\alpha t^{\alpha/2}\left(\frac{1}{\sigma^2 t}\right)^{\alpha/2} - \lambda\sigma\sqrt{t}\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(-\frac{\alpha}{2};\frac{1}{2};-\frac{t\mu^2}{2\sigma^2}\right)\right. \quad \text{(G.2)}$$
$$+ \frac{1}{\sqrt{2}\sigma}\mu\Gamma\left(\frac{\alpha}{2}+1\right)\left(\sigma^{\alpha+1}t^{\frac{\alpha}{2}+1}\left(\frac{1}{\sigma^2 t}\right)^{\frac{\alpha+1}{2}} + \sigma^\alpha t^{\frac{\alpha+1}{2}}\left(\frac{1}{\sigma^2 t}\right)^{\alpha/2}\right.$$
$$\left.\left. + 2\lambda\sigma t\sqrt{\frac{1}{\sigma^2 t}}\right) {}_1F_1\left(\frac{1-\alpha}{2};\frac{3}{2};-\frac{t\mu^2}{2\sigma^2}\right)\right)$$

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

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

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

Rewriting G.1

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

Result Conclusively, as can be seen in figures G.1 and G.2, second order effects cancel the statements made from "myopic" loss aversion. This doesn't mean that myopia doesn't have effects, rather that it cannot explain the "equity premium", not from the outside (i.e. the distribution might have different returns, but from the inside, owing to the structure of the Kahneman-Tversky value function $v(x)$.

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

Rewriting G.3 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).

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

Time preference under model error

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

This author once watched with a great deal of horror one Laibson [148] 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 [95], as economists temps to impart what seems to be a varying "discount rate" in a simplified model).[1]

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

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

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

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

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

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

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 4.1:

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

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

1 Farmer and Geanakoplos [88] have applied a similar approach to Hyperbolic discounting.

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

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

we get:

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

so

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

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

Further, fudging the distribution and normalizing it, when

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

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

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

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

Part VII

OPTION TRADING AND PRICING UNDER FAT TAILS

20 FINANCIAL THEORY'S FAILURES WITH OPTION PRICING†

LET US DISCUSS why option theory, as seen according to the so-called "neoclassical economics", fails in the real world. How does financial theory price financial products? The principal difference in paradigm between the one presented by Bachelier in 1900, [6] and the modern finance one known as Black-Scholes-Merton [24] and [169] lies in a few central assumptions by which Bachelier was closer to reality and the way traders do business and have done business for centuries.

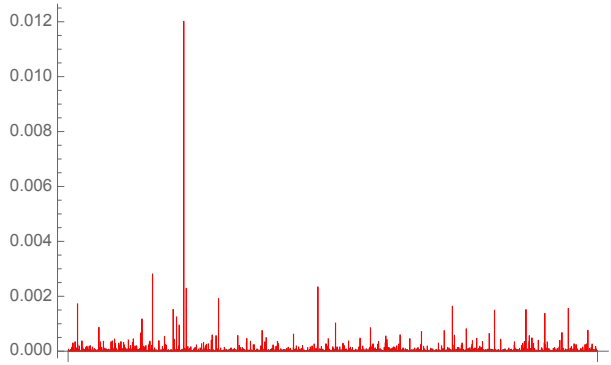


Figure 20.1: *The hedging errors for an option portfolio (under a daily revision regime) over 3000 days, under a constant volatility Student T with tail exponent $\alpha = 3$. Technically the errors should not converge in finite time as their distribution has infinite variance.*

20.1 BACHELIER NOT BLACK-SCHOLES

Bachelier's model is based on an actuarial expectation of final payoffs –not dynamic hedging. It means you can use any distribution! A more formal proof using

† Discussion chapter.

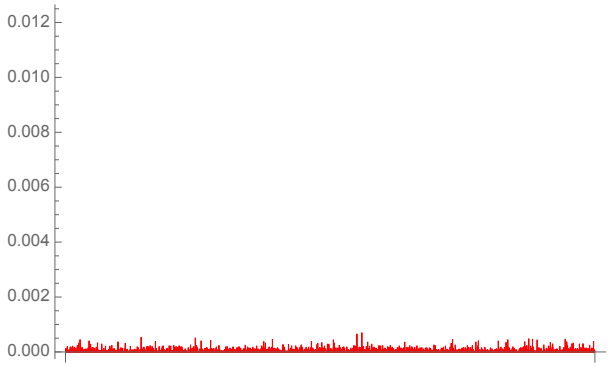


Figure 20.2: *Hedging errors for an option portfolio (daily revision) under an equivalent (rather fictional) "Black-Scholes" world.*

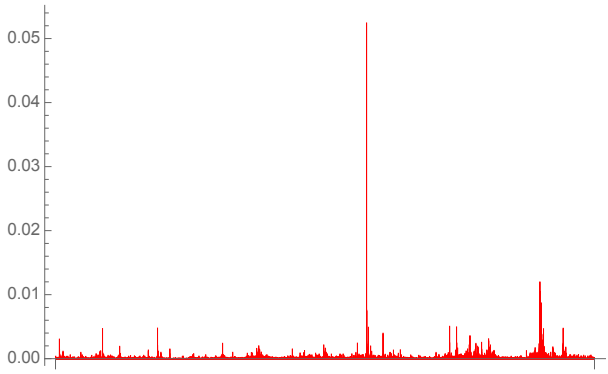


Figure 20.3: *Portfolio Hedging errors including the stock market crash of 1987.*

measure theory is provided in Chapter 21 so for now let us just get the intuition without too much mathematics.

The same method was later used by a series of researchers, such as Sprenkle [213] in 1964, Boness, [26] in 1964, Kassouf and Thorp, [248] in 1967, Thorp, [244] (only published in 1973).

They all encountered the following problem: how to produce a risk parameter –a risky asset discount rate – to make it compatible with portfolio theory? The Capital Asset Pricing Model requires that securities command an expected rate of return in proportion to their riskiness. In the Black-Scholes-Merton approach, an option price is derived from continuous-time dynamic hedging, and only in properties obtained from continuous time dynamic hedging –we will describe dynamic hedging in some details further down. Thanks to such a method, an option collapses into a deterministic payoff and provides returns independent of the market; hence it does not require any risk premium.

20.1.1 Distortion from Idealization

The problem we have with the Black-Scholes-Merton approach is that the requirements for dynamic hedging are extremely idealized, requiring the following strict

conditions. The operator is assumed to be able to buy and sell in a frictionless market, incurring no transaction costs. The procedure does not allow for the price impact of the order flow –if an operator sells a quantity of shares, it should not have consequences on the subsequent price. The operator knows the probability distribution, which is the Gaussian, with fixed and constant parameters through time (all parameters do not change). Finally, the most significant restriction: no scalable jumps. In a subsequent revision [Merton, 1976] allows for jumps but these are deemed to be Poisson arrival time, and fixed or, at the worst, Gaussian. The framework does not allow the use of power laws both in practice and mathematically. Let us examine the mathematics behind the stream of dynamic hedges in the Black-Scholes-Merton equation.

Assume the risk-free interest rate $r = 0$ with no loss of generality. The canonical Black-Scholes-Merton model consists in selling a call and purchasing shares of stock that provide a hedge against instantaneous moves in the security. Thus the portfolio π locally "hedged" against exposure to the first moment of the distribution is the following:

$$\pi = -C + \frac{\partial C}{\partial S} S \tag{20.1}$$

where C is the call price, and S the underlying security.

Take the change in the values of the portfolio

$$\Delta\pi = -\Delta C + \frac{\partial C}{\partial S} \Delta S \tag{20.2}$$

By expanding around the initial values of S, we have the changes in the portfolio in discrete time. Conventional option theory applies to the Gaussian in which all orders higher than $(\Delta S)^2$ and Δt disappears rapidly.

$$\Delta\pi = -\frac{\partial C}{\partial t}\Delta t - \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\Delta S^2 + O\left(\Delta S^3\right) \tag{20.3}$$

Taking expectations on both sides, we can see from (3) very strict requirements on moment finiteness: *all* moments need to converge. If we include another term, $-\frac{1}{6}\frac{\partial^3 C}{\partial S^3}\Delta S^3$, it may be of significance in a probability distribution with significant cubic or quartic terms. Indeed, although the n^{th} derivative with respect to S can decline very sharply, for options that have a strike K away from the center ot the distribution, it remains that the moments are rising disproportionately fast for that to carry a mitigating effect.

So here we mean all moments need to be finite and losing in impact –no approximation. Note here that the jump diffusion model (Merton,1976) does not cause much trouble since it has all the moments. And the annoyance is that a power law will have every moment higher than α infinite, causing the equation of the Black-Scholes-Merton portfolio to fail.

As we said, the logic of the Black-Scholes-Merton so-called solution thanks to Itô's lemma was that the portfolio collapses into a deterministic payoff. But let us see how quickly or effectively this works in practice.

20.1.2 The Actual Replication Process:

The payoff of a call should be replicated with the following stream of dynamic hedges, the limit of which can be seen here, between t and T

$$\lim_{\Delta t \to 0} \left(\sum_{i=1}^{n=T/\Delta t} \frac{\partial C}{\partial S} \Big|_{S=S_{t+(i-1)\Delta t}, t=t+(i-1)\Delta t,} \left(S_{t+i\Delta t} - S_{t+(i-1)\Delta t} \right) \right) \tag{20.4}$$

We break up the period into n increments Δt. Here the hedge ratio $\frac{\partial C}{\partial S}$ is computed as of time $t + (i\text{-}1)\,\Delta t$, but we get the nonanticipating difference between the price at the time the hedge was initiatied and the resulting price at $t + i\,\Delta t$.

This is supposed to make the payoff deterministic *at the limit of* $\Delta t \to 0$. In the Gaussian world, this would be an Itô-McKean integral.

20.1.3 Failure: How Hedging Errors Can Be Prohibitive.

As a consequence of the mathematical property seen above, hedging errors in an cubic α appear to be indistinguishable from those from an infinite variance process. Furthermore such error has a disproportionaly large effect on strikes away from the money.

In short: dynamic hedging in a power law world removes no risk.

NEXT

The next chapter will use measure theory to show why options can still be risk-neutral.

21 UNIQUE OPTION PRICING MEASURE WITH NEITHER DYNAMIC HEDGING NOR COMPLETE MARKETS‡

We present the proof that under simple assumptions, such as constraints of Put-Call Parity, the probability measure for the valuation of a European option has the mean derived from the forward price which can, but does not have to be the risk-neutral one, under any general probability distribution, bypassing the Black-Scholes-Merton dynamic hedging argument, and without the requirement of complete markets and other strong assumptions. We confirm that the heuristics used by traders for centuries are both more robust, more consistent, and more rigorous than held in the economics literature. We also show that options can be priced using infinite variance (finite mean) distributions.

21.1 BACKGROUND

Option valuations methodologies have been used by traders for centuries, in an effective way (Haug and Taleb, [124]). In addition, valuations by expectation of terminal payoff forces the mean of the probability distribution used for option prices to be that of the forward, thanks to Put-Call Parity and, should the forward be risk-neutrally priced, so will the option be. The Black Scholes argument (Black and Scholes, 1973, Merton, 1973) is held to allow risk-neutral option pricing thanks to dynamic hedging, as the option becomes redundant (since its payoff can be built as a linear combination of cash and the underlying asset dynamically revised through time). This is a puzzle, since: 1) Dynamic Hedging is not operationally feasible in financial markets owing to the dominance of portfolio changes resulting from jumps, 2) The dynamic hedging argument doesn't stand mathematically under fat tails; it requires a very specific "Black Scholes world" with many impossible assumptions, one of which requires finite quadratic variations, 3) Traders use the same Black-Scholes "risk neutral argument" for the valuation of options on as-

‡ Research chapter.

sets that do not allow dynamic replication, 4) Traders trade options consistently in domain where the risk-neutral arguments do not apply 5) There are fundamental informational limits preventing the convergence of the stochastic integral.[2]

There have been a couple of predecessors to the present thesis that Put-Call parity is sufficient constraint to enforce some structure at the level of the mean of the underlying distribution, such as Derman and Taleb (2005), Haug and Taleb (2010). These approaches were heuristic, robust though deemed hand-waving (Ruffino and Treussard, [204]). In addition they showed that operators need to use the risk-neutral mean. What this chapter does is

- It goes beyond the "handwaving" with formal proofs.
- It uses a completely distribution-free, expectation-based approach and proves the risk-neutral argument without dynamic hedging, and without any distributional assumption.
- Beyond risk-neutrality, it establishes the case of a unique pricing distribution for option prices in the absence of such argument. The forward (or future) price can embed expectations and deviate from the arbitrage price (owing to, say, regulatory or other limitations) yet the options can still be priced at a distibution corresponding to the mean of such a forward.
- It shows how one can *practically* have an option market without "completeness" and without having the theorems of financial economics hold.

These are done with solely two constraints: "horizontal", i.e. put-call parity, and "vertical", i.e. the different valuations across strike prices deliver a probability measure which is shown to be unique. The only economic assumption made here is that the forward exits, is tradable — in the absence of such unique forward price it is futile to discuss standard option pricing. We also require the probability measures to correspond to distributions with finite first moment.

Preceding works in that direction are as follows. Breeden and Litzenberger [31] and Dupire [71], show how option spreads deliver a unique probability measure; there are papers establishing broader set of arbitrage relations between options such as Carr and Madan [37][3].

However 1) none of these papers made the bridge between calls and puts via the forward, thus translating the relationships from arbitrage relations between options delivering a probability distribution into the necessity of lining up to the mean of the distribution of the forward, hence the risk-neutral one (in case the forward is arbitraged.) 2) Nor did any paper show that in the absence of second moment (say, infinite variance), we can price options very easily. Our methodology and proofs make no use of the variance. 3) Our method is vastly simpler, more direct, and robust to changes in assumptions.

2 Further, in a case of scientific puzzle, the exact formula called "Black-Scholes-Merton" was written down (and used) by Edward Thorp in a heuristic derivation by expectation that did not require dynamic hedging, see Thorpe [246].

3 See also Green and Jarrow [113] and Nachman [173]. We have known about the possibility of risk neutral pricing without dynamic hedging since Harrison and Kreps [121] but the theory necessitates extremely strong –and severely unrealistic –assumptions, such as strictly complete markets and a multiperiod pricing kernel

We make no assumption of general market completeness. Options are not redundant securities and remain so. Table 1 summarizes the gist of the paper.[4] [5]

21.2 PROOF

Define $C(S_{t_0}, K, t)$ and $P(S_{t_0}, K, t)$ as European-style call and put with strike price K, respectively, with expiration t, and S_0 as an underlying security at times t_0, $t \geq t_0$, and S_t the possible value of the underlying security at time t.

21.2.1 Case 1: Forward as risk-neutral measure

Define $r = \frac{1}{t-t_0}\int_{t_0}^{t} r_s \mathrm{d}s$, the return of a risk-free money market fund and $\delta = \frac{1}{t-t_0}\int_{t_0}^{t} \delta_s \mathrm{d}s$ the payout of the asset (continuous dividend for a stock, foreign interest for a currency).

We have the arbitrage forward price F_t^Q:

$$F_t^Q = S_0 \frac{(1+r)^{(t-t_0)}}{(1+\delta)^{(t-t_0)}} \approx S_0\, e^{(r-\delta)(t-t_0)} \tag{21.1}$$

by arbitrage, see Keynes 1924. We thus call F_t^Q the future (or forward) price obtained by arbitrage, at the risk-neutral rate. Let F_t^P be the future requiring a risk-associated "expected return" m, with expected forward price:

$$F_t^P = S_0(1+m)^{(t-t_0)} \approx S_0\, e^{m\,(t-t_0)}. \tag{21.2}$$

Remark: *By arbitrage, all tradable values of the forward price given S_{t_0} need to be equal to F_t^Q.*

"Tradable" here does not mean "traded", only subject to arbitrage replication by "cash and carry", that is, borrowing cash and owning the secutity yielding d if the embedded forward return diverges from r.

21.2.2 Derivations

In the following we take F as having dynamics on its own –irrelevant to whether we are in case 1 or 2 –hence a unique probability measure Q.

4 The famed Hakkanson paradox is as follows: if markets are complete and options are redudant, why would someone need them? If markets are incomplete, we may need options but how can we price them? This discussion may have provided a solution to the paradox: markets are incomplete *and* we can price options.

5 Option prices are not unique in the absolute sense: the premium over intrinsic can take an entire spectrum of values; it is just that the put-call parity constraints forces the measures used for puts and the calls to be the same and to have the same expectation as the forward. As far as securities go, options are securities on their own; they just have a strong link to the forward.

Table 21.1: *Main practical differences between the dynamic hedging argument and the static Put-Call parity with spreading across strikes.*

	Black-Scholes Merton	**Put-Call Parity with Spreading**
Type	Continuous rebalancing.	Interpolative static hedge.
Limit	Law of large numbers in time (horizontal).	Law of large numbers across strikes (vertical).
Market Assumptions	1) Continuous Markets, no gaps, no jumps. 2) Ability to borrow and lend underlying asset for all dates. 3) No transaction costs in trading asset.	1) Gaps and jumps acceptable. Possibility of continuous Strikes, or acceptable number of strikes. 2) Ability to borrow and lend underlying asset for single forward date. 3) Low transaction costs in trading options.
Probability Distribution	Requires all moments to be finite. Excludes the class of slowly varying distributions	Requires finite 1^{st} moment (infinite variance is acceptable).
Market Completeness	Achieved through dynamic completeness	Not required (in the traditional sense)
Realism of Assumptions	Low	High
Convergence	Uncertain; one large jump changes expectation	Robust
Fitness to Reality	Only used after "fudging" standard deviations per strike.	Portmanteau, using specific distribution adapted to reality

Define $\Omega = [0, \infty) = A_K \cup A_K^c$ where $A_K = [0, K]$ and $A_K^c = (K, \infty)$.

Consider a class of standard (simplified) probability spaces (Ω, μ_i) indexed by i, where μ_i is a probability measure, i.e., satisfying $\int_\Omega \mathrm{d}\mu_i = 1$.

Theorem 5

For a given maturity T, there is a unique measure μ_Q that prices European puts and calls by expectation of terminal payoff.

This measure can be risk-neutral in the sense that it prices the forward F_t^Q, but does not have to be and imparts rate of return to the stock embedded in the forward.

Lemma 21.1

For a given maturity T, there exist two measures μ_1 and μ_2 for European calls and puts of the same maturity and same underlying security associated with the valuation by expectation of terminal payoff, which are unique such that, for any call and put of strike K, we have:

$$C = \int_\Omega f_C \, \mathrm{d}\mu_1 \,, \tag{21.3}$$

and

$$P = \int_\Omega f_P \, \mathrm{d}\mu_2 \,, \tag{21.4}$$

respectively, and where f_C and f_P are $(S_t - K)^+$ and $(K - S_t)^+$ respectively.

Proof. For clarity, set r and δ to 0 without a loss of generality. By Put-Call Parity Arbitrage, a positive holding of a call ("long") and negative one of a put ("short") replicates a tradable forward; because of P/L variations, using positive sign for long and negative sign for short:

$$C(S_{t_0}, K, t) - P(S_{t_0}, K, t) + K = F_t^P \tag{21.5}$$

necessarily since F_t^P is tradable.

Put-Call Parity holds for all strikes, so:

$$C(S_{t_0}, K + \Delta K, t) - P(S_{t_0}, K + \Delta K, t) + K + \Delta K = F_t^P \tag{21.6}$$

for all $K \in \Omega$

Now a Call spread in quantities $\frac{1}{\Delta K}$, expressed as

$$C(S_{t_0}, K, t) - C(S_{t_0}, K + \Delta K, t),$$

delivers \$1 if $S_t > K + \Delta K$ (that is, corresponds to the indicator function $\mathbf{1}_{S>K+\Delta K}$), 0 if $S_t \leq K$ (or $\mathbf{1}_{S>K}$), and the quantity times $S_t - K$ if $K < S_t \leq K + \Delta K$, that is, between 0 and \$1 (see Breeden and Litzenberger, 1978[31]). Likewise, consider the converse argument for a put, with $\Delta K < S_t$.

At the limit, for $\Delta K \to 0$

$$\frac{\partial C(S_{t_0}, K, t)}{\partial K} = -P(S_t > K) = -\int_{A_K^c} \mathrm{d}\mu_1. \tag{21.7}$$

By the same argument:

$$\frac{\partial P(S_{t_0}, K, t)}{\partial K} = \int_{A_K} \mathrm{d}\mu_2 = 1 - \int_{A_K^c} \mathrm{d}\mu_2. \tag{21.8}$$

As semi-closed intervals generate the whole Borel σ-algebra on Ω, this shows that $\mu_1$and μ_2 are unique.

□

Lemma 21.2
The probability measures of puts and calls are the same, namely for each Borel set A in Ω, $\mu_1(A) = \mu_2(A)$.

Proof. Combining Equations 21.5 and 21.6, dividing by $\frac{1}{\Delta K}$ and taking $\Delta K \to 0$:

$$-\frac{\partial C(S_{t_0}, K, t)}{\partial K} + \frac{\partial P(S_{t_0}, K, t)}{\partial K} = 1 \tag{21.9}$$

for all values of K, so

$$\int_{A_K^c} \mathrm{d}\mu_1 = \int_{A_K^c} \mathrm{d}\mu_2, \tag{21.10}$$

hence $\mu_1(A_K) = \mu_2(A_K)$ for all $K \in [0, \infty)$. This equality being true for any semi-closed interval, it extends to any Borel set.

□

Lemma 21.3
Puts and calls are required, by static arbitrage, to be evaluated at same as risk-neutral measure μ_Q as the tradable forward.

Proof.

$$F_t^P = \int_\Omega F_t \, \mathrm{d}\mu_Q; \tag{21.11}$$

from Equation 21.5

$$\int_\Omega f_C(K) \, \mathrm{d}\mu_1 - \int_\Omega f_P(K) \, \mathrm{d}\mu_1 = \int_\Omega F_t \, \mathrm{d}\mu_Q - K \tag{21.12}$$

Taking derivatives on both sides, and since $f_C - f_P = S_0 + K$, we get the Radon-Nikodym derivative:

$$\frac{\mathrm{d}\mu_Q}{\mathrm{d}\mu_1} = 1 \tag{21.13}$$

for all values of K.

□

21.3 CASE WHERE THE FORWARD IS NOT RISK NEUTRAL

Consider the case where F_t is observable, tradable, and use it solely as an underlying security with dynamics on its own. In such a case we can completely ignore the dynamics of the nominal underlying S, or use a non-risk neutral "implied" rate linking cash to forward, $m^* = \frac{\log\left(\frac{F}{S_0}\right)}{t-t_0}$. the rate m can embed risk premium, difficulties in financing, structural or regulatory impediments to borrowing, with no effect on the final result.

In that situation, it can be shown that the exact same results as before apply, by remplacing the measure μ_Q by another measure μ_{Q^*}. Option prices remain unique [6].

21.4 COMMENT

We have replaced the complexity and intractability of dynamic hedging with a simple, more benign interpolation problem, and explained the performance of pre-Black-Scholes option operators using simple heuristics and rules, bypassing the structure of the theorems of financial economics.

Options can remain non-redundant and markets incomplete: we are just arguing here for a form of arbitrage pricing (which includes risk-neutral pricing at the level of the expectation of the probability measure), nothing more. But this is sufficient for us to use any probability distribution with finite first moment, which includes the Lognormal, which recovers Black Scholes.

A final comparison. In dynamic hedging, missing a single hedge, or encountering a single gap (a tail event) can be disastrous —as we mentioned, it requires a series of assumptions beyond the mathematical, in addition to severe and highly unrealistic constraints on the mathematical. Under the class of fat tailed distributions, increasing the frequency of the hedges does not guarantee reduction of risk. Further, the standard dynamic hedging argument requires the exact specification of the *risk-neutral* stochastic process between t_0 and t, something econometrically unwieldy, and which is generally reverse engineered from the price of options, as an arbitrage-oriented interpolation tool rather than as a representation of the process.

Here, in our Put-Call Parity based methodology, our ability to track the risk neutral distribution is guaranteed by adding strike prices, and since probabilities add up to 1, the degrees of freedom that the recovered measure μ_Q has in the gap area between a strike price K and the next strike up, $K + \Delta K$, are severely reduced, since the measure in the interval is constrained by the difference $\int_{A_K}^{c} \mathrm{d}\mu - \int_{A_{K+\Delta K}}^{c} \mathrm{d}\mu$. In other words, no single gap between strikes can significantly affect the probability measure, even less the first moment, unlike with dynamic hedging.

6 We assumed 0 discount rate for the proofs; in case of nonzero rate, premia are discounted at the rate of the arbitrage operator

In fact it is no different from standard kernel smoothing methods for statistical samples, but applied to the distribution across strikes.[7]

The assumption about the presence of strike prices constitutes a natural condition: conditional on having a *practical* discussion about options, options strikes need to exist. Further, as it is the experience of the author, market-makers can add over-the-counter strikes at will, should they need to do so.

ACKNOWLEDGMENTS

Peter Carr, Marco Avellaneda, Hélyette Geman, Raphael Douady, Gur Huberman, Espen Haug, and Hossein Kazemi.

7 For methods of interpolation of implied probability distribution between strikes, see Avellaneda et al.[4].

22 OPTION TRADERS NEVER USE THE BLACK-SCHOLES-MERTON FORMULA*,‡

OPTION TRADERS use a heuristically derived pricing formula which they adapt by fudging and changing the tails and skewness by varying one parameter, the standard deviation of a Gaussian. Such formula is popularly called " Black-Scholes-Merton" owing to an attributed eponymous discovery (though changing the standard deviation parameter is in contradiction with it). However, we have historical evidence that: (1) the said Black, Scholes and Merton did not invent any formula, just found an argument to make a well known (and used) formula compatible with the economics establishment, by removing the risk parameter through dynamic hedging, (2) option traders use (and evidently have used since 1902) sophisticated heuristics and tricks more compatible with the previous versions of the formula of Louis Bachelier and Edward O. Thorp (that allow a broad choice of probability distributions) and removed the risk parameter using put-call parity, (3) option traders did not use the Black-Scholes-Merton formula or similar formulas after 1973 but continued their bottom-up heuristics more robust to the high impact rare event. The chapter draws on historical trading methods and 19th and early 20th century references ignored by the finance literature. It is time to stop using the wrong designation for option pricing.

22.1 BREAKING THE CHAIN OF TRANSMISSION

For us, practitioners, theories should arise from practice [2]. This explains our concern with the "scientific" notion that practice should fit theory. Option hedging, pricing, and trading is neither philosophy nor mathematics. It is a rich craft with

* Research chapter.

2 For us, in this discussion, a "practitioner" is deemed to be someone involved in repeated decisions about option hedging, that is with a risk-P/L and skin in the game, not a support quant who writes pricing software or an academic who provides consulting advice.

traders learning from traders (or traders copying other traders) and tricks developing under evolution pressures, in a bottom-up manner. It is techne, not episteme. Had it been a science it would not have survived for the empirical and scientific fitness of the pricing and hedging theories offered are, we will see, at best, defective and unscientific (and, at the worst, the hedging methods create more risks than they reduce). Our approach in this chapter is to ferret out historical evidence of techne showing how option traders went about their business in the past.

Options, we will show, have been extremely active in the pre-modern finance world. Tricks and heuristically derived methodologies in option trading and risk management of derivatives books have been developed over the past century, and used quite effectively by operators. In parallel, many derivations were produced by mathematical researchers. The economics literature, however, did not recognize these contributions, substituting the rediscoveries or subsequent re formulations done by (some) economists. There is evidence of an attribution problem with Black-Scholes-Merton option formula , which was developed, used, and adapted in a robust way by a long tradition of researchers and used heuristically by option book runners. Furthermore, in a case of scientific puzzle, the exact formula called Black-Sholes-Merton was written down (and used) by Edward Thorp which, paradoxically, while being robust and realistic, has been considered unrigorous. This raises the following: 1) The Black-Scholes-Merton innovation was just a neoclassical finance argument, no more than a thought experiment [3], 2) We are not aware of traders using their argument or their version of the formula.

It is high time to give credit where it belongs.

22.2 INTRODUCTION/SUMMARY

22.2.1 Black-Scholes was an argument

Option traders call the formula they use the Black-Scholes-Merton formula without being aware that by some irony, of all the possible options formulas that have been produced in the past century, what is called the Black-Scholes-Merton formula (after Black and Scholes, 1973, and Merton, 1973) is the one the furthest away from what they are using. In fact of the formulas written down in a long history it is the only formula that is fragile to jumps and tail events.

First, something seems to have been lost in translation: Black and Scholes [25] and Merton [170] actually never came up with a new option formula, but only an theoretical economic argument built on a new way of deriving, rather re-deriving, an already existing and well known formula. The argument, we will see, is extremely fragile to assumptions. The foundations of option hedging and pricing were already far more firmly laid down before them. The Black-Scholes-Merton

3 Here we question the notion of confusing thought experiments in a hypothetical world, of no predictive power, with either science or practice. The fact that the Black-Scholes-Merton argument works in a Platonic world and appears to be elegant does not mean anything since one can always produce a Platonic world in which a certain equation works, or in which a rigorous proof can be provided, a process called reverse-engineering.

Figure 22.1: *Louis Bachelier, who came up with an option formula based on expectation. It is based on more rigorous foundations than the Black-Scholes dynamic hedging argument as it does not require a thin-tailed distribution. Few people are aware of the fact that the Black-Scholes so-called discovery was an argument to remove the expectation of the underlying security, not the derivation of a new equation.*

argument, simply, is that an option can be hedged using a certain methodology called dynamic hedging and then turned into a risk-free instrument, as the portfolio would no longer be stochastic. Indeed what Black, Scholes and Merton did was marketing, finding a way to make a well-known formula palatable to the economics establishment of the time, little else, and in fact distorting its essence.

Such argument requires strange far-fetched assumptions: some liquidity at the level of transactions, knowledge of the probabilities of future events (in a neoclassical Arrow-Debreu style) , and, more critically, a certain mathematical structure that requires thin-tails, or mild randomness, on which, later[4]. The entire argument is indeed, quite strange and rather inapplicable for someone clinically and observation-driven standing outside conventional neoclassical economics. Simply, the dynamic hedging argument is dangerous in practice as it subjects you to blowups; it makes no sense unless you are concerned with neoclassical economic theory. The Black-Scholes-Merton argument and equation flow a top-down general equilibrium theory, built upon the assumptions of operators working in full knowledge of the probability distribution of future outcomes in addition to a collection of assumptions that, we will see, are highly invalid mathematically, the main one being the ability to cut the risks using continuous trading which only works in the very narrowly special case of thin-tailed distributions. But it is not just these flaws that make it inapplicable: option traders do not buy theories , particularly speculative general equilibrium ones, which they find too risky for them and extremely lacking in standards of reliability. A normative theory is, simply, not good for

4 Of all the misplaced assumptions of Black Scholes that cause it to be a mere thought experiment, though an extremely elegant one, a flaw shared with modern portfolio theory, is the certain knowledge of future delivered variance for the random variable (or, equivalently, all the future probabilities). This is what makes it clash with practice the rectification by the market fattening the tails is a negation of the Black-Scholes thought experiment.

decision-making under uncertainty (particularly if it is in chronic disagreement with empirical evidence). People may take decisions based on speculative theories, but avoid the fragility of theories in running their risks.

Yet professional traders, including the authors (and, alas, the Swedish Academy of Science) have operated under the illusion that it was the Black-Scholes-Merton formula they actually used we were told so. This myth has been progressively reinforced in the literature and in business schools, as the original sources have been lost or frowned upon as anecdotal (Merton [172]).

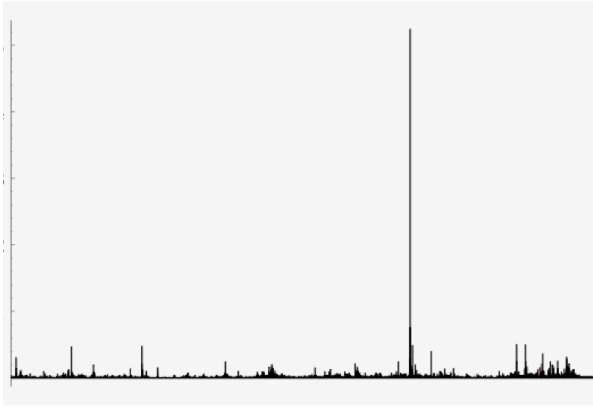


Figure 22.2: *The typical "risk reduction" performed by the Black-Scholes-Merton argument. These are the variations of a dynamically hedged portfolio (and a quite standard one). BSM indeed "smoothes" out variations but exposes the operator to massive tail events reminiscent of such blowups as LTCM. Other option formulas are robust to the rare event and make no such claims.*

This discussion will present our real-world, ecological understanding of option pricing and hedging based on what option traders actually do and did for more than a hundred years.

This is a very general problem. As we said, option traders develop a chain of transmission of techne, like many professions. But the problem is that the chain is often broken as universities do not store the acquired skills by operators. Effectively plenty of robust heuristically derived implementations have been developed over the years, but the economics establishment has refused to quote them or acknowledge them. This makes traders need to relearn matters periodically. Failure of dynamic hedging in 1987, by such firm as Leland O ′ ŹBrien Rubinstein, for instance, does not seem to appear in the academic literature published after the event (Merton, [172], Rubinstein,[202], Ross [200]); to the contrary dynamic hedging is held to be a standard operation [5].

There are central elements of the real world that can escape them academic research without feedback from practice (in a practical and applied field) can cause the diversions we witness between laboratory and ecological frameworks. This explains why some many finance academics have had the tendency to make smooth returns, then blow up using their own theories[6]. We started the other way around,

5 For instance how mistakes never resurface into the consciousness, Mark Rubinstein was awarded in 1995 the Financial Engineer of the Year award by the International Association of Financial Engineers. There was no mention of portfolio insurance and the failure of dynamic hedging.

6 For a standard reaction to a rare event, see the following: "Wednesday is the type of day people will remember in quant-land for a very long time," said Mr. Rothman, a University of Chicago Ph.D. who ran

first by years of option trading doing million of hedges and thousands of option trades. This in combination with investigating the forgotten and ignored ancient knowledge in option pricing and trading we will explain some common myths about option pricing and hedging. There are indeed two myths:

- That we had to wait for the Black-Scholes-Merton options formula to trade the product, price options, and manage option books. In fact the introduction of the Black, Scholes and Merton argument increased our risks and set us back in risk management. More generally, it is a myth that traders rely on theories, even less a general equilibrium theory, to price options.
- That we use the Black-Scholes-Merton options pricing formula. We, simply don't.

In our discussion of these myth we will focus on the bottom-up literature on option theory that has been hidden in the dark recesses of libraries. And that addresses only recorded matters not the actual practice of option trading that has been lost.

22.3 MYTH 1: TRADERS DID NOT PRICE OPTIONS BEFORE BSM

It is assumed that the Black-Scholes-Merton theory is what made it possible for option traders to calculate their delta hedge (against the underlying) and to price options. This argument is highly debatable, both historically and analytically.

Options were actively trading at least already in the 1600 as described by Joseph De La Vega implying some form of techneń, a heuristic method to price them and deal with their exposure. De La Vega describes option trading in the Netherlands, indicating that operators had some expertise in option pricing and hedging. He diffusely points to the put-call parity, and his book was not even meant to teach people about the technicalities in option trading. Our insistence on the use of Put-Call parity is critical for the following reason: The Black-Scholes-Merton Źs claim to fame is removing the necessity of a risk-based drift from the underlying security to make the trade risk-neutral. But one does not need dynamic hedging for that: simple put call parity can suffice (Derman and Taleb, 2005), as we will discuss later. And it is this central removal of the risk-premium that apparently was behind the decision by the Nobel committee to grant Merton and Scholes the (then called) Bank of Sweden Prize in Honor of Alfred Nobel: Black, Merton and Scholes made a vital contribution by showing that it is in fact not necessary to use any risk premium when valuing an option. This does not mean that the risk premium disappears; instead it is already included in the stock price. It is for having removed the effect of the drift on the value of the option, using a thought experiment, that their work was originally cited, something that was mechanically present by any form of trading and converting using far simpler techniques.

---

a quantitative fund before joining Lehman Brothers. "Events that models only predicted would happen once in 10,000 years happened every day for three days." One "Quant Sees Shakeout For the Ages – '10,000 Years" By Kaja Whitehouse,*Wall Street Journal* August 11, 2007; Page B3.

Options have a much richer history than shown in the conventional literature. Forward contracts seems to date all the way back to Mesopotamian clay tablets dating all the way back to 1750 B.C. Gelderblom and Jonker [103] show that Amsterdam grain dealers had used options and forwards already in 1550.

In the late 1800 and the early 1900 there were active option markets in London and New York as well as in Paris and several other European exchanges. Markets it seems, were active and extremely sophisticated option markets in 1870. Kairys and Valerio (1997) discuss the market for equity options in USA in the 1870s, indirectly showing that traders were sophisticated enough to price for tail events[7].

There was even active option arbitrage trading taking place between some of these markets. There is a long list of missing treatises on option trading: we traced at least ten German treatises on options written between the late 1800s and the hyperinflation episode[8].

22.4 METHODS AND DERIVATIONS

One informative extant source, Nelson [174], speaks volumes: An option trader and arbitrageur, S.A. Nelson published a book The A B C of Options and Arbitrage based on his observations around the turn of the twentieth century. According to Nelson (1904) up to 500 messages per hour and typically 2000 to 3000 messages per day were sent between the London and the New York market through the cable companies. Each message was transmitted over the wire system in less than a minute. In a heuristic method that was repeated in *Dynamic Hedging* [222] , Nelson, describe in a theory-free way many rigorously clinical aspects of his arbitrage business: the cost of shipping shares, the cost of insuring shares, interest expenses, the possibilities to switch shares directly between someone being long securities in New York and short in London and in this way saving shipping and insurance costs, as well as many more tricks etc.

7 The historical description of the market is informative until Kairys and Valerio [138] try to gauge whether options in the 1870s were underpriced or overpriced (using Black-Scholes-Merton style methods). There was one tail-event in this period, the great panic of September 1873. Kairys and Valerio find that holding puts was profitable, but deem that the market panic was just a one-time event :
"However, the put contracts benefit from the financial panic that hit the market in September, 1873. Viewing this as a one-time event, we repeat the analysis for puts excluding any unexpired contracts written before the stock market panic."
Using references to the economic literature that also conclude that options in general were overpriced in the 1950s 1960s and 1970s they conclude: "Our analysis shows that option contracts were generally overpriced and were unattractive for retail investors to purchase. They add: ÏEmpirically we find that both put and call options were regularly overpriced relative to a theoretical valuation model." These results are contradicted by the practitioner Nelson (1904): "The majority of the great option dealers who have found by experience that it is the givers, and not the takers, of option money who have gained the advantage in the long run."

8 Here is a partial list: Bielschowsky, R (1892): *Ueber die rechtliche Natur der Prämiengeschäfte*, Bresl. Genoss.-Buchdr; Granichstaedten-Czerva, R (1917): *Die Prämiengeschäfte an der Wiener Börse*, Frankfurt am Main; Holz, L. (1905) *Die Prämiengeschäfte*, Thesis (doctoral)–Universität Rostock; Kitzing, C. (1925):*Prämiengeschäfte : Vorprämien-, Rückprämien-, Stellagen- u. Nochgeschäfte ; Die solidesten Spekulationsgeschäfte mit Versicherg auf Kursverlust*, Berlin; Leser, E, (1875): *Zur Geschichte der Prämiengeschäfte*; Szkolny, I. (1883): *Theorie und praxis der prämiengeschäfte nach einer originalen methode dargestellt.*, Frankfurt am Main; Author Unknown (1925): *Das Wesen der Prämiengeschäfte*, Berlin : Eugen Bab & Co., Bankgeschäft.

Figure 22.3: *Espen Haug (coauthor of chapter) with Mandelbrot and this author in 2007.*

The formal financial economics canon does not include historical sources from outside economics, a mechanism discussed in Taleb (2007)[224]. The put-call parity was according to the formal option literature first fully described by Stoll [216], but neither he nor others in the field even mention Nelson. Not only was the put-call parity argument fully understood and described in detail by Nelson, but he, in turn, makes frequent reference to Higgins (1902) [127]. Just as an example Nelson (1904) referring to Higgins (1902) writes:

> It may be worthy of remark that calls are more often dealt than puts the reason probably being that the majority of punters in stocks and shares are more inclined to look at the bright side of things, and therefore more often see a rise than a fall in prices.
>
> This special inclination to buy calls and to leave the puts severely alone does not, however, tend to make calls dear and puts cheap, for it can be shown that the adroit dealer in options can convert a put into a call, a call into a putŹ, a call or more into a put-andcall, in fact any option into another, by dealing against it in the stock. We may therefore assume, with tolerable accuracy, that the call of a stock at any moment costs the same as the put of that stock, and half as much as the Put-and-Call.

The Put-and-Call was simply a put plus a call with the same strike and maturity, what we today would call a straddle. Nelson describes the put-call parity over many pages in full detail. Static market neutral delta hedging was also known at that time, in his book Nelson for example writes:

> Sellers of options in London as a result of long experience, if they sell a Call, straightway buy half the stock against which the Call is sold; or if a Put is sold; they sell half the stock immediately.

We must interpret the value of this statement in the light that standard options in London at that time were issued at-the-money (as explicitly pointed out by Nelson); furthermore, all standard options in London were European style. In London in- or out-of-the-money options were only traded occasionally and were known as fancy

options. It is quite clear from this and the rest of Nelson's book that the option dealers were well aware that the delta for at-the-money options was approximately 50%. As a matter of fact, at-the-money options trading in London at that time were adjusted to be struck to be at-the-money forward, in order to make puts and calls of the same price. We know today that options that are at-the-money forward and do not have very long time to maturity have a delta very close to 50% (naturally minus 50% for puts). The options in London at that time typically had one month to maturity when issued.

Nelson also diffusely points to dynamic delta hedging, and that it worked better in theory than practice (see Haug [123]. It is clear from all the details described by Nelson that options in the early 1900 traded actively and that option traders at that time in no way felt helpless in either pricing or in hedging them.

Herbert Filer was another option trader that was involved in option trading from 1919 to the 1960s. Filer(1959) describes what must be considered a reasonable active option market in New York and Europe in the early 1920s and 1930s. Filer mentions however that due to World War II there was no trading on the European Exchanges, for they were closed. Further, he mentions that London option trading did not resume before 1958. In the early 1900 Źs, option traders in London were considered to be the most sophisticated, according to [175]. It could well be that World War II and the subsequent shutdown of option trading for many years was the reason known robust arbitrage principles about options were forgotten and almost lost, to be partly re-discovered by finance professors such as Stoll.

Earlier, in 1908, Vinzenz Bronzin published a book deriving several option pricing formulas, and a formula very similar to what today is known as the Black-Scholes-Merton formula, see also Hafner and Zimmermann (2007, 2009) [116]. Bronzin based his risk-neutral option valuation on robust arbitrage principles such as the put-call parity and the link between the forward price and call and put options in a way that was rediscovered by Derman and Taleb (2005)[9]. Indeed, the put-call parity restriction is sufficient to remove the need to incorporate a future return in the underlying security it forces the lining up of options to the forward price[10].

Again, 1910 Henry Deutsch describes put-call parity but in less detail than Higgins and Nelson. In 1961 Reinach again described the put-call parity in quite some detail (another text typically ignored by academics). Traders at New York stock exchange specializing in using the put-call parity to convert puts into calls or calls into puts was at that time known as Converters. Reinach (1961) [195]:

9 The argument Derman Taleb(2005) [62] was present in [222] but remained unnoticed.

10 Ruffino and Treussard (2006) [201] accept that one could have solved the risk-premium by happenstance, not realizing that put-call parity was so extensively used in history. But they find it insufficient. Indeed the argument may not be sufficient for someone who subsequently complicated the representation of the world with some implements of modern finance such as "stochastic discount rates" while simplifying it at the same time to make it limited to the Gaussian and allowing dynamic hedging. They write that the use of a non-stochastic discount rate common to both the call and the put options is inconsistent with modern equilibrium capital asset pricing theory. Given that we have never seen a practitioner use stochastic discount rate, we, like our option trading predecessors, feel that put-call parity is sufficient & does the job.
The situation is akin to that of scientists lecturing birds on how to fly, and taking credit for their subsequent performance except that here it would be lecturing them the wrong way.

> Although I have no figures to substantiate my claim, I estimate that over 60 percent of all Calls are made possible by the existence of Converters.

In other words the converters (dealers) who basically operated as market makers were able to operate and hedge most of their risk by statically hedging options with options. Reinach wrote that he was an option trader (Converter) and gave examples on how he and his colleagues tended to hedge and arbitrage options against options by taking advantage of options embedded in convertible bonds:

Writers and traders have figured out other procedures for making profits writing Puts & Calls. Most are too specialized for all but the seasoned professional. One such procedure is the ownership of a convertible bond and then writing of Calls against the stock into which the bonds are convertible. If the stock is called converted and the stock is delivered.

Higgins, Nelson and Reinach all describe the great importance of the put-call parity and to hedge options with options. Option traders were in no way helpless in hedging or pricing before the Black-Scholes-Merton formula. Based on simple arbitrage principles they were able to hedge options more robustly than with Black-Scholes-Merton. As already mentioned static market-neutral delta hedging was described by Higgins and Nelson in 1902 and 1904. Also, W. D. Gann (1937) discusses market neutral delta hedging for at-the-money options, but in much less details than Nelson (1904). Gann also indicates some forms of auxiliary dynamic hedging.

Mills (1927) illustrates how jumps and fat tails were present in the literature in the pre-Modern Portfolio Theory days. He writes: "(...) distribution may depart widely from the Gaussian type because the influence of one or two extreme price changes".

22.4.1 Option formulas and Delta Hedging

Which brings us to option pricing formulas. The first identifiable one was Bachelier (1900) [5]. Sprenkle in 1961 [212] extended Bacheliers work to assume lognormal rather than normal distributed asset price. It also avoids discounting (to no significant effect since many markets, particularly the U.S., option premia were paid at expiration).

James Boness (1964) [26] also assumed a lognormal asset price. He derives a formula for the price of a call option that is actually identical to the Black-Scholes-Merton 1973 formula, but the way Black, Scholes and Merton derived their formula based on continuous dynamic delta hedging or alternatively based on CAPM they were able to get independent of the expected rate of return. It is in other words not the formula itself that is considered the great discovery done by Black, Scholes and Merton, but how they derived it. This is among several others also pointed out by Rubinstein (2006) [203]:

The real significance of the formula to the financial theory of investment lies not in itself, but rather in how it was derived. Ten years earlier the same formula had been derived by Case M. Sprenkle [212] and A. James Boness [26].

Samuelson (1969) and Thorp (1969) published somewhat similar option pricing formulas to Boness and Sprenkle. Thorp (2007) claims that he actually had an identical formula to the Black-Scholes-Merton formula programmed into his computer years before Black, Scholes and Merton published their theory.

Now, delta hedging. As already mentioned static market-neutral delta hedging was clearly described by Higgins and Nelson 1902 and 1904. Thorp and Kassouf (1967) presented market neutral static delta hedging in more details, not only for at-the-money options, but for options with any delta. In his 1969 paper Thorp is shortly describing market neutral static delta hedging, also briefly pointed in the direction of some dynamic delta hedging, not as a central pricing device, but a risk-management tool. Filer also points to dynamic hedging of options, but without showing much knowledge about how to calculate the delta. Another ignored and forgotten text is a book/booklet published in 1970 by Arnold Bernhard & Co. The authors are clearly aware of market neutral static delta hedging or what they name balanced hedge for any level in the strike or asset price. This book has multiple examples of how to buy warrants or convertible bonds and construct a market neutral delta hedge by shorting the right amount of common shares. Arnold Bernhard & Co also published deltas for a large number of warrants and convertible bonds that they distributed to investors on Wall Street.

Referring to Thorp and Kassouf (1967), Black, Scholes and Merton took the idea of delta hedging one step further, Black and Scholes (1973):

If the hedge is maintained continuously, then the approximations mentioned above become exact, and the return on the hedged position is completely independent of the change in the value of the stock. In fact, the return on the hedged position becomes certain. This was pointed out to us by Robert Merton.

This may be a brilliant mathematical idea, but option trading is not mathematical theory. It is not enough to have a theoretical idea so far removed from reality that is far from robust in practice. What is surprising is that the only principle option traders do not use and cannot use is the approach named after the formula, which is a point we discuss next.

22.5 MYTH 2: TRADERS TODAY USE BLACK-SCHOLES

Traders don'Źt do Valuation.

First, operationally, a price is not quite valuation. Valuation requires a strong theoretical framework with its corresponding fragility to both assumptions and the structure of a model. For traders, a price produced to buy an option when one has no knowledge of the probability distribution of the future is not valuation, but an expedient. Such price could change. Their beliefs do not enter such price. It can also be determined by his inventory.

This distinction is critical: traders are engineers, whether boundedly rational (or even non interested in any form of probabilistic rationality), they are not privy to informational transparency about the future states of the world and their probabilities. So they do not need a general theory to produce a price merely the avoidance of Dutch-book style arbitrages against them, and the compatibility with some standard restriction: In addition to put-call parity, a call of a certain strike K cannot trade at a lower price than a call $K + \Delta K$ (avoidance of negative call and put spreads), a call struck at K and a call struck at $K + 2\Delta K$ cannot be more expensive than twice the price of a call struck at $K + \Delta$ (negative butterflies), horizontal calendar spreads cannot be negative (when interest rates are low), and so forth. The degrees of freedom for traders are thus reduced: they need to abide by put-call parity and compatibility with other options in the market.

In that sense, traders do not perform valuation with some pricing kernel until the expiration of the security, but, rather, produce a price of an option compatible with other instruments in the markets, with a holding time that is stochastic. They do not need top-down science.

22.5.1 When do we value?

If you find traders operated solo, in a desert island, having for some to produce an option price and hold it to expiration, in a market in which the forward is absent, then some valuation would be necessary but then their book would be minuscule. And this thought experiment is a distortion: people would not trade options unless they are in the business of trading options, in which case they would need to have a book with offsetting trades. For without offsetting trades, we doubt traders would be able to produce a position beyond a minimum (and negligible) size as dynamic hedging not possible. (Again we are not aware of many non-blownup option traders and institutions who have managed to operate in the vacuum of the Black Scholes-Merton argument). It is to the impossibility of such hedging that we turn next.

22.6 ON THE MATHEMATICAL IMPOSSIBILITY OF DYNAMIC HEDGING

Finally, we discuss the severe flaw in the dynamic hedging concept. It assumes, nay, requires all moments of the probability distribution to exist[11].

Assume that the distribution of returns has a scale-free or fractal property that we can simplify as follows: for x large enough, (i.e. in the tails), $\frac{\mathbb{P}(X>nx)}{\mathbb{P}(X>x)}$ depends on n, not on x. In financial securities, say, where X is a daily return, there is no reason for P(X>20%)/P(X>10%) to be different from P(X>15%)/P(X>7.5%). This self-similarity at all scales generates power law, or Paretian, tails, i.e., above a crossover point, $\mathbb{P}(X > x) = Kx^{\alpha}$. It happens, looking at millions of pieces of

11 Merton (1992) seemed to accept the inapplicability of dynamic hedging but he perhaps thought that these ills would be cured thanks to his prediction of the financial world "spiraling towards dynamic completeness". Fifteen years later, we have, if anything, spiraled away from it.

data, that such property holds in markets all markets, baring sample error. For overwhelming empirical evidence, see Mandelbrot (1963), which predates Black-Scholes-Merton (1973) and the jump-diffusion of Merton (1976); see also Stanley et al. (2000), and Gabaix et al. (2003). The argument to assume the scale-free is as follows: the distribution might have thin tails at some point (say above some value of X). But we do not know where such point is we are epistemologically in the dark as to where to put the boundary, which forces us to use infinity.

Some criticism of these "true fat-tails" accept that such property might apply for daily returns, but, owing to the Central Limit Theorem , the distribution is held to become Gaussian under aggregation for cases in which α is deemed higher than 2. Such argument does not hold owing to the preasymptotics of scalable distributions: Bouchaud and Potters (2003) and Mandelbrot and Taleb (2007) argue that the presasymptotics of fractal distributions are such that the effect of the Central Limit Theorem are exceedingly slow in the tails in fact irrelevant. Furthermore, there is sampling error as we have less data for longer periods, hence fewer tail episodes, which give an in-sample illusion of thinner tails. In addition, the point that aggregation thins out the tails does not hold for dynamic hedging in which the operator depends necessarily on high frequency data and their statistical properties. So long as it is scale-free at the time period of dynamic hedge, higher moments become explosive, infinite to disallow the formation of a dynamically hedge portfolio. Simply a Taylor expansion is impossible as moments of higher order that 2 matter critically one of the moments is going to be infinite.

The mechanics of dynamic hedging are as follows. Assume the risk-free interest rate of 0 with no loss of generality. The canonical Black-Scholes-Merton package consists in selling a call and purchasing shares of stock that provide a hedge against instantaneous moves in the security. Thus the portfolio π locally "hedged" against exposure to the first moment of the distribution is the following:

$$\pi = -C + \frac{\partial C}{\partial S} S$$

where C is the call price, and S the underlying security. Take the discrete time change in the values of the portfolio

$$\Delta\pi = -\Delta C + \frac{\partial C}{\partial S} \Delta S$$

By expanding around the initial values of S, we have the changes in the portfolio in discrete time. Conventional option theory applies to the Gaussian in which all orders higher than ΔS^2 disappear rapidly.

Taking expectations on both sides, we can see here very strict requirements on moment finiteness: all moments need to converge. If we include another term, of order ΔS^3, such term may be of significance in a probability distribution with significant cubic or quartic terms. Indeed, although the nth derivative with respect to S can decline very sharply, for options that have a strike K away from the center of the distribution, it remains that the delivered higher orders of S are rising disproportionately fast for that to carry a mitigating effect on the hedges. So here

we mean all moments–no approximation. The logic of the Black-Scholes-Merton so-called solution thanks to Ito's lemma was that the portfolio collapses into a deterministic payoff. But let us see how quickly or effectively this works in practice. The Actual Replication process is as follows: The payoff of a call should be replicated with the following stream of dynamic hedges, the limit of which can be seen here, between t and T:

$$\lim_{\Delta t \to 0} \left(\sum_{i=1}^{n=T/\Delta t} \frac{\partial C}{\partial S} |_{S=S_{t+(i-1)\Delta t}, t=t+(i-1)\Delta t, \left(S_{t+i\Delta t} - S_{t+(i-1)\Delta t}\right)} \right) \tag{22.1}$$

Such policy does not match the call value: the difference remains stochastic (while according to Black Scholes it should shrink), unless one lives in a fantasy world in which such risk reduction is possible.

Further, there is an inconsistency in the works of Merton making us confused as to what theory finds acceptable: in Merton (1976) he agrees that we can use Bachelier-style option derivation in the presence of jumps and discontinuities, no dynamic hedging but only when the underlying stock price is uncorrelated to the market. This seems to be an admission that dynamic hedging argument applies only to some securities: those that do not jump and are correlated to the market.

22.6.1 The (confusing) Robustness of the Gaussian

The success of the formula last developed by Thorp, and called Black-Scholes-Merton was due to a simple attribute of the Gaussian: you can express any probability distribution in terms of Gaussian, even if it has fat tails, by varying the standard deviation σ at the level of the density of the random variable. It does not mean that you are using a Gaussian, nor does it mean that the Gaussian is particularly parsimonious (since you have to attach a σ for every level of the price). It simply mean that the Gaussian can express anything you want if you add a function for the parameter σ, making it a function of strike price and time to expiration.

This volatility smile, i.e., varying one parameter to produce $\sigma(K)$, or volatility surface, varying two parameter, $\sigma(S, t)$ is effectively what was done in different ways by Dupire (1994, 2005) [71, 72] and Derman [60, 63] see Gatheral (2006 [102]). They assume a volatility process not because there is necessarily such a thing only as a method of fitting option prices to a Gaussian. Furthermore, although the Gaussian has finite second moment (and finite all higher moments as well), you can express a scalable with infinite variance using Gaussian volatility surface. One strong constrain on the σ parameter is that it must be the same for a put and call with same strike (if both are European-style), and the drift should be that of the forward.

Indeed, ironically, the volatility smile is inconsistent with the Black-Scholes-Merton theory. This has lead to hundreds if not thousands of papers trying extend (what was perceived to be) the Black-Scholes-Merton model to incorporate stochastic volatility and jump-diffusion. Several of these researchers have been surprised that so few traders actually use stochastic volatility models. It is not a model that

says how the volatility smile should look like, or evolves over time; it is a hedging method that is robust and consistent with an arbitrage free volatility surface that evolves over time.

In other words, you can use a volatility surface as a map, not a territory. However it is foolish to justify Black-Scholes-Merton on grounds of its use: we repeat that the Gaussian bans the use of probability distributions that are not Gaussian whereas non-dynamic hedging derivations (Bachelier, Thorp) are not grounded in the Gaussian.

22.6.2 Order Flow and Options

It is clear that option traders are not necessarily interested in probability distribution at expiration time â given that this is abstract, even metaphysical for them. In addition to the put-call parity constrains that according to evidence was fully developed already in 1904, we can hedge away inventory risk in options with other options. One very important implication of this method is that if you hedge options with options then option pricing will be largely demand and supply based . This in strong contrast to the Black-Scholes-Merton (1973) theory that based on the idealized world of geometric Brownian motion with continuous-time delta hedging then demand and supply for options simply should not affect the price of options. If someone wants to buy more options the market makers can simply manufacture them by dynamic delta hedging that will be a perfect substitute for the option itself.

This raises a critical point: option traders do not estimate the odds of rare events by pricing out-of-the-money options. They just respond to supply and demand. The notion of implied probability distribution is merely a Dutch-book compatibility type of proposition.

22.6.3 Bachelier-Thorp

The argument often casually propounded attributing the success of option volume to the quality of the Black Scholes formula is rather weak. It is particularly weakened by the fact that options had been so successful at different time periods and places.

Furthermore, there is evidence that while both the Chicago Board Options Exchange and the Black-Scholes-Merton formula came about in 1973, the model was "rarely used by traders" before the 1980s (O 'Connell, 2001). When one of the authors (Taleb) became a pit trader in 1992, almost two decades after Black-Scholes-Merton, he was surprised to find that many traders still priced options sheets free, pricing off the butterfly, and off the conversion, without recourse to any formula.

Even a book written in 1975 by a finance academic appears to credit Thorpe and Kassouf (1967) – rather than Black-Scholes (1973), although the latter was present in its bibliography. Auster (1975):

Sidney Fried wrote on warrant hedges before 1950, but it was not until 1967 that the book Beat the Market by Edward O. Thorp and Sheen T. Kassouf rigorously, but simply, explained the short warrant/long common hedge to a wide audience.

We conclude with the following remark. Sadly, all the equations, from the first (Bachelier), to the last pre-Black-Scholes-Merton (Thorp) accommodate a scale-free distribution. The notion of explicitly removing the expectation from the forward was present in Keynes (1924) and later by Blau (1944) â and long a Call short a put of the same strike equals a forward. These arbitrage relationships appeared to be well known in 1904.

One could easily attribute the explosion in option volume to the computer age and the ease of processing transactions, added to the long stretch of peaceful economic growth and absence of hyperinflation. From the evidence (once one removes the propaganda), the development of scholastic finance appears to be an epiphenomenon rather than a cause of option trading. Once again, lecturing birds how to fly does not allow one to take subsequent credit.

This is why we call the equation Bachelier-Thorp. We were using it all along and gave it the wrong name, after the wrong method and with attribution to the wrong persons. It does not mean that dynamic hedging is out of the question; it is just not a central part of the pricing paradigm. It led to the writing down of a certain stochastic process that may have its uses, some day, should markets spiral towards dynamic completeness. But not in the present.

23 OPTION PRICING UNDER POWER LAWS: A ROBUST HEURISTIC*,‡

In this (research) chapter, we build a heuristic that takes a given option price in the tails with strike K and extends (for calls, all strikes $> K$, for put all strikes $< K$) assuming the continuation falls into what we define as "Karamata constant" or "Karamata point" beyond which the strong Pareto law holds. The heuristic produces relative prices for options, with for sole parameter the tail index α under some mild arbitrage constraints.

Usual restrictions such as finiteness of variance are not required.

The heuristic allows us to scrutinize the volatility surface and test theories of relative tail option mispricing and overpricing usually built on thin tailed models and modification of the Black-Scholes formula.

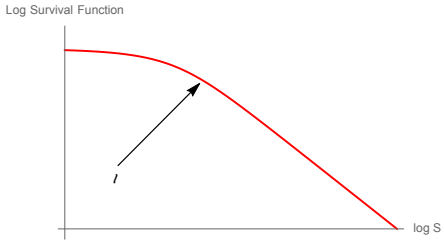


Figure 23.1: *The Karamata point where the slowly moving function is safely replaced by a constant $L(S) = l$. The constant varies whether we use the price S or its geometric return –but not the asymptotic slope which corresponds to the tail index α.*

Research chapter, with the Universa team: Brandon Yarckin, Chitpuneet Mann, Damir Delic, and Mark Spitznagel.

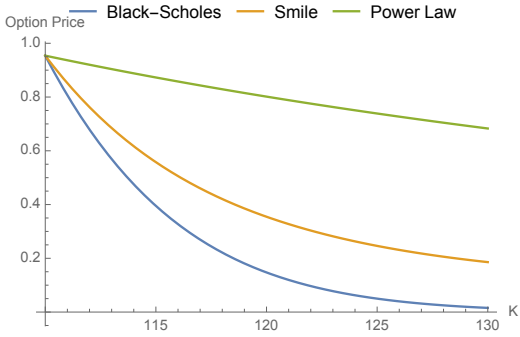


Figure 23.2: *We show a straight Black Scholes option price (constant volatility), one with a volatility "smile", i.e. the scale increases in the tails, and power law option prices. Under the simplified case of a power law distribution for the underlying, option prices are linear to strike.*

23.1 INTRODUCTION

The power law class is conventionally defined by the property of the survival function, as follows. Let X be a random variable belonging to the class of distributions with a "power law" right tail, that is:

$$\mathbb{P}(X > x) \sim L(x)\, x^{-\alpha} \tag{23.1}$$

where $L : [x_{\min}, +\infty) \to (0, +\infty)$ is a slowly varying function, defined as $\lim_{x\to+\infty} \frac{L(kx)}{L(x)} = 1$ for any $k > 0$ [22].

The survival function of X is called to belong to the "regular variation" class RV_α. More specifically, a function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is index varying at infinity with index ρ ($f \in RV_\rho$) when

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\rho.$$

More practically, there is a point where $L(x)$ approaches its limit, l, becoming a constant as in Fig. 23.1–we call it the "Karamata constant". Beyond such value the tails for power laws are calibrated using such standard techniques as the Hill estimator. The distribution in that zone is dubbed the strong Pareto law by B. Mandelbrot [160],[74].

23.2 CALL PRICING BEYOND THE KARAMATA CONSTANT

Now define a European call price $C(K)$ with a strike K and an underlying price S, $K, S \in (0, +\infty)$, as $(S - K)^+$, with its valuation performed under some probability measure $\mathbb{P}$, thus allowing us to price the option as $\mathbb{E}_P(S - K)^+ = \int_K^\infty (S - K) dP$. This allows us to immediately prove the following.

23.2.1 First approach, S is in the regular variation class

We start with a simplified case, to build the intuition. Let S have a survival function in the regular variation class RV_α as per 23.1. For all $K > l$ and $\alpha > 1$,

$$C(K) = \frac{K^{1-\alpha} l^\alpha}{\alpha - 1} \tag{23.2}$$

Remark 20

We note that the parameter l, when derived from an existing option price, contains all necessary information about the probability distribution below $S = l$, which under a given α parameter makes it unnecessary to estimate the mean, the "volatility" (that is, scale) and other attributes.

Let us assume that α is exogenously set (derived from fitting distributions, or, simply from experience, in both cases α is supposed to fluctuate minimally [236]). We note that $C(K)$ is invariant to distribution calibrations and the only parameters needed l which, being constant, disappears in ratios. Now consider as set the market price of an "anchor" tail option in the market is C_m with strike K_1, defined as an option for the strike of which other options are priced in relative value. We can simply generate all further strikes from $l = \left((\alpha - 1) C_m K_1^{\alpha - 1}\right)^{1/\alpha}$ and applying Eq. 23.2.

Result 1: Relative Pricing under Distribution for S

For $K_1, K_2 \geq l$,

$$C(K_2) = \left(\frac{K_2}{K_1}\right)^{1-\alpha} C(K_1). \tag{23.3}$$

The advantage is that all parameters in the distributions are eliminated: all we need is the price of the tail option and the α to build a unique pricing mechanism.

Remark 21: Avoiding confusion about L and α

The tail index α and Karamata constant l should correspond to the assigned distribution for the specific underlying. A tail index α for S in the regular variation class as as per 23.1 leading to Eq. 23.2 is different from that for $r = \frac{S - S_0}{S_0} \in RV_\alpha$. For consistency, each should have its own Zipf plot and other representations.

1. *If $\mathbb{P}(X > x) \sim L_a(x)\, x^{-\alpha}$, and $\mathbb{P}(\frac{X - X_0}{X_0} > \frac{x - X_0}{X_0}) \sim L_b(x)\, x^{-\alpha}$, the α constant will be the same, but the the various $L_{(.)}$ will be reaching their constant level at a different rate.*
2. *If $r_c = \log \frac{S}{S_0}$, it is not in the regular variation class, see theorem.*

The reason α stays the same is owing to the scale-free attribute of the tail index.

Theorem 6: Log returns

Let S be a random variable with survival function $\varphi(s) = L(s)s^{-\alpha} \in RV_\alpha$, where $L(.)$ is a slowly varying function. Let r_l be the log return $r_l = \log \frac{s}{s_0}$. $\varphi_{r_l}(r_l)$ is not in the RV_α class.

Proof. Immediate. The transformation $\varphi_{r_l}(r_l) = L(s)s^{-\frac{\log\left(\log^\alpha(s)\right)}{\log(s)}}$. □

We note, however, that in practice, although we may need continuous compounding to build dynamics [226], our approach assumes such dynamics are contained in the anchor option price selected for the analysis (or l). Furthermore there is no tangible difference, outside the far tail, between $\log \frac{S}{S_0}$ and $\frac{S-S_0}{S_0}$.

23.2.2 Second approach, S has geometric returns in the regular variation class

Let us now apply to real world cases where the returns $\frac{S-S_0}{S_0}$ are Paretian. Consider, for $r > l$, $S = (1+r)S_0$, where S_0 is the initial value of the underlying and $r \sim \mathcal{P}(l, \alpha)$ (Pareto I distribution) with survival function

$$\left(\frac{K - S_0}{lS_0}\right)^{-\alpha}, \quad K > S_0(1+l) \tag{23.4}$$

and fit to C_m using $l = \frac{(\alpha-1)^{1/\alpha} C_m^{1/\alpha} (K-S_0)^{1-\frac{1}{\alpha}}}{S_0}$, which, as before shows that practically all information about the distribution is embedded in l.

Let $\frac{S-S_0}{S_0}$ be in the regular variation class. For $S \geq S_0(1+l)$,

$$C(K, S_0) = \frac{(l\ S_0)^\alpha (K - S_0)^{1-\alpha}}{\alpha - 1} \tag{23.5}$$

We can thus rewrite Eq. 23.3 to eliminate l:

Result 2: Relative Pricing under Distribution for $\frac{S-S_0}{S_0}$

For $K_1, K_2 \geq (1+l)S_0$,

$$C(K_2) = \left(\frac{K_2 - S_0}{K_1 - S_0}\right)^{1-\alpha} C(K_1). \tag{23.6}$$

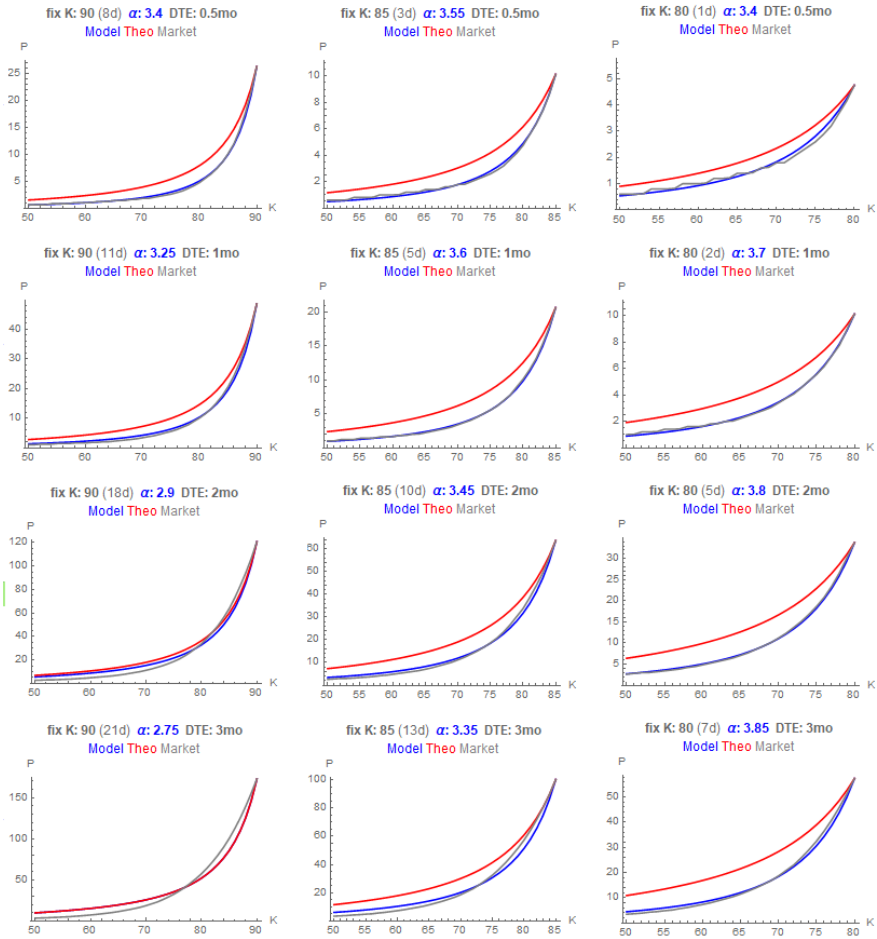


Figure 23.3: *Put Prices in the SP500 using "fix K" as anchor (from Dec 31, 2018 settlement), and generating an option prices using a tail index α that matches the market (blue) ("model), and in red prices for $\alpha = 2.75$. We can see that market prices tend to 1) fit a power law (matches stochastic volatility with fudged parameters), 2) but with an α that thins the tails. This shows how models claiming overpricing of tails are grossly misspecified.*

Remark 22

Unlike the pricing methods in the Black-Scholes modification class (stochastic and local volatility models, (see the expositions of Dupire, Derman and Gatheral, [73] [101], [59], finiteness of variance is not required for our model or option pricing in general, as shown in [226]. The only requirement is $\alpha > 1$, that is, finite first moment.

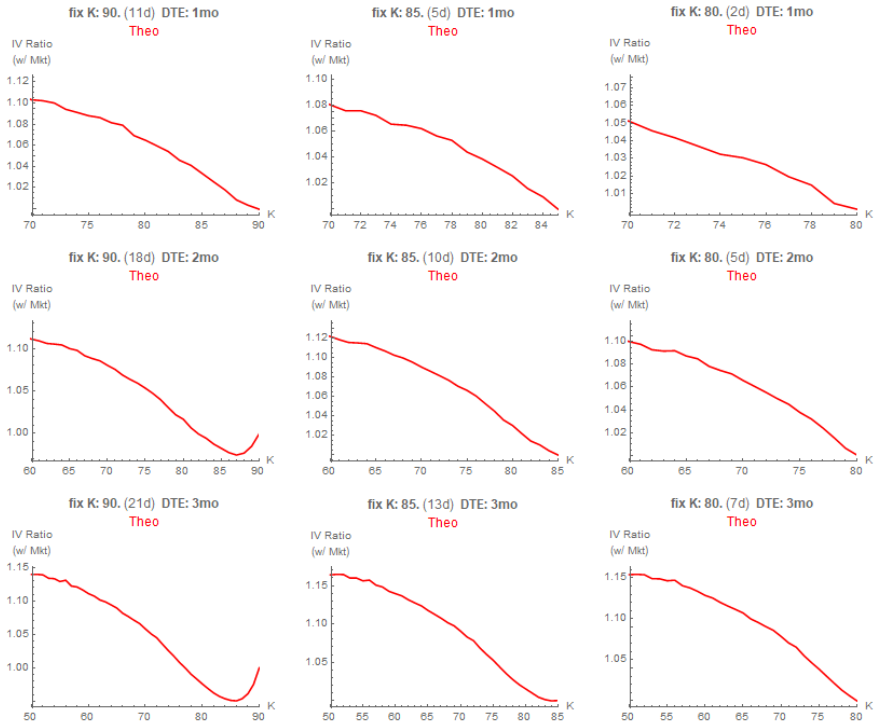


Figure 23.4: *Same results as in Fig 23.3 but expressed using implied volatility. We match the price to implied volatility for downside strikes (anchor 90, 85, and 80) using our model vs market, in ratios. We assume $\alpha = 2.75$.*

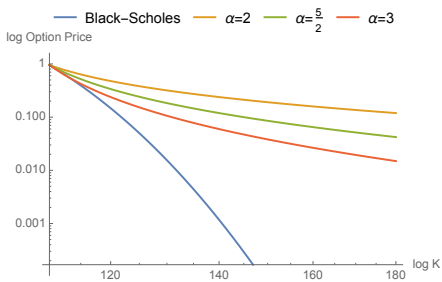


Figure 23.5: *The intuition of the Log log plot for the second calibration*

23.3 PUT PRICING

We now consider the put strikes (or the corresponding calls in the negative tail, which should be priced via put-call parity arbitrage). Unlike with calls, we can only consider the variations of $\frac{S-S_0}{S_0}$, not the logarithmic returns (nor those of S taken separately).

We construct the negative side with a negative return for the underlying. Let r be the rate of return $S = (1-r)S_0$, and Let $r > l > 0$ be Pareto distributed in the positive domain, with density $f_r(r) = \alpha\ l^\alpha r^{-\alpha-1}$. We have by probabilistic transformation and rescaling the PDF of the underlying:

$$f_S(S) = -\frac{\alpha\left(-\frac{S-S_0}{lS_0}\right)^{-\alpha-1}}{lS_0}\lambda \quad S \in [0, (1-l)S_0)$$

where the scaling constant $\lambda = \left(\frac{1}{(-1)^{\alpha+1}(l^\alpha-1)}\right)$ is set in a way to get $f_s(S)$ to integrate to 1. The parameter λ, however, is close to 1, making the correction negligible, in applications where $\sigma\sqrt{t} \leq \frac{1}{2}$ (σ being the Black-Scholes equivalent implied volatility and t the time to expiration of the option).

Remarkably, both the parameters l and the scaling λ are eliminated.

Result 3: Put Pricing

For $K_1, K_2 \leq (1-l)S_0$,

$$P(K_2) = P(K_1)\frac{(-1)^{1-\alpha}S_0^{-\alpha}\left((\alpha-1)K_2 + S_0\right) - (K_2 - S_0)^{1-\alpha}}{(-1)^{1-\alpha}S_0^{-\alpha}\left((\alpha-1)K_1 + S_0\right) - (K_1 - S_0)^{1-\alpha}} \tag{23.7}$$

23.4 ARBITRAGE BOUNDARIES

Obviously, there is no arbitrage for strikes higher than the baseline one K_1 in previous equations. For we can verify the Breeden-Litzenberger result [32], where the density is recovered from the second derivative of the option with respect to the strike $\frac{\partial^2 C(K)}{\partial K^2}|_{K\geq K_1} = \alpha K^{-\alpha-1}L^\alpha \geq 0$.

However there remains the possibility of arbitrage between strikes $K_1 + \Delta K$, K_1, and $K_1 - \Delta K$ by violating the following boundary: let $BSC(K, \sigma(K))$ be the Black-Scholes value of the call for strike K with volatility $\sigma(K)$ a function of the strike and t time to expiration. We have

$$C(K_1 + \Delta K) + BSC(K_1 - \Delta K) \geq 2\ C(K_1), \tag{23.8}$$

where $BSC(K_1, \sigma(K_1)) = C(K_1)$. For inequality 23.8 to be satisfied, we further need an inequality of call spreads, taken to the limit:

$$\frac{\partial BSC(K, \sigma(K))}{\partial K}|_{K=K_1} \geq \frac{\partial C(K)}{\partial K}|_{K=K_1} \tag{23.9}$$

Such an arbitrage puts a lower bound on the tail index α. Assuming 0 rates to simplify:

$$\alpha \geq \frac{1}{-\log(K-S_0)+\log(l)+\log(S_0)} \log\left(\frac{1}{2}\operatorname{erfc}\left(\frac{t\sigma(K)^2+2\log(K)-2\log(S_0)}{2\sqrt{2}\sqrt{t}\sigma(K)}\right) - \frac{\sqrt{S_0}\sqrt{t}\sigma'(K)K^{\frac{\log(S_0)}{t\sigma(K)^2}+\frac{1}{2}}\exp\left(-\frac{\log^2(K)+\log^2(S_0)}{2t\sigma(K)^2}-\frac{1}{8}t\sigma(K)^2\right)}{\sqrt{2\pi}}\right) \tag{23.10}$$

23.5 COMMENTS

As we can see in Fig. 23.5, stochastic volatility models and similar adaptations (say, jump-diffusion or standard Poisson variations) eventually fail "out in the tails" outside the zone for which they were calibrated. There has been poor attempts to extrapolate the option prices using a fudged thin-tailed probability distribution rather than a Paretian one –hence the numerous claims in the finance literature on "overpricing" of tail options combined with some psycholophastering on "dread risk" are unrigorous on that basis. The proposed methods allows us to approach such claims with more realism.

Finaly, note that our approach isn't about absolute mispricing of tail options, but relative to a given strike closer to the money.

ACKNOWLEDGMENTS

Bruno Dupire, Peter Carr, students at NYU Tandon School of Engineering.

24 FOUR MISTAKES IN QUANTITATIVE FINANCE*,‡

We discuss Jeff Holman's (who at the time was, surprisingly, a senior risk officer for a large hedge fund) comments in *Quantitative Finance* to illustrate four critical errors students should learn to avoid:

1. Mistaking tails (4th moment and higher) for volatility (2nd moment)
2. Missing Jensen's Inequality when calculating return potential
3. Analyzing the hedging results without the performance of the underlying
4. The necessity of a numéraire in finance.

The review of *Antifragile* by Mr Holman (Dec 4, 2013) is replete with factual, logical, and analytical errors. We will only list here the critical ones, and ones with generality to the risk management and quantitative finance communities; these should be taught to students in quantitative finance as central mistakes to avoid, so beginner quants and risk managers can learn from these fallacies.

24.1 CONFLATION OF SECOND AND FOURTH MOMENTS

It is critical for beginners not to fall for the following elementary mistake. Mr Holman gets the relation of the VIX (volatility contract) to betting on "tail events" backwards. Let us restate the notion of "tail events" in the *Incerto* (that is the four books on uncertainty of which *Antifragile* is the last installment): it means a disproportionate role of the tails in defining the properties of distribution, which, mathematically, means a smaller one for the "body".[2]

Discussion chapter.

2 The point is staring at every user of spreadsheets: kurtosis, or scaled fourth moment, the standard measure of fattailedness, entails normalizing the fourth moment by the square of the variance.

Mr Holman seems to get the latter part of the attributes of fattailedness in reverse. It is an error to mistake the VIX for tail events. The VIX is mostly affected by at-the-money options which corresponds to the center of the distribution, closer to the second moment not the fourth (at-the-money options are actually linear in their payoff and correspond to the conditional first moment). As explained about seventeen years ago in *Dynamic Hedging* (Taleb, 1997) (see appendix), in the discussion on such tail bets, or "fourth moment bets", betting on the disproportionate role of tail events of fattailedness is done by *selling* the around-the-money-options (the VIX) and purchasing options in the tails, in order to extract the second moment and achieve neutrality to it (sort of becoming "market neutral"). Such a neutrality requires some type of "short volatility" in the body because higher kurtosis means lower action in the center of the distribution.

A more mathematical formulation is in the technical version of the *Incerto* : fat tails means "higher peaks" for the distribution as, the fatter the tails, the more markets spend time between $\mu - \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma$ and $\mu + \sqrt{\frac{1}{2}\left(5-\sqrt{17}\right)}\sigma$ where σ is the standard deviation and μ the mean of the distribution (we used the Gaussian here as a base for ease of presentation but the argument applies to all unimodal distributions with "bell-shape" curves, known as semiconcave). And "higher peaks" means less variations that are not tail events, more quiet times, not less. For the consequence on option pricing, the reader might be interested in a quiz I routinely give students after the first lecture on derivatives: "What happens to at-the-money options when one fattens the tails?", the answer being that they should drop in value. [3]

Effectively, but in a deeper argument, in the QF paper (Taleb and Douady 2013), our measure of fragility has an opposite sensitivity to events around the center of the distribution, since, by an argument of survival probability, what is fragile is sensitive to tail shocks and, critically, should not vary in the body (otherwise it would be broken).

24.2 MISSING JENSEN'S INEQUALITY IN ANALYZING OPTION RETURNS

Here is an error to avoid at all costs in discussions of volatility strategies or, for that matter, anything in finance. Mr Holman seems to miss the existence of Jensen's inequality, which is the entire point of owning an option, a point that has been belabored in *Antifragile*. One manifestation of missing the convexity effect is a critical miscalculation in the way one can naively assume options respond to the VIX.

3 **Technical Point: Where Does the Tail Start?** As we saw in 4.3, for a general class of symmetric distributions with power laws, the tail starts at: $\pm\frac{\sqrt{\frac{5\alpha+\sqrt{(\alpha+1)(17\alpha+1)}+1}{\alpha-1}}s}{\sqrt{2}}$, 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 the heuristic definition of fragility as second order effect: the part of the distribution is convex to errors in the estimation of the scale. But in practice, because historical measurements of STD will be biased lower because of small sample effects (as we repeat fat tails accentuate small sample effects), the deviations will be $>$ 2-3 STDs.

> "A $1 investment on January 1, 2007 in a strategy of buying and rolling short-term VIX futures would have peaked at $4.84 on November 20, 2008 -and then subsequently lost 99% of its value over the next four and a half years, finishing under $0.05 as of May 31, 2013." [4]

This mistake in the example given underestimates option returns by up to... several orders of magnitude. Mr Holman analyzes the performance a tail strategy using investments in financial options by using the VIX (or VIX futures) as proxy, which is mathematically erroneous owing to second- order effects, as the link is tenuous (it would be like evaluating investments in ski resorts by analyzing temperature futures). Assume a periodic rolling of an option strategy: an option 5 STD away from the money [5] gains 16 times in value if its implied volatility goes up by 4, but only loses its value if volatility goes to 0. For a 10 STD it is 144 times. And, to show the acceleration, assuming these are traded, a 20 STD options by around 210,000 times[6]. There is a second critical mistake in the discussion: Mr Holman's calculations here exclude the payoff from actual in-the-moneyness.

One should remember that the VIX is not a price, but an inverse function, an index derived from a price: one does not buy "volatility" like one can buy a tomato; operators buy options correponding to such inverse function and there are severe, very severe nonlinearities in the effect. Although more linear than tail options, the VIX is still convex to actual market volatility, somewhere between variance and standard deviation, since a strip of options spanning all strikes should deliver the variance (Gatheral,2006). The reader can go through a simple exercise. Let's say that the VIX is "bought" at 10% -that is, the component options are purchased at a combination of volatilities that corresponds to a VIX at that level. Assume returns are in squares. Because of nonlinearity, the package could benefit from an episode of 4% volatility followed by an episode of 15%, for an average of 9.5%; Mr Holman believes or wants the reader to believe that this 0.5 percentage point should be treated as a loss when in fact second order un-evenness in volatility changes are more relevant than the first order effect.

24.3 THE INSEPARABILITY OF INSURANCE AND INSURED

One should never calculate the cost of insurance without offsetting it with returns generated from packages than one would not have purchased otherwise.

Even had he gotten the sign right on the volatility, Mr Holman in the example above analyzes the performance of a strategy buying options to protect a tail event without adding the performance of the portfolio itself, like counting the cost side of the insurance without the performance of what one is insuring that would not have been bought otherwise. Over the same period he discusses the market rose more than 100%: a healthy approach would be to compare dollar-for-dollar what

4 In the above discussion Mr Holman also shows evidence of dismal returns on index puts which, as we said before, respond to volatility not tail events. These are called, in the lingo, "sucker puts".

5 We are using implied volatility as a benchmark for its STD.

6 An event this author witnessed, in the liquidation of Victor Niederhoffer, options sold for $.05 were purchased back at up to $38, which bankrupted Refco, and, which is remarkable, without the options getting close to the money: it was just a panic rise in implied volatility.

an investor would have done (and, of course, getting rid of this "VIX" business and focusing on very small dollars invested in tail options that would allow such an aggressive stance). Many investors (such as this author) would have stayed out of the market, or would not have added funds to the market, without such an insurance.

24.4 THE NECESSITY OF A NUMÉRAIRE IN FINANCE

There is a deeper analytical error.

A barbell is defined as a bimodal investment strategy, presented as investing a portion of your portfolio in what is explicitly defined as a "numéraire repository of value" (*Antifragile*), and the rest in risky securities (*Antifragile* indicates that such numéraire would be, among other things, inflation protected). Mr Holman goes on and on in a nihilistic discourse on the absence of such riskless numéraire (of the sort that can lead to such sophistry as "he is saying one is safer on *terra firma* than at sea, but what if there is an earthquake?").

The familiar Black and Scholes derivation uses a riskless asset as a baseline; but the literature since around 1977 has substituted the notion of "cash" with that of a numéraire , along with the notion that one can have different currencies, which technically allows for changes of probability measure. A numéraire is defined as *the unit to which all other units relate*. (Practically, the numéraire is a basket the variations of which do not affect the welfare of the investor.) Alas, without numéraire, there is no probability measure, and no quantitative in quantitative finance, as one needs a unit to which everything else is brought back to. In this (emotional) discourse, Mr Holton is not just rejecting the barbell per se, but any use of the expectation operator with any economic variable, meaning he should go attack the tens of thousand research papers and the existence of the journal *Quantitative Finance* itself.

Clearly, there is a high density of other mistakes or incoherent statements in the outpour of rage in Mr Holman's review; but I have no doubt these have been detected by the *Quantitative Finance* reader and, as we said, the object of this discussion is the prevention of analytical mistakes in quantitative finance.

To conclude, this author welcomes criticism from the finance community that are not straw man arguments, or, as in the case of Mr Holmam, violate the foundations of the field itself.

24.5 APPENDIX (BETTING ON TAILS OF DISTRIBUTION)

From *Dynamic Hedging*, pages 264-265:

> *A fourth moment bet is long or short the volatility of volatility. It could be achieved either with out-of-the-money options or with calendars. Example: A ratio "backspread" or reverse spread is a method that includes the buying of out-of-the-money options in large amounts and the selling of smaller amounts of at-the-money but making sure the*

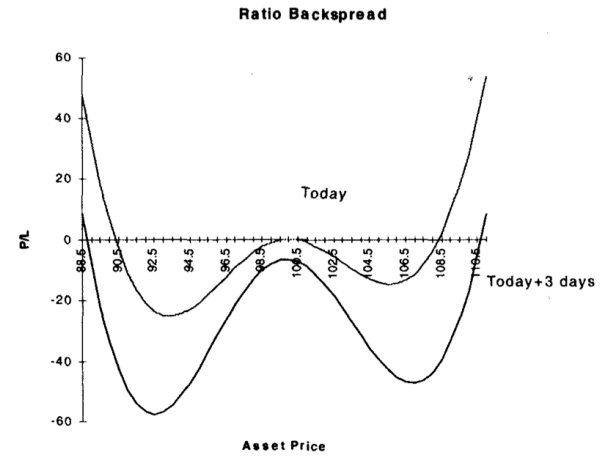


Figure 24.1: *First Method to Extract the Fourth Moment, from Dynamic Hedging, 1997.*

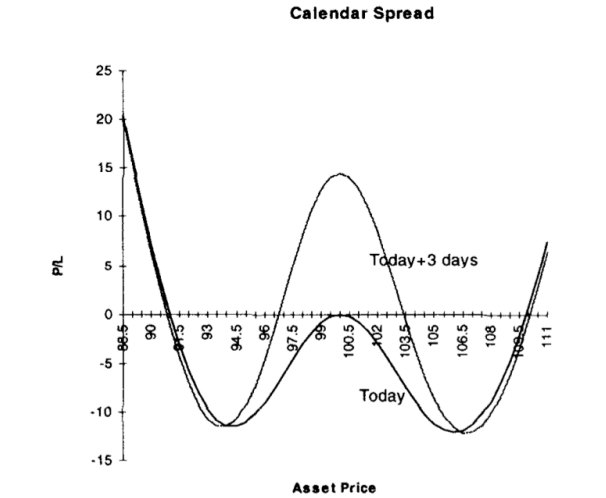


Figure 24.2: *Second Method to Extract the Fourth Moment , from Dynamic Hedging, 1997.*

trade satisfies the "credit" rule (i.e., the trade initially generates a positive cash flow). The credit rule is more difficult to interpret when one uses in-the-money options. In that case, one should deduct the present value of the intrinsic part of every option using the put-call parity rule to equate them with out-of-the-money.

The trade shown in Figure 1 was accomplished with the purchase of both out-of-the-money puts and out-of-the-money calls and the selling of smaller amounts of at-the-money straddles of the same maturity.

Figure 2 shows the second method, which entails the buying of 60- day options in some amount and selling 20-day options on 80% of the amount. Both trades show the position

benefiting from the fat tails and the high peaks. Both trades, however, will have different vega sensitivities, but close to flat modified vega.

See The Body, The Shoulders, and The Tails from section 4.3 where we assume tails start at the level of convexity of the segment of the probability distribution to the scale of the distribution.

25 TAIL RISK CONSTRAINTS AND MAXIMUM ENTROPY (W. D.& H. GEMAN)‡

ortfolio selection in the financial literature has essentially been analyzed under two central assumptions: full knowledge of the joint probability distribution of the returns of the securities that will comprise the target portfolio; and investors' preferences are expressed through a utility function. In the real world, operators build portfolios under risk constraints which are expressed both by their clients and regulators and which bear on the maximal loss that may be generated over a given time period at a given confidence level (the so-called *V*alue at Risk of the position). Interestingly, in the finance literature, a serious discussion of how much or little is known from a probabilistic standpoint about the multi-dimensional density of the assets' returns seems to be of limited relevance.

Our approach in contrast is to highlight these issues and then adopt throughout a framework of entropy maximization to represent the real world ignorance of the "true" probability distributions, both univariate and multivariate, of traded securities' returns. In this setting, we identify the optimal portfolio under a number of downside risk constraints. Two interesting results are exhibited: (i) the left- tail constraints are sufficiently powerful to override all other considerations in the conventional theory; (ii) the "barbell portfolio" (maximal certainty/ low risk in one set of holdings, maximal uncertainty in another), which is quite familiar to traders, naturally emerges in our construction.

25.1 LEFT TAIL RISK AS THE CENTRAL PORTFOLIO CONSTRAINT

Customarily, when working in an institutional framework, operators and risk takers principally use regulatorily mandated tail-loss limits to set risk levels in their

‡ Research chapter.

portfolios (obligatorily for banks since Basel II). They rely on stress tests, stop-losses, value at risk (VaR) , expected shortfall (–i.e., the expected loss conditional on the loss exceeding VaR, also known as CVaR), and similar loss curtailment methods, rather than utility. In particular, the margining of financial transactions is calibrated by clearing firms and exchanges on tail losses, seen both probabilistically and through stress testing. (In the risk-taking terminology, a stop loss is a mandatory order that attempts terminates all or a portion of the exposure upon a trigger, a certain pre-defined nominal loss. Basel II is a generally used name for recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision. The Value-at-risk, VaR, is defined as a threshold loss value K such that the probability that the loss on the portfolio over the given time horizon exceeds this value is ϵ. A stress test is an examination of the performance upon an arbitrarily set deviation in the underlying variables.) The information embedded in the choice of the constraint is, to say the least, a meaningful statistic about the appetite for risk and the shape of the desired distribution.

Operators are less concerned with portfolio variations than with the drawdown they may face over a time window. Further, they are in ignorance of the joint probability distribution of the components in their portfolio (except for a vague notion of association and hedges), but can control losses organically with allocation methods based on maximum risk. (The idea of substituting variance for risk can appear very strange to practitioners of risk-taking. The aim by Modern Portfolio Theory at lowering variance is inconsistent with the preferences of a rational investor, regardless of his risk aversion, since it also minimizes the variability in the profit domain –except in the very narrow situation of certainty about the future mean return, and in the far-fetched case where the investor can only invest in variables having a symmetric probability distribution, and/or only have a symmetric payoff. Stop losses and tail risk controls violate such symmetry.) The conventional notions of utility and variance may be used, but not directly as information about them is embedded in the tail loss constaint.

Since the stop loss, the VaR (and expected shortfall) approaches and other risk-control methods concern only one segment of the distribution, the negative side of the loss domain, we can get a dual approach akin to a portfolio separation, or "barbell-style" construction, as the investor can have opposite stances on different parts of the return distribution. Our definition of barbell here is the mixing of two extreme properties in a portfolio such as a linear combination of maximal conservatism for a fraction w of the portfolio, with $w \in (0, 1)$, on one hand and maximal (or high) risk on the $(1 - w)$ remaining fraction.

Historically, finance theory has had a preference for parametric, less robust, methods. The idea that a decision-maker has clear and error-free knowledge about the distribution of future payoffs has survived in spite of its lack of practical and theoretical validity –for instance, correlations are too unstable to yield precise measurements. It is an approach that is based on distributional and parametric certainties, one that may be useful for research but does not accommodate responsible risk taking. (Correlations are unstable in an unstable way, as joint returns for assets are not elliptical, see Bouchaud and Chicheportiche (2012) [42].)

There are roughly two traditions: one based on highly parametric decision-making by the economics establishment (largely represented by Markowitz [164]) and the other based on somewhat sparse assumptions and known as the Kelly criterion (Kelly, 1956 [140], see Bell and Cover, 1980 [15].) (In contrast to the minimum-variance approach, Kelly's method, developed around the same period as Markowitz, requires no joint distribution or utility function. In practice one needs the ratio of expected profit to worst-case return dynamically adjusted to avoid ruin. Obviously, model error is of smaller consequence under the Kelly criterion: Thorp (1969) [245], Haigh (2000) [118], Mac Lean, Ziemba and Blazenko [155]. For a discussion of the differences between the two approaches, see Samuelson's objection to the Kelly criterion and logarithmic sizing in Thorp 2010 [247].) Kelly's method is also related to left- tail control due to proportional investment, which automatically reduces the portfolio in the event of losses; but the original method requires a hard, non-parametric worst-case scenario, that is, securities that have a lower bound in their variations, akin to a gamble in a casino, which is something that, in finance, can only be accomplished through binary options. The Kelly criterion, in addition, requires some precise knowledge of future returns such as the mean. Our approach goes beyond the latter method in accommodating more uncertainty about the returns, whereby an operator can only control his left-tail via derivatives and other forms of insurance or dynamic portfolio construction based on stop-losses. (Xu, Wu, Jiang, and Song (2014) [261] contrast mean variance to maximum entropy and uses entropy to construct robust portfolios.) In a nutshell, we hardwire the curtailments on loss but otherwise assume maximal uncertainty about the returns. More precisely, we equate the return distribution with the *m*aximum entropy extension of constraints expressed as statistical expectations on the left-tail behavior as well as on the expectation of the return or log-return in the non-danger zone. (Note that we use *S*hannon entropy throughout. There are other information measures, such as Tsallis entropy [251] , a generalization of Shannon entropy, and Renyi entropy, [135] , some of which may be more convenient computationally in special cases. However, Shannon entropy is the best known and has a well-developed maximization framework.)

Here, the "left-tail behavior" refers to the hard, explicit, institutional constraints discussed above. We describe the shape and investigate other properties of the resulting so-called *m*axent distribution. In addition to a mathematical result revealing the link between acceptable tail loss (VaR) and the expected return in the Gaussian mean-variance framework, our contribution is then twofold: 1) an investigation of the shape of the distribution of returns from portfolio construction under more natural constraints than those imposed in the mean-variance method, and 2) the use of stochastic entropy to represent residual uncertainty.

VaR and CVaR methods are not error free –parametric VaR is known to be ineffective as a risk control method on its own. However, these methods can be made robust using constructions that, upon paying an insurance price, no longer depend on parametric assumptions. This can be done using derivative contracts or by organic construction (clearly if someone has 80% of his portfolio in numéraire securities, the risk of losing more than 20% is zero independent from all possible models of returns, as the fluctuations in the numéraire are not considered risky).

We use "pure robustness" or both VaR and zero shortfall via the "hard stop" or insurance, which is the special case in our paper of what we called earlier a "barbell" construction.

It is worth mentioning that it is an old idea in economics that an investor can build a portfolio based on two distinct risk categories, see Hicks (1939) [126]. Modern Portfolio Theory proposes the mutual fund theorem or "separation" theorem, namely that all investors can obtain their desired portfolio by mixing two mutual funds, one being the riskfree asset and one representing the optimal mean-variance portfolio that is tangent to their constraints; see Tobin (1958) [249], Markowitz (1959) [165], and the variations in Merton (1972) [168], Ross (1978) [199]. In our case a riskless asset is the part of the tail where risk is set to exactly zero. Note that the risky part of the portfolio needs to be minimum variance in traditional financial economics; for our method the exact opposite representation is taken for the risky one.

25.1.1 The Barbell as seen by E.T. Jaynes

Our approach to constrain only what can be constrained (in a robust manner) and to maximize entropy elsewhere echoes a remarkable insight by E.T. Jaynes in "How should we use entropy in economics?" [132]:

> "It may be that a macroeconomic system does not move in response to (or at least not solely in response to) the forces that are supposed to exist in current theories; it may simply move in the direction of increasing entropy as constrained by the conservation laws imposed by Nature and Government."

25.2 REVISITING THE MEAN VARIANCE SETTING

Let $\vec{X} = (X_1, ..., X_m)$ denote m asset returns over a given single period with joint density $g(\vec{x})$, mean returns $\vec{\mu} = (\mu_1, ..., \mu_m)$ and $m \times m$ covariance matrix Σ: $\Sigma_{ij} = \mathbb{E}(X_i X_j) - \mu_i \mu_j, 1 \leq i, j \leq m$. Assume that $\vec{\mu}$ and Σ can be reliably estimated from data.

The return on the portolio with weights $\vec{w} = (w_1, ..., w_m)$ is then

$$X = \sum_{i=1}^{m} w_i X_i,$$

which has mean and variance

$$\mathbb{E}(X) = \vec{w}\vec{\mu}^T, \quad V(X) = \vec{w}\Sigma\vec{w}^T.$$

In standard portfolio theory one minimizes $V(X)$ over all $\vec{w}$ subject to $\mathbb{E}(X) = \mu$ for a fixed desired average return μ. Equivalently, one maximizes the expected return

$\mathbb{E}(X)$ subject to a fixed variance $V(X)$. In this framework variance is taken as a substitute for risk.

To draw connections with our entropy-centered approach, we consider two standard cases:

(1) **N**ormal World: The joint distribution $g(\vec{x})$ of asset returns is multivariate Gaussian $N(\vec{\mu}, \Sigma)$. Assuming normality is equivalent to assuming $g(\vec{x})$ has maximum (Shannon) entropy among all multivariate distributions with the given first- and second-order statistics $\vec{\mu}$ and Σ. Moreover, for a fixed mean $\mathbb{E}(X)$, minimizing the variance $V(X)$ is equivalent to minimizing the entropy (uncertainty) of X. (This is true since joint normality implies that X is univariate normal for any choice of weights and the entropy of a $\mathcal{N}(\mu, \sigma^2)$ variable is $H = \frac{1}{2}(1 + \log(2\pi\sigma^2))$.) This is natural in a world with complete information. (The idea of entropy as mean uncertainty is in Philippatos and Wilson (1972) [185]; see Zhou –et al. (2013) [265] for a review of entropy in financial economics and Georgescu-Roegen (1971) [106] for economics in general.)

(2) **U**nknown Multivariate Distribution: Since we assume we can estimate the second-order structure, we can still carry out the Markowitz program, –i.e., choose the portfolio weights to find an optimal mean-variance performance, which determines $\mathbb{E}(X) = \mu$ and $V(X) = \sigma^2$. However, we do not know the distribution of the return X. Observe that *a*ssuming X is normally distributed $\mathcal{N}(\mu, \sigma^2)$ is equivalent to assuming the entropy of X is *m*aximized since, again, the normal maximizes entropy at a given mean and variance, see [185].

Our strategy is to generalize the second scenario by replacing the variance σ^2 by two left-tail value-at-risk constraints and to model the portfolio return as the maximum entropy extension of these constraints together with a constraint on the overall performance or on the growth of the portfolio in the non-danger zone.

25.2.1 Analyzing the Constraints

Let X have probability density $f(x)$. In everything that follows, let $K < 0$ be a normalizing constant chosen to be consistent with the risk-taker's wealth. For any $\epsilon > 0$ and $\nu_- < K$, the *v*alue-at-risk constraints are:

(1) **T**ail probability:

$$\mathbb{P}(X \leq K) = \int_{-\infty}^{K} f(x)\,\mathrm{d}x = \epsilon.$$

(2) **E**xpected shortfall (CVaR):

$$\mathbb{E}(X|X \leq K) = \nu_-.$$

Assuming (1) holds, constraint (2) is equivalent to

$$\mathbb{E}(XI_{(X\leq K)}) = \int_{-\infty}^{K} xf(x)\,\mathrm{d}x = \epsilon\nu_-.$$

Given the value-at-risk parameters $\theta = (K, \epsilon, \nu_-)$, let $\Omega_{var}(\theta)$ denote the set of probability densities f satisfying the two constraints. Notice that $\Omega_{var}(\theta)$ is convex: $f_1, f_2 \in \Omega_{var}(\theta)$ implies $\alpha f_1 + (1-\alpha) f_2 \in \Omega_{var}(\theta)$. Later we will add another constraint involving the overall mean.

25.3 REVISITING THE GAUSSIAN CASE

Suppose we assume X is Gaussian with mean μ and variance σ^2. In principle it should be possible to satisfy the VaR constraints since we have two free parameters. Indeed, as shown below, the left-tail constraints determine the mean and variance; see Figure 25.1. However, satisfying the VaR constraints imposes interesting restrictions on μ and σ and leads to a natural inequality of a "no free lunch" style.

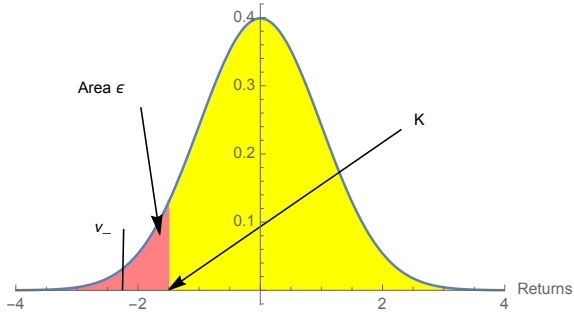


Figure 25.1: *By setting K (the value at risk), the probability ϵ of exceeding it, and the shortfall when doing so, there is no wiggle room left under a Gaussian distribution: σ and μ are determined, which makes construction according to portfolio theory less relevant.*

Let $\eta(\epsilon)$ be the ϵ-quantile of the standard normal distribution, –i.e., $\eta(\epsilon) = \Phi^{-1}(\epsilon)$, where Φ is the c.d.f. of the standard normal density $\phi(x)$. In addition, set

$$B(\epsilon) = \frac{1}{\epsilon \eta(\epsilon)} \phi(\eta(\epsilon)) = \frac{1}{\sqrt{2\pi}\epsilon\eta(\epsilon)} \exp\{-\frac{\eta(\epsilon)^2}{2}\}.$$

Proposition 25.1
If $X \sim N(\mu, \sigma^2)$ and satisfies the two VaR constraints, then the mean and variance are given by:

$$\mu = \frac{\nu_- + KB(\epsilon)}{1 + B(\epsilon)}, \quad \sigma = \frac{K - \nu_-}{\eta(\epsilon)(1 + B(\epsilon))}.$$

Moreover, $B(\epsilon) < -1$ and $\lim_{\epsilon \downarrow 0} B(\epsilon) = -1$.

The proof is in the Appendix. The VaR constraints lead directly to two linear equations in μ and σ:

$$\mu + \eta(\epsilon)\sigma = K, \quad \mu - \eta(\epsilon)B(\epsilon)\sigma = \nu_-.$$

Consider the conditions under which the VaR constraints allow a *positive* mean return $\mu = \mathbb{E}(X) > 0$. First, from the above linear equation in μ and σ in terms

of $\eta(\epsilon)$ and K, we see that σ increases as ϵ increases for any fixed mean μ, and that $\mu > 0$ if and only if $\sigma > \frac{K}{\eta(\epsilon)}$, –i.e., we must accept a lower bound on the variance which increases with ϵ, which is a reasonable property. Second, from the expression for μ in Proposition 1, we have

$$\mu > 0 \iff |\nu_-| > KB(\epsilon).$$

Consequently, the only way to have a positive expected return is to accommodate a sufficiently large risk expressed by the various tradeoffs among the risk parameters θ satisfying the inequality above. (This type of restriction also applies more generally to symmetric distributions since the left tail constraints impose a structure on the location and scale. For instance, in the case of a Student T distribution with scale s, location m, and tail exponent α, the same linear relation between s and m applies: $s = (K - m)\kappa(\alpha)$, where $\kappa(\alpha) = -\frac{i\sqrt{I^{-1}_{2\epsilon}\left(\frac{\alpha}{2},\frac{1}{2}\right)}}{\sqrt{\alpha}\sqrt{I^{-1}_{2\epsilon}\left(\frac{\alpha}{2},\frac{1}{2}\right)-1}}$, where I^{-1} is the inverse of the regularized incomplete beta function I, and s the solution of $\epsilon = \frac{1}{2} I_{\frac{\alpha s^2}{(k-m)^2+\alpha s^2}}\left(\frac{\alpha}{2}, \frac{1}{2}\right)$.)

25.3.1 A Mixture of Two Normals

In many applied sciences, a mixture of two normals provides a useful and natural extension of the Gaussian itself; in finance, the Mixture Distribution Hypothesis (denoted as MDH in the literature) refers to a mixture of two normals and has been very widely investigated (see for instance Richardson and Smith (1995) [197]). H. Geman and T. Ané (1996) [2] exhibit how an infinite mixture of normal distributions for stock returns arises from the introduction of a "stochastic clock" accounting for the uneven arrival rate of information flow in the financial markets. In addition, option traders have long used mixtures to account for fat tails, and to examine the sensitivity of a portfolio to an increase in kurtosis ("DvegaDvol"); see Taleb (1997) [222]. Finally, Brigo and Mercurio (2002) [34] use a mixture of two normals to calibrate the skew in equity options.

Consider the mixture

$$f(x) = \lambda N(\mu_1, \sigma_1^2) + (1 - \lambda)N(\mu_2, \sigma_2^2).$$

An intuitively simple and appealing case is to fix the overall mean μ, and take $\lambda = \epsilon$ and $\mu_1 = \nu_-$, in which case μ_2 is constrained to be $\frac{\mu - \epsilon\nu_-}{1-\epsilon}$. It then follows that the left-tail constraints are approximately satisfied for σ_1, σ_2 sufficiently small. Indeed, when $\sigma_1 = \sigma_2 \approx 0$, the density is effectively composed of two spikes (small variance normals) with the left one centered at ν_- and the right one centered at at $\frac{\mu - \epsilon\nu_-}{1-\epsilon}$. The extreme case is a Dirac function on the left, as we see next.

Dynamic Stop Loss, A Brief Comment One can set a level K below which there is no mass, with results that depend on accuracy of the execution of such a stop. The distribution to the right of the stop-loss no longer looks like the standard

Gaussian, as it builds positive skewness in accordance to the distance of the stop from the mean. We limit any further discussion to the illustrations in Figure **??**.

25.4 MAXIMUM ENTROPY

From the comments and analysis above, it is clear that, in practice, the density f of the return X is unknown; in particular, no theory provides it. Assume we can adjust the portfolio parameters to satisfy the VaR constraints, and perhaps another constraint on the expected value of some function of X (e.g., the overall mean). We then wish to compute probabilities and expectations of interest, for example $\mathbb{P}(X > 0)$ or the probability of losing more than $2K$, or the expected return given $X > 0$. One strategy is to make such estimates and predictions under the most unpredictable circumstances consistent with the constraints. That is, use the *m*aximum entropy extension (MEE) of the constraints as a model for $f(x)$.

The "differential entropy" of f is $h(f) = -\int f(x) \ln f(x)\, dx$. (In general, the integral may not exist.) Entropy is concave on the space of densities for which it is defined. In general, the MEE is defined as

$$f_{MEE} = \arg\max_{f \in \Omega} h(f)$$

where Ω is the space of densities which satisfy a set of constraints of the form $E\phi_j(X) = c_j, j = 1, ..., J$. Assuming Ω is non-empty, it is well-known that f_{MEE} is unique and (away from the boundary of feasibility) is an exponential distribution in the constraint functions, –i.e., is of the form

$$f_{MEE}(x) = C^{-1} \exp\left(\sum_j \lambda_j \phi_j(x)\right)$$

where $C = C(\lambda_1, ..., \lambda_M)$ is the normalizing constant. (This form comes from differentiating an appropriate functional $J(f)$ based on entropy, and forcing the integral to be unity and imposing the constraints with Lagrange multıipliers.) In the special cases below we use this characterization to find the MEE for our constraints.

In our case we want to maximize entropy subject to the VaR constraints together with any others we might impose. Indeed, the VaR constraints alone do not admit an MEE since they do not restrict the density $f(x)$ for $x > K$. The entropy can be made arbitrarily large by allowing f to be identically $C = \frac{1-\epsilon}{N-K}$ over $K < x < N$ and letting $N \to \infty$. Suppose, however, that we have adjoined one or more constraints on the behavior of f which are compatible with the VaR constraints in the sense that the set of densities Ω satisfying all the constraints is non-empty. Here Ω would depend on the VaR parameters $\theta = (K, \epsilon, \nu_-)$ together with those parameters associated with the additional constraints.

25.4.1 Case A: Constraining the Global Mean

The simplest case is to add a constraint on the mean return, –i.e., fix $\mathbb{E}(X) = \mu$. Since $\mathbb{E}(X) = \mathbb{P}(X \leq K)\mathbb{E}(X|X \leq K) + \mathbb{P}(X > K)\mathbb{E}(X|X > K)$, adding the mean constraint is equivalent to adding the constraint

$$\mathbb{E}(X|X > K) = \nu_+$$

where ν_+ satisfies $\epsilon\nu_- + (1-\epsilon)\nu_+ = \mu$.

Define

$$f_-(x) = \begin{cases} \frac{1}{(K-\nu_-)} \exp\left[-\frac{K-x}{K-\nu_-}\right] & \text{if } x < K, \\ 0 & \text{if } x \geq K. \end{cases}$$

and

$$f_+(x) = \begin{cases} \frac{1}{(\nu_+-K)} \exp\left[-\frac{x-K}{\nu_+-K}\right] & \text{if } x > K, \\ 0 & \text{if } x \leq K. \end{cases}$$

It is easy to check that both f_- and f_+ integrate to one. Then

$$f_{MEE}(x) = \epsilon f_-(x) + (1-\epsilon)f_+(x)$$

is the MEE of the three constraints. First, evidently

1. $\int_{-\infty}^{K} f_{MEE}(x)\,\mathrm{d}x = \epsilon$;
2. $\int_{-\infty}^{K} x f_{MEE}(x)\,\mathrm{d}x = \epsilon\nu_-$;
3. $\int_{K}^{\infty} x f_{MEE}(x)\,\mathrm{d}x = (1-\epsilon)\nu_+$.

Hence the constraints are satisfied. Second, f_{MEE} has an exponential form in our constraint functions:

$$f_{MEE}(x) = C^{-1} \exp\left[-(\lambda_1 x + \lambda_2 I_{(x\leq K)} + \lambda_3 x I_{(x\leq K)})\right].$$

The shape of f_- depends on the relationship between K and the expected shortfall ν_-. The closer ν_- is to K, the more rapidly the tail falls off. As $\nu_- \to K$, f_- converges to a unit spike at $x = K$ (Figures 25.3 and 25.4).

25.4.2 Case B: Constraining the Absolute Mean

If instead we constrain the absolute mean, namely

$$E|X| = \int |x| f(x)\,\mathrm{d}x = \mu,$$

then the MEE is somewhat less apparent but can still be found. Define $f_-(x)$ as above, and let

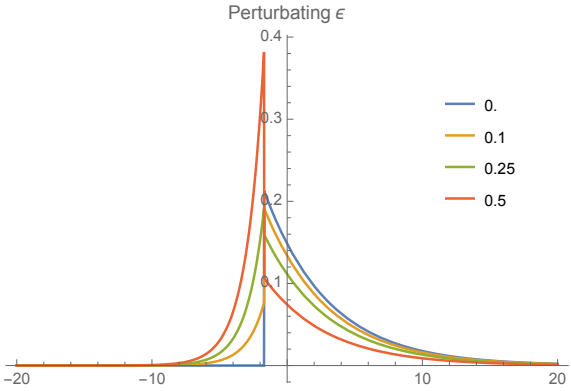


Figure 25.3: *Case A: Effect of different values of ϵ on the shape of the distribution.*

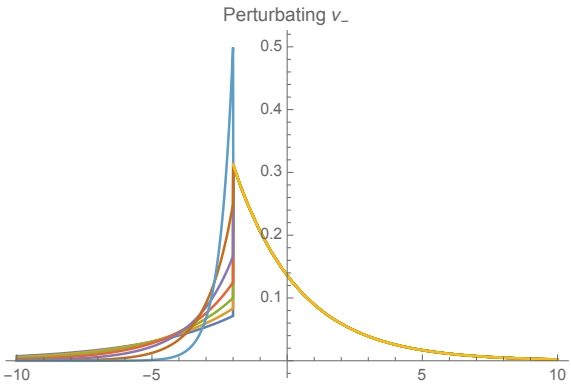


Figure 25.4: *Case A: Effect of different values of ν_- on the shape of the distribution.*

$$f_+(x) = \begin{cases} \frac{\lambda_1}{2-\exp(\lambda_1 K)} \exp(-\lambda_1 |x|) & \text{if } x \geq K, \\ 0 & \text{if } x < K. \end{cases}$$

Then λ_1 can be chosen such that

$$\epsilon \nu_- + (1-\epsilon) \int_K^\infty |x| f_+(x)\, \mathrm{d}x = \mu.$$

25.4.3 Case C: Power Laws for the Right Tail

If we believe that actual returns have "fat tails," in particular that the right tail decays as a Power Law rather than exponentially (as with a normal or exponential density), than we can add this constraint to the VaR constraints instead of working with the mean or absolute mean. In view of the exponential form of the MEE, the density $f_+(x)$ will have a power law, namely

$$f_+(x) = \frac{1}{C(\alpha)} (1+|x|)^{-(1+\alpha)}, x \geq K,$$

for $\alpha > 0$ if the constraint is of the form

$$E\left(\log(1+|X|)|X > K\right) = A.$$

Moreover, again from the MEE theory, we know that the parameter is obtained by minimizing the logarithm of the normalizing function. In this case, it is easy to show that

$$C(\alpha) = \int_K^\infty (1+|x|)^{-(1+\alpha)}\,\mathrm{d}x = \frac{1}{\alpha}(2-(1-K)^{-\alpha}).$$

It follows that A and α satisfy the equation

$$A = \frac{1}{\alpha} - \frac{\log(1-K)}{2(1-K)^\alpha - 1}.$$

We can think of this equation as determining the decay rate α for a given A or, alternatively, as determining the constraint value A necessary to obtain a particular Power Law α.

The final MEE extension of the VaR constraints together with the constraint on the log of the return is then:

$$f_{MEE}(x) \quad = \quad \epsilon I_{(x\le K)}\frac{1}{(K-\nu_-)}\exp\left[-\frac{K-x}{K-\nu_-}\right] + (1-\epsilon)I_{(x>K)}\frac{(1+|x|)^{-(1+\alpha)}}{C(\alpha)},$$

(see Figures 25.5 and 25.6).

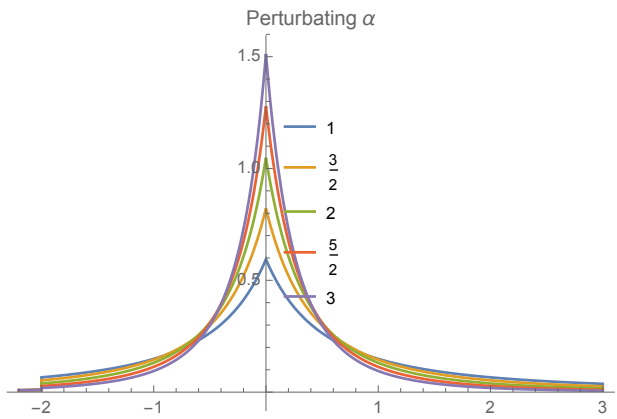


Figure 25.5: *Case C: Effect of different values of on the shape of the fat-tailed maximum entropy distribution.*

25.4.4 Extension to a Multi-Period Setting: A Comment

Consider the behavior in multi-periods. Using a naive approach, we sum up the performance as if there was no response to previous returns. We can see how Case A approaches the regular Gaussian, but not Case C (Figure 25.7).

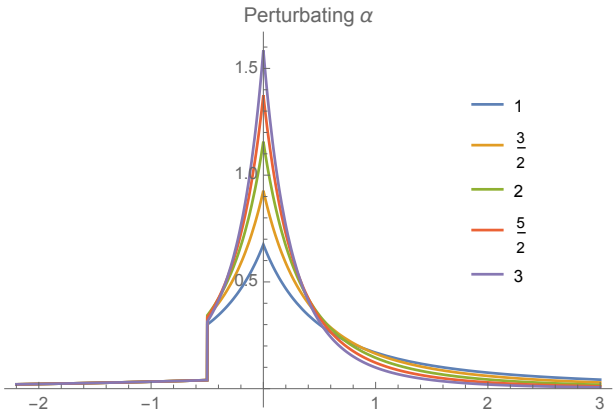


Figure 25.6: *Case C: Effect of different values of on the shape of the fat-tailed maximum entropy distribution (closer K).*

For case A the characteristic functioncan be written:

$$\Psi^A(t) = \frac{e^{iKt}(t(K - \nu_-\epsilon + \nu_+(\epsilon - 1)) - i)}{(Kt - \nu_- t - i)(-1 - it(K - \nu_+))}$$

So we can derive from convolutions that the function $\Psi^A(t)^n$ converges to that of an n-summed Gaussian. Further, the characteristic function of the limit of the average of strategies, namely

$$\lim_{n\to\infty} \Psi^A(t/n)^n = e^{it(\nu_+ + \epsilon(\nu_- - \nu_+))}, \tag{25.1}$$

is the characteristic function of the Dirac delta, visibly the effect of the law of large numbers delivering the same result as the Gaussian with mean $\nu_+ + \epsilon(\nu_- - \nu_+)$.

As to the Power Law in Case C, convergence to Gaussian only takes place for $\alpha \geq 2$, and rather slowly.

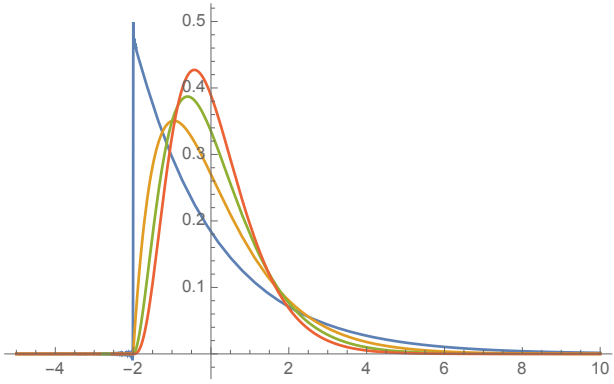


Figure 25.7: *Average return for multiperiod naive strategy for Case A, that is, assuming independence of "sizing", as position size does not depend on past performance. They aggregate nicely to a standard Gaussian, and (as shown in Equation (25.1)), shrink to a Dirac at the mean value.*

25.5 COMMENTS AND CONCLUSION

We note that the stop loss plays a larger role in determining the stochastic properties than the portfolio composition. Simply, the stop is not triggered by individual components, but by variations in the total portfolio. This frees the analysis from focusing on individual portfolio components when the tail –via derivatives or organic construction– is all we know and can control.

To conclude, most papers dealing with entropy in the mathematical finance literature have used minimization of entropy as an optimization criterion. For instance, Fritelli (2000) [97] exhibits the unicity of a "minimal entropy martingale measure" under some conditions and shows that minimization of entropy is equivalent to maximizing the expected exponential utility of terminal wealth. We have, instead, and outside any utility criterion, proposed entropy maximization as the recognition of the uncertainty of asset distributions. Under VaR and Expected Shortfall constraints, we obtain in full generality a "barbell portfolio" as the optimal solution, extending to a very general setting the approach of the two-fund separation theorem.

25.6 APPENDIX/PROOFS

Proof of Proposition 1: Since $X \sim N(\mu, \sigma^2)$, the tail probability constraint is

$$\epsilon = \mathbb{P}(X < K) = \mathbb{P}(Z < \frac{K-\mu}{\sigma}) = \Phi(\frac{K-\mu}{\sigma}).$$

By definition, $\Phi(\eta(\epsilon)) = \epsilon$. Hence,

$$K = \mu + \eta(\epsilon)\sigma \tag{25.2}$$

For the shortfall constraint,

$$\begin{aligned}
\mathbb{E}(X; X < k) &= \int_{-\infty}^{K} \frac{x}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} \, \mathrm{d}x \\
&= \mu\epsilon + \sigma \int_{-\infty}^{(K-\mu)/\sigma)} x\phi(x) \, \mathrm{d}x \\
&= \mu\epsilon - \frac{\sigma}{\sqrt{2\pi}} \exp -\frac{(K-\mu)^2}{2\sigma^2}
\end{aligned}$$

Since, $\mathbb{E}(X; X < K) = \epsilon v_-$, and from the definition of $B(\epsilon)$, we obtain

$$v_- = \mu - \eta(\epsilon)B(\epsilon)\sigma \tag{25.3}$$

Solving (25.2) and (25.3) for μ and σ^2 gives the expressions in Proposition 1.

Finally, by symmetry to the "upper tail inequality" of the standard normal, we have, for $x < 0$, $\Phi(x) \leq \frac{\phi(x)}{-x}$. Choosing $x = \eta(\epsilon) = \Phi^{-1}(\epsilon)$ yields $\epsilon = \mathbb{P}(X < \eta(\epsilon)) \leq$

$-\epsilon B(\epsilon)$ or $1 + B(\epsilon) \leq 0$. Since the upper tail inequality is asymptotically exact as $x \to \infty$ we have $B(0) = -1$, which concludes the proof.

BIBLIOGRAPHY AND INDEX

BIBLIOGRAPHY

[1] Inmaculada B Aban, Mark M Meerschaert, and Anna K Panorska. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277, 2006.

[2] Thierry Ané and Hélyette Geman. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance*, 55(5):2259–2284, 2000.

[3] Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. The promise of prediction markets. *Science*, 320(5878):877, 2008.

[4] Marco Avellaneda, Craig Friedman, Richard Holmes, and Dominick Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Mathematical Finance*, 4(1):37–64, 1997.

[5] L. Bachelier. *Theory of speculation in: P. Cootner, ed., 1964, The random character of stock market prices,*. MIT Press, Cambridge, Mass, 1900.

[6] Louis Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 1900.

[7] Kevin P Balanda and HL MacGillivray. Kurtosis: a critical review. *The American Statistician*, 42(2):111–119, 1988.

[8] August A Balkema and Laurens De Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974.

[9] August A Balkema and Laurens De Haan. Limit distributions for order statistics. i. *Theory of Probability & Its Applications*, 23(1):77–92, 1978.

[10] August A Balkema and Laurens de Haan. Limit distributions for order statistics. ii. *Theory of Probability & Its Applications*, 23(2):341–358, 1979.

[11] Shaul K Bar-Lev, Idit Lavi, and Benjamin Reiser. Bayesian inference for the power law process. *Annals of the Institute of Statistical Mathematics*, 44(4):623–639, 1992.

[12] Nicholas Barberis. The psychology of tail events: Progress and challenges. *American Economic Review*, 103(3):611–16, 2013.

[13] Jonathan Baron. *Thinking and deciding, 4th Ed.* Cambridge University Press, 2008.

[14] Norman C Beaulieu, Adnan A Abu-Dayya, and Peter J McLane. Estimating the distribution of a sum of independent lognormal random variables. *Communications, IEEE Transactions on*, 43(12):2869, 1995.

[15] Robert M Bell and Thomas M Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, 5(2):161–166, 1980.

[16] Shlomo Benartzi and Richard Thaler. Heuristics and biases in retirement savings behavior. *Journal of Economic perspectives*, 21(3):81–104, 2007.

[17] Shlomo Benartzi and Richard H Thaler. Myopic loss aversion and the equity premium puzzle. *The quarterly journal of Economics*, 110(1):73–92, 1995.

[18] Shlomo Benartzi and Richard H Thaler. Naive diversification strategies in defined contribution saving plans. *American economic review*, 91(1):79–98, 2001.

[19] Sergei Natanovich Bernshtein. Sur la loi des grands nombres. *Communications de la Société mathématique de Kharkow*, 16(1):82–87, 1918.

[20] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[21] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[22] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27. Cambridge university press, 1989.

[23] Giulio Biroli, J-P Bouchaud, and Marc Potters. On the top eigenvalue of heavy-tailed random matrices. *EPL (Europhysics Letters)*, 78(1):10001, 2007.

[24] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. 81:637–654, May–June 1973.

[25] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[26] A.J. Boness. Elements of a theory of stock-option value. 72:163–175, 1964.

[27] Jean-Philippe Bouchaud, Marc Mézard, Marc Potters, et al. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256, 2002.

[28] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[29] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[30] George Bragues. Prediction markets: The practical and normative possibilities for the social production of knowledge. *Episteme*, 6(1):91–106, 2009.

[31] D. T. Breeden and R. H. Litzenberger. Price of state-contigent claimes implicit in option prices. 51:621–651, 1978.

[32] Douglas T Breeden and Robert H Litzenberger. Prices of state-contingent claims implicit in option prices. *Journal of business*, pages 621–651, 1978.

[33] Henry Brighton and Gerd Gigerenzer. Homo heuristicus and the bias–variance dilemma. In *Action, Perception and the Brain*, pages 68–91. Springer, 2012.

[34] Damiano Brigo and Fabio Mercurio. Lognormal-mixture dynamics and calibration to market volatility smiles. *International Journal of Theoretical and Applied Finance*, 5(04):427–446, 2002.

[35] Peter Carr. Bounded brownian motion. *NYU Tandon School of Engineering*, 2017.

[36] Peter Carr, Hélyette Geman, Dilip B Madan, and Marc Yor. Stochastic volatility for lévy processes. *Mathematical finance*, 13(3):345–382, 2003.

[37] Peter Carr and Dilip Madan. Optimal positioning in derivative securities. 2001.

[38] Lars-Erik Cederman. Modeling the size of wars: from billiard balls to sandpiles. *American Political Science Review*, 97(01):135–150, 2003.

[39] Bikas K Chakrabarti, Anirban Chakraborti, Satya R Chakravarty, and Arnab Chatterjee. *Econophysics of income and wealth distributions*. Cambridge University Press, 2013.

[40] David G Champernowne. A model of income distribution. *The Economic Journal*, 63(250):318–351, 1953.

[41] Shaohua Chen, Hong Nie, and Benjamin Ayers-Glassey. Lognormal sum approximation with a variant of type iv pearson distribution. *IEEE Communications Letters*, 12(9), 2008.

[42] 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.

[43] 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.

[44] Pasquale Cirillo. Are your data really pareto distributed? *Physica A: Statistical Mechanics and its Applications*, 392(23):5947–5962, 2013.

[45] Pasquale Cirillo and Nassim Nicholas Taleb. Expected shortfall estimation for apparently infinite-mean models of operational risk. *Quantitative Finance*, pages 1–10, 2016.

[46] Pasquale Cirillo and Nassim Nicholas Taleb. On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, 452:29–45, 2016.

[47] Pasquale Cirillo and Nassim Nicholas Taleb. What are the chances of war? *Significance*, 13(2):44–45, 2016.

[48] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[49] Rama Cont and Peter Tankov. *Financial modelling with jump processes*, volume 2. CRC press, 2003.

[50] Harald Cramér. *On the mathematical theory of risk*. Centraltryckeriet, 1930.

[51] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[52] Camilo Dagum. Inequality measures between income distributions with applications. *Econometrica*, 48(7):1791–1803, 1980.

[53] Camilo Dagum. *Income distribution models*. Wiley Online Library, 1983.

[54] Anirban DasGupta. *Probability for statistics and machine learning: fundamentals and advanced topics*. Springer Science & Business Media, 2011.

[55] Herbert A David and Haikady N Nagaraja. Order statistics. 2003.

[56] Bruno De Finetti. Probability, induction, and statistics. 1972.

[57] Bruno De Finetti. *Philosophical Lectures on Probability: collected, edited, and annotated by Alberto Mura*, volume 340. Springer Science & Business Media, 2008.

[58] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

[59] Kresimir Demeterfi, Emanuel Derman, Michael Kamal, and Joseph Zou. A guide to volatility and variance swaps. *The Journal of Derivatives*, 6(4):9–32, 1999.

[60] Kresimir Demeterifi, Emanuel Derman, Michael Kamal, and Joseph Zou. More than you ever wanted to know about volatility swaps. *Working paper, Goldman Sachs*, 1999.

[61] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2007.

[62] E. Derman and N. Taleb. The illusion of dynamic delta replication. *Quantitative Finance*, 5(4):323–326, 2005.

[63] Emanuel Derman. The perception of time, risk and return during periods of speculation. *Working paper, Goldman Sachs*, 2002.

[64] Marco Di Renzo, Fabio Graziosi, and Fortunato Santucci. Further results on the approximation of log-normal power sum via pearson type iv distribution: a general formula for log-moments computation. *IEEE Transactions on Communications*, 57(4), 2009.

[65] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[66] Persi Diaconis and Sandy Zabell. Closed form summation for classical distributions: variations on a theme of de moivre. *Statistical Science*, pages 284–302, 1991.

[67] Cornelius Frank Dietrich. *Uncertainty, calibration and probability: the statistics of scientific and industrial measurement*. Routledge, 2017.

[68] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.19 of 2018-06-22. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[69] Daniel Dufresne. Sums of lognormals. In *Proceedings of the 43rd actuarial research conference*. University of Regina, 2008.

[70] Daniel Dufresne et al. The log-normal approximation in financial and other computations. *Advances in Applied Probability*, 36(3):747–773, 2004.

[71] Bruno Dupire. Pricing with a smile. 7(1), 1994.

[72] Bruno Dupire. Exotic option pricing by calibration on volatility smiles. In *Advanced Mathematics for Derivatives: Risk Magazine Conference*, 1995.

[73] Bruno Dupire et al. Pricing with a smile. *Risk*, 7(1):18–20, 1994.

[74] Danny Dyer. Structural probability bounds for the strong pareto law. *Canadian Journal of Statistics*, 9(1):71–77, 1981.

[75] Iddo Eliazar. Inequality spectra. *Physica A: Statistical Mechanics and its Applications*, 469:824–847, 2017.

[76] Iddo Eliazar. Lindy's law. *Physica A: Statistical Mechanics and its Applications*, 486:797–805, 2017.

[77] Iddo Eliazar and Morrel H Cohen. On social inequality: Analyzing the rich–poor disparity. *Physica A: Statistical Mechanics and its Applications*, 401:148–158, 2014.

[78] Iddo Eliazar and Igor M Sokolov. Maximization of statistical heterogeneity: From shannon's entropy to gini's index. *Physica A: Statistical Mechanics and its Applications*, 389(16):3023–3038, 2010.

[79] Iddo I Eliazar and Igor M Sokolov. Gini characterization of extreme-value statistics. *Physica A: Statistical Mechanics and its Applications*, 389(21):4462–4472, 2010.

[80] Iddo I Eliazar and Igor M Sokolov. Measuring statistical evenness: A panoramic overview. *Physica A: Statistical Mechanics and its Applications*, 391(4):1323–1353, 2012.

[81] Paul Embrechts. *Modelling extremal events: for insurance and finance*, volume 33. Springer, 1997.

[82] Paul Embrechts and Charles M Goldie. On convolution tails. *Stochastic Processes and their Applications*, 13(3):263–278, 1982.

[83] Paul Embrechts, Charles M Goldie, and Noël Veraverbeke. Subexponentiality and infinite divisibility. *Probability Theory and Related Fields*, 49(3):335–347, 1979.

[84] M Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):247–271, 1909.

[85] Michael Falk et al. On testing the extreme value index via the pot-method. *The Annals of Statistics*, 23(6):2013–2035, 1995.

[86] Michael Falk, Jürg Hüsler, and Rolf-Dieter Reiss. *Laws of small numbers: extremes and rare events*. Springer Science & Business Media, 2010.

[87] Kai-Tai Fang. Elliptically contoured distributions. *Encyclopedia of Statistical Sciences*, 2006.

[88] Doyne James Farmer and John Geanakoplos. Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates. 2009.

[89] J Doyne Farmer and John Geanakoplos. Power laws in economics and elsewhere. In *Santa Fe Institute*, 2008.

[90] William Feller. 1971an introduction to probability theory and its applications, vol. 2.

[91] William Feller. An introduction to probability theory. 1968.

[92] Baruch Fischhoff, John Kadvany, and John David Kadvany. *Risk: A very short introduction*. Oxford University Press, 2011.

[93] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press, 1928.

[94] Andrea Fontanari, Pasquale Cirillo, and Cornelis W Oosterlee. From concentration profiles to concentration maps. new tools for the study of loss distributions. *Insurance: Mathematics and Economics*, 78:13–29, 2018.

[95] 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.

[96] David A Freedman. Notes on the dutch book argument ". *Lecture Notes, Department of Statistics, University of Berkley at Berkley, http://www. stat. berkeley. edu/~ census/dutchdef. pdf*, 2003.

[97] Marco Frittelli. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Mathematical finance*, 10(1):39–52, 2000.

[98] Xavier Gabaix. Power laws in economics and finance. Technical report, National Bureau of Economic Research, 2008.

[99] Xavier Gabaix. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30(1):185–206, 2016.

[100] Armengol Gasull, Maria Jolis, and Frederic Utzet. On the norming constants for normal maxima. *Journal of Mathematical Analysis and Applications*, 422(1):376–396, 2015.

[101] Jim Gatheral. *The Volatility Surface: a Practitioner's Guide*. John Wiley & Sons, 2006.

[102] Jim Gatheral. *The Volatility Surface: A Practitioner's Guide*. New York: John Wiley & Sons, 2006.

[103] Oscar Gelderblom and Joost Jonker. Amsterdam as the cradle of modern futures and options trading, 1550-1650. *William Goetzmann and K. Geert Rouwenhorst*, 2005.

[104] Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

[105] Donald Geman, Hélyette Geman, and Nassim Nicholas Taleb. Tail risk constraints and maximum entropy. *Entropy*, 17(6):3724, 2015.

[106] Nicholas Georgescu-Roegen. The entropy law and the economic process, 1971. *Cambridge, Mass*, 1971.

[107] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

[108] Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, New York, 1999.

[109] Corrado Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[110] BV Gnedenko and AN Kolmogorov. *Limit Distributions for Sums of Independent Random Variables (1954)*.

[111] Charles M Goldie. Subexponential distributions and dominated-variation tails. *Journal of Applied Probability*, pages 440–442, 1978.

[112] 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.

[113] Richard C Green, Robert A Jarrow, et al. Spanning and completeness in markets with contingent claims. *Journal of Economic Theory*, 41(1):202–210, 1987.

[114] Emil Julius Gümbel. Statistics of extremes. 1958.

[115] Laurens Haan and Ana Ferreira. Extreme value theory: An introduction. *Springer Series in Operations Research and Financial Engineering (*, 2006.

[116] Wolfgang Hafner and Heinz Zimmermann. Amazing discovery: Vincenz bronzin's option pricing models. 31:531–546, 2007.

[117] Torben Hagerup and Christine Rüb. A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308, 1990.

[118] John Haigh. The kelly criterion and bet comparisons in spread betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(4):531–539, 2000.

[119] Peter Hall. On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16(2):433–439, 1979.

[120] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

[121] J Michael Harrison and David M Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic theory*, 20(3):381–408, 1979.

[122] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction, springer series in statistics, 2009.

[123] Espen G. Haug. *Derivatives: Models on Models*. New York: John Wiley & Sons, 2007.

[124] 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.

[125] Friedrich August Hayek. The use of knowledge in society. *The American economic review*, 35(4):519–530, 1945.

[126] John R Hicks. *Value and capital*, volume 2. Clarendon press Oxford, 1939.

[127] Leonard R. Higgins. *The Put-and-Call*. London: E. Wilson., 1902.

[128] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[129] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[130] HM James Hung, Robert T O'Neill, Peter Bauer, and Karl Kohne. The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, pages 11–22, 1997.

[131] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[132] E.T. Jaynes. How should we use entropy in economics? 1991.

[133] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[134] Hedegaard Anders Jessen and Thomas Mikosch. Regularly varying functions. *Publications de l'Institut Mathematique*, 80(94):171–192, 2006.

[135] Petr Jizba, Hagen Kleinert, and Mohammad Shefaat. Rényi's information transfer between financial time series. *Physica A: Statistical Mechanics and its Applications*, 391(10):2971–2989, 2012.

[136] Valen E Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

[137] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[138] Joseph P Kairys Jr and NICHOLAS VALERIO III. The market for equity options in the 1870s. *The Journal of Finance*, 52(4):1707–1723, 1997.

[139] Ioannis Karatzas and Steven E Shreve. Brownian motion and stochastic calculus springer-verlag. *New York*, 1991.

[140] John L Kelly. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.

[141] Gideon Keren. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273, 1991.

[142] Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*, volume 470. John Wiley & Sons, 2003.

[143] Andrei Nikolaevich Kolmogorov. On logical foundations of probability theory. In *Probability theory and mathematical statistics*, pages 1–5. Springer, 1983.

[144] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[145] Samuel Kotz and Norman Johnson. *Encyclopedia of Statistical Sciences*. Wiley, 2004.

[146] VV Kozlov, T Madsen, and AA Sorokin. Weighted means of weakly dependent random variables. *MOSCOW UNIVERSITY MATHEMATICS BULLETIN C/C OF VESTNIK-MOSKOVSKII UNIVERSITET MATHEMATIKA*, 59(5):36, 2004.

[147] Jean Laherrere and Didier Sornette. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[148] David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

[149] Deli Li, M Bhaskara Rao, and RJ Tomkins. The law of the iterated logarithm and central limit theorem for l-statistics. Technical report, PENNSYLVANIA STATE UNIV UNIVERSITY PARK CENTER FOR MULTIVARIATE ANALYSIS, 1997.

[150] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. In *Decision making and change in human affairs*, pages 275–324. Springer, 1977.

[151] Sarah Lichtenstein, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory*, 4(6):551, 1978.

[152] Michel Loève. *Probability Theory. Foundations. Random Sequences*. New York: D. Van Nostrand Company, 1955.

[153] Filip Lundberg. *I. Approximerad framställning af sannolikhetsfunktionen. II. Återförsäkring af kollektivrisker. Akademisk afhandling... af Filip Lundberg,...* Almqvist och Wiksells boktryckeri, 1903.

[154] HL MacGillivray and Kevin P Balanda. Mixtures, myths and kurtosis. *Communications in Statistics-Simulation and Computation*, 17(3):789–802, 1988.

[155] LC MacLean, William T Ziemba, and George Blazenko. Growth versus security in dynamic investment analysis. *Management Science*, 38(11):1562–1585, 1992.

[156] Dhruv Madeka. Accurate prediction of electoral outcomes. *arXiv preprint arXiv:1704.02664*, 2017.

[157] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

[158] Spyros Makridakis and Nassim Taleb. Decision making and planning under low levels of predictability, 2009.

[159] Benoit Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by ha simon. *Information and Control*, 2(1):90–99, 1959.

[160] Benoit Mandelbrot. The pareto-levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.

[161] Benoit Mandelbrot. The stable paretian income distribution when the apparent exponent is near two. *International Economic Review*, 4(1):111–115, 1963.

[162] Benoit B Mandelbrot. New methods in statistical economics. In *Fractals and Scaling in Finance*, pages 79–104. Springer, 1997.

[163] Benoît B Mandelbrot and Nassim Nicholas Taleb. Random jump, not random walk, 2010.

[164] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

[165] Harry M Markowitz. *Portfolio selection: efficient diversification of investments*, volume 16. Wiley, 1959.

[166] RARD Maronna, Douglas Martin, and Victor Yohai. *Robust statistics*. John Wiley & Sons, Chichester. ISBN, 2006.

[167] R. Mehera and E. C. Prescott. The equity premium: a puzzle. *Journal of Monetary Economics*, 15:145–161, 1985.

[168] Robert C Merton. An analytic derivation of the efficient portfolio frontier. *Journal of financial and quantitative analysis*, 7(4):1851–1872, 1972.

[169] Robert C. Merton. The relationship between put and call prices: Comment. 28(1):183–184, 1973.

[170] Robert C. Merton. Theory of rational option pricing. 4:141–183, Spring 1973.

[171] Robert C. Merton. Option pricing when underlying stock returns are discontinuous. 3:125–144, 1976.

[172] Robert C Merton and Paul Anthony Samuelson. Continuous-time finance. 1992.

[173] David C Nachman. Spanning and completeness with options. *The review of financial studies*, 1(3):311–328, 1988.

[174] S. A. Nelson. *The A B C of Options and Arbitrage.* The Wall Street Library, New York., 1904.

[175] S. A. Nelson. *The A B C of Options and Arbitrage*. New York: The Wall Street Library., 1904.

[176] Hansjörg Neth and Gerd Gigerenzer. Heuristics: Tools for an uncertain world. *Emerging trends in the social and behavioral sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 2015.

[177] Donald J Newman. *A problem seminar*. Springer Science & Business Media, 2012.

[178] Hong Nie and Shaohua Chen. Lognormal sum approximation with type iv pearson distribution. *IEEE Communications Letters*, 11(10), 2007.

[179] John P Nolan. Parameterizations and modes of stable distributions. *Statistics & probability letters*, 38(2):187–195, 1998.

[180] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[181] T. Mikosch P. Embrechts, C. Kluppelberg. *Modelling Extremal Events*. Springer, 2003.

[182] Vilfredo Pareto. La courbe des revenus. *Travaux de Sciences Sociales*, pages 299–345, 1896 (1964).

[183] O. Peters and M. Gell-Mann. Evaluating gambles using dynamics. *Chaos*, 26(2), 2016.

[184] T Pham-Gia and TL Hung. The mean and median absolute deviations. *Mathematical and Computer Modelling*, 34(7-8):921–936, 2001.

[185] George C Philippatos and Charles J Wilson. Entropy, market risk, and the selection of efficient portfolios. *Applied Economics*, 4(3):209–220, 1972.

[186] Charles Phillips and Alan Axelrod. *Encyclopedia of Wars:(3-Volume Set)*. Infobase Pub., 2004.

[187] James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.

[188] Thomas Piketty. Capital in the 21st century, 2014.

[189] Thomas Piketty and Emmanuel Saez. The evolution of top incomes: a historical and international perspective. Technical report, National Bureau of Economic Research, 2006.

[190] Iosif Pinelis. Characteristic function of the positive part of a random variable and related results, with applications. *Statistics & Probability Letters*, 106:281–286, 2015.

[191] Steven Pinker. *The better angels of our nature: Why violence has declined*. Penguin, 2011.

[192] Dan Pirjol. The logistic-normal integral and its generalizations. *Journal of Computational and Applied Mathematics*, 237(1):460–469, 2013.

[193] EJG Pitman. Subexponential distribution functions. *J. Austral. Math. Soc. Ser. A*, 29(3):337–347, 1980.

[194] Svetlozar T Rachev, Young Shin Kim, Michele L Bianchi, and Frank J Fabozzi. *Financial models with Lévy processes and volatility clustering*, volume 187. John Wiley & Sons, 2011.

[195] Anthony M. Reinach. *The Nature of Puts & Calls*. New York: The Bookmailer, 1961.

[196] Lewis F Richardson. Frequency of occurrence of wars and other fatal quarrels. *Nature*, 148(3759):598, 1941.

[197] Matthew Richardson and Tom Smith. A direct test of the mixture of distributions hypothesis: Measuring the daily flow of information. *Journal of Financial and Quantitative Analysis*, 29(01):101–116, 1994.

[198] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[199] Stephen A Ross. Mutual fund separation in financial theory—the separating distributions. *Journal of Economic Theory*, 17(2):254–286, 1978.

[200] Stephen A Ross. *Neoclassical finance*. Princeton University Press, 2009.

[201] Francesco Rubino, Antonello Forgione, David E Cummings, Michel Vix, Donatella Gnuli, Geltrude Mingrone, Marco Castagneto, and Jacques Marescaux. The mechanism of diabetes control after gastrointestinal bypass surgery reveals a role of the proximal small intestine in the pathophysiology of type 2 diabetes. *Annals of surgery*, 244(5):741–749, 2006.

[202] Mark Rubinstein. *Rubinstein on derivatives*. Risk Books, 1999.

[203] Mark Rubinstein. *A History of The Theory of Investments*. New York: John Wiley & Sons, 2006.

[204] Doriana Ruffino and Jonathan Treussard. Derman and taleb's 'the illusions of dynamic replication': a comment. *Quantitative Finance*, 6(5):365–367, 2006.

[205] Harold Sackrowitz and Ester Samuel-Cahn. P values as random variables—expected p values. *The American Statistician*, 53(4):326–331, 1999.

[206] Gennady Samorodnitsky and Murad S Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[207] D Schleher. Generalized gram-charlier series with application to the sum of log-normal variates (corresp.). *IEEE Transactions on Information Theory*, 23(2):275–280, 1977.

[208] Jun Shao. *Mathematical Statistics*. Springer, 2003.

[209] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[210] SK Singh and GS Maddala. A function for size distribution of incomes: reply. *Econometrica*, 46(2), 1978.

[211] Didier Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization, and disorder: concepts and tools*. Springer, 2004.

[212] C.M. Sprenkle. Warrant prices as indicators of expectations and preferences. *Yale Economics Essays*, 1(2):178–231, 1961.

[213] C.M. Sprenkle. *Warrant Prices as Indicators of Expectations and Preferences: in P. Cootner, ed., 1964, The Random Character of Stock Market Prices,*. MIT Press, Cambridge, Mass, 1964.

[214] AJ Stam. Regular variation of the tail of a subordinated probability distribution. *Advances in Applied Probability*, pages 308–327, 1973.

[215] Stephen M Stigler. Stigler's law of eponymy. *Transactions of the New York academy of sciences*, 39(1 Series II):147–157, 1980.

[216] Hans R Stoll. The relationship between put and call option prices. *The Journal of Finance*, 24(5):801–824, 1969.

[217] Cass R Sunstein. Deliberating groups versus prediction markets (or hayek's challenge to habermas). *Episteme*, 3(3):192–213, 2006.

[218] Giitiro Suzuki. A consistent estimator for the mean deviation of the pearson type distribution. *Annals of the Institute of Statistical Mathematics*, 17(1):271–285, 1965.

[219] E. Schechtman S.Yitzhaki. *The Gini Methodology: A primer on a statistical methodology*. Springer, 2012.

[220] N N Taleb and R Douady. Mathematical definition, mapping, and detection of (anti) fragility. *Quantitative Finance*, 2013.

[221] Nassim N Taleb and G Martin. The illusion of thin tails under aggregation (a reply to jack treynor). *Journal of Investment Management*, 2012.

[222] Nassim Nicholas Taleb. *Dynamic Hedging: Managing Vanilla and Exotic Options*. John Wiley & Sons (Wiley Series in Financial Engineering), 1997.

[223] Nassim Nicholas Taleb. *Incerto: Antifragile, The Black Swan , Fooled by Randomness, the Bed of Procrustes, Skin in the Game*. Random House and Penguin, 2001-2018.

[224] Nassim Nicholas Taleb. Black swans and the domains of statistics. *The American Statistician*, 61(3):198–200, 2007.

[225] Nassim Nicholas Taleb. Errors, robustness, and the fourth quadrant. *International Journal of Forecasting*, 25(4):744–759, 2009.

[226] Nassim Nicholas Taleb. Finiteness of variance is irrelevant in the practice of quantitative finance. *Complexity*, 14(3):66–76, 2009.

[227] Nassim Nicholas Taleb. *Antifragile: things that gain from disorder*. Random House and Penguin, 2012.

[228] Nassim Nicholas Taleb. Four points beginner risk managers should learn from jeff holman's mistakes in the discussion of antifragile. *arXiv preprint arXiv:1401.2524*, 2014.

[229] Nassim Nicholas Taleb. The meta-distribution of standard p-values. *arXiv preprint arXiv:1603.07532*, 2016.

[230] Nassim Nicholas Taleb. Stochastic tail exponent for asymmetric power laws. *arXiv preprint arXiv:1609.02369*, 2016.

[231] Nassim Nicholas Taleb. Election predictions as martingales: an arbitrage approach. *Quantitative Finance*, 18(1):1–5, 2018.

[232] Nassim Nicholas Taleb. How much data do you need? an operational, pre-asymptotic metric for fat-tailedness. *International Journal of Forecasting*, 2018.

[233] Nassim Nicholas Taleb. *Skin in the Game: Hidden Asymmetries in Daily Life*. Penguin (London) and Random House (N.Y.), 2018.

[234] Nassim Nicholas Taleb. *Technical Incerto, Vol 1: The Statistical Consequences of Fat Tails, Papers and Commentaries*. Monograph, 2019.

[235] Nassim Nicholas Taleb. Common misapplications and misinterpretations of correlation in social" science. *Preprint, Tandon School of Engineering, New York University*, 2020.

[236] Nassim Nicholas Taleb. *The Statistical Consequences of Fat Tails*. STEM Academic Press, 2020.

[237] Nassim Nicholas Taleb, Elie Canetti, Tidiane Kinda, Elena Loukoianova, and Christian Schmieder. A new heuristic measure of fragility and tail risks: application to stress testing. *International Monetary Fund*, 2018.

[238] Nassim Nicholas Taleb and Raphael Douady. On the super-additivity and estimation biases of quantile contributions. *Physica A: Statistical Mechanics and its Applications*, 429:252–260, 2015.

[239] Nassim Nicholas 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.

[240] Nassim Nicholas Taleb and George A Martin. How to prevent other financial crises. *SAIS Review of International Affairs*, 32(1):49–60, 2012.

[241] Nassim Nicholas Taleb and Avital Pilpel. I problemi epistemologici del risk management. *Daniele Pace (a cura di)" Economia del rischio. Antologia di scritti su rischio e decisione economica", Giuffre, Milano*, 2004.

[242] Nassim Nicholas Taleb and Constantine Sandis. The skin in the game heuristic for protection against tail events. *Review of Behavioral Economics*, 1:1–21, 2014.

[243] Jozef L Teugels. The class of subexponential distributions. *The Annals of Probability*, 3(6):1000–1011, 1975.

[244] Edward Thorp. A corrected derivation of the black-scholes option model. *Based on private conversation with Edward Thorp and a copy of a 7 page paper Thorp wrote around 1973, with disclaimer that I understood Ed. Thorp correctly.*, 1973.

[245] Edward O Thorp. Optimal gambling systems for favorable games. *Revue de l'Institut International de Statistique*, pages 273–293, 1969.

[246] Edward O Thorp. Extensions of the black-scholes option model. *Proceedings of the 39th Session of the International Statistical Institute, Vienna, Austria*, pages 522–29, 1973.

[247] Edward O Thorp. Understanding the kelly criterion. *The Kelly Capital Growth Investment Criterion: Theory and Practice', World Scientific Press, Singapore*, 2010.

[248] Edward O. Thorp and S. T. Kassouf. *Beat the Market*. New York: Random House, 1967.

[249] James Tobin. Liquidity preference as behavior towards risk. *The review of economic studies*, pages 65–86, 1958.

[250] Jack L Treynor. Insights-what can taleb learn from markowitz? *Journal of Investment Management*, 9(4):5, 2011.

[251] Constantino Tsallis, Celia Anteneodo, Lisa Borland, and Roberto Osorio. Nonextensive statistical mechanics and economics. *Physica A: Statistical Mechanics and its Applications*, 324(1):89–100, 2003.

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

[253] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[254] Willem Rutger van Zwet. *Convex transformations of random variables*, volume 7. Mathematisch centrum, 1964.

[255] SR Srinivasa Varadhan. *Large deviations and applications*, volume 46. SIAM, 1984.

[256] SR Srinivasa Varadhan. *Stochastic processes*, volume 16. American Mathematical Soc., 2007.

[257] José A Villaseñor-Alva and Elizabeth González-Estrada. A bootstrap goodness of fit test for the generalized pareto distribution. *Computational Statistics & Data Analysis*, 53(11):3835–3841, 2009.

[258] Eric Weisstein. *Wolfram MathWorld*. Wolfram Research www.wolfram.com, 2017.

[259] 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.

[260] Heath Windcliff and Phelim P Boyle. The 1/n pension investment puzzle. *North American Actuarial Journal*, 8(3):32–45, 2004.

[261] Yingying Xu, Zhuwu Wu, Long Jiang, and Xuefeng Song. A maximum entropy method for a robust portfolio problem. *Entropy*, 16(6):3401–3415, 2014.

[262] Yingying Yang, Shuhe Hu, and Tao Wu. The tail probability of the product of dependent random variables from max-domains of attraction. *Statistics & Probability Letters*, 81(12):1876–1882, 2011.

[263] Jay L Zagorsky. Do you have to be smart to be rich? the impact of iq on wealth, income and financial distress. *Intelligence*, 35(5):489–501, 2007.

[264] 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.

[265] Rongxi Zhou, Ru Cai, and Guanqun Tong. Applications of entropy in finance: A review. *Entropy*, 15(11):4909–4931, 2013.

[266] Vladimir M Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Soc., 1986.

[267] VM Zolotarev. On a new viewpoint of limit theorems taking into account large deviationsr. *Selected Translations in Mathematical Statistics and Probability*, 9:153, 1971.