Roger Bowden

The Information Theory of Comparisons

With Applications to Statistics and the Social Sciences

Springer

The Information Theory of Comparisons

Roger Bowden

The Information Theory of Comparisons

With Applications to Statistics and the Social Sciences

Springer

Roger Bowden
Kiwicap Research Ltd.
Kelburn, Wellington
New Zealand

ISBN 978-981-13-1549-7 ISBN 978-981-13-1550-3 (eBook)
https://doi.org/10.1007/978-981-13-1550-3

Library of Congress Control Number: 2018947773

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Every now and then, research interests that have hitherto been tucked away in the recesses of the mind suddenly converge and come to the fore. So, it is with the present contribution. Any social scientist has constantly to face up to the problem of comparing distributions, whether across people, across nations, over time or how they might impact on the survival of a financial entity. And the audience being what it typically is, one has to present such comparisons and contrasts with as much impact as possible, but also with as much brevity as limited time, and even more limited attention (in the media for instance), typically allows. It had always struck me how little the standard metrics really inform such comparisons.

Likewise, I had been reasonably familiar with information theory, from the time of my Manchester Ph.D. thesis on the spectral domain and later, from the work on income distribution of scholars like Henri Theil. But it did seem to me at the time that the standard entropy metric, as it stood, told one remarkably little. Yet, the fascination remained, just as I think it still does with so many researchers across many fields.

Then, in more recent times, the epochal rise in income and wealth inequality started to focus everyone's attention on just how to measure and report such trends. The thought occurred to me that instead of imposing this or that metric as an external commentator, perhaps there was another way of looking at things, namely from the point of view of the income earners themselves, and their collective aggregate.

Things started to converge at this point with some earlier work on distribution shifting, which in turn had been initiated in the context of financial risk management. I had in addition been lucky enough to attend a series of research workshops moderated by Guenther Loeffler of Ulm University in Germany. At one of these, in the relaxing environment of Heligkreutztal abbey, I attended an informative talk by Matthias Boehm, at the time a Ph.D. student, on subjectivist probability. It seemed to me that once again, there was effectively an example of distribution shifting.

At this point, convergence was more or less complete and the present project got underway as such. It was helped along by a Gambrinus Research Fellowship at Dortmund Technical University where I worked in conjunction with Peter Posch

and Daniel Ullman. Daniel's energy and thoroughness has informed much of the empirical work reported in our joint papers. Walter Kraemer has also been very supportive; he knows much more about income distribution and related measures than I will ever know.

In the latter stages of preparation, global warming has probably helped the project along. Already famous as the windiest capital city in the world, Wellingon's extreme winter over the past 2 years has meant little else was possible save to cower inside and work. However, in this respect, I was fortunate to have the encouragement of Stephen Jones, regional editor of Springer Nature. Thanks also go to Sanjiev Mathiazhagan for keeping me on the ball, as production editor, also Komala Jaishankar. A special vote of thanks goes to two reviewers of the manuscript, who have provided the independent perspectives and detail that can be so valuable to any author.

Kelburn, New Zealand Roger Bowden

Contents

Introduction

In the study that follows, the objects of comparison take the form of probability or frequency distributions that share a contextual relevance. There is a range of such contexts in the social sciences: over time, across different countries or social groups, alternative investments, or of opinions. In many cases, the social importance of comparisons calls for special attention to their meaning, with a requirement that any proposed methodology should encompass any inherent complexity in ways that standard textbook measures and metrics may fall short. Thus, a first and primary agenda is to bring together an emerging body of work on measures and metrics for the comparison of distributions, measures that in a very broad sense could be regarded as sufficient statistics for complexity.

As to complexity in itself, a suitable vantage point is complexity as it is defined and developed in informational entropy. In this respect, the classic Shannon entropy metric is often reported for this or that distribution, but is rarely used; or if used, is subject to a range of objections rooted in context. One agenda is therefore to show that well-motivated directional decompositions of total entropy can indeed yield more meaningful distributional comparisons. In turn, the metrics that result have meaning in their own right, even without the need to call on a dedicated entropic interpretation. In a social science context, they amount to a different perspective, in which the focus shifts to how the subjects themselves might view things.

In this respect, an initial commentary on generic distributions may be in order. A beginning student of statistics could be forgiven for thinking of the normal and related distributions as canonical, applying not only to asymptotic sampling distributions but to the realities of physical phenomena and of economic and social data analysis. Especially, in the latter contexts, however, a normal distribution tends to be an exception rather than the rule. Empirical distributions are often characterised by long tails in one or both directions, over time scales that may vary from the very short run to the longer. Indeed, there would hardly be any point in studying income distributions if a perfect bell-shaped curve was always evident.

The structural mechanisms that generate these asymmetries and temporal changes may or may not be known. Either way, there are often welfare outcomes associated with long tails, especially if these are more pronounced on one side. It is

the comparison of such outcomes over different distributions that is the core agenda of the present study.

A complete welfare evaluation could in theory be carried out if a well-defined preference function existed between alternative distribution functions. But this itself is a relative rarity, fraught with its own difficulties. Thus, there may not be universal agreement between users as to what such a function should look like. Indeed, Kenneth Arrow showed very early on that a social welfare function to reconcile such individual preferences does not in general exist.

Likewise, textbooks of finance still exposit a portfolio theory based on an assumption of risk aversion, meaning here a concave utility function for money. But there is abundant behavioural evidence (as well as simple common sense) that many investors think more in terms of a Friedman–Savage type utility function that is concave downwards for losses, and convex upwards for gains.

In the face of variation both in context and in observer attitudes, a recourse is to fall back on distribution metrics that summarise and convey as much information as possible about the given distribution; and do so in brief, informative metrics. Conventional metrics as the mean, the median, or every higher order moments fall short in this respect. There is indeed no particular reason why metrics based on linear or polynomial functions should convey anything except the most basic indications as to location or shape. In particular, they fall short in indicating the complexity dimension inherent in long tails.

In theory, complexity is captured by the Shannon or differential entropy of the observed distribution, which is the expected value of the log of the density. Pretty much any listing of the properties of distributions in common use will specify its Shannon entropy, but more as a standalone item, with little if any further commentary. However, this does not in itself tell us much about the way that entropic complexity is distributed over the range space of the distribution. A first objective of the present study is therefore to supplement such standard listings with more informative metrics.

Informative, in this context, means that such metrics are derived using arguments based on entropy. But to do that requires a further look at entropic complexity, and just how this is generated or distributed over different regions of the range space. In itself, Shannon entropy is certainly worth paying attention to in socioeconomic work, for a distribution with higher value for its Shannon complexity is one with a wider range of outcomes that can be either beneficial or on the other hand troubling. But from a user point of view, the Shannon entropy tells us nothing about just where the complexity arises—is it in a good zone or bad zone of the range space? A financial risk manager will be concerned about complexity in the left-hand tail region, of bad outcomes. A social commentator will likely be concerned about complexity in the right-hand tail region, referring in the context to a wider range of higher incomes or wealth.

In this respect, there is a difference between tail complexity and tail probability as such. To say that the tail probability is 10% is a summary judgment that does not tell us whether or not the remaining 10% decays very quickly thereafter. The difference can be important in contexts such as financial risk management. Even if

one sets a lower 5% limit as the tolerable probability for losses, what remains beyond that can be either moderately uncomfortable or extremely damaging.

Partition entropy is designed to address this problem on a general level. It takes the form of a function over the range space, which in the first instance tells the observer just where the complexity is generated, breaking this down into that associated with outcomes greater than or less than a given point. It is in effect a binary outcome, but manifesting as a function over the range space.

The partition entropy function itself is of very simple form, nothing very remarkable as such. But in addition to its inherent indicative value, it lends itself to the derivation of metrics that do have quite general descriptive value and a wide variety of potential applications. For the associated metrics, in particular those for asymmetry and spread, turn out to have intuitive appeal in their own right. But the entropic background completes the picture, in the process generating distributional transformation of independent interest, notably left and right shifting of the subject distribution. These can be used to generate an entire spectrum of distributional shapes. In turn, the moments of these entropically shifted distributions are associated with a double smoothing of those of original distribution.

It is this property that generates the recently developed entropic metrics for spread and asymmetry. Given an arbitrary value in the domain of the distribution, one can assess its position relative to other points in the domain in terms of the conditional expected values above and below. Taking differences allowing for sign gives a net advantage or disadvantage function. Taking the absolute differences gives a function for spread. And taking the expected values over all possible points in the domain gives the entropic metrics for asymmetry and spread. In turn, these can be expressed very simply in terms of the moments of the left and right entropic shifts of the original distribution function. A unique feature is that a simple internal change in sign converts the asymmetry metric to the spread measure. In this sense, the entropic asymmetry and spread metrics are dual to one another.

An ultimate test of any proposed metric is the scope of its application. In this respect, the entropic measures of asymmetry and spread must in themselves have substantive meaning in their respective contexts. The present study builds on a framework of applications in economics and social science for which this is the case.

The development sequence commences with Chap. 1, which establishes the basic ideas of partition entropy and the relationship with classic Shannon entropy. Computational aspects follow with specific reference to discrete distributions and histograms. Applications include grade scaling in education and tail complexity and market prices in finance.

Chapter 2 introduces the left and right entropic shifting of a given distribution, interpreted in terms of their relationship with partition entropy. Extensions encompass partial shifting and alternative formulations of the shifting algorithm. Perspectives on mixture distributions follow, which can in turn be applied to financial data and financial risk management. Entropic shifts can be used to provide scalable smoothing kernels in graphical work, together with end correction procedures. A topical application is to climate change data. Dynamic representations are applied to the outcome of opinion polling over time.

Chapter 3 contains the core development of the double smoothing process and consequent distributional metrics for spread and asymmetry. This extends to the relationship with Gini's mean absolute difference and to stochastic dominance. An application is to the dynamic bias in executive remuneration.

Further applications flow in Chap. 4. Income distribution has been a perennial topic of academic discussion, but never more so than in recent years, with the explosion of incomes at the upper end of the scale contrasting with an overpopulated lower end of the scale. The first part of the chapter shows how to interpret and apply the new metrics for asymmetry and spread to this context. This is followed by stock market performance in finance together with an application to the actuarial science of survival and age distributions.

Chapter 5 resumes conceptual development with a deeper look at entropic complexity and its metric implications. Supplementary measures for distributional spread and asymmetry can be derived in terms of the effective concentration of entropic mass, such that the given distribution is formally equivalent to bipolar outcomes, for example, as 'good' or 'bad'. One can think of this as a transformation of the original probability measure to a new one that simplifies decision-making. Applications continue to investment fund performance and to actuarial work.

Extension to bivariate and multivariate contexts follows in Chap. 6. Interpretations of double smoothing in two or more dimensions are exposited, together with the resulting entropic measures for spread and asymmetry. In turn, this can be applied to economic welfare measures. A more substantive application extends to financial market efficiency, via a conceptual concordance with the ordered mean difference, which corresponds to a co-smoothing where a subject variate is smoothed according to progressive values of a benchmark. The ordered mean difference construction can be employed to examine the issue of whether high profile offerings such as hedge funds add value, and if so whether this is aligned with their advertised purpose.

Chapter 7 concludes with a broad-ranging discussion as to further perspectives. This commences with the interpretation of partition entropy and associated metrics in terms of risk constructs. Financial risk management, in particular, calls for special attention to tail probabilities, with the left-hand tail of particular importance. Discussion continues with a review of the generics of comparison, and whether all the foregoing exposition should belong to the respective domains of ordinal versus cardinal metric evaluations.

A further topic concerns how to handle comparisons that may involve subjectivist probability. Information-based rescaling represents a way of reconciling decision rules in such a framework with those based on mainstream decision theory. The broader perspective is rounded off with the more conjectural discussion on organisational complexity and its possible reconciliation with coding complexity.

Computation throughout is fairly straightforward and can be handled with generalist packages such as Microsoft Excel. An exception arises in the work in Chap. 2 on data smoothing and end correction. In this respect, the appendix contains some source code in VBA that can be embedded as functions into an Excel environment.

Nearly, every topic studied in their respective chapters has its own associated literature, which while informative is often very extensive. In order not to interrupt

exposition with constant literature references, dedicated literature notes appear as the final section in each chapter. The aim is to set its substantive content within the general context. The literature notes are oriented to this purpose and should not be regarded as comprehensive accounts of their respective bibliographies. In some cases, the relevant literature is either (or both) classic or very large. In such cases, the references are often limited to established textbooks rather than to the historical original journal articles.

It could also be noted that there is an alternative line of thought, and accompanying literature, that deals with the information theory of model comparisons: of two alternative models, which is the better fit for the given body of data? The present study does not address issues of this kind, though there is some reference to it in the literature notes to Chap. 1.

Finally, this book is a study in general statistical methodology. The early part of each chapter develops themes of more or less universal applicability, with chapters following on from each other as the underlying theory is progressively exposited. The later sections of each chapter enlarge on illustrative application of the theory to that point. Chapter 4 is a standalone departure from this model, dealing as it does with applications of substantial economic or social importance. Readers with a primary interest in the general body of statistical theory or techniques, as distinct from this or that applied context, may find it helpful on first reading to skip this chapter and indeed some of the more specific applications embedded in the later part of each chapter.

As a general area, however, statistics finds many applications, even within the embracing field of economics and social science. Specialists with a primary interest in this or that particular discipline may find it helpful to have a guide to chapters and sections of specific interest. Table 1 provides such a guide. Sections not listed under their own headings are those that develop the general statistical methodology that underpins such specialist areas.

Table 1 A reference breakdown of specialist interests

Specialist area	*Chapter and section*
Actuarial science	4.4, 4.5
Climatology	2.4
Data analysis	2.4
Demography	4.4
Economics	3.4, 3.5, 4.1, 4.2, 7.3, 7.4
Education	1.6
Finance	1.7, 2.6, 3.6, 5.4, 6.3, 7.2
Political Science	2.5
Psychology	7.4
Management Science	7.5

Chapter 1
Partition Entropy

1.1 Introduction

From the practical point of view, many of the statistical measures developed in this book can be read and understood without any formal reference to entropy; they have intuitive or contextual relevance just as they stand. However a fuller understanding of these and other operational outcomes derives from their origin in entropy.

The idea of entropy, as a formalisation of information theory, has a long history in several different contexts. That of specific relevance is the general notion of complexity. Although embedded in more or less traditional probability, this adds a dimension of its own. A distribution which is more diffuse or has longer tails is inherently more complex than one of narrower focus. Complexity as a general notion has several connotations, technically distinct although with a common theme of unpredictability. The literature notes to this chapter summarise major interdisciplinary variants. However the context in what follows is specifically with the idea of entropic complexity. In what follows, Sect. 1.1 is an informal and elementary exposition of the idea of entropic complexity, and how this relates to the conventional Shannon (or differential) measure of total entropy.

Shannon entropy as such is a single metric; often reported among standard distribution metrics, but with rather little idea conveyed as to just what it refers to and how to use it. Hence Sect. 1.2 considers how to decompose the standard entropy metric into different regions of the domain. Thus one could consider two alternative distributions both with the same right hand tail probability at a given confidence level, but one with a much longer tail from that point on. The difference might be of considerable practical significance, and instances are reviewed in this and the chapters that follow.

The notion of partition entropy represents such a decomposition, taking the form of a function that differentiates between areas of high and low complexity. Two distributions might have exactly the same value for Shannon entropy; but in itself

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_1

this provides virtually no information as to how complexity is distributed along the domain axis. The partition entropy function fills this gap. As a function, it measures uncertainty in binary form, with its value at any given point capturing the indecision in assigning either an outcome greater than that point (the 'up') or less ('down'). Section 1.3 elaborates in terms of complexity, making also a connection with the principle of maximum likelihood. Complexity in terms of the 'ups' versus 'downs' is the stuff of professional success or failure for the financial investment community, so a brief digression follows, which has its own entertainment value. A further extension to the time series frequency domain illustrates the importance of the complexity arising from asymmetry.

Substantive development resumes in Sect. 1.4. A further connection of some importance is with the log odds function, often invoked as a description of how gamblers behave in contexts such as horse racing. The odds of winning versus losing reflect a binary outcome: 'up' for win or 'down' for lose. The partition entropy at any point can be viewed in terms of the conditional expected log odds to that point.

The core theory of partition entropy and associated metrics finds a natural expression with continuous valued random variables. For expositional convenience, much of the theoretical development is therefore exposited with distribution functions over a continuous domain. In practice, however, many variables of interest are defined over a discrete valued range space. The same basic ideas adapt readily to the discrete case, but computational issues do arise, notably how to deal with the log of zero. Section 1.5 turns to computational exigencies that result. One solution is to treat the distribution function $F(x)$ and its complement $1 - F(x)$ symmetrically, which in practice means basing the computation of the partiton entropy function $h(x)$ on the average of the two.

The last two substantive sections develop applications of the basic partition entropy function. Section 1.6 deals with a problem often experienced (but more rarely acknowledged) by educators, namely how to manage exam or coursework marks when the grade distribution is seen as suboptimal. This is a practical exercise in resolving operational uncertainty, which is of particular concern around the pass-fail boundary – again, the decision as to an 'up' or 'down' as a desirable outcome. The partition entropy function yields a scaling algorithm of more relevance than traditional polynomial based formulas.

Section 1.7 invokes a principle earlier referred to, that long tails carry more operational implications than just the tail probability as such. Thus in a financial market context, an investor with a very high reservation price relative to the current market price might well be prepared to buy a lot more of the stock, so that the ultimate clearing price would reflect the operational weight of potential investment. In turn, this is better captured by the partition entropy function in that region.

Section 1.8 concludes with the literature notes. Many research areas are at least cognate to the ideas of the present chapter, some of substantive importance. For reasons of space the literature review is of necessity a limited summary of what amounts to a classic body of established work in applied probability and statistics.

1.2 Information, Complexity and Entropy

In classical statistics information is commonly identified, directly or indirectly, with variance or variance reduction. An estimator of a given parameter contains more information than another if its sampling variance is smaller; or if bias exists, the mean square error is less. In a multivariate context, R.A. Fisher's information matrix, consisting in the limiting second derivatives of the log likelihood function, establishes a lower bound for the covariance matrix of any consistent estimator. However, it will be a thesis of subsequent chapters that variance in itself is an insufficient approach to problems of distributional uncertainty. While extensions in the form of higher order moments for asymmetry and kurtosis can add further insight, the classical moment approach rests on algebraic powers that need have no intrinsic connection with more systematically developed notions of uncertainty and risk. Moment based measures can indeed be useful for specific purposes, and indeed much of what follows in later chapters can be justified on more or less intuitive grounds with little or no reference to a formal framework in a theory of information. But they do carry extra conviction where such a framework exists and can be called upon.

Information in the structured form of entropy provides such a reference frame. The parent notion originated in statistical mechanics, with nineteenth century authors such as Ludwig Boltzmann and Willard Gibbs. The latter author, in particular, provided a formula for the entropy of a thermodynamic system as a probability weighted sum over microstates i of the corresponding log probabilities $\sum_i p_i \ln p_i$. It was a form of this kind that provided the basis for the formal theory of communication and coding developed by Claude Shannon, in the late 1940s, with subsequent contributions by other prominent authors such as John Tukey. In turn, this drew on the pioneering work of John Von Neumann in the theory of computation that underpinned the development of modern computers.

In this context, codes are a way of representing symbols in common use in a form that can be recognised and acted upon by computing machines. Complexity, and metrics that recognise complexity, are an important outcome that feature in the present study. So for this reason a short non-technical review of coding may be useful at this point.

The symbols in question take the form of a generalised alphabet. An example of the latter is the ASCI code for English, which basically consists of all the symbols on the standard laptop keyboard, 128 of them. The basic symbol set is often referred to as a generalised alphabet. Quantum computers possibly excepted, mainstream computers, or Von Neumann machines, operate via elementary switches, commonly represented as states '0' or '1' for 'off' or 'on', though alternatives such as 'U' (up) or 'D'(down) can also be useful for interpretive purposes. A binary code is then a mapping from the set of symbols to a set of code words, each of which is an assemblage of 0,1 bits.

So suppose an alphabet list of just three symbols or states: $S = \{a, b, c\}$. A suitable binary code might be of the form

$$\begin{aligned} a &\to 0 \\ b &\to 10 \\ c &\to 11 \end{aligned} \tag{1.1}$$

A fourth symbol in such a series would require three bits, so d might be allocated 101, and so on. A larger symbol set will require longer code words. Thus the ASCI set requires a seven bit binary code: $2^7 = 128$.

A binary code is then a mapping that takes the set of symbols in S to a set of code words. Thus consider two messages each of length 10 symbols:

$$\text{a b a a c a b b c a} \to 0100011010101 10 \quad (15 \text{ bits}) \tag{1.2a}$$

$$\text{a a a a b a a a a c} \to 0000100000010 \quad (13\text{bits}). \tag{1.2b}$$

The binary code (1) is actually a prefix code (or 'Huffman' code), which means that there is is no codeword that is the initial segment of any other codeword in the message set. In the above example the presence of contiguous two zero bits automatically means a new symbol with meaning of 'a'. The prefix property means that a special delimit marker between codewords is not required. It is more economical in form, meaning that any given message can be coded into a shorter message.

It will further be noted that a simpler message, as in (1.2b), can be coded with a shorter coded length. Indeed a given message can be regarded as a probability distribution of the constituent symbols, defined on S. Some symbols recur with greater probability than others. That is why the letter a in Scrabble is worth only 1 point; while z, being much rarer, is worth 10.

A symbol length function L allocates its associated number of bits, according to a given scheme; thus in the above example $L(a) = 1$; $L(b) = 2$; $L(c) = 2$. A convenient algorithm sets the length of each symbol code x in inverse proportion to its probability $p(x)$ of occurring, either in any single message or in general usage. In particular, for any symbol $x \in S$ a prefix code can always be found with

$$L(x) = [-\log_2 p(x)], \tag{1.3a}$$

where the square brackets refer to the nearest integer greater than the given number. Thus suppose the symbol a occurs with probability ½ in a given message, while b and c occur with probabilities ¼ each. This would give $L(a) = 1$; $L(b) = 2$; $L(c) = 2$ bits. The expected code length for any given message can then be obtained as

$$E[L] = \sum_x p(x)L(x). \tag{1.3b}$$

For some purposes or circumstances, it is convenient to disregard the integral part operator in expression (1.3a), so that (1.3b) becomes the non discretised form

$$E[L] = -\sum_{x} p(x)\log_2 p(x), \tag{1.4}$$

Expression (1.4) is the form generally quoted for discrete valued entropy. Its interpretive value is as a reference for the complexity of any given message as the expected number of bits required to code it. On such a role, the message set (1.2a) would have entropy value of 1.486, compared with 0.923 with message (1.2b), correctly capturing the eyeballing impression that the former is more complex than the latter. Entropy, in other words, can be taken as a prima facie indicator of the complexity of a given message.

The intuitive insights of complexity extend to random variables X on a discrete valued domain. Symbols can be identified with values $X = x_i$ in that domain and a given realisation of values x_i as a message that requires binary (0,1) type coding. Leptokurtic (peaked density) distributions have a smaller range of higher probability values with shorter code lengths. As in example (1.2a), encoding a random sample of values will lead to many smaller bit lengths and just a few longer, the latter with very low probabilities that contribute little to the expected code length for the random sample. Platykurtic densities endow more complexity to the random sample, with a range of more frequently occurring probability values; example (1.2b) is analogous. The difference can be seen in the case of a binomial distribution $B(r_i; n, p)$ taking values $r_i = 1, 2, \ldots, n$ with p the probability of success at each trial. So on $n = 20$ trials, maximum entropy of 3.208 arises with $p = 0.5$. Setting p at 0.9 results in a negatively skewed and leptokurtic density, with entropy at the lower value of 2.405.

Extension to random variables over a compact domain R takes the form of an integral

$$\kappa = -\int_{R} f(x)\log_2 f(x)dx \tag{1.5}$$

with $f(x), F(x)$ the density and distribution functions at x. The values of these integrals are widely reported as more or less standard properties of the respective distribution functions, commonly cast in terms of the natural logarithms as $\ln f(x)$, which differs from expression (1.5) only by the factor $1/\ln(2)$. The integral (1.5) is often referred to as differential entropy, or sometimes just as Shannon entropy where the context is undestood.

However, the extension to the continuous domain is not without problems. The resulting integral can end up negative. The uniform distribution is often cited in this respect with U(0, 1/2) an example. Over extended domains it may not even exist (converge) as $x \to \pm\infty$. And indeed the concept of an infinite alphabet list (set of symbols x) does not naturally correspond with the process of bit coding, since there would have to be symbol codes of infinite length ($n : 2^n = \infty$ has no solution).

A recourse is to decompose the total differential entropy into components less subject to such difficulties.

1.2.1 Entropic Decompositions

Shannon entropy is just a number, with no reference as to where uninformative regions lie. The same value of κ could equally result from a left or right skewed distribution or again, from a perfectly symmetric one. In the course of the present study several directional decompositions will prove useful. These enable a breakdown of total entropy into the comparative uncertainty associated with subregions of the domain. Thus a positively skewed distribution can be expected to generate more of its total uncertainty in the right hand tail, relative to the left hand tail or medial region.

A first such directional decomposition is obtained using the entropies associated with the conditional lower and upper distributions. Fix a given marker value $X = x$. Relative to this marker point, densities of the lower and upper conditionals are given by

$$f(X|X \leq x) = \frac{f(X)}{F(x)}; \quad f(X|X > x) = \frac{f(X)}{1 - F(x)}$$

with values zero outside their respective conditional domains.

Each of these will have its own Shannon entropy:

$$\kappa_d(x) = -\int_{-\infty}^{x} \ln\left(\frac{f(X)}{F(x)}\right) dX = \ln F(x) - E\left[\ln f(X)|X \leq x\right]. \tag{1.6a}$$

$$\kappa_u(x) = -\int_{x}^{\infty} \ln\left(\frac{f(X)}{1 - F(x)}\right) \frac{f(X)}{1 - F(x)} dX = \ln(1 - F(x)) - E\left[\ln f(X)|X > x\right]. \tag{1.6b}$$

It is then easy to show that total differential entropy divides into two parts:

$$k = h(x) + [F(x)\kappa_d(x) + (1 - F(x))\kappa_u(x)]. \tag{1.7}$$

where

$$h(x) = -(F(x)\ln F(x) + (1 - F(x)\ln(1 - F(x)). \tag{1.8}$$

The conditional entropy components (1.6a, 1.6b) are explored further in Sect. 3.2, also in Sect. 7.1. The function defined by expression (1.8) will be called the partition entropy at value or marker point x. Its meaning and properties will be explored in the sections that follow. Expression (1.7) indicates that Shannon

differential entropy can be decomposed at any given marker value into the partition entropy at that point plus the lower and upper conditional entropies weighted by their respective conditional probabilities.

Expression (1.7) is not the only possible decomposition of total entropy. Chapter 7 utilises a decomposition into just two parts $\kappa = \kappa_L(x) + \kappa_U(x)$ associated with the respective upper or right hand, and lower left hand divisions of the domain. These are referred to in that context as the lower and upper directional entropies. In the present chapter expression (1.7), and in particular the partition entropy component (1.8), forms the basis of ensuing discussion.

As a general note, transformations of the above kinds are often effected via Radon-Nikodym (R-N) derivatives operating on the original distribution. The present study contains several such contexts. The R-N derivative of one distribution $F_q(x)$ with respect to *F(x)* as another is conventionally denoted by $\frac{dF_q}{dF} = \xi_q$. If:

(i) The associated function $\xi_q(x)$ is nonnegative over its domain; and
(ii) has an expected value of unity $\int_* \xi_q(x) dF(x) = 1$,

then the respective densities are related by

$$f_q(x) = \xi_q(x) f(x). \tag{1.9}$$

No general analytical expression exists connecting the respective distribution functions F_q, F as such, but there are some`. Thus if we can write

$$F_q(x) = (1 + \xi_q(x))F(x),$$

then the densities must be related by expression (1.9) with $\xi_q(x) = -\ln F(x)$. In general, the R-N derivative must be specific to the given distribution F. However in some cases the functional form of $\xi_q(x)$ can originate in another distribution function with its parameters adjusted to ensure properties (i, ii) with respect to the distribution function under consideration.

1.3 Partition Entropy

The simplest possible entropic content arises with a variable that takes just two values, so that in effect its alphabet code list has dimension just 2. By transferring primary attention to its distribution function, any random variable X can be described in such terms. For any given value x within the domain, we can consider two events, $X \leq x$ and the complement $X > x$. Correspondingly, the support space of a given distribution function $F(x)$ can be divided into subspaces $L(x) = \{X : X \leq x\}$ and its complement $R(x) = \{X : X > x\}$. For brevity the symbol D ('down') will often be used for the first event and U ('up') for its complement.

On repeated drawings of X with the partition value x fixed, the resulting message list takes the form of a random sequence of U's and D's with probabilities $1 - F(x)$ and $F(x)$. The entropy is then the expected alphabet length as

$$h(x) = -[F(x)\ln(F(x)) + (1 - F(x))\ln(1 - F(x))]. \tag{1.10}$$

For the purposes of the present section it will be assumed that X is a continuous valued random variable. Section 1.5 considers the discrete case.

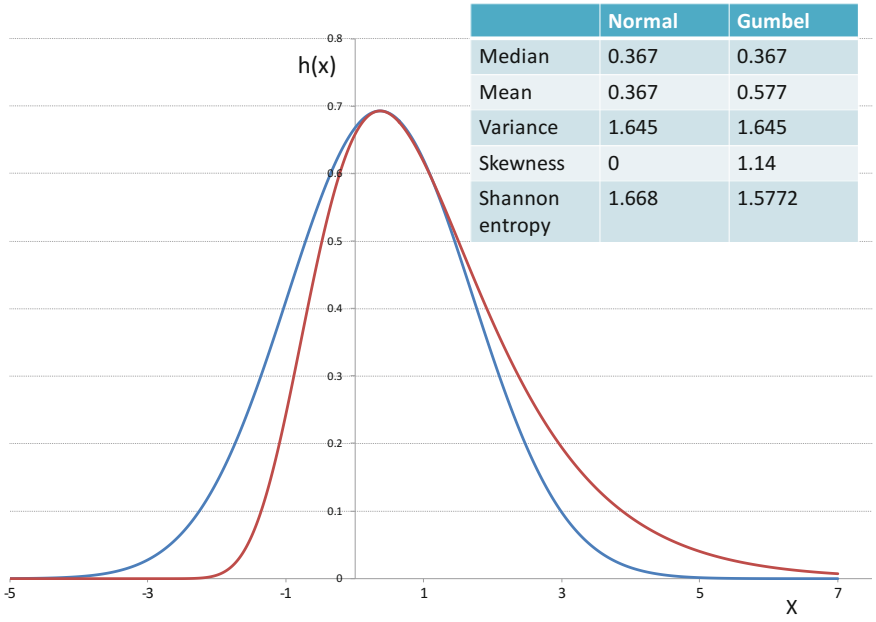
The binary entropy quantity defined by expression (1.10) will be referred as the partition entropy at x. It can be regarded as a measure of the information about the value of the random variable X derived from knowing whether $X \leq x$ *or* $X > x$. Expressed using natural logs to base 2, $h_2(x) = (1/\ln 2)h(x)$. A potential notational clash with the hazard function should be noted at this point (see end of chapter literature review).

In terms of standard entropy theory, the partition entropy corresponds to the mutual information between variable X and a regime indicator variable for the partition into either *L(x)* or *R(x)*. An alternative interpretation from classical statistics arises in connection with a Probit variable, which takes just two values ('U' and 'D'), with respective probabilities $p_d = prob(D;\theta)$; $p_u = prob(U;\theta) = 1 - p_d$ for a parameter θ of interest. Over a sample sequence of independent observations $i = 1, 2, \ldots, n$, the observed number of 'ups' might be n_U and of downs as n_D. The sample likelihood would then be $prob(D;\theta)^{n_d} prob(U;\theta)^{n_u}$, with the sample average log likelihood as $\frac{n_D}{n}\ln prob(D;\theta) + \frac{n_U}{n}\ln prob(U;\theta)$. With suitable regularity conditions this tends almost surely to $p_u \ln p_u + p_d \ln p_d$ as $n \to \infty$. In the current context, $p_u = F(x)$; $p_d = 1 - F(x)$ for fixed partition point x, and with parameters θ understood. Referring back to definition (1.15) the value $-h(x)$ can thus be interpreted as the limiting log likelihood function of a Probit experiment outcome.

Partition entropy is specific to the chosen value of x. Over the entire range of the random variable it therefore takes the form of a function rather than a scalar. As earlier noted, Shannon differential entropy $h = -E[\ln(f(x)]$ can be decomposed into the partition or locational entropy *h(x)* at any specific point x in the domain of X, plus the conditional differential entropies of the truncated distributions for $X \leq x, X > x$, weighted by their respective probabilities of occurring *F(x), 1 − F(x)* An analogy is with the between-group and within-group sums of squares in the analysis of variance.

If X is a continuous random variable over a range R, basic properties of the partition entropy function can be summarised as follows:

(i) If R is the entire real line, $\lim_{x\to\pm\infty} h(x) = 0$;
(ii) *h (x)* has maximum value $\ln 2$ at the median $x = x_m$ of the distribution (or value 1 if *h(x)* is expressed in logs to base 2);
(iii) The average value $E[h(x)] = \int_R h(x)f(x)dx = \frac{1}{2}$.



	Normal	Gumbel
Median	0.367	0.367
Mean	0.367	0.577
Variance	1.645	1.645
Skewness	0	1.14
Shannon entropy	1.668	1.5772

Fig. 1.1 Partition entropy functions, symmetric and asymmetric

Over the collection of marker points x, the partition entropy function shares the same skewness tendencies as the parent density function. Figure 1.1 illustrates for a Gumbel distribution compared with a normal distribution.

The Normal is the distribution that maximises differential entropy for a given variance. But relative to the Gumbel, the distribution of entropy uncertainty along the x-axis is quite different, with higher code values apparent for the positively skewed Gumbel in the right hand tail.

The general point is worth further elaboration. Figure 1.2 depicts the density $f(x)$ and partition entropy function $h(x)$ for a Gompertz distribution (actually reversed from the usual to show a longer right hand tail). The points $x = R$ and $x = L$ are equidistant from the median m.

With $x = m$ (point M) a sequence of 10 realisations (as a'message') would have equal proportions of U and D in sequences like U D U D U U D U D D. A corresponding sequence with x set at R would still contain a significant number of U's; perhaps something like D D D U D D D D U D. The former is evidently a more complex message. By way of comparison, suppose we set $x = L$. The corresponding message would be consist predominantly of U's, with very few D's. So the ranking in terms of complexity would be M > R> L. The partition entropy function h captures this insight: $h(M) > h(R) > h(L)$.

Technically speaking, the symbol set for a single realisation would be very simple: $U \rightarrow 1$; $D \rightarrow 0$, with no real reference to the probability of either symbol (U or D) in a given message. However, one could consider drawing repeated

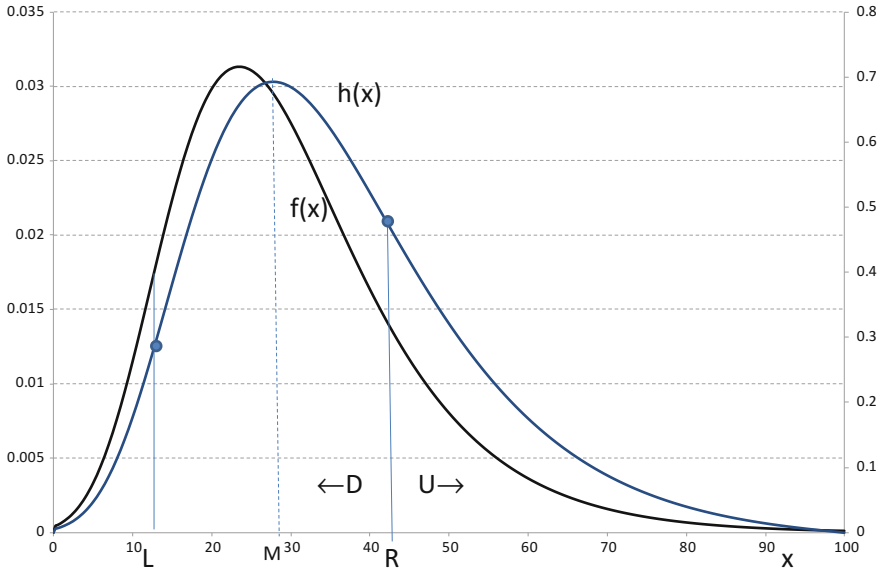


Fig. 1.2 Ups and downs for the partition entropy function versus the density

samples of a given size n, in effect establishing a sampling distribution. This would result in an extended message set. Thus for a sample of size $n = 3$, the symbol set would consist of the eight elements DDD, DDU, DUD, DUU, UDD, UDU, UUD, UUU. Given probabilities of D, U, one can assign probabilities to each element of the symbol set as $p(DDD), p(DDU), \ldots, p(UUU)$ and using (4) the expected non discretised code length of the message. One can show that for arbitrary sample size n, the expected code length is $n \times h(x)$.

The partition entropy function $h(x)$ can be regarded as capturing complexity, with specific reference as to how this varies along the range of the underlying random variable. The issue to be explored in this and subsequent chapters is how this insight can be exploited in decision making, and in metrics that facilitate the judgements necessary to do this.

1.3.1 Complexity and the Psychology of Investment Decisions

Preoccupations with ups and downs, and their relative distribution over ranges of possible outcomes, are the very stuff of the psychology of investment decisions. This is very evident in the context of recent or new company floats, in any form of emergent technologies or new market opportunities. Indeed, there are investors with an appetite for risk who specialise in this niche, either as principals in their own

right or as unitholders in dedicated managed funds, of which there are now a fair number.

To see the sort of decision process involved, suppose that the current share price of a newly hatched company is P_0. The first decision element is whether it will go up (U) or down (D), as a very basic judgement at the current price. But the same sort of judgement process is extended to all future prices. Thus the investor will be trying to assess the relative probabilities of U's and D's at every price $P \geq P_0$. Seen from the perspective of the current price, an assessment must be made of tail complexity from then on. Assessments of this kind are repeated at every subsequent point in time, with a judgement call as to whether sufficient upside potential remains to justify maintaining a long position. Long tail perceptions drive this form of investment.

As always, investment decisions in general may or may not be well informed or soundly based. In this respect, program trading provides chapter and verse. Program traders in financial markets are characterised by paying more attention to stock price patterns, as distinct from underlying fundamentals.

An early and still popular instance is Fibonacci trading. The Fibonacci numbers refer to the sequence 0, 1, 1, 2, 3, 5, 8, 13, 21, 34…. where each number is the sum of the previous two. They then have the fixed ratio property that to a reasonable approximation each number is 1.618 times the preceding or 61.8% larger. In turn, this property generates an entire sequence of such; for example, if we divide each number by the number two places to the right we get 38.2%, and so on.

The starting point of Fibonacci trading is to take a recent experience of a transition from a lower to a higher price. Then subdivide its range into intervals based on the Fibonacci numbers. Starting with the midpoint of the low-high range as zero, mark in the successively higher Fibonacci intervals as 14.5, 23.6, 38.2, 61.8 and 76.4% of the upper half interval. These numbers serve as markers for program trading based on subsequent prices. Particular attention is given to the 61.8% marker. It is claimed that this is commonly a reversal or 'retracement' point, and would be so even without any reflexive contribution from Fibonacci traders. Once the price has reached this point, sell signals are triggered and the price subsequently declines, even if this is only local and the upward movement is eventually resumed.

Figure 1.3 depicts an idealised example that serves also to illustrates one (of the many) potential flaws in the standard procedure, namely that it is based on an arithmetic decomposition of the low-high interval and not on the underlying distribution of prices within that interval. The partition entropy functions of two distributions A and B are depicted based on an identical range space. The popular upper 61.8% sell point is marked in as the vertical line, as is the recommendation to sell once this has been breached. This might make sense in the case of distribution A. But for the platykurtic distribution B, one would be selling when considerable upside complexity can still occur. There are still plenty of U's as well as D's.

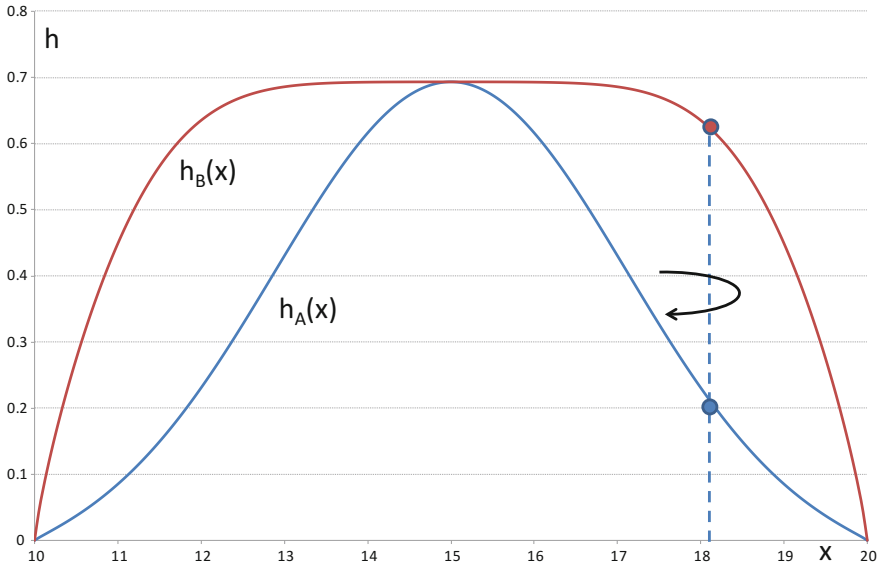


Fig. 1.3 Fibonacci retracement illustrated

1.3.2 An Extension: The Frequency Domain

Finally, it is worth pointing out that contextual extensions exist by a process of analogy. Time series analysis provides one such context. A regular zero mean stationary time series (y_t) with $Var(y) = \sigma^2$ can be decomposed into frequency specific components via the Cramer representation:

$$y_t = \int_{-\pi}^{\pi} e^{i\omega t} dy(w),$$

where $E[|dy(w)|^2] = \sigma^2 g(\omega) dw$ with $g(w)$ as the spectral density of the process. The latter gives the effective distribution of power (interpreted as variance) over the frequency range. A more complex series is one for which this power is spread more over the frequency axis. A process that consists of a single sinusoid would have zero entropic complexity, in effect just a deterministic sine wave. At the other extreme, the power spectrum for white noise, which has no autocorrelation structure, is flat.

Now there is nothing stop one from regarding the spectral density function $g(w)$ in the light of a probability density for the distribution of spectral power over the interval $(-\pi, \pi)$. The Shannon entropy is then an aggregate measure of the complexity of the process. But as before this tells us nothing about how the spectral power is distributed over the axis; is it higher at low frequencies (close to $w = 0$) or higher towards the extremes? The corresponding partition entropy function can be

used to address issues of this kind. Likewise it can be employed in relative welfare evaluations (e.g. in optimal control) of the effect of a proposed time series filter.

For the purposes of subsequent development, however, tangential extensions of this kind remain as background possibilities, as do other constructions of the idea of complexity such as those indicated in the literature notes. Thus in what follows, the context is taken to be a more or less standard probability or frequency distribution function, and complexity is interpreted as entropic complexity.

1.4 Partition Entropy and the Log Odds Function

In applications such as investing in dotcom company start-ups or horse racing, losses are typically limited to the amount of the initial investment. Motivation for the investment is the blue sky potential in the right hand tail area. Tail length, and how one reacts to this becomes the focus. At any given point x, the remaining tail length in the usual probability metric is $1 - F(x)$, effectively the right hand critical probability. In the metric generated by the partition entropy entailed, it is simply the value $h(x)$. The latter decays at a slower rate than does the critical probability.

Figure 1.4 depicts the two for a Gumbel distribution and for a normal distribution with its standard deviation chosen to result in the same 10% tail probability as that of the Gumbel. The gaps between the survival function $1 - F(x)$ and the partition entropy function are depicted as double headed arrows for the two distributions (marked as G for Gumbel and N for normal). The proportionate gap between the two becomes wider as x increases along the right hand tail, indicating that partition entropy decays less rapidly, and is moreover more responsive to longer tails than is the distribution function as such. In this sense, gamblers can be regarded as responding more to complexity than to strict probability. On this particular issue, Sect. 7.4 reviews more formal rationales in terms of subjectivist probability and behavioural economics.

An insight that reinforces the role of entropic complexity in tail behaviour is the relationship with the odds function. As before, for a given a marker value x, define two subsets of the support space $L(x) = \{X \leq x\}$, $R(x) = \{X > x\}$ for corresponding realisations of the random variable X. The logarithm of the odds that a point taken at random will lie in *L(x)* relative to *R(x)* is then defined by

$$\lambda(x) = \ln\left(\frac{F(x)}{1 - F(x)}\right). \tag{1.11}$$

The complementary function $\bar{\lambda}(x) = -\lambda(x)$ refers to the odds of being in the upper zone *R(x)* relative to *L(x)*. The case where $\lambda(x)$ is a linear function of x leads to the logistic distribution, so for this particular case the log odds function is a straight line with the median as crossing point. The logistic density is leptokurtic, with relatively short right hand tails. In contrast the Gumbel log odds is steeply

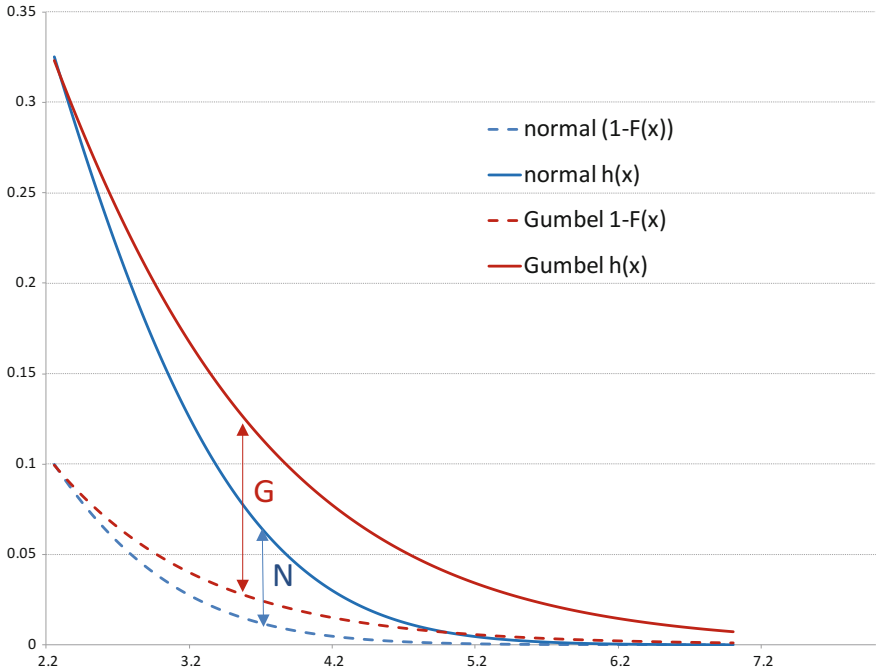


Fig. 1.4 Right hand tail decay

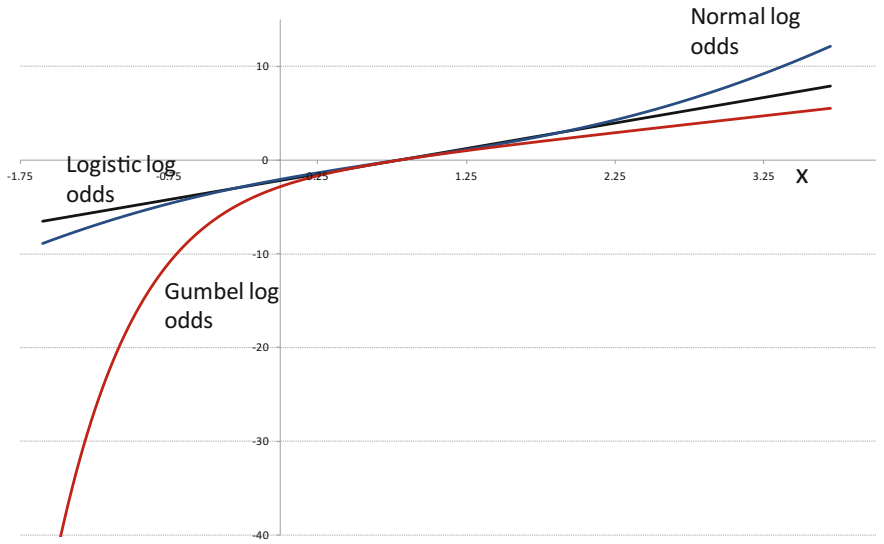


Fig. 1.5 Log odds functions, logistic, gumbel and normal

rising in the left hand, rising only more slowly in the longer right hand tail. Figure 1.5 illustrates, taking the normal and logistic distribution as comparators, with respective parameters chosen to result in the same Shannon entropy, in this case unity.

The log-odds function (1.11) is related to the density by $f(x) = \lambda'(x)F(x)(1 - F(x)$, and to total Shannon (differential) entropy by $H = -E[\ln(f(x))] = 2 - E[\ln(\lambda'(x))]$. In terms of the partition entropy function,

$$h'(x) = -f(x)\lambda(x); \;\; h(x) = -\int_{-\infty}^{x} \lambda(s)f(s)ds.$$

For any given marker value x, partition entropy can be written in terms of the conditional expectation of the log odds function as

$$h(x) = -F(x)E[\lambda(X)|X \leq x] = (1 - F(x))E[\lambda(X)|X > x],$$

where the log odds function $\lambda(.)$ is that of the unconditional distribution function F. A more complex right hand tail is one in which the log odds function decays less rapidly, relative to the remaining tail probability as such.

1.5 Discrete Valued Random Variables and Histograms

Turning to discrete valued random variables, the partition entropy can be defined for any given value $X = x_i$ as:

$$h(x_i) = -[F(x_i)\ln(F(x_i)) + (1 - F(x_i))\ln(1 - F(x_i))]; \; i = 1, 2, \ldots, N.$$

In working with the function $h(x)$ as a whole, the indeterminacy of $\ln(1 - F(x))$ at the terminal point x_N is not in itself a problem, as the product with $1 - F(x)$ ensures their product as zero. In computing one simply sets the last element as $h_N(x) = 0$.

However two further problems arise. The first is that if the density $f(x)$ is symmetric, one would like the resulting partition entropy function $h(x)$ to also be symmetric. Given a discrete domain, this does not happen with the tabulation of F as it stands. The latter starts from $F(x_1) = p(x_1)$, giving $h(x_1) = -[p(x_1)\ln p(x_1) + (1 - p(x_1))\ln(1 - p(x_1))] \neq 0$. However at the final point x_N we have $F(x_N) = 1$ and we set $h(x_N) = 0$. The resulting partition entropy function would as a consequence not be symmetric. The second point is related, and arises in symmetry based calculations that build on the resulting $h(x)$; examples are given in the chapters that follow.

One way of resolving such problems is to realise that the formula for h is itself symmetric in nature. It is written in terms both of the original distribution function $F(x)$ (e.g. 'mortality') and the complementary function $\Phi(x) = 1 - F(x)$ (e.g.

'survival'). A natural tabulation of the latter is to start at x_N and accumulate backwards, so that

$$\Phi(x_N) = p(x_N); \Phi(x_{N-1}) = \Phi(x_N) + p(x_{N-1}); \ldots \Phi(x_1) = \Phi(x_2) + p(x_1).$$

The implied forward version for F, written as $F_b(x)$, is recovered as $F_b(x) = 1 - \Phi(x)$ leading to the equivalent forward recursion as

$$F_b(x_i) = F_b(x_{i-1}) + p(x_{i-1});\ i = 2, 3, \ldots, N,$$

with $F_b(x_1) = 0$, compared with $F(x_1) = p(x_1)$ as in the usual forward tabulation.

The final step is to combine both versions as a centred discrete tabulation $F_c(x_i) = 0.5 * (F(x_i) + F_b(x_i))$.

This results in the sequence

$$F_c(x_1) = 0.5p(x_1);$$

$$F_c(x_i) = F_c(x_{i-1}) + 0.5(p(x_i) + p(x_{i-1})).$$

Using this version results in $F_c(x_1) = 0.5p(x_1)$ and $1 - F_c(x_N) = 0.5 * p(x_N)$.

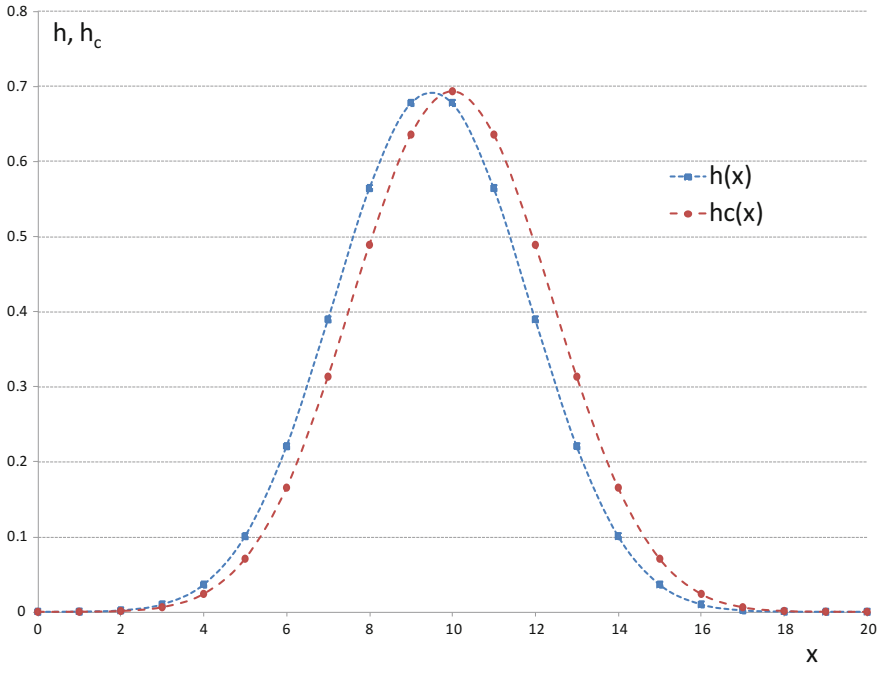


Fig. 1.6 Discrete partition entropy: centred versus non centred intervals

Thus if the density is symmetric, it will follow that $h(x_1) = h(x_N)$, and similarly for other points in the x domain. In all cases involving discrete computation, it is necessary to normalise so that the resulting densities sum to exactly unity.

Figure 1.6 illustrates with a binomial density with p = 0.5 and 20 trials. The centred version $h_c(x)$ is symmetric and correctly centred on the median.

There are alternative ways to resolve the bidirectional problem, referring to the need to accommodate both the forward (e.g. probability of death) and backward (survival) accumulations. In some circumstances it may be justifiable to treat the discrete data as a sample from a continuous distribution, derive a smoothed version of *F(x)* and thence the computed *h(x)* in the form of a continuous function. Demography provides a context of this kind. The applications that follow in this and further chapters make use of both approaches, depending upon context.

1.6 Resolving Grade Uncertainty

Scoring systems occur in many contexts that involve individual assessment with an element of subjectivity. They are most familiar in educational imperatives, where students have to be assigned marks or grades for reporting purposes. However, less formal contexts have also become pervasive in commercial life and social media.

Most scoring systems have some element of subjectivity. Even multiple choice tests can have embedded elements of personal judgement as to question definition or the most appropriate box to tick. Scoring consistency is a further general problem. This can arise where multiple assessors are required for a given application, or where results have to be standardised over time and over different social or geographic contexts. Likewise, formal qualification systems generally try for score equivalence across different subjects. Examples are the GSCE Uniform Marking System in the UK or the Universities Admission Centre in Australia.

On a less formal level, the economic and personal consequences of a fail can be serious and often expensive for students. Most experienced teachers have had to face up to problems of assessment consistency, and even doubts as to their own judgment, in situations where remarking may not be feasible or even a solution. Many will have their own customary scaling algorithm. The adjustment across lower to higher raw scores may be linear monotonic. More commonly it is modal, utilising zero and 100 as benchmarks, with remaining polynomial parameters chosen to generate a required mean and variance.

As noted, there may be administrative and related reasons to scale marks. But as a matter of educational philosophy, the informational content of assessment scores is important. For instance, there may be good reasons to argue against automatic monotonicity, where the lower the score, the higher the upward adjustment. A very low mark such as 5% has signalling value to the student that he or she is not suited to the subject – scaling it up to 25% might send a wrong signal that repeating the course could lead to success. Likewise, if the raw median falls well short of the pass mark, this is providing information that the test is likely too difficult. In such cases,

the informational context of the decision to scale is implicitly being recognised. By way of contrast, polynomial based scaling systems are linked more to an idealised mark distribution, rather than one which imbeds information about the candidate or the exam.

Partition entropy enables a scaling approach that explicitly recognises the informational content of raw scores, which can concern not only the candidate (particular raw score) but the scoring criterion or test itself. A realisation near the median carries more informational value from knowing whether other marks are likely to be greater than, or less than, the given value. The teacher should be correspondingly less confident about a decision to fail such a student and to give the benefit of the doubt. By way of contrast, an extreme mark such as 5 or 95% has little entropic information value, though plenty of personal signalling content for the candidate! The assessor can be confident that such students should either fail with an E or be awarded an A^+.

The informational approach therefore suggests a scaling algorithm that assigns heavier weight to the marks that carry more information, which typically means marks closer to the median than those further away. However, if for some reason monotonic scaling is preferred, an informational theory approach based on entropic shifting is also available (Chap. 2).

Let the original or raw scores on a given assessment be $x_1, x_2, \ldots, x_N$. To see how the partition entropy function $h(x_i)$ can be used to scale scores, let x_m be the median score, which is taken as the central reference (in place of the arithmetic mean sometimes used). Further, let x_m^* be a desired or target median, and define the shortfall as the difference:

$$shortfall = x_m^* - x_m.$$

In the basic scaling algorithm, the adjusted mark for candidate i is of the form

$$x_i^* = x_i + \left(\frac{h(x_i)}{\ln 2}\right)^{\theta} \times shortfall\,; \quad \theta > 0.$$

At the raw score median, $h(x_m)/\ln 2 = 1$ so the desired median x_m^* is achieved. The user defined parameter θ adjusts the incidence or degree of the scaling, with $\theta = 1$ as the benchmark. Here $\theta < 1$ will correspond to a more liberal treatment, with $\theta > 1$ as more conservative. In either case the maximum scaling effect is achieved at the median of the raw scores; the adjustment factor determines how rapidly this falls off away from the median.

The scaling algorithm is computationally straightforward. For the generalist user, it can be executed using Excel worksheet functions such as *Countif* to establish distribution functions directly from the alphabetical class list scores.

If it is desired that the scaling incidence is more generous short of the median, this can be achieved with user defined parameters θ_0, θ_1 such that

$$\theta_i = \theta_0 - \theta_1 \times I_-(x_i);\ 0 < \theta_1 < \theta_0,$$

where the unit left hand sign function $I_-(x_i)$ can be obtained in Excel as

$$I_-(x_i) = -\frac{1}{2}[sign(x_i - x_m) - abs(sign(x_i - x_m))].$$

Criteria for choosing flexible scaling parameters are considered below.

Figure 1.7 compares scaled and unscaled Excel histograms for N = 38 student examination marks, assembled to be representative of the decision exigencies often met with. With a raw median mark of 45%, and a pass mark of 50%, the test has been revealed as too tough for such students. The scaled marks have a median of 55 and a pass rate of 65.8%. However very low marks remain that way, continuing to convey a signal to the students concerned that perhaps this subject is not for them.

1.7 Tail Complexity and Market Clearing Prices

The general theme of complexity extends to comparative complexity as between upper and lower tails of a given distribution. A potential application is in understanding the clearing of market prices in financial asset trading, as in the share price from day to day or second by second, on a stock market. Two subthemes arise: the first concerns the actual determination of a clearing price, and the second concerns market stability. The context in both cases is where hold, sell or buy intentions are not uniform.

This is indeed the case in equity markets. Even the most casual perusal of investment analyst recommendations reveals quite radical differences of opinion as to whether a given stock is a good buy or on the other hand, a sell. There may be relative uniformity at certain times and at others, quite fundamental differences. Figure 1.8 illustrates with a range of analyst recommendations for JPMorgan Chase & Co, one of the world's major banks. A 'buy' zone would indicate that some analysts think that the current price errs on the side of conservatism, while a smaller number of analysts think quite the opposite.

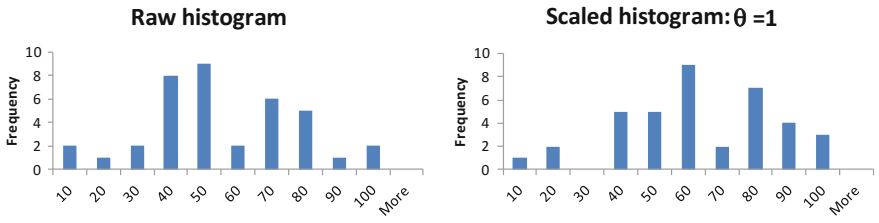


Fig. 1.7 Raw versus scaled test marks

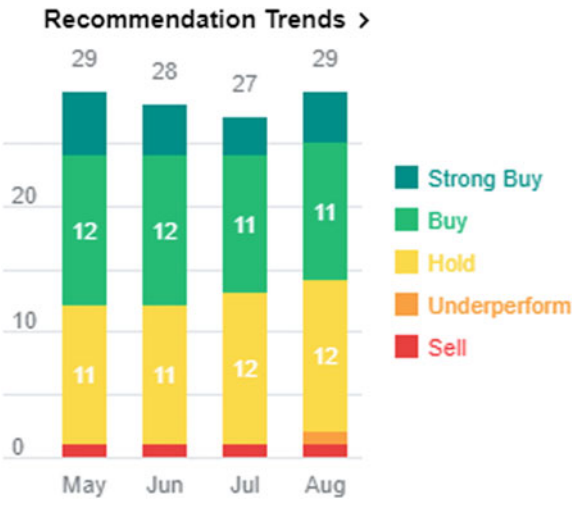


Fig. 1.8 Analyst recommendations: JP Morgan Chase & Co

A simple model of market clearing balances up the amounts demanded (the buy side) and the amount demanded (sell side). At any instant of time, let p be a current quoted price, and for simplicity imagine that this is a very liquid stock, so bid-offer gaps are ignored. For any prospective market participant, denote by x_i his or her break even ('reservation') price such that an order amount to buy or sell is proportional to the difference $x_i - p$, and suppose there are $n(x_i)$ such agents.

Specifically, the pressure from the buy side is $n(x_i)(x_i - p)_+$ and that from the sell side is $n(x_i)(p - x_i)_+$. Let $f(x_i) = \frac{n(x_i)}{N}$ refer to a probability density of agent breakeven prices. The market clearing price p is then given by

$$\sum_{i=1}^{N} f(x_i)(x_i - p)_+ = \sum_{i=1}^{N} f(x_i)(p - x_i)_+ . \tag{1.12a}$$

Anticipating notation of a later chapter, let

$$\mu_l(p) = E_F[x|x \leq p] \text{ and } \mu_r(p) = E_F[x|x > p].$$

Then expression (1.12a) becomes

$$F(p)(p - \mu_l(p)) = (1 - F(p))(\mu_r(p) - p) \tag{1.12b}$$

The solution to (1.12b) is $p = \mu$, where $E[x_i] = \mu$ is the average breakeven point over the different agents. In this simple linear demand/supply scenario, the market price clears to as the expected value of the respective break even points. An even simpler model would be where each agent has the capacity to buy or sell just the one unit of the stock. The market clearing price would then be the median $p = x_m$.

However, over different times and trading scenarios there is a wide range of both intentions and capacities to either buy or sell, or on the other hand simply to hold. And as the flow of news continues from trading period to the next, the distributions of buy or sell intentions will change, both as to location and spread. Figure 1.9 contrasts the partition entropy functions for two alternative distributions (A, B) of investor breakeven points, one of them with a much longer long tail in the optimist direction. The point marked as p is a current market price. Under scenario A there is good reason to think that the equilibrium price will autocorrect to the common mean and median x_m.

Scenario B distribution shares x_m as median, so a common maximum for the partition entropy function. But in this case the current price p has a better chance of being an equilibrium. The market would see a flow of both sell orders (D for down) and buy orders (U for up). At price p a broker would see very few U's coming through in the case of distribution A; but a mix of D's as well as U's in the case of distribution B, reflecting in the case the greater entropic complexity of the right hand tail. As earlier noted, the mean $\mu_B > \mu_A$ is a possible equilibrium clearing price. But there is a case that such a long positive tail would also tend to suggest that some of the more optimistic investors would put in proportionately larger buy orders at the current price p. Thus the equilibrium price might exceed even the mean μ_B. The issue of just what might result is taken up in Chap. 6, in connection with the centre of entropy as a distribution metric.

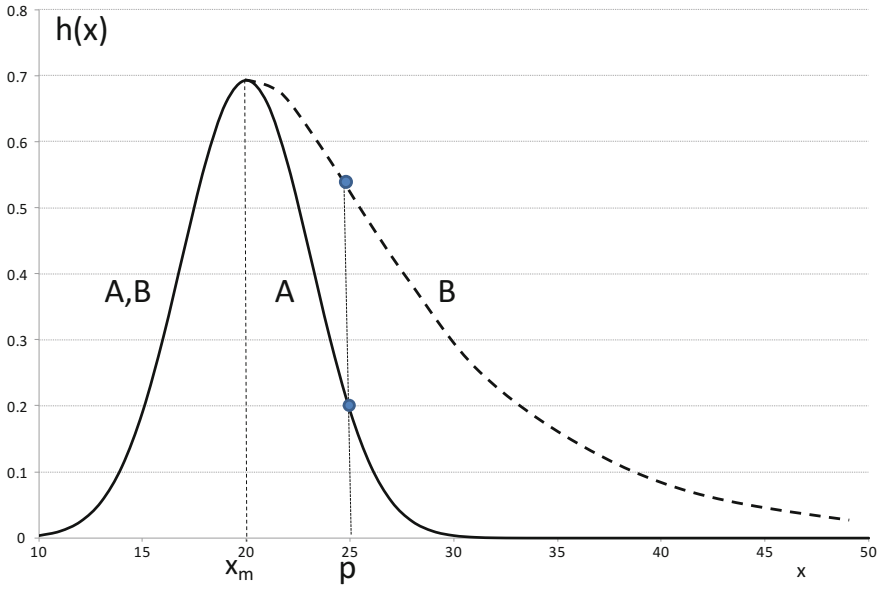


Fig. 1.9 Market clearing equilibria as shape dependent

1.8 Literature Notes

The characterisation of exactly what is meant by complexity has taken place along several lines in different literatures. Closest in character to the entropic approach is Kolmogorov/Solomonoff complexity. Given a string of characters, each character having its own allocated number of bits, one seeks the shortest possible description of the given string in some programmable language. The length of this description is then the total number of bits needed to code it. Kolmogorov complexity refers to a lower bound for the total bit length among all possible programmable language. Li and Vitanyi (1997) is a useful reference.

An alternative approach is that of Lyapunov, in the context of a time series x_t. Plotted in phase space (e.g. as $x_{t+1}\, v.\ x_t$ or as $\Delta x_t\, v.\ x_t$) the Lyapunov exponents characterise the progressive volume of the phase space as reflected in the tangents, with the object of determining whether the system converges to an attractor point. Chaotic systems do not converge in this sense, so are considered to be more complex. There are numerous applied studies in fields such as economics or finance which have reference in one way or another to this line of complexity.

A less formal approach to complexity has been attempted in organisation theory, spanning such issues as the optimum coexistence of a system's formal structure together with the independent or self organised actions of those who have to work within it. However formal measures or metrics of organisational complexity have not yet found common acceptance. Section 7.5 briefly addresses this topic.

Resuming the thread of Sect. 1.1, there are many web discussions on entropy, though some are within the framework of thermodynamics or fluid dynamics, and do not enter into the complexity aspect as such. Some useful texts that do are Sethna (2006), Von Baeyer (2003), and Dugdale (1996). Applied to statistics, there are a number of treatises and texts. Useful examples are Pinsker (1964) and Kullback (1968).

A related topic is Kullback-Leibler information divergence. Given two alternative distributions $F(x),\ G(x)$ defined over a range * for the same apparent outcome x, this is defined as

$$\int_* g(x) \log \frac{g(x)}{f(x)} dx,$$

or the corresponding weighted sum if x is discrete valued. This is the information gain if G is used instead of F to model the data. In complexity terms, it is the expected code length that would result from using G instead of the code that would have been optimal for F, the object in both cases being to minimise the expected code length.

Extensions to bivariate contexts are considered in Chap. 6, which includes further relevant literature notes on entropy in general. The Wikipedia articles on entropy, and related topics, such as information divergence and the Akaike

information criterion for model selection, are in the main quite comprehensive and well informed.

In economics, Georgescu-Roegen (1971) proposed to explain the course of economic life as a progressive and irreversible transformation of 'low entropy natural resources' into 'high entropy' economic outcomes. The usage of energy resources is of this character, being dissipated ultimately as waste heat. Further references to entropy in economics are in Chap. 4 in connection with income distribution. But these do not deal explicitly with the complexity dimension, casting discussion in terms of the conventional formula for Shannon entropy as a starting point.

R.A. Fishers' information matrix is formally defined by the limit in probability (almost sure etc. as applicable) of the second derivative of the log likelihood function. As such, its inverse gives the lower bound for the asymptotic variance of any other consistent estimator. Any textbook of advanced statistics will have a coverage, while Norden (1972, 1973) supplies an extensive bibliography. Its inverse is indirectly a measure of the information content of the sampling distribution of parametric estimators.

Partition entropy was introduced in Bowden (2012), under the name of locational entropy, which remains a synonym. The source contains the basic properties of $h(x)$ listed in Sect. 1.3, together with proofs.

With reference to Sect. 1.2, the range decomposition of total entropy has been considered by Di Crescenzo and Longobardi (2002), also Asadi et al. (2005, 2006).

Fibonacci trading is exposited by many enthusiastic, but on the whole informal, sources on the web. No formal studies have ever confirmed its empirical merit.

On the use of scaling in educational assessment, Manly (1988), Broydon (1983), Krzanowski et al. (1985) are cognate references.

A potential notational clash is noted at this point. The notation *h(x)* is commonly also used for the hazard function or survival function (the inverse is the Mills ratio) $h(x) = \frac{f(x)}{1-F(x)}$, which is such that $h(x) = \frac{-d\log(1-F(x))}{dx}$.

References

Asadi, M., Ebrahimi, N., & Soofi, E. S. (2005). Dynamic generalized information measures. *Statistics and Probability Letters., 71,* 85–98.

Asadi, M., Ebrahimi, N., Hamedani, G. G., & Soofi, E. S. (2006). Information measures for Pareto distributions and order statistics. In N. Balakrishnan, E. Castillo, & J. M. Sarabia (Eds.), *Advances in distribution theory, order statistics, and inference (In honor of Barry Arnold)* (pp. 207–223). Boston: Birkhäuser.

Bowden, R. J. (2012). Information, measure shifts and distribution metrics. *Statistics, A Journal of Theoretical and Applied Statistics, 46,* 249–262.

Broydon, C. G. (1983). A mark-scaling algorithm. *Computer Journal, 26,* 109–112.

Di Crescenzo, A., & Longobardi, M. (2002). Entropy based measure of uncertainty in past lifetime distributions. *Journal of Applied Probability, 39,* 434–440.

Dugdale, J. S. (1996). *Entropy and its physical meaning* (2nd ed.). UK: Taylor and Francis; US: CRC. ISBN 0-7484-0569-0.

Georgescu-Roegen, N. (1971). *The entropy law and the economic process*. Harvard University Press. ISBN 0-674-25781-2.

Krzanowski, W. J., Mead, R., & Thorne, S. (1985). On the use of average marks in degree examination data. *Applied Statistics, 34,* 219–226.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

Li, M., & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). Berlin: Springer. ISBN 0-387-94868-6.

Manly, B. F. J. (1988). The comparison and scaling of student assessment marks in several subjects. *Journal of the Royal Statistical Society Series C: Applied Statistics, 37,* 385–395.

Norden, R. H. (1972). A survey of maximum likelihood estimation: Part 1. *International Statistical Review, 40,* 329–354.

Norden, R. H. (1973). A survey of maximum likelihood estimation: Part 2. *International Statistical Review, 41,* 39–58.

Pinsker, M. S. (1964). *Information and information stability of random variables and processes*. San Francisco: Holden Day.

Sethna, J. P. (2006). *Statistical mechanics: Entropy, order parameters, and complexity*. Oxford: University Press.

Von Baeyer, C. H. (2003). *Information – the new language of science*. Harvard University Press. ISBN 0-674-01387-5.

Chapter 2
Entropic Shifting Perspectives and Applications

2.1 Introduction

With the passage of time or circumstance probability distributions change, and as they do so, the allocation of entropic complexity likewise changes. The present chapter develops a systematic way of modelling such developments. In doing so it continues with the groundwork for the origination of the entropic metrics of later chapters, which in themselves have value even where the underlying distributions remain stationary.

Partition entropy provides a useful framework for distribution plasticity in its own right. Starting from the idea of unit left and right entropic shifts (Sect. 2.1), a calculus of partial left and right shifts, concentrators and spreaders, is developed in Sect. 2.2.

Section 2.3 generalises to the context of distributional mixtures. The mixing weights for such compound distributions collectively comprise a kernel function, which is applied to a hypothesised underlying generator distribution, the latter remaining stable as the mixing weights vary over time. The objective is then to back out the generator, which could be identified as a long run stationary distribution. Entropic spreading or concentration describes how distributions such as stock market returns can end up with the long tailed property observed over longer horizons.

Section 2.4 takes up a very common problem in applied time series analysis, specifically its graphical presentation and interpretation. In data smoothing the need is to filter out underlying trends from an overlay of more transitional influences and disturbances. The partition entropy function yields a particularly convenient filtering kernel, with the flexibility of ready adjustment to filter windows of differing lengths.

An operational payoff is to the problem of end correction, where the fixed window length has to be shortened in a systematic way as the last available data point is approached. Density shifting, using the right entropic shift, redistributes

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_2

kernel weight to cope with this problem in a systematic way. The topical application is to climate change. In this case, special weight attaches to more recent observations as supporting (or otherwise) the continuation of tends that have become apparent with the complete kernel available with more historical observations.

Section 2.5 turns to dynamics over time, as where preferences or social attitudes may change quite quickly. Partial left and right shifts or their generalisations can be a convenient way of modelling distributions of opinion as difference or differential equations. A remarkable instance is the sea change of attitudes towards gay marriage in the US.

A financial theme is continued in Sect. 2.6, which deals with risk management constructs used in bank prudential management and insurance. The problem addressed here is how to construct risk management limits to cope with the long tailed property as it is generated over different horizons. Section 2.7 is the literature review.

2.2 Left and Right Entropic Shifts

Partition entropy can be established as the difference between the values of two distribution functions derived from the original distribution function by a process of directed shifting. If the range space of the random variable is unidimensional then the shifts are directed to the left and right of the original, respectively. These can be regarded and referred to as unit shifts of the parent distribution function. Higher dimensional range spaces are considered in Chap. 6, but the directional character is preserved.

In all cases the new distribution functions are defined on the same range space as the original, so the new random variable has the same domain as the random variable from which it is derived. It is simply a change of probability measure. The shifts are accomplished via a process that corresponds to a change of measure accomplished via appropriately left and right oriented Radon-Nikodym derivatives, earlier introduced in Sect. 1.2. In the present context, where the primary focus is on specific unit shifts, the measure theoretic interpretations are not essential, but it is useful to develop things more generally, as other types of shifts can also be contemplated within the same framework.

Thus let P denote the original probability measure so that if B is any Borel set (bounded and closed) in the domain, then the probability value attached to B under the new measure is $P(B)$. Then the transformation to the new measure is accomplished by a function $\xi(x)$ such that $Q(B) = E_P[I_B\xi]$; the transformation is often written as $\xi = \frac{dQ}{dP}$.

In the one dimensional and continuous case, if we start with $F(x)$ to end up with $Q(x)$ then $\xi(x) = \frac{dQ}{dP} = \frac{dQ/dx}{dP/dx}$, so that $\xi(x)$ specifies the ratio of the two densities at any given point. For this to be a valid measure shift two things are required:

(i) $\xi(x) \geq 0$ for all x, i.e. ξ must be a nonnegative function; and (ii) $E_P[\xi(x)] = 1$. The function $\xi(x)$ is referred to as a Radon Nikodym derivative. The new density is given by $q(x) = \xi(x)p(x)$.

In the present context, the starting point is the natural probability measure and P is identified with an original distribution function $F(x)$. The latter is taken as continuous; the discrete case is reserved for a later section. Two kinds of shift will be explored, corresponding to unit left and right shifts of the original distribution function. These are respectively defined by the R-N derivatives

$$\xi_L(x) = -\ln F(x)$$
$$\xi_R(x) = -\ln(1 - F(x)).$$

Nonnegativity is apparent, while integration by parts shows that $E_F[\xi_L(x)] = E_F[\xi_R(x)] = 1$. Thus ξ_L, ξ_R qualify as R-N derivatives for changes of measure. Figure 2.1 depicts the left and right shift functions for the normal distribution, which is symmetric, and the Gumbel, which is skewed to the right.

The derivatives ξ_L, ξ_R are of independent interest. In a reliability or mortality context, they correspond to the left and right hand hazard functions:

$$\xi_L'(x) = -\frac{f(x)}{F(x)}; \quad \xi_R'(x) = \frac{f(x)}{1 - F(x)}$$

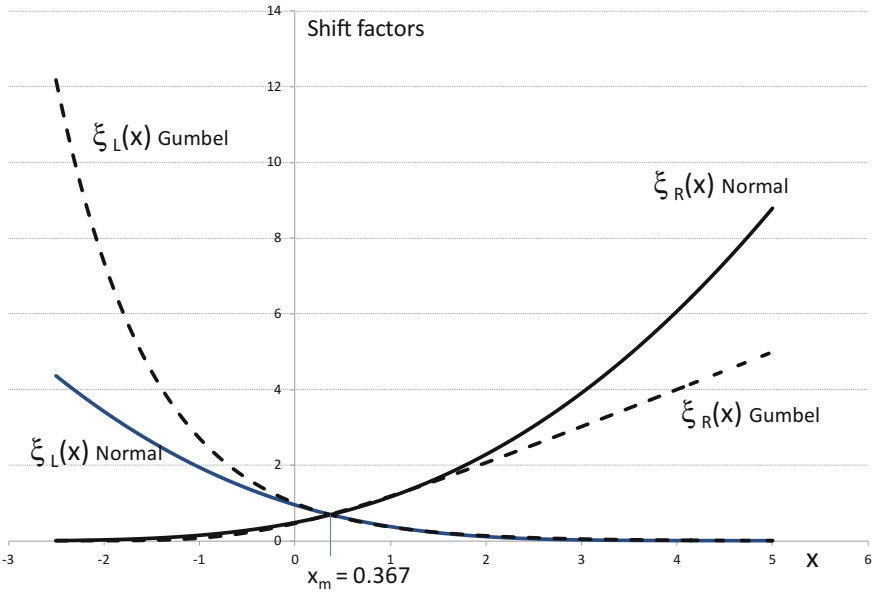


Fig. 2.1 Shift factors for some common distributions

Thus $\xi_R(x)$ would refer to the log of the survival function at time x.

The specific interest for present purposes arises in connection with directional distribution shifts. Given an original density $f(x)$, the new densities, defined on the same range space, are given by

$$f_L(x) = f(x)\xi_L(x) = -f(x)\ln(F(x)); \tag{2.1a}$$

$$f_R(x) = f(x)\xi_R(x) = -f(x)\ln(1 - F(x)) \tag{2.1b}$$

The corresponding distribution functions $F_L(x)$, $F_R(x)$ can be written in the form

$$F_L(x) = F(x)(1 + \xi_L(x));$$

$$(1 - F_R(x)) = (1 - F(x))(1 + \xi_R(x)).$$

The right shifted version $F_R(x)$ is therefore homologous in form with the left $F_L(x)$, except that it is more naturally cast in terms of the 'survival function' complements $1 - F(x)$ and $1 - F_R(x)$. Explicit expressions are

$$F_L(x) = F(x)(1 - \ln F(x)); \tag{2.2a}$$

$$F_R(x) = 1 - (1 - F(x))(1 - \ln(1 - F(x))). \tag{2.2b}$$

The two distribution functions and their densities are respectively referred to as the unit left and right shifts; partial shifts are discussed at a later point. Figure 2.2a, b illustrate their effect for the Gumbel distribution. Figure 2.2a is the density. A density that is already positively skewed, as in the original, becomes bunched up to the left under L. The right shifted density assumes a less asymmetric shape. For a symmetric density such as the normal, the left and right shifted densities are anti-symmetric about the median, though no longer symmetric.

Figure 2.2b depicts the effects on the distribution function in the case of a unit right shift. Vertical translation is invariant at the original median m of F; for any distribution, $F_L(m) = \frac{1}{2}(1 + \ln(2))$ while $F_R(m) = \frac{1}{2}(1 - \ln(2))$. Horizontal translation to points of equal cumulative probabilities requires the solution to equations such as $F_L(y) = F(x)$ and must in general be solved numerically. In Fig. 2.2b the point $x = 1.0$ with $F(1) = 0.692$ translates to y = 2.357 for $F_R(y) = F(x)$.

The unit left and right shifted distribution functions $F_L(x)$, $F_R(x)$ are directly connected with the partition entropy function $h(x)$ of Chap. 1. Inserting their respective definitions as in expressions (2.2a, 2.2b) gives

$$F_L(x) - F_R(x) = h(x).$$

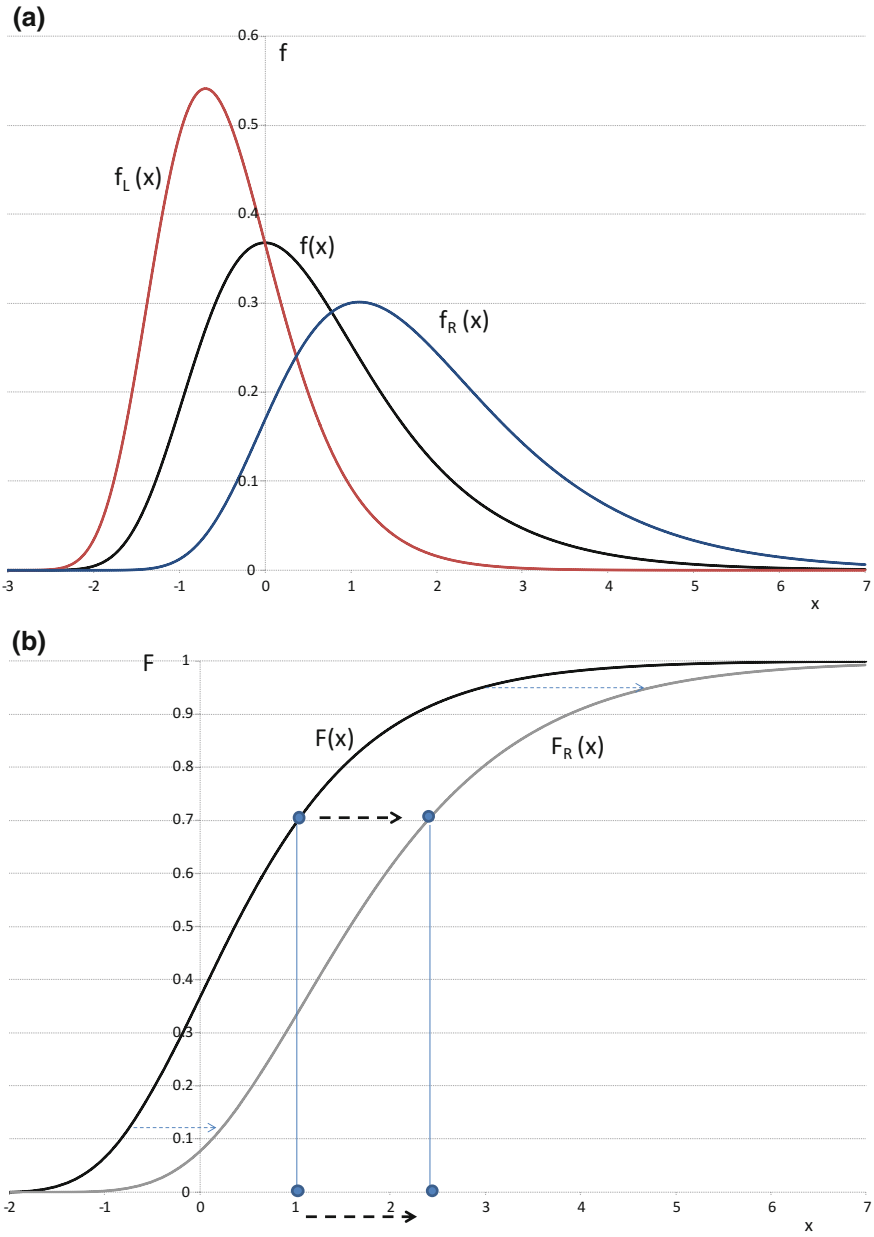


Fig. 2.2 **a** Left and right shifted densities, **b** left and right shifted distribution functions

A distribution function F exhibiting a big vertical spread between F_L, F_R at a given point x has therefore a high entropic spread at that point. Or if the difference is systematic over x, then the partition entropy function has itself a wider spread.

In terms of the respective densities,

$$h'(x) = f_L(x) - f_R(x).$$

So the difference between the left and right shifted densities is identical to the derivative of the partition entropy function. It follows from the density definitions (1) that at the median m of F,

$$f_L(m) = f_R(m) = f(m)(1 + \ln 2).$$

Thus the left and right shifted densities cross at the median.

Shifts can be iterated, e.g. as F_{LL} for further leftward translocations. Note, however, that $F_{LR} \neq F$, so that a leftward shift followed by a right does not restore the original. Instead, the inversion of a given distribution F resulting from a left shift is given by the distribution function $F*$ that satisfies

$$F_L^*(x) = F*(x)(1 - \ln(F*x)) = F(x);\ all\ x.$$

The formal solution to this functional equation is $F*(x) = e^{W[\frac{(-F(x))}{e} + 1]}$, where $W(.)$ is the Lambert W or log product function. Similarly, $F*(x) = 1 - e^{W(\frac{(1-F(x))}{e} + 1)}$ for the right shift inverse.

The Lambert W function has a Taylor series expansion $W(x) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} x^n$, but the radius of convergence (*1/e*) is not suited to the current context and it can in any case be very slow. Instead, basic numerical methods such as *Solver* in an Excel-VBA environment can be used if such inversions are required. It is necessary to constrain the iteration away from values respectively F = 0 or 1 (as ln(1-F)), to avoid the log of zero.

2.3 Shift Kernels: Concentrators and Spreaders

With the unit left and right shifts in place, a menu of partial and combined shifts becomes available.

(a) Partial right (similarly left) shifts can be achieved with the Radon-Nikodym derivative

$$\xi_{\lambda R}(x) = (1 - \lambda) + \lambda \xi_R(x);\ 0 < \lambda < 1.$$

This leads to

$$F_{\lambda R}(x) = (1-\lambda)F(x) + \lambda F_R(x),$$

with a corresponding partial shift of the density.

(b) Convex combinations of the unit shifts can be employed, of the form

$$\xi_\theta(x) = \theta\xi_L(x) + (1-\theta)\xi_R(x); \quad 0<\theta<1.$$

Such combinations give rise to density and distribution functions of the form

$$f_\theta(x) = \xi_\theta(x)f(x) = \theta f_L(x) + (1-\theta)f_R(x); \tag{2.3a}$$

$$F_\theta(x) = \theta F_L(x) + (1-\theta)F_R(x)\ . \tag{2.3b}$$

The most important special case is θ = ½, which will be called the 'centred' shift. In this case,

$$\xi_c(x) = \frac{1}{2}(\xi_L(x) + \xi_R(x)) = -\frac{1}{2}\ln(F(x)(1-F(x)).$$

The resulting density $f_c(x) = \frac{1}{2}(f_L(x) + f_R(x))$ will intersect the natural density *f (x)* at points where $\xi_c(x) = 1$. (It should be noted that this is not the same construct as the centred discrete histogram of Sect. 1.5). In addition to its use in measures of spread and asymmetry, the centred shift is related directly to partition entropy by the censored expectations

$$h(x) = 2F(x)(E[\xi_c(X)|X \le x] - 1) = 2(1-F(x))(1 - E[\xi_c(X)|X > x]).$$

Combinations of type (b) can be thought of as entropic spreaders, as they increase the entropy value relative to the original, resulting in a more platykurtic distribution. Entropic spreaders find a number of applications in the chapters that follow. The central version, in particular, provides a useful vantage point for the entropic measures of asymmetry in Chap. 5.

(c) The probabilities in mixture shifts can alternatively be based on the parent distribution *F(x)*. To accomplish this, the partition entropy function *h(x)* can itself act as a Radon Nikodym derivative via $\xi_h(x) = 2h(x)$. The associated density and distribution functions represent probability weighted combinations of the elementary left and right hand shifts:

$$f_h(x) = \xi_h(x)f(x) = 2[F(x)f_L(x) + (1-F(x))f_R(x)];$$

$$F_h(x) = F(x)F_L(x) + (1-F(x))F_R(x).$$

Mixtures of this general type have the effect of diminishing the entropy values relative to the original *F*, resulting in a more peaked (leptokurtic) distribution; they can be referred to as concentrators. The mixture distribution function shifts F_c and

F_h have the same median as the parent $F(x)$, but $f_c(x_m) < f(x_m) < f_h(x_m)$, so they cross the parent distribution function from different directions.

Figure 2.3a depicts some spreaders and concentrators with a $N(1.05, 0.5)$ density as the original, while Fig. 2.3b depicts the corresponding partition entropy functions. The meaning of spreaders versus concentrators becomes evident, if we take the area beneath the partition entropy function as an indicator of spread. Readers familiar with the theory of stochastic dominance may note that the concentrator F_h is second order stochastic dominant over F, which is in turn dominant over the spreader version F_c.

Mixtures of the preceding types (a), (b), (c) together with further variants can be generically described in terms of a weighting function that can be referred to as a mixing kernel. Given a specified mixing weight $w(\lambda)$ defined over the extended unit interval $-1 \leq \lambda \leq 1$, define a kernel weighted distribution function as

$$F(x) = \int_{-1}^{1} w(\lambda)\tilde{\xi}(\lambda; x)F*(x)d\lambda, \tag{2.4}$$

where the kernel $\tilde{\xi}(\lambda; x)$ is anti-symmetric in the parameter λ:

$$\begin{aligned}\tilde{\xi}(\lambda; x) &= \xi_L(-\lambda; x);\ \lambda \leq 0 \\ &= \xi_R(\lambda; x);\ \lambda > 0\,.\end{aligned}$$

Specifying the range of λ in this way encompasses both left and right entropic shifts. Thus for the linear partial entropic shifts

$$\xi_L(-\lambda; x) = (1+\lambda) - \lambda\xi_L^*(x)\,; \lambda \leq 0, \text{ and}$$

$$\xi_R(\lambda; x) = (1-\lambda) + \lambda\xi_R^*(x)\,; \lambda > 0.$$

Combinations of this kind can be repeated as multiple left or right hand shifts. Section 2.3 interprets distribution mixtures within a framework of this kind. An application to risk management follows in Sect. 2.6.

Finally, multi-step shifting with unit left and right shift at each step can be accomplished via the recursion

$$F_n^L(x) = F_{n-1}^L(x)(1 - \ln(F_{n-1}^L(x))); \ n = 1, 2, \ldots$$

$$f_n^L(x) = -f_{n-1}^L(x)\ln(F_{n-1}^L(x)),$$

with a similar recursion for the right shifts based on expression (2.2b) or the partial shifts.

Sequential shifting can be a useful way to generate new distributional shapes. Thus starting from a symmetric distribution like the logistic, one can generate sequential right shifts that become more and more skewed to the right, with progressively longer right hand tails. However this is not a universal property; the right

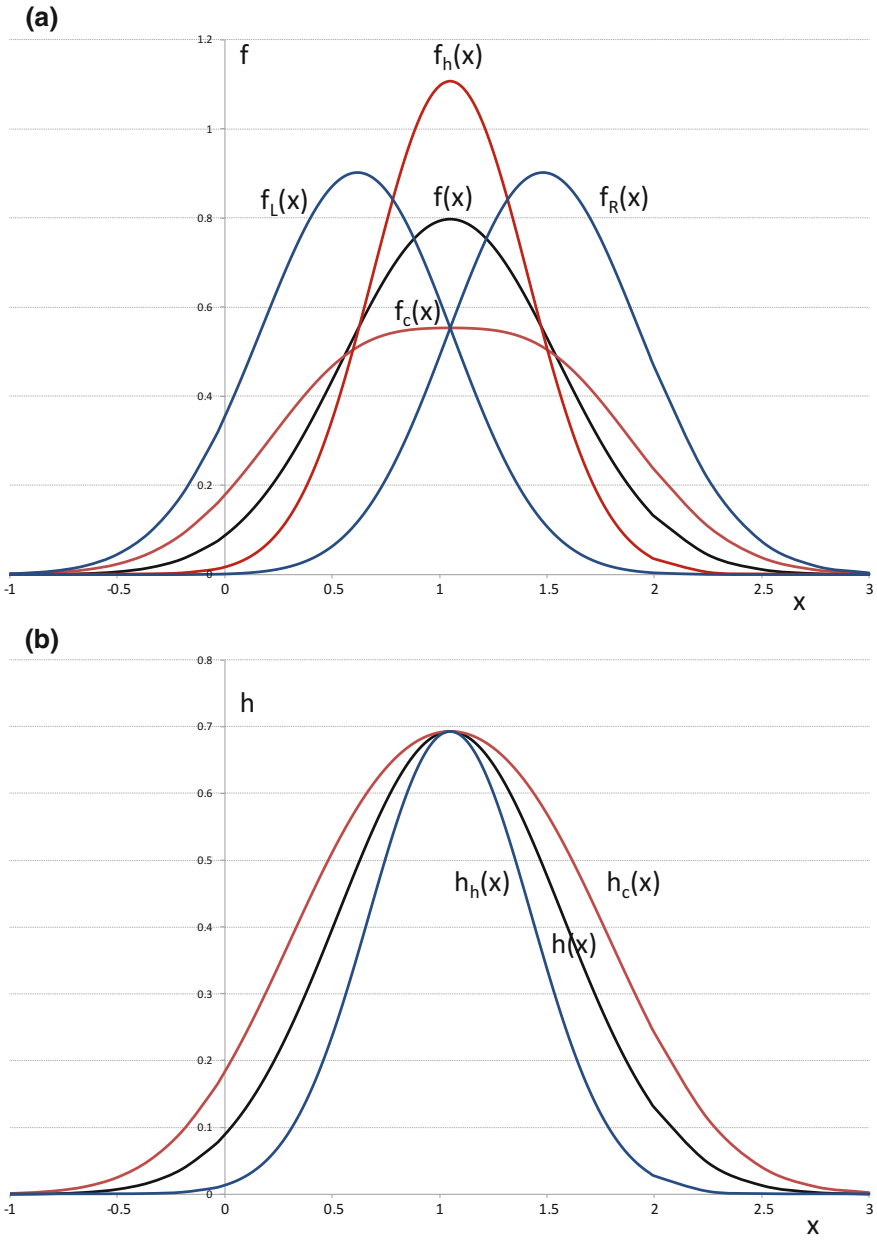


Fig. 2.3 **a** Left and right shifted densities, extended scope, **b** Central and entropic kernel shifted partition entropy functions

shifted normal densities continue to be symmetric, with a linear envelope. In general, non trivial stationary solutions do not exist for the above recursions. Section 2.6 is an extended discussion as an application in financial risk management

2.3.1 Some Extensions

(a) The gamma alternative

The foregoing family does not exhaust the possible specifications for mixing kernels. An alternative formulation is based on the gamma function, with the R-N derivative as

$$\xi_\alpha(x) = \frac{1}{\Gamma(\alpha+1)}(-\ln(F(x))^\alpha;\ \alpha \geq 1.$$

The corresponding density and distribution functions are obtained as

$$f_\alpha(x) = \xi_\alpha(x) f(x); \quad F_\alpha(x) = \frac{1}{\Gamma(\alpha+1)}\Gamma(\xi_L(x); \alpha+1),$$

where $\Gamma(y; \alpha+1)$ is the upper incomplete gamma function. This version removes the limitation that single stage shifts have to be within the unit interval. However, it is computationally more demanding, as well as less intuitive in nature, and will not be used in the applied work of the present study.

(b) Bayesian perspectives

Operational perspectives sometimes entail a process of learning about distributions. An instance arises in finance, where investors have to learn about emergent companies, those newly listed on the stock exchange or are developing technology that has yet to be proved. The opinions of other investors, who may include some perhaps more in the know, are reflected in the current price p_t, which thereby acquires value as an informational signal.

In a more general context, let the current time t information set be of the form $I_t = (I_{t-1}, p_t)$. An underlying mixing model such as expression (2.3a, 2.3b) is thought to be involved, with f_L, f_R as the left and right shifted densities of an underlying kernel specified as $f * (p)$.

Price mediated learning about a mixing parameter such as θ as in expression (2.3a, 2.3b) might then take the form of a Bayesian update process, where $w(\theta|I_t)$ represents the probability density of the unknown parameter θ. Bayesian learning would be generated as

$$w(\theta|I_t) = kf(p|I_{t-1}, \theta)w(\theta|I_{t-1}),$$

where $k = \int_\theta f(p|I_{t-1}, \theta)w(\theta|I_{t-1})d\theta$.

An optimistic scenario would be one where the mixing distribution $w(\theta|I_t)$ is perceived to the shifting towards the right.

2.4 Perspectives on Mixture Distributions

Many empirical distributions appear to differ quite markedly over shorter or longer dated horizons. Financial data are subject to this problem. Over a longer horizon, the longer tails of daily or monthly equity returns, i.e. leptokurtosis, is more manifest than over shorter periods. In turn, the short horizon returns can be skewed one way or the other. In bad times, local distributions may be skewed towards the left (negative skew metrics), indicating a higher probability of bigger losses, while in good times the reverse might be the case. That the longer run might be viewed as a mixture of time-varying shorter run distributions finds explicit expression in econometric models such as Garch-m where the mean and variance are modelled as time dependent.

However modelled, the long versus short run horizon distinction creates problems for financial risk management. Value at Risk (VaR) seeks to limit the probability of losses by selecting portfolios that will lose more than a preassigned capital with probability less than a critical point (commonly 1% for daily returns, 5% for annual). But the probability is assessed using a historical distribution of portfolio returns, raising the issue of just what the horizon should be.

A related prudential measure called Conditional Value at Risk (CVaR) does the same thing, estimating expected capital loss given that the VaR critical point has been violated. But in practice, practitioner back testing of risk management models such as VaR frequently fails, in the sense that the risk limits as determined with reference to the distribution over the past year are violated *ex post* more often than budgeted.

The underlying problem can be regarded as akin to a mixture distribution scenario. The resulting long horizon unconditional could be regarded as a mixture distribution of successive short run distributions. A prior distinction to be resolved is that between parametric mixtures, where the constituent distribution forms are the same but the parameters change, and environments where the shorter run distributions may be of quite different form. Historical data on share returns, for instance, encompass good times and bad times, as well as more normal or 'business as usual' times. In bad times, local distributions may be skewed towards the left (negative skew metrics), indicating a higher probability of bigger losses, while in good times the reverse might be the case.

To the extent that shorter horizon distributions can exhibit changes in shape as well as in parameters, the mixture problem becomes compounded in its tractability, especially where relatively little is known about the precise data generation process. Risk managers may be better advised to budget for more adverse outcomes than suggested by the historical distribution while not exclusively relying on the

immediate past. This would be of the nature of a minimax rule for potential loss exposure.

The development that follows establishes a basis for decision rules of the latter kind by utilising mixtures of non parametric entropic transformations. Such transformations can encompass shifts to the left or right, spreading, or on the other hand concentration. Thus in bad times, the local distributions could be regarded as derived with reference to more normal stable times but shifted to the left; or to the right in better times. Given an observed historical distribution, it becomes possible to back out an originating distribution (the 'generator'), given a hypothesised mixture kernel.

For purposes of risk management, the mixture kernel itself embodies user priors as to the relative frequency of normal 'business as usual', good, and bad times. It is itself semi parametric in nature, the effective choice of parameters reducing to just two, the good and bad state probabilities. As mixing kernels, both the uniform and its entropic scale independent variant fall within this framework.

More specifically, an observed long run observed distribution F is assumed to be generated in terms of a mixing distribution, or mixing kernel, defined over the extended unit interval and applied to a generator F* that is to be recovered. This can then be utilised in risk modelling to establish different scenarios; notably to a leftward shift of the generator.

The mixing kernel itself can have a dual interpretation. A positivist view is that it represents the user's best judgement as to the relative frequency of different states of the world. Facilitating the estimation problem is a result in this section indicating that the choice can be made in terms of just two parameters. A second possible interpretation is normative in nature: the mixing kernel represents a risk manager's certainty equivalents for the states of nature. The latter interpretation might be expected to overweight the probability of bad states.

A general kernel weighted distribution function can be defined in terms of expression (2.4) of Sect. 2, reproduced here as

$$F(x) = \int_{-1}^{1} w(\lambda)\tilde{\xi}(\lambda; x)F*(x)d\lambda,$$

with the mixing weights as $w(\lambda); 0 < \lambda < 1$. In this context, the R-N shift kernel is given by

$$\begin{aligned}\tilde{\xi}(\lambda; x) &= \xi_L(-\lambda; x);\ \lambda \leq 0\\ &= \xi_R(\lambda; x);\ \lambda > 0\,.\end{aligned}$$

The objective in what follows will be to back out the generator function *F*(x)* from (4), knowing *F(x)* and *w(λ)*. The inversion problem has been referred to as such in Sect. 2.1; the present formulation is more general.

Implementations of the mixing density can be facilitated in their interpretation by defining conditional weight functions $w_L(\lambda), w_R(\lambda)$ such that

$w(\lambda) = w_L(\lambda)\theta_L;\ \lambda \le 0; \quad w(\lambda) = w_L(\lambda)\theta_R;\ \lambda > 0,$ with θ_L, θ_R interpreted as conditional probabilities such that $\theta_L, \theta_R \ge 0; \quad \theta_L + \theta_R = 1$. The respective conditional means are

$$\mu_L = \int_{-1}^{0} \lambda w_L(\lambda) d\lambda \le 0 \ \text{ and } \ \mu_R = \int_{0}^{1} \lambda w_R(\lambda) d\lambda > 0.$$

For convenience, define

$$\phi_L = -\theta_L \mu_L;\ \phi_R = \theta_R \mu_R;\ \phi_0 = 1 - \phi_L - \phi_R.$$

These will be referred to as the 'mixing coefficients'. In this notation, the associated R-N derivative reduces to

$$\xi(x) = \phi_0 + \phi_L \xi_L^*(x) + \phi_R \xi_R^*(x).$$

This is nonnegative (because every term is so) and has unit expectation. The associated distribution function can therefore be reduced to a three way mixture involving the generator F^* and its unit left and right entropic shifts $F,\ F_L^*,\ F_R^*$:

$$F(x) = \phi_0 F_*(x) + \phi_L F_L^*(x) + \phi_L F_R^*(x).$$

If the weighting kernel $w(\lambda)$ is symmetric about zero, then $\theta_L = \theta_R = \frac{1}{2}; \quad \mu_R = -\mu_L = \mu > 0$ so $\phi_L = \phi_R = \frac{\mu}{2}$, and F reduces to

$$F(x) = (1 - \mu) F * (x) + \mu F_c(x),$$

where F_c is the simple centred shift of Sect. 2.2. The centred shift F_c is itself a limiting case $\mu \to 1$ where all the mixing weight shifts to the ends of the range.

More usually, weighting is heavier towards the centre, though it may also be useful for risk management purposes to assume that it is evenly distributed throughout the range. Two weighting kernels that have proved useful in the context of time series smoothing are as follows:

(a) The uniform kernel represents maximum *ex ante* ignorance, equivalently the most informative Shannon entropy:

$$w_{un}(\lambda) = \frac{1}{2};\ W_{un}(\lambda) = \frac{1}{2}(1 + \lambda);\ -1 \le \lambda \le 1. \tag{2.5a}$$

For the uniform kernel, $\mu = \frac{1}{2}$ and the smoothed density reduces to

$$F_{un} = 0.5F * + 0.25F_L^* + 0.25F_R^*. \tag{2.5b}$$

(b) A derived variant that emphasises more the middle ground can be obtained by normalising the locational entropy function of the above uniform distribution to give

$$w_{en}(\lambda) = \ln 2 - \frac{1}{2}[(1+\lambda)\ln(1+\lambda) + (1-\lambda)\ln(1-\lambda)];\ -1 < \lambda < 1; \qquad (2.6a)$$

$$W_{en}(\lambda) = \lambda \ln 2 - \frac{1}{4}[(1+\lambda)^2 \ln(1+\lambda) - (1-\lambda)^2 \ln(1-\lambda) - 2(1+\lambda)]$$

This will be referred to as the entropic uniform density. It is strictly concave throughout its domain; the amount of weight progressively declines towards the edges. In this case, $\mu = \frac{1}{3}(\frac{1}{2} + \ln 2) \approx 0.4$, resulting in

$$F_{en}(x) = 0.6F * (x) + 0.2F_L^*(x) + 0.2F_R^*(x). \qquad (2.6b)$$

Expressions such as (2.5b), (2.6b) could be taken as the long run historical distributions.

In real time, the weight density $w(\lambda)$ and mixing weights φ might themselves shift to give the current time t distribution (e.g. as good or bad times). However, the present concern is not with transition as much as the long run limiting behaviour.

Thus any given time, let $\phi_t' = [\phi_{L,t},\ \phi_{0,t},\ \phi_{R,t}]$ be a vector of the mixing coefficients, resulting in

$$F_t(x) = \phi_{0,t}F_*(x) + \phi_{L,t}F_L^*(x) + \phi_{L,t}F_R^*(x).$$

Over a long period of time T one might have $\frac{1}{T}\sum_{t=1}^{T} \phi_t \to^p \phi$, uniformly in x. The resulting average $\frac{1}{T}F_t(x)$ will converge in distribution to the historical distribution F as in (2.5b, 2.6b). In bad times $\phi_{L,t} > \phi_L$, and the time t density will be shifted to the left relative to the long run. Conversely in good times where $\phi_{R,t} > \phi_R$.

In applications, the object will be to back out the implied generator F^* from a given historical distribution mixture F. Specified shifts of the generator can then be used to examine what might happen over distributions shorter periods, especially those that might encompass more adverse states of the world. The backing out process itself has to be done numerically, given an assumed long run mixing distribution, such as $\phi = 0.5,\ 0.25,\ 0.25$ for the uniform mixture as in (2.5b) or $\phi = 0.6,\ 0.2,\ 0.2$ for the entropic mixture in (2.6b). Section 2.6 provides an illustration.

2.5 The Entropic Kernel: Data Smoothing and End Correction

The partition entropy function $h(x)$ has up to now been exposited as a function that measures the entropic uncertainty attached to each point in the range of a random variable. The entropic values associated with any given point could also be

interpreted as providing information about neighbouring points, to an extent measurable by their respective partition entropy values. If the partition entropy function is sharply peaked, then a point $x = x_i$ to the right of the median will contain little information about points closer to the median. The very low value of $h(x_i)$ resolves little of the indeterminacy associated with ups versus downs, relative to the median, on any subsequent realisation.

Considerations of this kind arise in the smoothing of time series observations. A subject time series is supposed to be generated by a systematic part, denoted by y_t*, with the observed series overlain by random noise ε_t. The systematic part or 'signal' is assumed to be slowly evolving, so that knowing the values of y_t* would be very informative about the next value y_{t+1}*. All that can be observed, however, are the realised values y_t. This will likely be main source of information about the signal y_t*. However Bayes theorem tells us that the next realisation y_{t+1} will also contain information about y_i*, containing as it does information about y_{t+1}*.

It is this insight that forms the basis of kernel smoothing in time series analysis. More structured procedures such as the Kalman filter address the problem as one of filtering out the underlying signal with a recursive model of how the underlying values are supposed to be generated. The Hodrick-Prescott filter is a variant in which the signal is specified as an auto-regression. Both entail recursive recovery of an estimated underlying signal, but with a specific generating structure assumed.

More informal smoothing uses local windows about a given data point. This is commonly executed with kernel functions that attach diminishing weight to observations indexed further from the current data point, taken as the window centre. A related context is non parametric regression, where the index set consists of sequential values of an independent variable, and the object is to estimate the conditional expectation of a dependent variable. The kernel approach is less structured than formal filter procedures. But it may be more robust to a lack of a prior knowledge about the underlying data generation process, or an incorrect specification of the underlying data structure.

A variety of kernel specifications are in use, with profiles often based on common density functions, including the uniform as a kernel representation of a fixed window unweighted moving average. The Epanechnikov kernel, which is based on a quadratic function, is widely cited as an efficiency standard, as it minimises the asymptotic mean integrated square error in the particular case where the data are independently drawn from a common underlying probability distribution. It will be employed as a comparison standard in the development that follows, though it does make rather specific assumptions about the data generation process, with noise generated as independent and identically distributed random variables.

The Epanechnikov kernel is concave in form, while continuous kernels originating from continuous distributions, such as the Gaussian, have mixed concave-convex profiles and points of inflexion. In more structured contexts, optimality properties are commonly established in terms of loss functions adapted to specific data generation structures. In data smoothing, on the other hand, the disturbance properties are commonly unknown, and the objective may simply be to

aid pattern comprehension ('eyeballing'), much as in the lower order details of wavelet analysis.

The general kernel smoothing problem can be set up in operational terms as follows. Denote the given series of discrete time observations by $y_t; t = 1, 2, \ldots, T$. A time t-centred window of length B, where B is an odd number less than T, will be taken to refer to the observations $y_{t+\tau}; \tau \in [-\frac{B-1}{2}, \frac{B-1}{2}]$, interpreted as an integer set. Thus $B = 7$ for daily data would span the current observation (at time t) together with the three days before and the three days following. A span of two weeks (14 days) before and after the current observation would constitute a 29 day window. A fixed bandwidth weighting scheme will be defined as a set of non-negative weights k_τ that are zero for observations $y_{t+\tau}$ outside the given window.

The general discrete time smoothing formula is then of the form

$$\widehat{y}_t = \sum_{\tau=-\infty}^{\infty} k_\tau y_{t+\tau}; t \in \left[\frac{B+1}{2}, T - \frac{B-1}{2}\right], \tag{2.7}$$

with the weights summing to unity: $k_\tau \geq 0, \sum_\tau k_\tau = 1$.

The popular Epanechnikov weights can be defined in the present context as proportional to the function

$$K(\tau) = \frac{3}{4}\left(1 - \left(\frac{\tau^2}{\lambda}\right)\right),\ |\tau| < \lambda;\ k(\tau)\ = 0\ \textit{otherwise},$$

where the bandwidth λ is a parameter chosen by the user. Its basis is therefore a user defined local quadratic with the peak at the current data point y_t.

An alternative entropic approach to smoothing is described in what follows. This derives the window weights k_τ as proportional to the partition entropy of an underlying probability distribution. The latter as starting point can be interpreted as a Bayesian type prior for the degree of influence of neighbouring observation points $\{y_{t+\tau}\}$ on the given point y_t. A convenient starting point is to assume the uniform density as an uninformative prior. Converting to its partition entropy function then captures a desired effect that the edge of the window contains less information than the centre.

Given the above notation conventions, the operational discrete time entropy kernel can be summarised in the following steps.

1. Set a nominal bandwidth (or smoothing window) length B, e.g. B = 7 or 29 days. The current data point is taken as the midpoint of the bandwidth.
2. Construct a uniform distribution over $\tau \in [-\frac{B+1}{2}, \frac{B+1}{2}]$ such that for interior points,

$$f(\tau) = \frac{1}{B+1},\ F(\tau) = \frac{1}{2} + \frac{\tau}{B+1}\ ,$$

with end points $F(-\frac{B+1}{2}) = 0$ and $F(\frac{B+1}{2}) = 1$.

3. For each point strictly inside the assigned bandwidth, compute the partition entropy as

$$h_\tau = -[F \ln F + (1-F)\ln(1-F)]; F = F(\tau).$$

Applied to the uniform distribution this gives

$$h_\tau = -\left[\left(\frac{1}{2} + \frac{\tau}{B+1}\right)\ln\left(\frac{1}{2} + \frac{\tau}{B+1}\right) + \left(\frac{1}{2} - \frac{\tau}{B+1}\right)\ln\left(\frac{1}{2} - \frac{\tau}{B+1}\right)\right].$$

For the notional end points of the bandwidth extension, where the logs are undefined, set $h_\tau = 0$ *if* $\tau = -\frac{B+1}{2}, \frac{B+1}{2}$. There is zero partition entropy at such points.

4. For interior points $\tau \in [-(\frac{B-1}{2}), (\frac{B-1}{2})]$, set $k_\tau = h_\tau / \sum_\tau h_\tau$ to ensure that the kernel weights sum to 1.

The proposed entropic weights for $\tau \in [-\frac{B-1}{2}, \frac{B-1}{2}]$ are defined by

$$k_\tau = h_\tau / \sum\nolimits_\tau h_\tau \text{ , where} \tag{2.8}$$

$$h_\tau = -\left[\left(\frac{1}{2} + \frac{\tau}{B+1}\right)\ln\left(\frac{1}{2} + \frac{\tau}{B+1}\right) + \left(\frac{1}{2} - \frac{\tau}{B+1}\right)\ln\left(\frac{1}{2} - \frac{\tau}{B+1}\right)\right].$$

Utilising the weights (2.8) in expression (2.7) constitutes the entropic moving average.

With regard to the range, special attention must be paid to beginning and end values, to avoid taking the log of zero. The procedure as suggested under point 3 is to notionally extend the window by one unit at each end. For example, suppose the assigned bandwidth is 7 days, meaning the index runs as −3, −2…2, 3. We notionally extend the range to 9 days, with weights of zero assigned to the new end points −4,4.

A scalable kernel is one that adjusts automatically with the desired window width, so that there is no need to change parameters beyond the width decision. The uniform based entropic kernel is a one parameter family, standardised by the chosen bandwidth B. It is therefore scalable, automatically so once the window width is chosen. By way of contrast, the Epanechnikov kernel scales up with its bandwidth parameter λ, but not automatically so with the desired window width, so that a further decision must be made as to the value to set for λ if one changes e.g. from a monthly window to a yearly one. From this point of view, the entropic kernel is more user friendly.

Apart from this, the entropic and Epanechnikov kernels are both strictly concave in form. The entropic version allocates more weight away from the centre, reflecting

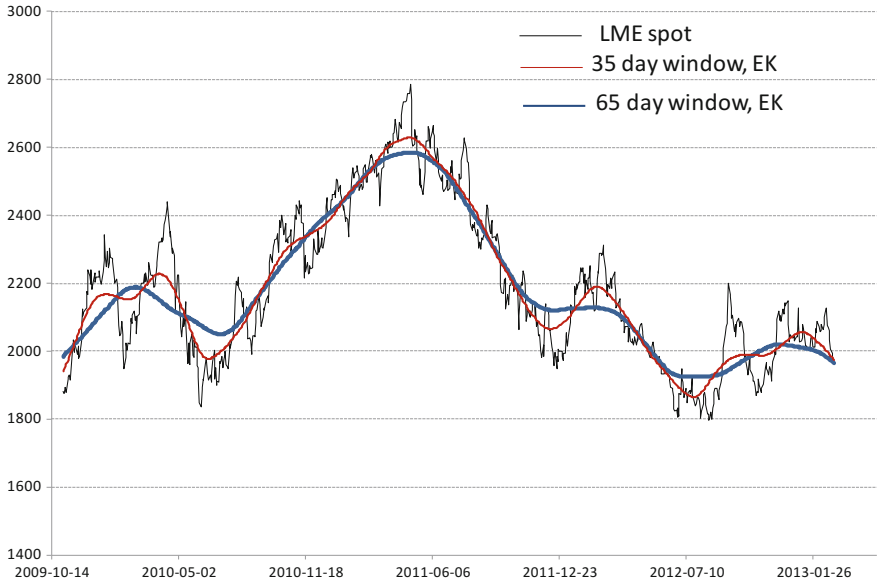


Fig. 2.4 London Aluminium prices over the GFC, original and smoothed with the entropic kernel

the greater degree of uncertainty or agnosticism inherent in the entropic approach. However, the log function implies it has greater curvature towards the edges of the window, which are progressively penalised by more than for the quadratic Epanechnikov kernel.

Figure 2.4 illustrates the entropic kernels in the context of commodity markets, which are often volatile over different time scales. The daily data are of the London Metal Exchange (LME) spot price for aluminium over a period that captures the flight to safety via commodities that occurred at the time of the global financial crisis and the reaction that subsequently set in with economic recovery. The 65 day window shows that the market fall starting in 2011 was more gradual than was the case for world equity prices; it also took hold a year or two later.

2.5.1 End Correction with Kernel Compression

An exigency that affects all time series kernels is the end or edge correction problem. If a 100 day kernel is used, then at current time t the complete 100 day bandwidth cannot be utilised for any observation later than t-*50*. This is a serious limitation in a real time context, where specific interest may lie with what current data are revealing about trends. Climate change data are an instance. Conclusions about underlying global warming should certainly not be based on a current warm year. But on the other hand, a suitably long kernel window cannot be used to draw

even provisional or preliminary conclusions as to future trends that adequately encompass the years of the immediate past. Concerns of this kind are likewise implicit in moving average based trading rules in financial markets.

A variety of existing techniques are used for edge (or end) correction. Some assume a local data generation process. Others employ data reflection, assuming the unavailable future observations that would be required to complete the current window are the same as those of recent real time. Alternatively they might utilise jackknifing, which amounts to a randomised version of the same thing. In practice, a common recourse is simply to progressively shorten the length of the window. The problem with this is that it discards earlier observations that might provide information, and would have otherwise have been included.

Figure 2.5a illustrates, and will help in setting notational conventions. Real time denoted as t is measured from right to left, with $t = T$ as the last available observation. Kernel time τ is measured from left to right. The smoothing window is 9 periods, which means that the last complete window is centred at real time T − 4. But if the focus is now to find a smoothed value for time T − 2, one that correctly emphasises T − 2 as the centre, we would need observations at T + 1, T + 2, which are unavailable. In Fig. 2.5b the kernel has been entropically shifted to the left. It cannot completely compensate for the missing observations, but it does place more emphasis on the new centre at time T − 2.

The process can be made systematic by allowing sequential partial shifts of the kernel $w(\tau)$ treating the latter as a density with distribution function $W(\tau)$. At each stage k, the left shift can be weighted according to a parameter $0 < \mu_k \leq 1$. The corresponding kernel weight function is given by

$$w_k(\tau) = \xi_k(\tau) w_{k-1}(\tau); \quad W_k(\tau) = W_{k-1}(\tau)(1 + \xi_k(\tau)),$$

with the partial shift parameters in the sequence

$$\xi_k(\tau) = (1 - \mu_k) + \mu_k \xi(\tau); \quad \xi(\tau) = -\ln(W_{k-1}(\tau).$$

In turn, this will require a criterion for choosing the shift parameter μ at each stage k. In the work that follows, μ_k is chosen so that the expected value of the kernel weights, regarded as a distributed lag, corresponds to the current smoothing point. Thus if the current focus is at the real time point $T - m + k$, we require the sequences μ_k to be set such that

$$\sum_{\tau=0}^{2m} \tau w^k(\tau) = m - k.$$

Figure 2.6 illustrates with a kernel window of 131 days centred at the midpoint m = 65. The sequence of shifts is illustrated up to the first 10 leftward shifts. As the focus approaches the terminal time T, the shifted kernels gradually change to eventually resemble exponential distributed lags, but at a more moderate pace than is the case for full unit shifts at each stage. However the effective width, measured as Shannon entropy of the kernels, does narrow rapidly after a certain point; thus for

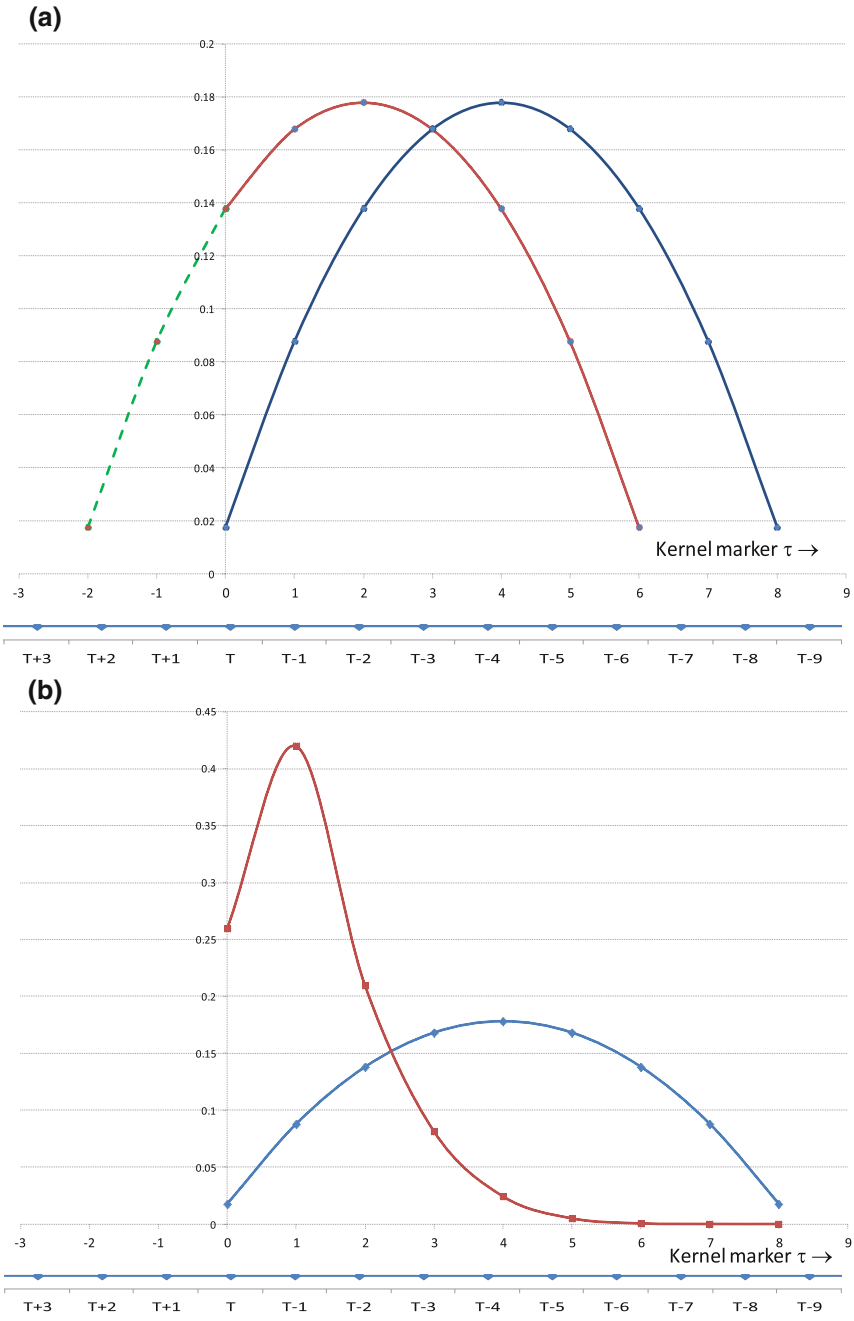


Fig. 2.5 **a** Incomplete smoothing window, **b** Smoothing window completed via entropic shifting

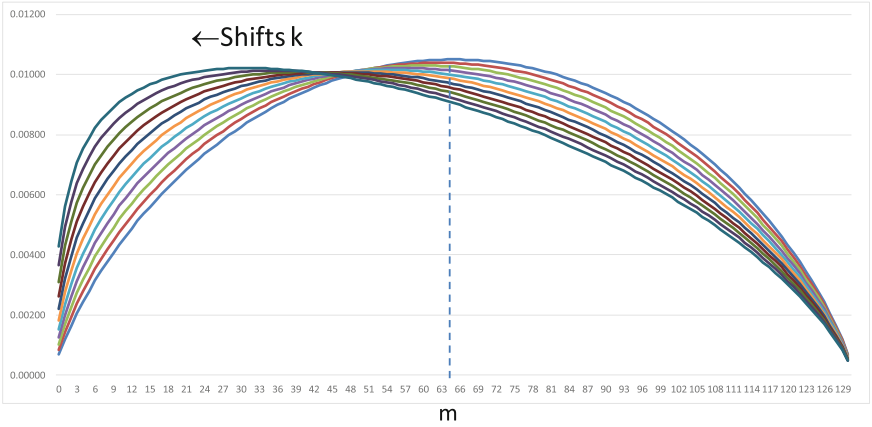


Fig. 2.6 Progressive entropic completion for end correction

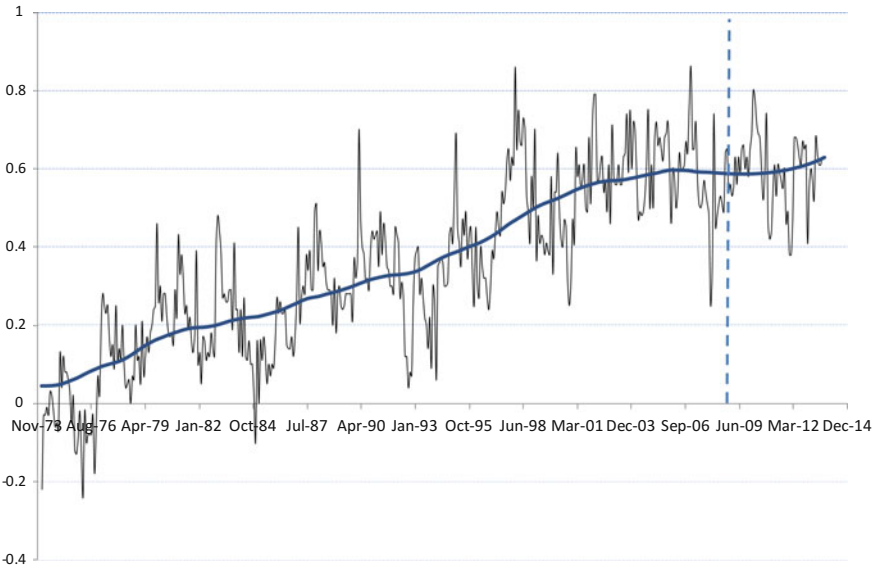


Fig. 2.7 Application to climate change data Source: Blaesche et al. (2014)

a kernel 50 periods, it narrows to just 20% of its base width after 40 compressions. A more structured metric for window width in the form of the effective entropic spread is considered in Chap. 5. Excel- VBA code to accomplish the complete sequence of shifts via VBA functions is reproduced in Appendix 1.

Figure 2.7 is an application to global temperature data, namely the GHCN monthly average sea and land surface temperature anomalies, prepared by the US National Climatic Data Centre/National Oceanic and Atmospheric Administration.

These are deviations from the long term average global temperature; consolidation of northern and southern hemispheres helps to dampen purely seasonal effects. The raw data is monthly, January 1880–October 2013. The choice m = 60 for the entropic smoothing kernel corresponds to the commonly used 10 year moving average, with entropic bandwidth of 64. The last complete window is then centred at October 2008, marked in as the hatched vertical line in Fig. 2.7.

Beyond that date, end correction methods become necessary. The last available observation was at October 2013 and with a window half width $m = 60$ the last complete window is centred at October 08. At the time it was claimed by some observers that there had been a pause in global warming, reflected in the apparent levelling out of the smoothed series after 2005. However, when the kernel centres are progressively extended with kernel compression after that date, the indications from the figure suggested that warming might have been proceeding anew. As of January 2012 the effective bandwidth is less than 75% relative to the complete window, diminished to 50% by December 2012. Thus a tentative finding that warming had resumed from 2012 has to be tempered with the reliability factor of less than 75%. In the event, the resumption of warming since 2014 has since proved the earlier indications to be correct. The 'pause' was only temporary.

2.6 The Social Dynamics of Opinion

It has long been a feature of social and political life that attitudes, as reflected in opinion polls and electoral outcomes, can change rapidly within a relatively short period of time. A striking recent example concerns U.S. attitudes to gay marriage, where over the space of just two to three years from 2010, a clear majority against was transformed in successive Gallup and similar polls to an almost equally clear majority in favour (Fig. 2.8).

Primary causal influences were more or less in common with earlier instances, even if more pronounced in context. They can include public exhortations by influential figures such as movie stars and politicians, together with implicit or explicit media advocacy. Or there can be a sympathetic reaction to the plight of friends or public figures whose personal circumstances have forced a change in views. Influences of this kind may be characterised as external or autonomous inputs. Augmenting such inputs are social feedbacks, where knowledge of the attitudes of peer groups influence the way that individuals will respond or vote. Bandwagon effects are commonly cited in this respect—individuals find comfort in the crowd.

In practice, attention is commonly focussed on a scalar outcome of public or policy interest, such as the proportion in favour. However even if the ultimate outcome is a scalar such as the magical 50%, causal mechanisms cannot be modelled as a more or less simple dynamics of scalar metrics. To understand phenomena such as political polls requires attention to the entire probability distribution of attitudes and how this might evolve over time. In turn, modelling

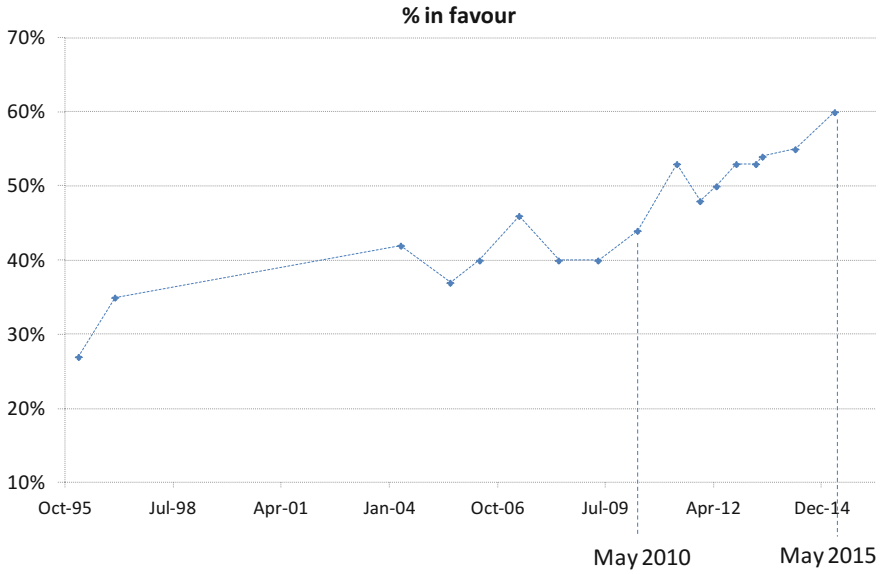


Fig. 2.8 Gallup poll outcomes, U.S. attitudes to gay marriage

progressive poll results as distribution shifting should entail both autonomous inputs (e.g. news items or celebrity endorsements) and reflexive impacts such as feedback from published poll results. The dynamics of partition entropy can be utilised to shed light on the way that opinion distributions as a whole can change over time.

It will be convenient to adopt a model of liberal versus conservative attitudes, as a single dimension of variation. The scale adopted is the unit interval [0,1] ranked from conservative to liberal, with the conservative end against, and liberals for, the given policy measure. This has the convenience of correspondence to a voting scale, with $x = x_c$ taken as a critical point for some proposed measure to succeed. A common choice would be 50%, i.e. $x_c = 0.5$, but in other contexts, such as effective political or financial control, it may be as little as 30%.

A first point concerns the importance of distributional spread. Figure 2.9a depicts a high spread versus a low spread density $f(x)$, both modelled as beta distributions with the indicated parameters. The median is the same in both cases, but at 38.6% is not sufficient to pass the proposed measure, which in this example will require a simple majority: $x_c = 50\%$. Figure 2.9b utilises the corresponding distribution functions, depicted as the proportion in favour $1 - F$. The two distributions (high spread versus low spread) have had the same shift factor applied. But while this is now sufficient for the proposed measure to pass (point A), it is not the case for the low spread distribution (point B).

Over time and with the flow of information, the shifting of attitudes and their distribution is progressive. At each stage, the new distribution is obtained as an

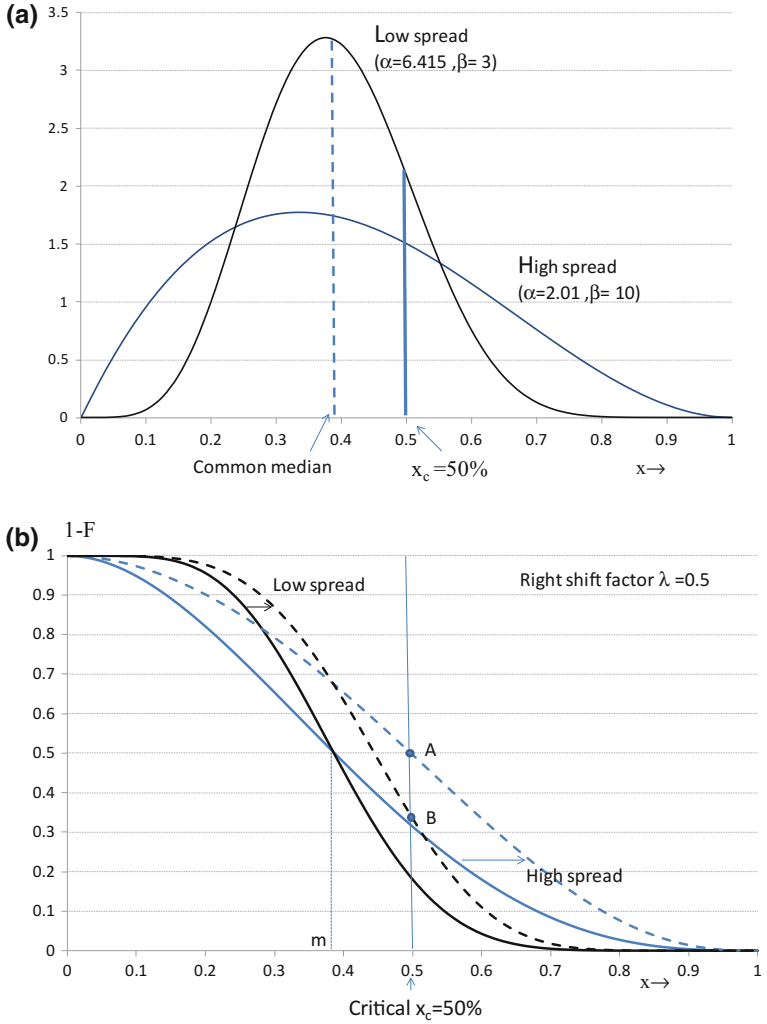


Fig. 2.9 **a** High versus low spread densities, **b** Importance of spread in response to entropic shifts

endogenous transformation of the previous. The drivers for the relevant shift factors can include poll data derived from the distribution as it currently stands. Two features motivate the use of entropic shifting in the dynamics of public opinion:

(a) With a finite attitudinal scale (here zero to unity), the shifts have maximum impact towards the original median. Thus the middle ground can be more easily swayed one way or the other, but hard core opinions at the extreme ends are little affected.

(b) Entropic shifts have more impact where the parent distribution has a higher spread. Intuitively a lower spread indicates that opinions are more compact, and in this sense more decidedly held overall. By way of contrast, higher spread might imply a greater number of people who would at the outset be prepared to favour the proposed measure, with a more substantial core of activism.

2.6.1 The Momentum of Opinion

The illustrative context will be one of repeated right hand shifting of opinions, modelled as a sequence of right entropic shifts as in Sect. 2.2. Repeated rightward shifting, for any given value x, generates a sequence defined in discrete time as

$$F_t(x) = (1 - \lambda_t)F_{t-1}(x) + \lambda_t F_{R,t-1}(x);\ t = 1, 2.3.\ldots. \tag{2.9}$$

Here $F_{R,t-1}(x)$ is the unit right entropic shift of $F_{t-1}(x)$, and $\lambda_t : 0 < \lambda_t \leq 1$ is a specified process defined on the unit interval.

Consistent with the context, it will be convenient to define the distribution complement as $\tilde{F}(x) = 1 - F(x)$. Thus at $x = x_c$, $\tilde{F}(x_c)$ is the proportion that would vote in favour of the proposed measure and a primary objective is to study how this develops over time.

With this convention, Eq. (2.9) simplifies to the following difference equation:

$$\Delta\tilde{F}_t(x) = -\lambda_t\tilde{F}_{t-1}(x)\ln\tilde{F}_{t-1}(x)\ \text{(discrete time)};\ \text{or}$$

$$\frac{d}{dt}\ln\tilde{F}_t(x) = -\lambda(t)\ln\tilde{F}_t(x)\ \text{(continuous time)}.$$

Thus the distribution function at any given point x evolves according to a logarithmic diffusion process, depending on the specification of the shift factor λ_t.

In a homogenous or non reflexive process, λ_t is an exogenous function of time. Influences of this sort might be advocacy pronouncements from influential entertainment or political personalities, media advocacy, or court rulings. In continuous time,

$$\ln(\tilde{F}_t(x)) = \ln(\tilde{F}_0(x))e^{-\int_0^t \lambda(s)ds},$$

with the discrete time approximation

$$\ln(\tilde{F}_t(x)) \approx \ln(\tilde{F}_0(x))\Pi_{s=1}^t(1 - \lambda_s).$$

In either case, the log vertical displacement ratio $\ln(\tilde{F}_t(x)) : \ln(\tilde{F}_0(x))$ is the same for all evaluation points x, motivating the description of case (a) as a

homogenous shift. As an extension, the shift factor λ_t could itself be stochastic, generated as a non-negative diffusion process.

In a reflexive process, the shift parameter λ_t is also a function of $\tilde{F}_{t-1}(x)$. This occurs where individuals are influenced by the opinions of others, in the form of published poll results. A leading instance would be where $\tilde{F}_{t-1}(x_c)$ has moved closer to the critical point for the measure to be passed, e.g. as the difference $(0.5 - F_{t-1}(x_c))_+$. If people think that others are in favour, this will likely induce personal attitude shifts in conformity—the comfort of the crowd. There may also be acceleration effects from changes in this term, a momentum effect. The bandwagon effect is often cited in this context. An illustrative specification with both influences is:

$$\begin{aligned} z_t =& [\pi_t(1 - \gamma_1(F_{t-1}(x_c) - 0.5)_+) + \gamma_2(\tilde{F}_{t-1}(x_c) - \tilde{F}_{t-2}(x_c)), 0)]_+; \\ & \gamma_1 > 0,\ \gamma_2 > 0, \end{aligned} \tag{2.10}$$

followed by $\lambda_t = 1 - \mathrm{e}^{-\theta z_t}$.

In expression (2.10), the exogenous factor (media, celebrity advocacy etc.) is represented by the factor π_t, on a scale of zero to unity. The remaining terms incorporate reflexive elements. The first term incorporates a negative interaction: $\frac{\partial \lambda_t}{\partial \gamma_1} < 0$. If the current proportion in favour ($\tilde{F}_t = 1 - F_t$) is well away from a majority, the media advocacy has less impact. In specification (2.10) the effect cuts out once a majority is reached, though this is not a necessary feature.

The second term $\tilde{F}_{t-1}(x_c) - \tilde{F}_{t-2}(x_c)$ with $\frac{\partial \lambda_t}{\partial \gamma_2} > 0$ incorporates the momentum or bandwagon effect. If the polls show increasing support, this sways existing fence sitters in favour.

For the full reflexive case closed form solutions do not in general exist, requiring an iterative solution. Figure 2.10 depicts the output from such a sequence. As illustrated, the autonomous element π_t is assumed to taper off quite early, with the reflexive drivers taking over thereafter. The base reflexive element is the proportion in favour, incorporated in expression (2.10) as the coefficient γ_1. However, even where a majority is reached and this cuts out, the bandwagon effect (γ_2) continues to operate, driving the outcome beyond the bare majority. A full response to the autonomous input, represented in Fig. 2.10 as a distributed impulse, takes time to work through the reflexive feedback loops.

The role of opinion diversity is a common thread that runs through the preceding discussion. Higher spread indicates that society as a whole is not acting like a collective. A very low spread may mean that opinions held more in common reflect a more coherent prevailing social mores, with less for the feedback dynamics to build on. As Fig. 2.9a, b suggest, the effect is felt not just at the critical margin; a higher spread distribution generates uniformly higher opinion shifts across the entire spectrum of attitudes. Figure 2.10 illustrates an outcome. By way of contrast, a corresponding dynamics for the low spread distribution falls short of ever achieving an ultimate majority—reflexive feedback has insufficient leverage. In

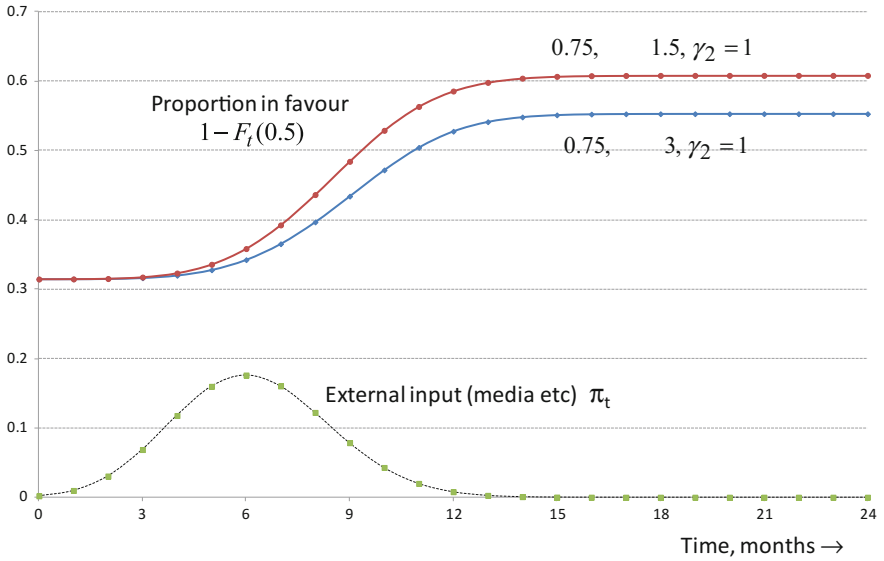


Fig. 2.10 Simulation outcome

predicting ultimate outcomes it might therefore assist to have opinion polls supplemented with a range of attitude boxes, ranging from 'strongly against' to 'strongly in favour'.

In the foregoing analysis each individual is assumed to occupy a single point along the attitudinal scale. An alternative is to imagine that respondents may themselves feel subjective uncertainty about just where they stand as to the strength of their opinions. To the extent that individuals have different subjective profiles, the aggregate then becomes a mixed distribution problem. The framework of Sects. 2.2, 2.3 continues to apply. However, there is likely to be even more scope for the shifting of the middle ground in response to feedback effects.

Over the years, it would seem that social information has been of increasing importance in generating feedback effects. Formal polls are just one potential source. The rise of online social media is another, where participants can readily discover the opinions of their peers and be influenced by them. To be sure, sites of this kind are usually age and culture specific. The virtue of a formal opinion poll, one claiming to be scientific, is that it samples a more complete spectrum of ages and social class. But there may be interactions between the two. Thus younger people, traditionally more influenced by the opinions of their peers, might find it easier to gauge what their own feedback response to a poll should be, by logging on to communal discussion among their peers. If conjectures of this kind prove to be correct, it presages even faster and more pronounced feedback shifts in public opinion.

2.7 Application of Mixture Theory to Value at Risk

Value at risk is a very basic form of risk metric used extensively in contexts such as bank management and insurance. The concept itself is based on nothing more than a one tailed lower critical point, with a pre-set allowable probability, commonly 1% over a daily interval or 5% over a month. The idea is to make sure that the institution's portfolio of risky prospects does not lose capital with a downside probability that exceeds this prudential limit. As such, value at risk is one of the underpinnings of the widely accepted (or effectively imitated) Basle prudential regime for bank management.

In its origins, value at risk (VaR) and related prudential risk metrics were always supposed to refer to the portfolio risk in normal (usual) states of the world. From the prudential of view, however, it is arguably the exposure to more stressful times that should be the focus of attention, for such a time could well be within the frame of the forthcoming accounting period. On the other hand, there could arise times such as the global financial crisis, at which no reasonable portfolio allocation short of cash could cope. Thus VaR and related risk metrics can never encompass all states of the world. All that is required of a risk manager should be to make better provision for more adverse short run outcomes than what can be computed from the overall historical record.

As a statistical regularity, VaR is not without its own empirical problems. An apparent failure of historical back testing for moving sample frames suggests that prudentials such as value at risk could better be informed by regarding the historical record as a mixture of shorter dated frequency functions. The development that follows utilises the mixture framework of Sect. 2.3 to construct more flexible decision rules as to the reference distribution for value at risk. Once a mixture weighting pattern has been decided on, the generator distribution can be backed out by means of the inversion process as in Sect. 2.3. The generator distribution then offers the manager a menu of choice as to just where to set the VaR limit: over the entire history, or with respect to the left shifted extreme as a suitably pessimistic outcome for the coming period, even if not the absolute worst case. The discussion that follows elaborates.

The portfolio is taken for illustrative purposes as the regimen of the S&P500 US equity index. The raw data takes the form of daily returns spanning May 1994–May 2014, with 5218 observations. The value at risk period is taken as the one day exposure, a common choice for dealing rooms. The frequency histogram over the entire history is fat tailed, with a slight suggestion of a negative skew ($\gamma_3 = -0.056$, or see below for entropic asymmetry). The median is at 0.00041. The historical 5% VaR point over the complete history of one day returns is −1.83%.

With respect to the assumed mixing kernel, a very risk averse manager might assume a uniform mixture as in (2.5a) giving more weight to the bad or good short run elements. A somewhat less risk averse variant is the entropic uniform mixture (2.6a). In the event, the inversion outcomes from the two alternatives are very similar.

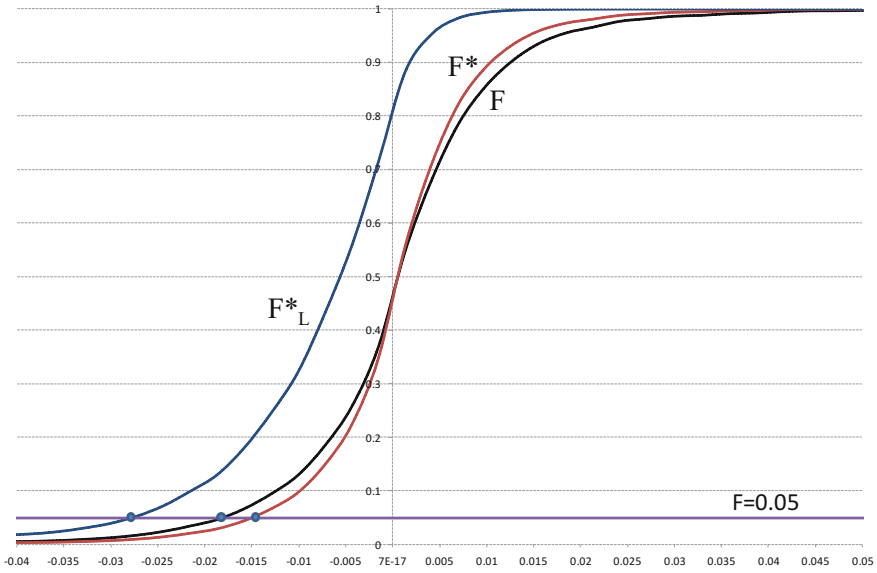


Fig. 2.11 Normal and abnormal times for the S&P500: entropic uniform mixture

The resulting generator F^* corresponding to the entropic uniform mixture is depicted in Fig. 2.11, along with the historical distribution F. The unit left shifted generator is included in the same diagram.

The generator F^* can be regarded as a benchmark distribution for normal times. Its left shift can be regarded as an adverse state that inherits a reasonable probability of occurring without falling into the complete disaster zone. Setting prudential limits with respect to this distribution would yield margin of safety over the distribution for more usual times, and indeed with respect even to the entire history.

Turning to VaR and risk management, the historical lower 5% point is at $F = -0.0183$; that for the generator is $F^* = -0.0150$; and for the unit left hand generator shift the lower 5% point is at $F_L^* = -0.0277$. A risk averse manager who wants to play safe over shorter horizons might consider an existing portfolio that would otherwise satisfy the historical 5% point (−0.0183) as though it should be exposed to an adverse shorter run bad state, as indicated by the 5% point for F_L^*.

The expected shortfall is the conditional expected loss given that the VaR point has been is triggered. At their respective lower 5% points the expected losses from then on are 0.0024 for the historical F; 0.0011 for the generator F^*; and 0.0061 for its unit left shift F_L^*. The rankings are consistent with the mixing construct, namely that the observed long(ish) tail of the historical distribution is the result of mixing together 'business as usual' with distributions associated with periods of unusually good or bad states, the latter embodied by the respective left and right shifts of the generator. Even if the lower 5% point for F_L^* was acceptable, management could

rule that the expected shortfall of 0.0061 was not, and call for a more conservative portfolio.

2.8 Literature Notes

With reference to Sect. 2.1, and the use of R-N derivatives, general treatments of measure changes can be found in Shilov and Gurevich (1977) and Billingsley (1986). Later chapters in the present work utilises further motivations and contexts for the use of R-N derivatives and change of measure.

The left and right entropic shifts, together with basic variants and proofs of properties, were developed in Bowden (2012).

The corresponding distribution functions have recently found application in survival analysis, where the concern is with the residual uncertainty of lifetime, given that age x has been reached. If $\bar{F}(x) = 1 - F(x)$ is the survival function from age x, an overall measure of survival uncertainty was proposed by Rao et al. (2004) as cumulative residual entropy, defined as $CRE = -\int_0^\infty \bar{F}(x) \ln(\bar{F}(x))dx$. In present notation this becomes $\int_0^\infty (F(x) - F_R(x))dx$. For extensions see e.g. Di Crescendo and Toomaj (2015), Kapodistria and Psarrakos (2012) and Tahmasebi et al. (2017). Sections 4.3, 5.5 of the present work also have an actuarial context, though not based on the CRE.

The Epanechnikov kernel utilised in Sect. 2.3 was developed in a multivariate context in Epanechnikov (1969), which is a translation of an earlier Russian paper. For its optimality conditions see e.g. Wand and Jones (1995). Different applications to estimate the conditional expectation are discussed and utilised in a regression context, in versions such as the kernels of Nadaraya-Watson (Nadaraya 1964; Watson 1964), Priestley-Chao (1972) and Gasser-Muller (Gasser and Muller 1984). The entropic kernel of Sect. 2.3 is based upon Bowden (2013).

Turning to end or edge correction, existing techniques might assume a local data generation process of a given class e.g. local linearity Fan and Gijbels (1984); utilise jacknifing (Rice 1984); or employ data reflection (Boneva et al. 1971; Schuster 1985; Silverman 1986; Ghosh and Huang 1992). Jones (1993) is a unified account of such edge correction methods with further variants, while (Wand and Jones 1995) provide a general survey in the density estimation context. The data generation assumptions underpinning such techniques are arguably less appropriate for financial markets, where the preoccupation is more of the nature of filtering than smoothing, with an emphasis on real time.

The partial shift methodology based on the entropic kernel was developed in Blaesche et al. (2014).

For different aspects of the dynamics of social opinion and the importance of feedback in Sect. 2.4, see e.g. Asch (1955), Bowden (1987, 1988), Bikhchandani et al. (1992), Nadean et al. (1993), and Brewer (2014a,b).

Value at risk is covered by most textbooks in financial institutions management. Examples are Jorion (2003) and Saunders and Cornett (2006).

References

Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193,* 31–35.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100,* 992–1026.

Billingsley, P. (1986) *Probability and measure* (3rd ed.). Hoboken N.Y.: Wiley.

Blaesche, T., Bowden R. J., & Posch, P. N. (2014). Data smoothing and end correction using entropic kernel compression. *Stat: International Statistics Association,* 3, 250–257.

Boneva, L. I., Kendall, D. G., & Stefanov, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data analyst. *Journal of the Royal Statistical Society Series B, 33,* 1–70.

Bowden, R. J. (1987). Repeated sampling in the presence of publication effects. *Journal of the American Statistical Association*, 82, with commentaries by P.C. Ordeshook and L.S. Cahoon, 476-484/91.

Bowden, R. J. (1988). *Statistical games and human affairs.* New York: Cambridge University Press; electronic ed. 2009.

Bowden, R. J. (2012). Information, measure shifts and distribution metrics. *Statistics, A Journal of Theoretical and Applied Statistics, 46,* 249–262.

Bowden, R. J. (2013). Entropic kernels for data smoothing. *Statistics & Probability Letters, 83,* 916–922.

Brewer, P. R. (2014a). Public opinion about gay rights and gay marriage. *International Journal of Public Opinion Research, 26,* 279–282.

Brewer, P. R. (ed.) (2014b). Special issue: Public opinion on gay rights and marriage. *International Journal of Public Opinion Research*, *26*(3), 283–300.

Di Crescenzo, A., & Toomaj, A. (2015). Extension of the past lifetime and its connection to the cumulative residual entropy. *Journal of Applied Probability, 52,* 1156–1174.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theory of Probability and Applications, 14,* 153–158.

Gasser, T., & Muller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics, 11,* 171–185.

Ghosh, B. K., & Huang, W.-M. (1992). Optimum bandwidths and kernels for estimating certain discontinuous densities. *Annals of the Institute of Statistical Mathematics, 44,* 563–577.

Jones, M. C. (1993). Simple boundary correction for kernel density estimations. *Statistics & Computing, 3,* 135–146.

Jorion, P. (2003). *Financial risk manager handbook* (2nd ed.). New York: Wiley.

Kapodistria, S., & Psarrakos, G. (2012). Some extensions of the residual lifetime and its connection to the cumualtive residual entropy. *Probability in the Engineering and Informational Sciences, 26,* 129–146.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Applications, 9,* 141–142.

Priestley, M. B., & Chao, M. T. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society, B34,* 385–392.

Rao, M., Chen, Y., Vemuri, B. C., & Wang, F. (2004). Cumulative residual entropy: A new measure of information. *IEEE Transaction Information Theory*, *50*, 1220–1228.

Rice, J. A. (1984). Boundary modification for kernel regression. *Communications in Statistics: Theory & Methods, 13,* 893–900.

Saunders, A., & Cornett, M. M. (2006). *Financial institutions management: A risk management perspective* (5th ed.). New York: McGraw-Hill.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics: Theory & Methods, 14,* 1123–1136.

Shilov, G. E., & Gurevich, B. L. (1977). *Integral, measure and derivative: A unified approach*. New York: Dover.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Tahmasebi, S., Eskandarzadeh, M., & Ali Akbar, J. (2017). An extension of generalized cumulative residual entropy. *Journal of Statistical Theory and Applications, 16,* 165–177.

Wand, M. P., & Jones, M. C. (1995). Kernel smoothing. *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Watson, G. C. (1964) Smooth regression analysis. *Sankhya A*. 359–372.

Chapter 3
Moments, Measures and Metrics

3.1 Introduction

Metrics of one kind or another, all potentially important in practice, are the subject of the present chapter. While their substance and usefulness can be established with no specific connection to entropy as such, it turns out that they have a derivational connection with the unit left and right entropic shifts. Therefore Sect. 3.1 commences with a brief review of the moments of the respective left and right shifted distributions.

Section 3.2 resumes discussion in the context of the original distribution, initiating the progressive left and right conditional moving average functions for the natural distribution. These are then related, as special cases of a general result, to the unconditional expected values of the entropically shifted distributions.

Sections 3.3 and 3.4 consolidate this work and add further concepts and perspectives. Based on the conditional left and right moving average functions, spread and asymmetry functions are introduced, with their domain as the original range space. Taking expected values, one ends up with new spread and asymmetry metrics for the original distribution, related in a very simple way to the first moments of the left and right entropically shifted distributions. Indeed, a single internal sign change converts the asymmetry metric to that for spread. In this sense they are dual to one another.

The foregoing results turn out to be pivotal for much that follows in this and subsequent chapters. The quest for meaning is a general agenda behind the specifics. Conventional textbook metrics are typically based on polynomial functions such as cubic for asymmetry, and fourth order for kurtosis; with the normal distribution employed as benchmark, so a specific choice of comparator distribution. In the case of asymmetry, there are many functions that are antisymmetric about a designated distribution median. The cubic polynomial, with its switch from downside concavity to upside convexity, might be well motivated for a Friedman-Savage type utility function but not for others. Many alternative distribution metrics

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_3

that have been proposed are specific to their particular context with little, if any, applicability in general.

Entropic complexity does provide a general point of reference for the development that follows in this and other chapters. But in the social science context, the measures that result for asymmetry and spread can also be regarded as representations of how the subjects might view their own relativity to others. It is this internal perspective, as well as the entropic complexity, that provides the required measure of generality to the proposed metrics.

Section 3.5 pauses to consider the relationship between these metrics and Gini's mean absolute difference, historically used in contexts of income distributions, together with further generic metrics.

Section 3.6 resumes the main thread of the discussion on measures. The applied context is the socially fraught debate on what many see as excessive and unnecessary executive pay awards, and how pay packages of this size might arise and become pervasive. The left and right conditional moving averages provide a useful framework to explore the dynamics of executive remuneration. Specific interest attaches to conditions under which the resulting evolution is upwardly unstable.

Section 3.7 explores the relationship of the entropic spread and asymmetry functions to the topic of stochastic dominance, foreshadowing work on a similar comparative theme in later chapter.

Section 3.8 is the literature review.

3.2 Moments for the Entropic Shifts

In terms of the framework for left and right shifting developed in Sect. 2.1, the expected values for the unit left and right unit shifted distribution for a continuous range space (below often denoted simply as *) and density are respectively given by

$$\mu_L = \int_* x f_L(x)dx = -\int_* x f(x)\ln F(x)dx$$

$$\mu_R = \int_* x f_R(x)dx = -\int_* x f(x)\ln(1-F(x))dx$$

Formulas of this kind extend naturally to densities with jumps, and also to a discrete valued range space. It will be assumed in what follows that the range of the left and right shifted distributions remains that of the parent distribution. Convergence of the relevant integrals can also be a problem for exceptionally long tailed densities such as the Cauchy.

It is always the case for a continuous density that $\mu_L < \mu < \mu_R$. There is in general no exact analytical relationship between the left and right means μ_L, μ_R and the parent mean μ. One exception is the uniform distribution. For a uniform distribution

defined over a finite range [0,A], $\mu = \frac{1}{2}A,\ \mu_L = \frac{1}{4}A,\ \mu_R = \frac{3}{4}A$ so that the respective means partition up the interval equally.

Apart from this, it is necessary to evaluate $\mu_L,\ \mu_R$ numerically, although this poses little difficulty in practice. In the case of discrete distributions the centred version $F_c(x)$ should be used as outlined in Sect. 1.5.

If the parent distribution is symmetric then the left and right unit shifts are reflections of each other about the parent mean, so $\mu - \mu_L = \mu_R - \mu$. Scale independence of the parent is preserved in the left and right entropic means. Thus suppose there are location and scale parameters $\alpha,\ \beta$ such that $F(x;\alpha,\beta) = F(\tilde{x};\ 0,\ 1)$ with $\tilde{x} = (x-\alpha)/\beta$. The normal, logistic and Gumbel are examples of scalable distributions. In such cases $\mu = \beta\tilde{\mu} + \alpha$ and likewise $\mu_L = \beta\tilde{\mu}_L + \alpha$, similarly for $\mu_R\,, \tilde{\mu}_R$.

An important property, much used in what follows, is that the difference $\mu_R - \mu_L$ is equal to the area, denoted d, beneath the partition entropy function $h(x)$:

$$\mu_R - \mu_L = \int_* h(x)dx = d. \tag{3.1}$$

The result can be proved in several ways. Perhaps the simplest is to differentiate by parts and us the fact that $h'(x) = f_L(x) - f_R(x)$. A regularity condition is that the product $xh(x) \to 0$ as $x \to \pm\infty$.

The property (3.1) could be taken as indicating that for high spread distributions, for which the partition entropy function is relatively diffuse, with a higher value for the area d, the left and right shifted densities are correspondingly far apart.

Anticipating later development, the measure d will provisionally be referred to as the entropic spread. Thus in the case of a scalable distribution where β is commonly regarded as a dispersion indicator, a change in the integration variable to $\tilde{x} = (x-\alpha)/\beta$ leads to $d = \beta\,\tilde{d}$, so the entropic spread scales up commensurately. An analytical solution for d is available for some distributions. For a uniform distribution over the range $[0, N]$, $d = N/2$.

In other contexts a useful general relationship, which follows from the definitions of $f_L(x), f_R(x)$, is:

$$d = \mathrm{cov}(x, \lambda(x)),$$

where $\lambda(x) = \ln\left(\frac{F(x)}{1-F(x)}\right)$ is the log odds function that $X \le x$ versus $X > x$. Thus for a unit scale logistic distribution, the log odds function is linear, resulting in $d = \pi^2/3 \approx 3.289$, in this case equal to the variance.

Apart from such special cases, a numerical integration of the function $h(x)$ can be used to find the entropic spread d. This can be approached from different computational directions, but in order to ensure consistency with the primary use purposes, it is best to commence by first tabulating the left and right unit shift means $\mu_L,\ \mu_R$, then taking their difference to find d.

3.3 The Double Smoothing Property

The double smoothing property refers to a very useful relativity property that arises indirectly from taking logarithms of the original distribution function in the process of deriving the unit shifted distributions and their respective parameters. The following result gives a general relationship between expected values based on the unit shifts $F_L,\ F_R$ and those based on the original distribution function F.

As before, suppose X is a random variable of interest with realisations $X = x$. Let $g = g(x)$ be a measurable function such that $E_F[g(x)] < \infty$. Also let $\phi_L(x)$ be the conditional expectation function defined by $\phi_L(x) = E[g(X)|X \leq x]$. Similarly, let $\phi_R(x) = E[g(X)|X > x]$, the conditional expectation over values to the right of x. It is then true that

$$E_F[\phi_L(x)] = E_{F_L}[g(x)]; \tag{3.2a}$$

$$E_F[\phi_R(x)] = E_{F_R}[g(x)]. \tag{3.2b}$$

An easy proof is to integrate by parts the respective right hand sides, leading to an integration with respect to $f_L(x),\ f_R(x)$.

Expressions (3.2a, 3.2b) indicate that an expectation with respect to the unit left shifted distribution can be regarded as a second layer of progressive accumulation with respect to the original distribution function $F(x)$. Having averaged the values up to $X = x$, we then take the progressive average of the results.

This can be thought of as a double smoothing process. Starting from the lower bound of the range, the first layer is a progressive average of X values up to a preassigned point x and the second is to average the results over all points x. Or starting from the right hand end of the range, we first smooth the X values down to a given value x, then take the average of all such values.

A caveat is that the function to be smoothed at each stage should not itself depend upon the current market point x, as an additional argument to the dummy integration variables X. In other words, the result does not automatically apply to a function of the form $g(X, x)$.

The double smoothing property finds a number of applications in the present study. As an instance, the expected values of the regime conditional entropies (3.6a, 3.6b) of Sect. 1.2 can be evaluated using $g(x) = \ln f(x)$ to give

$$E[\kappa_d(x)] = 1 - E_L[\ln f(x)]; \quad E[\kappa_u(x)] = 1 - E_R[\ln f(x)].$$

For current purposes, however, the most important special case is $g(x) = x$. Here $\phi(x) = E[X|X \leq x] = \mu_l(x)$, the conditional mean up to point x. The mean of the unit left shifted distribution is then given by

$$\mu_L = E_{F_L}[x] = E_F[\mu_l(x)]. \tag{3.3a}$$

It follows directly that $\mu_L < \mu$.

A complementary development exists for the right hand or upper conditional expectation $\mu_r(x) = E[X|X > x]$. In this case,

$$\mu_R = E_{F_R}[x] = E_F[\mu_r(x)], \tag{3.3b}$$

with $\mu_R > \mu$. The upper case subscript in each case refers to the unconditional mean of the left and right shifts; thus μ_L as defined in (3.3a) should be distinguished from μ_l, which refers to the conditional mean function $\mu_l(x) = E[X|X \leq x]$.

3.4 Asymmetry and Spread Functions

The local smoothing average functions $\mu_l(x) = E[X|X \leq x]$ and $\mu_r(x) = E[X|X > x]$ are the starting point for the development that follows. Their evolution can be expressed in recursive form as

$$\mu_l'(x) = \frac{f(x)}{F(x)}(x - \mu_l(x)); \quad \mu_r'(x) = \frac{f(x)}{1 - F(x)}(\mu_r(x) - x). \tag{3.4}$$

Both are increasing with x: $\mu_l(x)$ is concave in form rising towards the unconditional mean μ, while $\mu_r(x)$ is increasing convex, rising from μ.

The proposed metric for asymmetry or skewness derives from the function

$$v(x) = (\mu_r(x) - x) - (x - \mu_l(x)). \tag{3.5a}$$

An intuitive rationale might run in terms of a penalty function for directional departures from symmetry. Thus if the current point x lies well to the left of the median, the weight of comparative expected values will be greater above than below, so $v(x) > 0$. Conversely, if x is well to the right of the median, $v(x) < 0$. Such aspects are explored more fully in what follows.

Sign apart, the function $v(x)$ shares some general properties with both the cubic penalty function $\propto (x - \mu)^3$ of the Pearson symmetry metric, and the mean absolute deviation $E[|x - \mu|]$. Thus for any value x,

$$F(x)\mu_l(x) + (1 - F(x))\mu_r(x) = \mu.$$

It follows that at the distribution median, $v(x_m) = 2(\mu - x_m)$, so that if the distribution is symmetric, $v(x_m) = 0$. A positively skewed distribution typically has $\mu > x_m$, so

$v(x_m) > 0$. Also

$$v''(x_m) = 2f'(x_m)d(x_m) + 8f^2(x_m)v(x_m),$$

where $d(x) = \mu_r(x) - \mu_r(x)$ is a nonnegative spread function explored below. In particular, if the density $f(x)$ is symmetric, then $v''(x_m) = 0$ and $v(x)$ has a point of inflection at the common mean and median.

At the extremes, the function $v(x)$ in most cases is asymptotically linear, though for very long tails it may rise more steeply, while the scaling constant may differ as between upper and lower regimes. The logistic case is illustrated in Fig. 3.1. Thus the proposed penalty function shares the sigmoid property of the cubic in the vicinity of the mean, but the linear penalty of the mean absolute deviation at the extremes. The sign convention is chosen to ensure that the final metric gives the same convention as to positive and negative skewness as the Pearson metric. Thus purely as a penalty function, $v(x)$ could be regarded as a hybrid between the Pearson and the mean absolute deviation metrics.

A function dual to $v(x)$ that embodies a dispersion or spread dimension can be originated by simply changing the sign in expression (3.5a) to give

$$d(x) = (\mu_r(x) - x) + (x - \mu_l(x)). \tag{3.5b}$$

This could be viewed as an isolation or separation indicator. For any given value of x, it averages the gap to the mean of observations above as $\mu_r(x)$ and those below

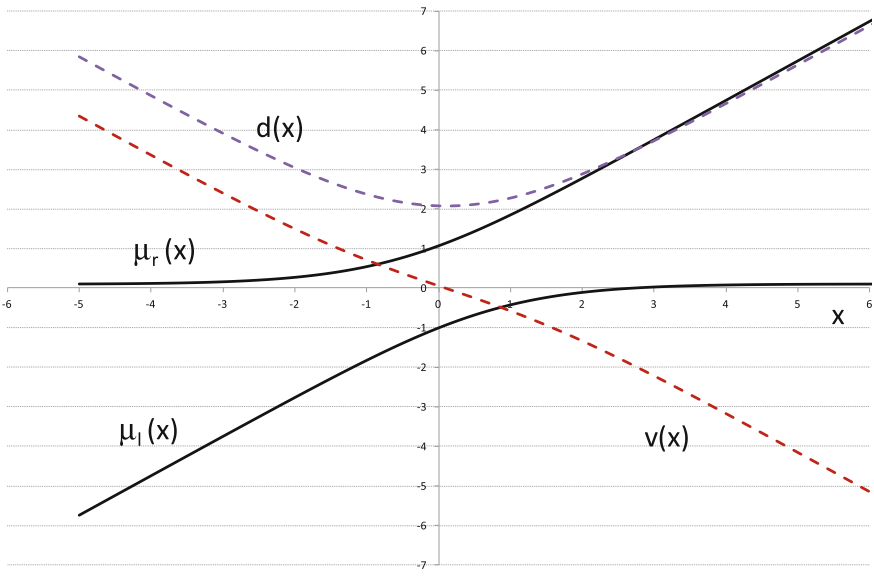


Fig. 3.1 Logistic moving average, asymmetry and spread functions

as $\mu_l(x)$. The dispersion penalty function $d(x)$ is nonnegative and convex ($d''(x) > 0$). Figure 3.1 illustrates.

Making use of expression (3.4) gives

$$d'(x_m) = 4f(x_m)(\mu - x_m).$$

If $\mu = x_m$ the minimum value of $d(x)$ will lie at the common mean and median, with $d(\mu) = 2\mu_r(\mu) > 0$. Otherwise, the minimum will usually lie to the left of the median for positively skewed distributions, and to the right of the median for negative skewness.

Figure 3.1 illustrates for a logistic distribution with the common mean and median at $\mu = 1$, and scale parameter $\beta = 0.75$. For the logistic distribution,

$$\mu_l(x) = x + \frac{\beta}{F(x)}\ln(1 - F(x))$$

so that $\mu_r(x) = \frac{\mu - F(x)\mu_l(x)}{1-F(x)}$. The ratio $\frac{\ln(1-F(x))}{F(x)}$ approaches the value (-1) as $x \to -\infty$. Hence the left conditional expected value $\mu_l(x)$ rises from asymptotic to the line $y = x - \beta$ to the value $\mu_l(x) \to \mu$ as $x \to \infty$. Conversely, the right hand moving average $\mu_r(x)$ starts at the unconditional mean μ and becomes asymptotic to the line $x + \beta$ as $x \to \infty$. The effect of a higher spread parameter β is to distance the approach to the asymptote. Also depicted are the asymmetry and spread generators $v(x),\ d(x)$.

3.5 Summary Metrics for Asymmetry and Spread

Summary metrics for asymmetry and spread can be derived by taking expected values of the respective penalty functions together with expressions (3.3a, 3.3b). This results in two metrics:

$$v = E_F[v(x)] = (\mu_R - \mu) - (\mu - \mu_L) = 2\left[\frac{1}{2}(\mu_L + \mu_R) - \mu\right]; \tag{3.6a}$$

$$d = E_F[d(x)] = (\mu_R - \mu) + (\mu - \mu_L) = \mu_R - \mu_L. \tag{3.6b}$$

Expression (3.6a) in terms of the differences $(\mu_R - \mu)$ versus $(\mu - \mu_L)$ reinforces the idea of asymmetry as differential displacement of the respective left and right shifted distributions. A symmetric distribution will shift equally relative to the common mean and median, resulting in a zero asymmetry metric. A positively skewed distribution will shift more to the right than it does to the left, resulting in $\mu_R - \mu > \mu - \mu_L$ and $v > 0$. Expression (3.6b) for the spread adds the two positive differences.

Supplementary rationales can be derived by using the result that

$$\text{cov}(x, \log F(x)) = \mu - \mu_L;\ \text{cov}(x, \log(1 - F(x)) = \mu - \mu_R.$$

This leads to

$$v = \text{cov}(x, \log F(x)) + \text{cov}(x, \log(1 - F(x)).$$

Similarly,

$$d = \text{cov}(x, \log F(x)) - \text{cov}(x, \log(1 - F(x));\ \text{or}$$

$$d = \text{cov}(x, \lambda(x)),$$

where $\lambda(x) = \log(F(x)/(1 - F(x)))$ is the log odds of $X \leq x$ versus $X > x$.

The measure d always exists for any distribution of finite range. However the relevant area integral may not converge for exceptionally long tailed densities with infinite range; the Cauchy distribution is an instance.

Assuming existence, it is straightforward to show that the metric d satisfies elementary requirements for a dispersion metric. Thus it is invariant with respect to translation ($x \rightarrow y = x + c$); homogenous with respect to scale ($y = cx;\ c > 0$); and unaffected by sign reversal ($y = -x$).

In addition, if distribution B is a median (m) preserving spread of A (so $F_B(x) > F_A(x);\ x < m$ and $F_B(x) < F_A(x);\ x > m$), then $d_B > d_A$. Adopting the stochastic dominance point of view, if the distribution function F_A is second order dominant over F_B, with the same mean, then distribution B has a greater d metric. Section 3.7 enlarges on the stochastic dominance angle.

Similarly, the proposed asymmetry metric v obeys most, if not all, of the standard requirements for a measure of asymmetry. Thus for any symmetric density it has value zero; it is homogenous to linear transformation of x; and if x is replaced by $x + z(x)$ with $z'(x) > 0$, then the resulting v metric remains consistent in sign with the original.

As a final observation, the sampling theory of the metrics v or d is essentially that of the sample mean of the compound random variables $g(x_i) = x_i \ln \hat{F}(x_i);\ i = 1, 2 \ldots, n$ with n as the sample size. In forming successive $g(x_i)$, the empirical distribution of $F(x_i)$ must be formed as the sample proportion less than or equal to the given element x_i. On the other hand, the classic central limit theorems assume that the random variables $g(x_i)$ are independent, which is not going to be the case. There may in addition be boundedness problems arising from the logarithm of F, so that each element of the sample sum cannot be assumed to have a small individual effect on the sample mean. Thus the sampling theory, including limits in probability or the asymptotic distribution, remains to be fully investigated. In purely descriptive contexts, such as income distribution, problems of this nature do not usually arise.

3.6 Gini's Mean Absolute Difference and Welfare Variants

Gini's mean absolute difference (MD) can be defined as the weighted absolute metric distance between two randomly drawn observations from the same distribution function. Given a discrete valued empirical frequency function f (e.g. for incomes, or wealth),

let x_i, y_j denote drawings from the bivariate density defined by $\phi(x_i, y_j) = f(x_i)f(y_j)$. Then the Gini mean absolute difference is conventionally given by

$$MD = \sum_i \sum_j |x_i - y_j| f(x_i) f(y_j).$$

Several alternative formulas for the mean absolute difference exist, for computation or other purposes. An instance is

$$MD = 4Cov(x, F(x)), \tag{3.7}$$

where $F(x)$ is the distribution function corresponding to $f(x)$.

If μ is the arithmetic mean, the value $\frac{1}{2}MD/\mu$ gives the familiar Gini coefficient, widely quoted in income distribution studies as measuring the extent to which the cumulative proportion of total income is matched by the cumulative proportion of people. Chapter 4 is a more extended treatment of such aspects.

The Gini mean difference as such has had a long history of discovery and rediscovery, but even in recent times remains just an alternative method for computing the Gini coefficient in studies of income distribution. Part of the difficulty in its more widespread acceptance lies with the nature of the representation, as the expected value of all pairwise absolute differences. This makes it less than transparent in pattern recognition, as to biases, one way or the other, that might contribute to the interpretation of the resulting value. It also becomes difficult to relate the MD to other measures of spread or asymmetry, or indeed to decide which of the two contributes more to the MD.

The development that follows explores such issues by utilising the upper and lower conditional means at any given value of the subject variable. Cast in such terms, the MD metric belongs to a generalised mean difference family, which includes the measures d and v for distribution spread and asymmetry. In such terms the MD is revealed as having contributions from both spread (generalised dispersion) and asymmetry; so in this sense, it is not a pure measure for spread as such, despite its formal appearance.

It will be convenient for expository purposes to assume a continuous range space. Extension to discrete or histogram data is straightforward. It will also help to adopt a common symbol for the subject variable (income, etc.), with values x, X in place of x,y. The convention will then be that X is drawn first, conditional upon a given value x, followed by variation in x. In such terms,

$$MD = \int_x \int_X |x - X| f(X|x) f(x) dX dx. \tag{3.8a}$$

Now

$$|X - x| = (X - x)_+ - (X - x)_-, \tag{3.8b}$$

where $(X - x)_- = X - x$ if $X < x$, and zero otherwise.

Splitting up the integral (3.8a) conformably with (3.8b) results in

$$MD = E_x\{E_X[(X - x)|X > x)] \times (1 - F(x)) + E_X[(x - X)|X \leq x] \times F(x)\}.$$

This can now be put into a form involving the conditional mean functions $\mu_l(x),\ \mu_r(x)$. Let

$$d_G(x) = (\mu_r(x) - x) \times (1 - F(x)) + (x - \mu_l(x)) \times F(x). \tag{3.9}$$

Then

$$MD\ = E[d_G(x)] = E_x[(\mu_r(x) - x) \times (1 - F(x)) + (x - \mu_l(x)) \times F(x)]$$

provides an expression for MD in progressive mean difference terms, referring to the differential behaviour of the upper and lower conditional means.

In the present context, primary interest attaches to a comparison of MD with the entropic spread. From the definition of $d_G(x)$ it follows that $MD < d$. Substituting expressions for $d(x),\ v(x)$ as in Chap. 3 results in

$$d_G(x) = \frac{1}{2} d(x) - v(x)\left(F(x) - \frac{1}{2}\right).$$

Taking expected values, together with $E[F(x)] = 1/2$, this can be written

$$MD = d_G = \frac{1}{2} d - Cov(v(x), F(x)). \tag{3.10a}$$

Now $v'(x) < 0$ with $v(x) = 0$. It follows that $Cov(v(x), F(x)) < 0$, and $MD > \frac{1}{2} d$.

In summary, $d > MD > d/2$ provides general upper and lower bounds for MD.

If $v(x)$ is linear, then simple proportionality holds as between MD and d. Thus a uniform distribution with $F(x) = \frac{x}{N};\ 0 \leq x \leq N$ has $v(x) = \frac{N}{2} - x$ with $d = \frac{N}{2}$. Making use of expression (3.10a) results in $MD = \frac{2}{3} d = \frac{N}{3}$. The gap between MD and its lower bound of $d/2$ is equivalent to 17% of the mean income.

In general, however, the relationship between MD and the entropic d will depend upon the asymmetry of the distribution. A positively skewed distribution will have a heavier loading in the area of lower incomes, so the steeper changes in $F(x)$ in this region will interact with higher values of the asymmetry metric; the gap

between MD and $\tilde{d} = \frac{d}{2}$ will be larger. Chap. 4 contains a more extended discussion of income distribution as such.

As an extension, comparison of expressions (3.9) and (3.5b) shows that the functions underlying the MD and $\tilde{d} = d/2$ metrics are of the generic form

$$d_w(x) = (\mu_r(x) - x)(1 - w(x)) + (x - \mu_l(x))w(x);\ 0 \leq w(x) \leq 1,$$

with $w(x) = F(x)$ for the MD and $w(x) = \frac{1}{2}$ for the entropic spread $\tilde{d}$. So in each case $E[w(x)] = \frac{1}{2}$.

Equivalently,

$$d_w(x) = \frac{1}{2}d(x) - v(x)\left(w(x) - \frac{1}{2}\right).$$

Taking expected values, this result is consistent with Eq. (3.10a), with $F(x)$ replaced by the weighting function $w(x)$:

$$d_w = \frac{1}{2}d - Cov(v(x), w(x)). \tag{3.10b}$$

However, the covariance involved can now be either positive or negative, depending on the welfare weights assigned to lower versus higher values of the envy function $v(x)$.

In this respect, the general form can be adapted to context, perhaps with $w(x)$ interpreted as a penalty function. In an investment context, the risk free rate ρ could be viewed as a benchmark for funds management. If the realised return is x, a client would view such an outcome as disadvantageous where the expected exceedance $\mu_r(x) - x > 0$. This might especially be the case where $F(\rho) = prob(x < \rho)$ is appreciable. For in that event, the client experiences a double dose of regret: the expected exceedance is high, and in any case the client could have invested in the risk free rate. A relevant performance metric might therefore be of the form

$$d_\rho(x) = (x - \mu_l(x))(1 - F(\rho)) - (\mu_r(x) - x)F(\rho).$$

This has expected value $d_\rho = (\mu - \mu_L) - (\mu_R - \mu_L)F(\rho)$, with the second term representing the effective penalty of being outperformed when the risk free rate is positioned higher relative to the general distribution of returns $F(x)$.

3.7 The Economic Dynamics of Remuneration Relativities

Remuneration relativities between parties that are judged to be comparable in some way are pervasive in the economic environment. In recent years, the particular context of executive remuneration has received a lot of recent media attention. This

is hardly surprising in the view of their soaring pay packages, which for large North American corporations are now of the order of 200 times that of the median worker. Notwithstanding a substantial economic and legal literature, it remains a puzzle as to why otherwise undistinguished people, in many cases no more than recent succession managers, can end up getting multimillion dollar packages. A common defence is to claim that the package in question has been researched by specialist pay consultants, reporting to the board's remuneration subcommittee. Relativities, but of the firm's CEO with others in the same or related industries, are an important input into their recommendations.

Whether the remuneration consultants are truly independent has itself been seen as an issue. As the sceptic story commonly goes, the Board have appointed the CEO and wish to have the ongoing pay recommendations support their judgement. They may in addition supply or choose the benchmark criterion to the consultant, which increasingly incorporate nonfinancial outcomes, such diversity targets. To ensure a continuing relationship, the consultants in turn will err on the side of favouring or reinforcing the Board's judgement. The resulting 'Lake Woebegone' effect, where 'all the men are above average', has been highlighted by a number of academic authors; the literature notes refer.

The analytic framework of the present chapter can be used to study the dynamics of remuneration relativities. In this context, a remuneration consultant would research how the CEO of a subject firm compares with the means of those above and below. An adjustment is then made, with parameters that may differ from the upward comparison to the downward. Taking the average over all such comparators gives the mean shift of the average package value. Whether this is greater or less than the original value depends on the skewness of otherwise of the starting distribution, as well as the relative up or down adjustment parameters. Disturbances that originate as idiosyncratic to a given firm can spread quite rapidly via the ensuing dynamics. Resetting the pay of succession managers is an important consideration for aggregate remuneration stability.

In a CEO context, 'pay' or 'income' typically refers to a package of base salary together with incentive bonuses, stock options, retirement or takeover provisions, some or all of which may be deferred as to final vestment. To combine these as a scalar figure (henceforth 'income'), one can imagine an annualised present value, utilising a real options framework for contingent outcomes, together with time discounting to cover future lodgements.

Benchmark criteria often govern the package value for a given company, to be taken into account by remuneration consultants. These commonly encompass asset value, annual company earnings, number of employees, market capitalisation, and other criteria specific to the industry. Thus if a given manager has a nominal annual package remuneration of y (dollars, euros etc.), this may be determined in relation to a combination $w'z$ where the w_i are the weights given to a collection of indicators z_i. The standardised value for firm i will be taken as $x_i = y_i/w'z_i$; in what follows, this forms the basis for comparison by remuneration consultants for that industry.

Note that this assumes that all consultants employ the same set of evaluation weights. If this is not the case, comparability distortions can arise. Thus suppose the

consultant for firm A standardises against criterion $w_a'z$, using $x_a = y/w_a'z$ as a yardstick for determining nominal income y_a. But firm B's consultant, in reviewing firm A for comparative purposes, uses different weights $w_b'z$, setting the standardised value $x_a = y_a/w_b'z$ for the same nominal income y_a. Effectively this would result in two possible recommendations for the nominal income for the CEO of firm B. A sympathetic remuneration committee might choose the remuneration consultant to favour the higher outcome. In what follows, however, a consistent set of weights w will be employed across different firms in the same industry, while acknowledging a strategic set of attribute weights as a possible distortion.

With this proviso, let $\{x(i,t)\}$ constitute the standardised set of CEO incomes at time t for a comparator set of firms $i = 1, 2, \ldots, N$. For historical or other reasons (see below), the standardised incomes might differ between firms at any given time. However, it will be in the mind of CEO's and their firm's consultants that they may be better or worse off relative to their peers.

A simple remedy might be to adjust according to whether CEO for firm i is above or below the mean:

$$x(i,t+1) = x(i,t) + \lambda_+ [\mu_t - x(i,t)]_+ - \lambda_- [x(i,t) - \mu_t]_+; \quad 0 \le \lambda_+, \lambda_- \le 1, \tag{3.11}$$

where $\mu_t = \frac{1}{N}\sum_{i=1}^{N} x(i,t)$ is the current average standardised income. The special case $\lambda_- = 0$ would indicate the difficulty in getting any CEO currently above the mean to accept a fall in remuneration. Downward inflexibility aside, simple mean regression as in (3.11) is arguably unrealistic on more general grounds. It would indicate that any CEO currently paid just a little less than the mean would be completely indifferent to the existence of a range of what may be significantly higher incomes, such as with a bimodal income distribution centred at the mean.

A more realistic approach would be to adjust on a graduated basis according to the mass above and below any given current income. So let $\mu_r(x)$ denote the conditional mean of standardised incomes above x, i.e. $\mu_r(x) = E[X|X > x)$, and likewise $\mu_l(x) = E[X|X \le x]$ the conditional mean of incomes less than the given x. The dynamic specification (3.11) would be replaced by

$$x(i.t+1) = x(i,t) + v(x(i,t);\lambda), \text{ where} \tag{3.12a}$$

$$v(x(i,t);\lambda) = \lambda_+ [\mu_r(x(i,t)) - x(i,t)] - \lambda_- [x(i,t) - \mu_l(x(i,t))].$$

For any given income x, the outcome is now a balance between a push up in the direction of higher comparative incomes (the first right hand term in the expression for $v(.)$), and a pull down towards the lower incomes (the second right hand term). In contrast to the simple mean correction (3.11), all comparator incomes feed into the relativity adjustment. A regularity condition $v'(x) > -1$ can be added to ensure that the result does not reverse rankings; so if $x_1 = x + dx$, then $x_1 + v(x_1) > x + v(x)$, between any two times $t, t + dt$.

Equation (3.12a) is the relativity equation, expressing the pure comparability effects. To this can be added equations for nominal income determination:

$$y*(i,t+1) = x(i,t+1)w'z_{t+1} \tag{3.12b}$$

$$y(i.t+1) = y(i,t) + \mu\,[y*(i,t+1) - y(i,t)] + \varepsilon(i,t); \quad 0 < \mu \leq 1. \tag{3.12c}$$

Following a reset of the income basis according to comparability, the indicated target nominal income would be given by (3.12b). However, Eq. (3.12c) indicates that this adjustment can itself be partial in any period. An idiosyncratic zero mean disturbance $\varepsilon(i,t)$ is added, which can encompass managerial performance adjudged to be superior, the outcome of consultant insecurity with respect to that company, or any other individual effects.

The dynamic process is ongoing. Once the nominal outcome is determined for firm i, including the disturbance ε_i, this feeds into the next period's measurement for its comparator relativity. There is therefore a channel for nominal outcomes to feed into outcomes based in relativities.

Relativities impact on managerial incomes indirectly via the nominal (y) effects, as well as via comparator standardisation. But a full understanding starts with regularity conditions under which pure comparator relativities do, or do not, lead to stable outcomes. For this it will suffice to consider the comparative effect of the standardised incomes x as in Eq. (3.12a). Thus if $\mu_t = E_i[x(i,t)]$ denotes the average standardised income at any given time, then the stability conditions relate to whether or not $\mu_{t+1} = \mu_t$ versus $\mu_{t+1} > \mu_t$.

Taking expected value of both sides of (3.12a), we obtain the change in mean standardised income between the two periods as

$$\Delta\mu = \lambda_+(\mu_R - \mu) - \lambda_-(\mu - \mu_L). \tag{3.13}$$

Any change in mean incomes between the two periods this depends upon the adjustment coefficients λ_+, λ_- and the skewness or otherwise of the initial income distribution. In all cases, neutrality can only be achieved if $\lambda_- > 0$; some acceptance of downward flexibility is necessary.

From expression (3.13), mean stationarity as between times *t, t + 1* requires $\lambda_+(\mu_R - \mu) = \lambda_-(\mu - \mu_L)$. Setting $v = (\mu_R - \mu) - (\mu - \mu_L)$, mean stationarity requires

$$\frac{\lambda_+}{\lambda_-} = \frac{\mu - \mu_L}{v + (\mu - \mu_L)} = \frac{1}{1 + v/(\mu - \mu_L)}.$$

Now $\frac{v}{\mu - \mu_L} \geq -1$. In the context of expression (3.13) it is therefore the case that

(a) If the initial distribution is symmetric, mean neutrality is ensured only if $\lambda_+ = \lambda_-$;
(b) If $v > 0$ (positive skewness), then mean neutrality can only be achieved if $\lambda_+ < \lambda_-$;

(c) If $v<0$ (negative skewness), then mean neutrality requires $\lambda_+ > \lambda_-$.

The most favourable outcome for mean stability is evidently where the initial distribution of comparator incomes is negatively skewed and where managers on higher incomes are more relaxed about a small prospective drop in their own incomes. Even this seems at first sight to be a tall order. However, it could occur where their individual firm results are bad, which might make managers more resigned to a finding of lower relativities in the next round.

It is also significant that the incomes are specific to the firm, and that individual CEO's come and go. Thus downward flexibility could and should occur in the course of a managerial succession, as distinct from an outside appointment with an established track record. An instance rose in the 2017 US Congressional hearing into the Mylan epipen pricing debacle, where the then recent internal successor CEO Heather Bresch characterised her $18 million package as 'middle of the road'. That this has evidently not been the case is one of the more disturbing features of some recent instances, suggesting that multimillion dollar outcomes have become institutionalised. Remuneration resets can have positive socioeconomic outcomes.

3.8 Relationship with Stochastic Dominance

Distributional comparisons are often cast in terms of stochastic dominance. Distribution A is first order dominant over distribution B if $F_A(x) \leq F_B(x)$. Thus in a financial context with x as the rate of return, investors would always prefer A to B as an investment prospect. This would result in forcing up the price of security A and lowering its rate of return. In such a context, first order stochastic dominance (FSD) is understandably rare.

Second order stochastic dominance is a relationship between the cumulated distribution values. Define

$$\Phi_l(x) = \int_{-\infty}^{x} F(X)dX.$$

A complementary function accumulates the survival function:

$$\Phi_r(x) = \int_{x}^{\infty} (1 - F(X))dX.$$

For a logistic distribution $\Phi_l(x) = -\beta \ln(1 - F(x))$, $\Phi_r(x) = -\beta \ln F(x)$.

Distribution A is second order dominant (SSD) over distribution B if and only if $\Phi_{A,l}(x) \leq \Phi_{B,l}(x)$. An operational meaning is that risk averse investors would prefer A to B as an investment outcome. In this context, investor risk aversion amounts to a utility function for money (here, x) that is globally concave, which implies that $E[u(x)] < u(E[x])$. Risk neutrality would amount to equality in the latter

relationship: investors would make decisions solely on the basis of expected outcomes, effectively an affine linear utility function for money.

A sufficient condition for SSD is the 'once cross over' rule: if $F_A(x)$ crosses $F_B(x)$ just once from beneath, then FSD does not hold but distribution A is SSD over B. Intuitively, risk averse investors are willing to sacrifice a prospect of higher returns in favour of better protection against losses.

A connection exists between the defining functions for SSD and the progressive left and right conditional mean functions of the earlier sections. The left conditional average at $X = x$ is given by

$$\mu_l(x) = \frac{1}{F(x)} \int_{-\infty}^{x} X dF(X).$$

Integrating by parts results in

$$\mu_l(x) = x - \frac{1}{F(x)} \Phi_l(x).$$

Similarly, the right conditional average can be written as

$$\mu_r(x) = -\frac{1}{(1 - F(x))} \int_{x}^{\infty} X d(1 - F(X)),$$

giving

$$\mu_r(x) = x + \frac{1}{1 - F(x)} \Phi_r(x).$$

The two can be combined as

$$\Phi_l(x) = F(x)(x - \mu_l(x))$$

$$\Phi(x) = (1 - F(x))(\mu_r(x) - x).$$

An alternative is to express the SSD accumulations in terms of the spread and asymmetry functions $d(x), v(x)$. This leads to

$$v(x) = -\frac{1}{F(x)} \Phi_l(x) + \frac{1}{1 - F(x)} \Phi_r(x)$$

$$d(x) = \frac{1}{F(x)} \Phi_l(x) + \frac{1}{1 - F(x)} \Phi_r(x),$$

with the inverse relationship as:

$$\Phi_l(x) = \frac{1}{2}F(x)(d(x) - v(x))$$

$$\Phi_r(x) = \frac{1}{2}(1 - F(x))(d(x) + v(x)).$$

Second order stochastic dominance criteria can therefore be cast in alternative forms such that for every value x,

$$F_A(x)(x - \mu_{l,A}(x)) \leq F_B(x)(x - \mu_{l,B}(x)) \; ; \tag{3.14a}$$

$$F_A(x)(d_A(x) - v_A(x)) \leq F_B(x)(d_B(x) - v_B(x)). \tag{3.14b}$$

Expression (3.14a) captures the essentially conservative nature of the SSD concept. If the two medians coincide at $x = m$, a necessary condition for A to be SSD over B is that $\mu_{l,A}(m) > \mu_{l,B}(m)$. The relationship between stochastic dominance and the spread and asymmetry functions is revisited in Chap. 5.

3.9 Literature Notes

The entropic spread and asymmetry metrics and functions were developed in Bowden (2016a, b), Bowden (2017).

A general set of conditions thought to be suitable for a dispersive ordering can be found in Jeon et al. (2006). In such terms, the proposed entropic spread metric d qualifies as a dispersive ordering. It could further be noted that dispersion metrics have been proposed that are based respectively on Shannon entropy and on the Fisher information matrix. Kostal et al. (2013) is a reference.

Turning to skewness, many measures of distribution asymmetry have been proposed. The most basic are cast in terms of comparing the mean, median and mode, although the third order moment features in classical statistics. On measures of this type, the Wikipedia article on nonparametric skewness is as good a starting point as any. Other recent measures have been proposed by Groeneveld and Meeden (1984, 2009). Earlier, Van Zwet (1964) proposed a set of conditions relevant to the ordering of skewness across different distributions. Kraemer (1998) is a treatment in the context of income inequality. More recently, Kraemer and Dette (2016) investigate the consistency of the envy metric v with such a set of axioms.

The Gini coefficient has had a long history of discovery and rediscovery (David 1968). Kraemer (1998) reviews the use of this and other measures in the general context of income distribution. A further review on this topic is contained in Sect. 4.1.

On executive compensation, Carpenter and Yermack (1999), also Balsam (2002), are book length accounts. The labour economics chapter by K. Murphy (1999) is a comprehensive account, *inter alia* covering principles and practices of pay setting in relation to performance. Two sceptical articles by Bebchuk and Fried (2003, 2004) make a good read, while a more formal contribution was Frydman and Jenter (2010).

Finally, second order stochastic dominance was linked to risk aversion and investor decisions under risk in Rothschild and Stiglitz (1970).

References

Balsam, S. (2002). *An introduction to executive compensation*. San Diego: Elsevier Academic Press.

Bebchuk, L. A., & Fried, J. M. (2003). Executive compensation as an agency problem. *Journal of Economic Perspectives, 17,* 71–92.

Bebchuk, L. A., & Fried, J. E. (2004). *Pay without performance: The unfulfilled promise of executive compensation*. Boston: MIT.

Bowden, R.J. (2016a) Giving Gini direction: An asymmetry metric for economic disadvantage. *Economics Letters*, *138*, 96–99.

Bowden, R. J. (2016b). Dual spread and asymmetry distribution metrics based in partition entropy. Kiwicap Research. Retrieved from http://www.wellesley.org.nz/papers_public.asp.

Bowden, R. J. (2017). Distribution spread and location metrics using entropic separation. *Statistics and Probability Letters, 124,* 148–153.

Carpenter, J., & Yermack, D. (1999). *Executive compensation and shareholder value: Theory and evidence*. Boston: Kluwer Academic.

David, H. A. (1968). Gini's mean difference rediscovered. *Biometrika*, *55*, 573–575.

Frydman, C., & Jenter, D. (2010). CEO Compensation. *Annual Review of Financial Economics, 2,* 75–102.

Groeneveld, R., & Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician* 33, 391–399.

Groeneveld, R., & Meeden, G. (2009). An improved skewness measure. *Metron - International Journal of Statistics* 67, 325–327.

Jeon, J., Kochar, S., & Park, C. G. (2006). Dispersive ordering -some applications and examples. *Statistical Papers, 47,* 227–247.

Kostal, L., Lansky, P., & Pokora, O. (2013). Measures of statistical dispersion based on Shannon and Fisher information concepts. *Information Sciences, 235,* 214–223.

Kraemer, W. (1998). Measurement of inequality. *Handbook of applied economic statistics* (Vol. 39). CRC Press.

Kraemer, W., & Dette, H. (2016). Beyond inequality: A novel measure of skewness and its properties. Working paper, Institute for Economic and Social Statistics, Dortmund Technical University. Retrieved from https://www.statistik.tu-dortmund.de/kraemer.html.

Murphy, K. J. (1999). Executive compensation. *Handbook of labor economics* (Vol. 3 Part B). New York: Elsevier.

Rothschild, M., & Stiglitz, J. E. (1970). Increasing risk: I. *A definition. Journal of Economic Theory, 2,* 225–243.

Van Zwet, W. R. (1964). *Convex transformations of random variables. Volume 7 of mathematics centre tract*. Amsterdam: Mathematisch Centrum.

Chapter 4
Information Comparisons in Practice

4.1 Introduction

With the basic theory of the spread and asymmetry measures now in hand, a range of applications can be explored, some of considerable topical interest in their own right.

The first of these is income distribution. Section 4.1 outlines existing approaches to the measurement of the nature and degree of inequality, in the process consolidating some of the content in Sect. 3.5. Still the best known of these is the Gini coefficient, which refers to the concordance, or lack if it, between the progressive proportion of people and that of their total incomes. But in principle, a given Gini coefficient can arise either as a result of too many people on low incomes; or again, too many people on higher incomes. By way of contrast, social commentary is more often concerned with the difference between positive and negative skewness. The former is regarded as more of a concern. Positive skewness means that the weight of the distribution has been pushed to the left, which means too many people on lower incomes accompanied by a long tail to the right of fewer people, some with very high incomes.

Of the existing inequality metrics that seek to rectify the shortcomings of the Gini coefficient, the more influential could be regarded as imposing observer-calibrated welfare parameters as benchmarks for what is, or is not, excessive inequality. The present agenda in Sect. 4.2 is not prescriptive in this sense. The perspective is instead internal, seen as it were from the point of view of the subjects themselves. Subjects look to incomes above theirs compared to incomes below. Aggregating over all subjects, as a form of double smoothing, results in the dual v and d metrics of Chap. 3. The former measures directed inequality and the latter the spread or dispersion dimension. From there it is a simple matter to plot the one against the other over time as a dynamic phase plane. Illustrations are included for the US and Europe.

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_4

Section 4.3 takes up quite a different socioeconomic tack, this time in terms of measuring stock market performance as seen by investors. Here there is typically a benchmark (such as the general S&P500 stock market index) and the returns on a subject equity or portfolio are to be assessed relative to the benchmark. Existing excess return indexes, such as the Sharpe index, do not adequately encompass the kind of asymmetry that investors value, namely that associated with positive skewness. Once again, the entropic ν metric can be added to the basic Sharpe metric to better reflect the potential asymmetry of investment returns.

Actuarial uncertainty is revisited in Sect. 4.4. The preoccupation here is with the survival probability for a given population. Conditional life expectancy, which refers to the expected survival time given the current age, can be cast into correspondence with the right conditional mean function of Chap. 3. This leads to an overall welfare measure as to the expected longevity in a given population in term of the right hand entropic mean relative to the original mean itself. Subpopulations can be compared in this respect e.g. males versus females.

Section 4.5 concludes with the literature notes.

4.2 Income Distribution

Income inequality has been a perennial topic of economic and social interest, but never more so than the present, where dramatic changes have followed within just a short span of time. Performance driven management rewards that have been seen as excessive have attracted much public attention and are discussed in Chap. 3. But other influences have been at least as pervasive; and arguably more important in their implications for middle workers in particular (the 'lumpen proletariat', as Karl Marx dismissively referred to them). Technological displacement for middle management, import competition from cheaper emerging countries, free trade agreements, adverse fallouts from public spending bubbles, commodity price reversals, the global financial crisis, are just some of the causal influences, combining as the perfect storm in their fallouts for remuneration and employment down the line.

In the most general sense, metrics for income distribution are part of a wider body of knowledge into social welfare functions developed and debated over many years. But establishing a consensus as to an optimal measure for income equality has not been easy, for theoretical work in economics has indicated that it is not in general possible for society to ever agree on a consistent ordinal social utility function. The outcome is that it would not be possible to establish universal agreement among the subjects themselves as to a single best metric for income inequality.

The empirics have therefore focused upon metrics for income inequality that appeal in designated ways to the observer's own preconceptions as to fairness. Still the best known and most widely accepted metric of this kind is the Gini coefficient, which measures the non-alignment of the accumulated percentage of income with

the progressive numbers of the people enjoying it. The Gini coefficient has already figured in Sect. 3.5 in connection with the mean absolute difference, but it is necessary at the present juncture to consider its motivation and from there, its potential shortcomings.

Figure 4.1 illustrates. Income earners in a society or specified group are first ordered lowest to highest. Then the Lorenz curve, as it is called, graphs the cumulative percentage of total income to the cumulative percent of people earning it. If everybody enjoyed the same income, the outcome would be the 45° line. The Gini coefficient G measure the extent to which this is not true, as the area between the curve and the 45° line (for country B the shaded area) as a fraction of the area under the 45° line itself. For a continuous distribution of incomes y, with distribution function F(y), this reduces to

$$G = \frac{1}{\mu}\int_0^1 F(y)(1 - F(y))dy.$$

Alternative expressions in terms of the Gini mean difference are given in Chap. 3. Most developed countries have their Gini coefficients in the 60–70% range, a figure which has been gradually increasing over time.

The Gini coefficient has found a number of uses in contexts other than economics. However, the discussion that follows is in terms of income or wealth distribution, which was indeed the original context.

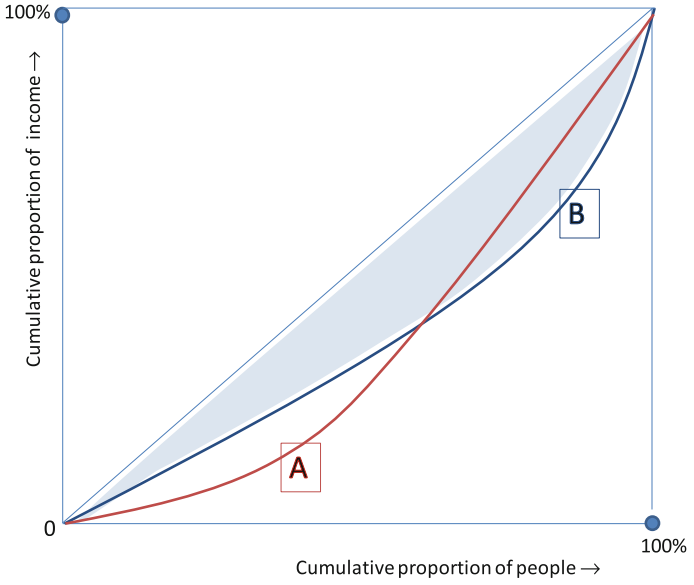


Fig. 4.1 The ambiguity of the Gini coefficient

Although of venerable origin, dating as far back as 1912 with the work of Corrado Gini, the Gini coefficient has a number of intrinsic problems. In Fig. 4.1, the two countries A and B would have just the same value of G. But the social welfare implications are quite different. In country A, people of low incomes share proportionately less of the total income take to any given point than for country B.

In other words, the Gini index lacks direction as to the source of the inequality; it does not properly pick up the kind of inequality that would concern most observers, namely positive distribution skewness, meaning too many people on low incomes. As suggested in Sect. 3.5, the Gini coefficient has more of the character of a generalised spread metric. Of course, the spread of incomes is indeed of interest as such. But it does need to be supplemented with considerations relating to the degree of skewness in *F*.

An agenda in the latter case has been to find income asymmetry metrics that have economic meaning, as distinct from the textbook third order moment. One general approach is to compare distributions over time or location in terms of a user assigned inequality aversion parameter (ε). Results are commonly tabulated against different values of ε, with higher ε values as a focus for social concerns about the share of lower income groups. Implicit in this is a user chosen social welfare function.

Metrics based on Shannon entropy metric have figured in another line of development. It is not immediately apparent just why entropy, viewed as total informational complexity, should correspond to any social welfare function. However it can be shown that some of the inequality parameter approaches reduce to a metric that in turn maps into the Shannon entropy function. But so far as entropy in general is concerned, partition entropy might constitute a better starting point, as it is explicitly concerned with the 'more' or 'less' dimension that is the focus of attention.

As a general comment, such approaches on the choice of metric could be regarded as imposing value judgements on the part of an external observer, who would set inequality aversion parameters such as ε. However it is also possible to imagine a different thought experiment that seeks to aggregate in some meaningful way how each subject thinks about his or her own income in comparison to that of others. This could be referred to as an internal observer approach.

A simple way to do this is via a linear expected utility scale in which each individual derives positive utility to the extent that his or her income exceeds the conditional expected income below; and negative to the extent that it falls short of the conditional expected income above. The net difference is then aggregated over the relative number in each income band, i.e. the density of the income distribution. A negative index means that on the average people think that others are better off than themselves; so the proposed 'v-index' or metric could evoke net divergence, disadvantage or even envy. As a supplement, the metric enables an external observer to tell at a glance whether a higher Gini arises from spread or positive asymmetry. This approach to inequality does indeed have an entropic reference, but in terms of the left and right unit entropic shifts of the income distribution function.

A more complete approach of this kind would also have to pay attention to two further welfare aspects. One is the dispersion of the income distribution, as a 'noticeability' property. Wide dispersion attracts more attention to very high or very low incomes, in this context relative to my own as an internal observer. A second is the average or median income itself. If times are good, higher incomes in others attract less social opprobrium. There is a correspondence here with managerial remuneration practices. If the firm is doing well, and stockholder rewards are good, then a hike in executive remuneration is carried, even with acclaim, at the company AGM.

4.3 Implementation as Entropic Asymmetry and Spread Metrics

To see just how the internal observer approach works, suppose my income is y, and I have perfect knowledge of all the incomes both above and below mine, which for expositional purposes are assumed to be a continuum in the interval $0 \leq y < \infty$.

I first look at the average income below me: $E[Y|Y \leq y] = \mu_l(y)$. Relative to this group I am better off to the extent of the difference $(y - \mu_l(y))$. Then I look at the average income above me: $\mu_r(y) = E_F[Y|Y > y]$. Relative to this group I am worse off according to the difference $(\mu_r(y) - y)$.

My net envy or subjective divergence is measured as the difference

$$v(y) = (\mu_r(y) - y) - (y - \mu_l(y)).$$

Over the entire distribution of incomes, the aggregate net envy or divergence is

$$v = E[v(y)] = \int_0^\infty f(y)v(y)dy.$$

The same sort of argument could be mounted in support of a spread measure. With my income as y, I look to see how many people and above me, and on the average how far. This can be proxied by the distance $\mu_r(y) - y$. Now I look to the left, as to how many people have incomes below mine, and on the average how much. A proxy is $y - \mu_l(y)$. The sum of the two, or their average, can then be treated as a proxy for my relative isolation from each side, giving rise to the complementary function

$$d(y) = (\mu_r(y) - y) + (y - \mu_l(y)).$$

Over the entire distribution of incomes, this averages to

$$d = E[d(y)] = \int_0^\infty f(y)d(y)dy.$$

As in Chap. 3 it follows that

$$v = \mu_L + \mu_R - 2\mu$$

$$d = \mu_L - \mu_R,$$

where the means are those of the left and right entropic shifts $F_L,\ F_R$ of the original income distribution function F.

The two quantities v, d have been validated in Chap. 3 as metrics for distribution asymmetry and spread. But in addition the metric v satisfies a number of conditions that have been held up as necessary for any economically meaningful inequality measure. It is scale independent: the index for a set $\{x\}$ is scaled up proportionately for $\{\lambda x\}$ with λ a positive scalar. It is decomposable: the index for a consolidated collection of countries or regions can be decomposed into the sum of their respective index plus a further index computed from a collection of their mean incomes. It also satisfies the 'Robin Hood transfer principle': if a dollar (or any sum less than their difference) is transferred from a rich person to a poor one, the resulting metric v is smaller.

In some contexts it is useful to normalise the inequality index by dividing by the spread index d as the ratio v/d. This is then unit free, and in such cases complete scale independence would apply (i.e. scale absolutely independent of λ as above). In general, however, it is desirable to present both v and d, as they offer the different distributional perspectives earlier referred to, encompassing both directional inequality and pure spread. Finally, it may be useful to normalise both metrics by the mean income, in order to abstract from temporal changes in incomes as a whole, or alternatively cross country comparisons.

Where the v-metric is more pronounced, there are welfare implications in indicating the scale of correction required to restore symmetry. A value of 25% of the mean income would suggest that 25% of the mean income could be redistributed to households below the mean to restore net envy to zero. The criterion does not in itself specify the pattern of any such notional redistribution, which would have to ensure continuity around the mean; so that $y = \mu - \$1$ gets very little extra to avoid disturbing the relativity with $y = \mu + \$1$. But it can serve to indicate the scale of the redistribution required.

Figure 4.2 depicts a modified US income histogram for 2013. The range has been truncated by omitting the highest income band. The latter is open beyond $200,000, meaning that insufficient information is publicly available to be able to calculate a meaningful distribution beyond this figure. But even as is, the histogram indicates fairly pronounced positive skewness. The (truncated) mean household income is $60,181. Gini for the truncated data is 43.01. The summary numbers supplied by the Bureau for the complete sample indicate a Gini of 47.6, with a

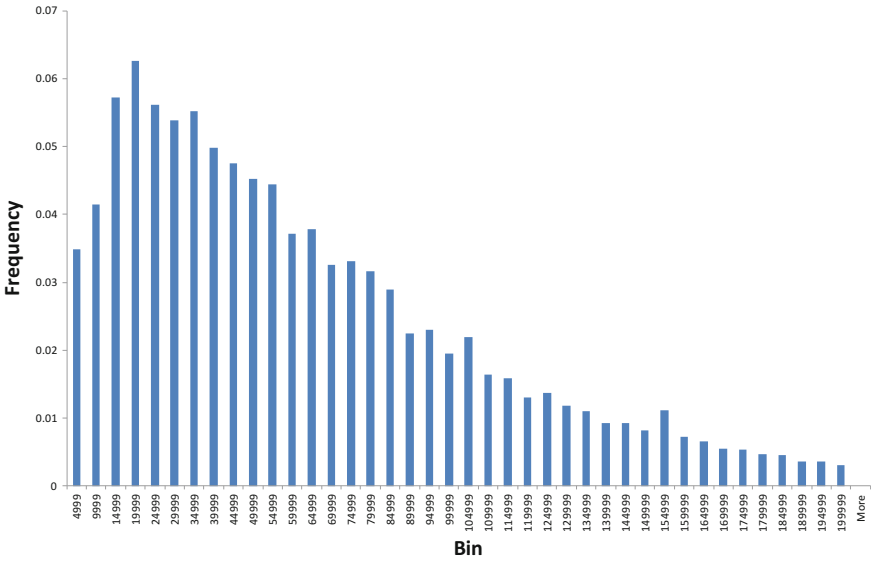


Fig. 4.2 Truncated US household income 2013 (Data source: US Bureau of the Census: table HINC-01: sample of 68,000 households.)

comprehensive (non truncated) mean of $72,641. The difference in the Gini coefficients indicates that the truncated v-metric is going to be an underestimate.

For the publicly available sample, the entropic asymmetry and spread metrics can be calculated as v = $6,850 which corresponds to 11.38% of the mean income, and d = $72,154. Scaling up by the ratio of the complete to the truncated Gini suggests a redistribution of the order of $7,581 as a first approximation once all higher incomes are taken into account.

4.3.1 Social Welfare Aspects

The three metrics, namely v for asymmetry, d for spread, and μ (or the median) for central tendency, provide potential cues for thinking about how the average person might react to their publication. A positive skewness indicator v will indicate that the average person is net disadvantaged relative to those below and above on the income scale. In turn, this will be more noticeable where the distribution has a wider spread d. And in both cases the reaction will be moderated when all incomes are higher. If I have enough for my daily needs and annual holiday, I am less likely to be obsessed with what the senior public servant down the road earns.

The foregoing suggests an ordinal social welfare function, geared to the average worker. This might be of the general form

$$SW = \psi(\mu, v, d), \tag{4.1a}$$

with derivatives

$$\phi_1 = \frac{\partial \psi}{\partial \mu} > 0; \;\; \phi_2 = \frac{\partial \psi}{\partial v} < 0; \;\; \phi_3 \; = \frac{\partial \psi}{\partial d} < 0.$$

Thus social welfare is specified as increasing with average income, but decreasing with asymmetry and spread.

There might also be interactions between v and d, such that $\frac{\partial^2 \psi}{\partial d \partial v} < 0$. The latter would indicate that people are more concerned about asymmetry when the underlying spread is larger.

As a further aid to interpretation, suppose that ψ is separable between μ on the one hand, and (v, d) on the other, of the general form

$$SW = \psi(\mu, \phi(v, d)). \tag{4.1b}$$

This would certainly be the case if the parent function (4.1a) was homogenous in its three arguments, but it is not necessary to introduce such a restriction.

The form (4.1b) could be interpreted as saying that the average worker looks first at the spread and asymmetry in relation to his or her own income, then modifies any reaction if the personal income is higher or lower during any given year. For any given income there is therefore a set of indifference curves (level surfaces) as between spread and skewness. However, these may not be uniformly concave or convex. For as skewness (v) becomes more positive, it requires progressively lower dispersion to materially lessen the envy. By way of contrast, if v becomes more negative, it might require a progressively higher spread to preserve the same social utility. The social indifference curves relating skewness to spread might therefore be sigmoid in shape.

4.3.2 Dynamics: The v-d Phase Plane

Over time the metrics for spread and asymmetry change, especially so if the economy is impacted by substantial external events or structural changes. This was the case for many European economies over an interval of time spanning the period before, during, and after the Global Financial Crisis (GFC), with further impacts from the fall in the price of oil. Countries were in fact differentially affected, with some emerging much better than others.

A standout example of the latter was Norway, which weathered its own price oil shock very well and took policy measures that amounted to a social redistribution of incomes. Figure 4.3 for Norway illustrates with income histograms for three different years, 2005, 2010, 2013. The v and d metrics are listed in the insert.

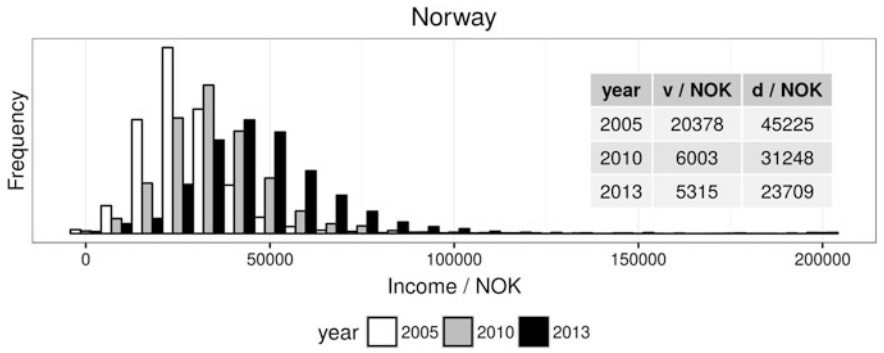


Fig. 4.3 Norway: monthly disposable income (1 NOK ~ 0.12 USD) (Data source: OECD income statistics)

Evidently Norway's income distribution became less asymmetric, but the spread diminished with the transition from a positively skewed to a more symmetric distribution.

A convenient way to portray temporal changes in the social context of incomes is to graph the v and d metrics together in the form of a directed phase plane over time, using the same database as for Fig. 4.3. To compensate for accompanying changes in the average incomes, it is convenient to scale as v/μ versus d/μ.

Figure 4.4 illustrates for France. The adverse social impact of the GFC is apparent, with a substantial rise in both spread and inequality in 2007-8. By 2013 this was still not fully restored to the pre GFC period. This could correspond to a radical downward shift in the social welfare function, especially so as the average income in fact declined over that period. By way of contrast, other European countries fared much better over the same period, some actually diminishing their inequality and spread metrics.

4.4 Application to Stock Market Performance

Performance metrics for equities or managed funds are established tools of the finance industry, objects of media reporting and investor assessment. The metrics in question are usually comparative in nature against some benchmark. For example, the Sharpe index compares the mean of the security return against the risk free rate, effectively the expected return on a portfolio long in the security return, short in the risk free asset. Even short of this, however, there are some hidden assumptions in the use of standard metrics. In the case of the Sharpe index, it is tacitly assumed that the distribution of equity returns is symmetric. But this is not necessarily true, even in relatively normal times. If returns are asymmetric over any given interval, then extraordinary exposures to loss, or else opportunities for gain, can arise. In attempts to capture such gains or losses, a number of authors have proposed incorporating a

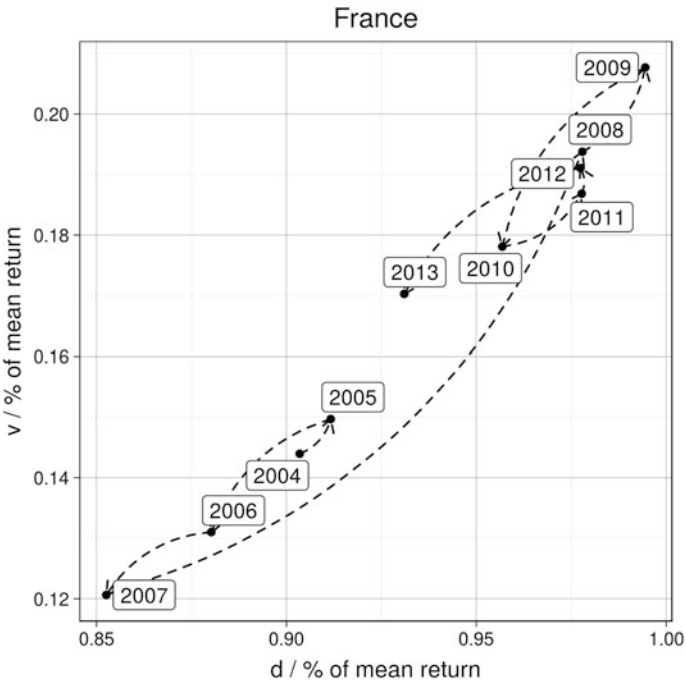


Fig. 4.4 Income distribution changes in France over the GFC (*Source* Bowden et al. 2018)

contribution from skewness, the later measured as the conventional third order moment.

It could be argued, however, that any such diagnostic metric should be contextual in nature, with reference to the investor's gain or loss, in a way that the textbook third order moment for asymmetry cannot. Moreover, there is room for exploration as to just what should be assumed, even tacitly, about the underlying utility function. A case could be made that the standard concave (risk-averse) utility function is not a comprehensive representation for investor motivation. More appropriate might be a utility function that is concave on the downside and convex on the upside. A form of this kind is often referred to as a Friedman-Savage utility function, for which there is some corroborating support from more recent studies in behavioural economics. Thus a fund manager would be averse to losses, the more so with the prospect of employment termination if losses are heavy. On the other hand, the same manager will be motivated by a more generous performance bonus for progressive gains on the upside.

To incorporate the possible asymmetry of returns, one can draw on the v-metric, or on the function $v(x)$ that underpins it. However in an investment context, where r refers to returns, things have to be reversed. On any given day suppose the return is r. Now I win to the extent that this exceeds the average return below (a better day), and I lose to the extent that it falls short of the average above (other days are

better). This indicates that a more appropriate welfare function for the finance context is the negative of the net economic disadvantage function, i.e.

$$w(r) = -v(r) = (r - \mu_l(r) - (\mu_r(r) - r)),$$

the 'win' function. The function $w(r)$ is concave below (but convex above) a break-even value on the return axis. As returns r become more negative, the win function (actually a loss in this zone) $w(r)$ becomes asymptotic from above to the 45 line, and from below as $r \to \infty$. The expected value of the win function is therefore

$$w = 2\left(\mu - \frac{1}{2}(\mu_L + \mu_R)\right), \tag{4.2}$$

where the respective means are those of returns under the original and unit shifted distributions. The spread d, as derived from the function $d(r)$ can be taken as unchanged.

The standard Sharpe performance metric is written as

$$S = \frac{\mu_r - r_f}{\sigma},$$

where μ_r is the expected return, σ its standard deviation and r_f is a risk free rate. Defining $\tilde{r} = r - r_f$ as the excess return, with distribution function $\tilde{F}$, the Sharpe ratio becomes

$$S = \frac{E_{\tilde{F}}[\tilde{r}]}{\tilde{\sigma}} = \frac{\tilde{\mu}_r}{\tilde{\sigma}},$$

Here and elsewhere the superscript tilde refers to excess returns. Thus $\tilde{\sigma}$ denotes the standard deviation of excess returns.

The proposed W-metric is an alternative to the Sharpe ratio. Because it explicitly invokes an asymmetry consideration, the benchmark is taken as the median of returns instead of the mean. The performance metric is defined as

$$W = \frac{\tilde{w} + \tilde{r}_m}{\tilde{d}}, \tag{4.3}$$

where $\tilde{r}_m$ is the median of the excess return distribution and the asymmetry metric $\tilde{w}$ is defined as in expressions (4.2) applied to the excess return. The denominator $\tilde{d}$ of the W metric is the nonnegative spread metric d applied to excess returns. For standardisable distributions such as the normal, the logistic or the Gumbel, the metrics d and the standard deviation σ are proportional via the scale parameter. In such cases there is little effective difference between using either $\tilde{d}$ or $\tilde{\sigma}$ for the denominator.

The numerator of the metric (4.3) splits into two terms: $\tilde{r}_m$ is the median excess return, while $\tilde{w}$ captures the asymmetry of the excess returns distribution. Rewriting the numerator of W results in:

$$\tilde{w} + \tilde{r}_m = (\mu_r - r_f) + [\tilde{w} + r_m - \mu_r].$$

Thus the W-metric and the Sharpe ratio's numerator differ in the term $[\tilde{w} + r_m - \mu_r]$.

In this context, consider the following cases:

(a) The distribution of excess returns $\tilde{r}$ is symmetric. In this case $\tilde{w} = 0$, $r_m = \mu_r$, and the numerators of the W-metric and Sharpe are identical.
(b) The distribution of excess returns is positively skewed. In this case $\tilde{w} < 0$, and it is likely that $r_m - \mu_r < 0$. Together this means that $W < S$.
(c) The distribution of excess returns is negatively skewed. In this case, $\tilde{w} > 0$, and it is likely that $r_m - \mu_r > 0$. Together this means $W > S$.

Cases (b) and (c) reflect the implied Friedman-Savage investor utility basis: investors would like a positively skewed distribution; increasing marginal utility in the higher zone. But they might back away from a negatively skewed one; too much weight in the low zone, the area of more negative marginal utility. The difference between W and S is generated by the implicit underlying utility functions, and the way that these are responsive to the distribution of returns. The Sharpe metric S tacitly assumes a linear utility function, while the W metric is more responsive to the mixed concave-convex Friedman-Savage type utility function.

Adaptations of the measure W can be devised for other benchmarks. Instead of using the risk free rate, the comparator could be the market return R, so $\tilde{r} = r - R$ is a compound return, long in the subject security and short in the market. Alternatively, if a CAPM model is thought to apply, a generalisation of Jensen's alpha can be defined with the comparator return as the market, scaled by the security's beta.

To illustrate the new measure, daily log returns of Ford Motors Company are employed, together with the S&P 500 index as a market proxy. Following convention, the 10 year US treasury yield is employed as a proxy for the risk free rate from 1990 to 2015, Technically, even a 10 year US bond rate is not risk free over the unit holding period, but on the other hand can be taken as simply a returns benchmark.

The Sharpe ratio and the W metrics are compared in Fig. 4.5. For each year, the distribution of the daily returns is created, then apply the two measures to these distributions. The risk free rate is utilised for a benchmark, as in the classic Sharpe ratio. The dashed bar corresponds to W in Eq. (4.3) and the solid bar to the Sharpe ratio. The two measures correlate with a coefficient of $\rho_{S,W} = 0.496$ which is statistically significant at 1%. However, there are several occurrences (1990, 1991 and 2007) where not only the absolute values between the Sharpe and the W-measure differ, but also the signs.

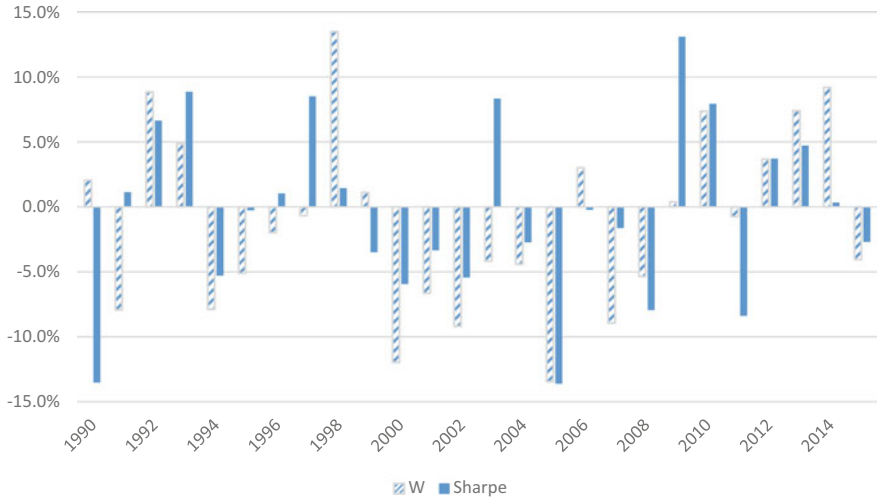


Fig. 4.5 Comparison of the *W* metric and the Sharpe ratio using the risk-free rate as benchmarks *Source* Bowden et al. (2017)

4.5 Actuarial Life Expectancy

The functions and metrics for asymmetry and spread can be adapted to the exigencies of actuarial science, which deals with survival and mortality rates of human populations. The calculus of the latter provides the formal underpinning for the assessment of premiums in the life insurance industry. Extensions exist by a process of analogy to contexts such as equipment failure in industry, and even to biological population assessments.

Actuarial science has its own notation to describe the probabilities of survival that are needed in pricing life insurance and related products, which can be quite involved. For present purposes, discussion will focus on survival probabilities, where l_x denotes the number of people who survive (live) to age x. This is conventionally calibrated off an initial 100,000 births, with $l_0 = 100{,}000$. Survival between age x and age $x + 1$ is described in terms of transitional probabilities of death in that interval. So if $p_x, q_x = 1 - p_x$ denote transitional probabilities that someone aged x will survive (p_x) or die (q_x) in that age interval, then

$$l_{x+1} = l_x p_x = l_x(1 - q_x).$$

Tabulation takes two general forms, with variants on each:

(a) A cohort life table focuses on people born in a specific year. Thus a cohort table for the year 1940 would start with 100,000 people born in that year and trace their survival history since then.

(b) A period life table is a snapshot at one particular point in time covering all age groups as of that particular point. Table 4.1 illustrates for the US.

Making use of such data often entails specific assumptions about the survival or death probabilities. Thus if I am now aged x, the probability that I will survive for t more years and then die within the following k years is given by

$$\frac{l_{x+t} - l_{x+t+k}}{l_x}.$$

However, this refers to the future, which has not yet transpired to someone born in my year. So to obtain such survival probabilities, it is conventional to use transitional probabilities $p_x, p_{x+1} \ldots p_{x+t+k-1}$ that are calibrated off the recorded survival probabilities (or proportions) of people senior to me.

Table 4.1 US life table (period life table) 2014

	Males			Females		
Exact age	Death probability	Number of lives	Life expectancy	Death probability	Number of lives	Life expectancy
0	0.006322	100,000	76.33	0.005313	100,000	81.11
1	0.000396	99,368	75.81	0.000346	99,469	80.54
2	0.000282	99,328	74.84	0.000221	99,434	79.57
3	0.000212	99,300	73.86	0.000162	99,412	78.59
4	0.000186	99,279	72.88	0.000131	99,396	77.6
5	0.000162	99,261	71.89	0.000116	99,383	76.61
6	0.000144	99,245	70.9	0.000106	99,372	75.62
7	0.000129	99,231	69.91	0.000098	99,361	74.63
8	0.000114	99,218	68.92	0.000091	99,351	73.64
9	0.0001	99,206	67.93	0.000086	99,342	72.64
10	0.000093	99,197	66.94	0.000084	99,334	71.65
…	…	…	…	…	…	…
…	…	…	…	…	…	…
110	0.568528	3	1.2	0.539881	17	1.27
111	0.596954	1	1.13	0.572274	8	1.18
112	0.626802	1	1.05	0.606611	3	1.09
113	0.658142	0	0.98	0.643007	1	1.01
114	0.691049	0	0.92	0.681588	0	0.93
115	0.725602	0	0.86	0.722483	0	0.86
116	0.761882	0	0.79	0.761882	0	0.79
117	0.799976	0	0.74	0.799976	0	0.74
118	0.839975	0	0.68	0.839975	0	0.68
119	0.881973	0	0.63	0.881973	0	0.63

Source US OACT: https://www.ssa.gov/oact/STATS/table4c6.html

As it stands, this can be an imperfect estimate at times when rapid advances in public health are taking place. However, a convention of this sort is used where necessary to complete official life tables. Thus the US cohort life table for the birth year 1940 extends to age 119 and is calibrated of the existing recorded transition probabilities for extreme old age, i.e. people born prior to 1940.

For present purposes it will be convenient to recast discussion so that the base reference is effectively just one person, which in turn means that conventional probabilities can be invoked. With this convention,

$$l_x = 1 - F(x),$$

so that $F(x)$ is the probability of death by age x, with $1 - F(x)$ as the probability of survival. These are depicted for US males in Fig. 4.6, with $f(x)$ as the density interpreted as the probability of death within each age interval $x, x+1$.

In such terms, the probability that I will survive exactly t more years (and die within the subsequent year) is given by the ratio

$$\frac{(1 - F(x+t)) - (1 - F(x+t+1))}{1 - F(x)}.$$

An operational version is given by

$$\frac{f(x+t)}{1 - F(x)}. \tag{4.4}$$

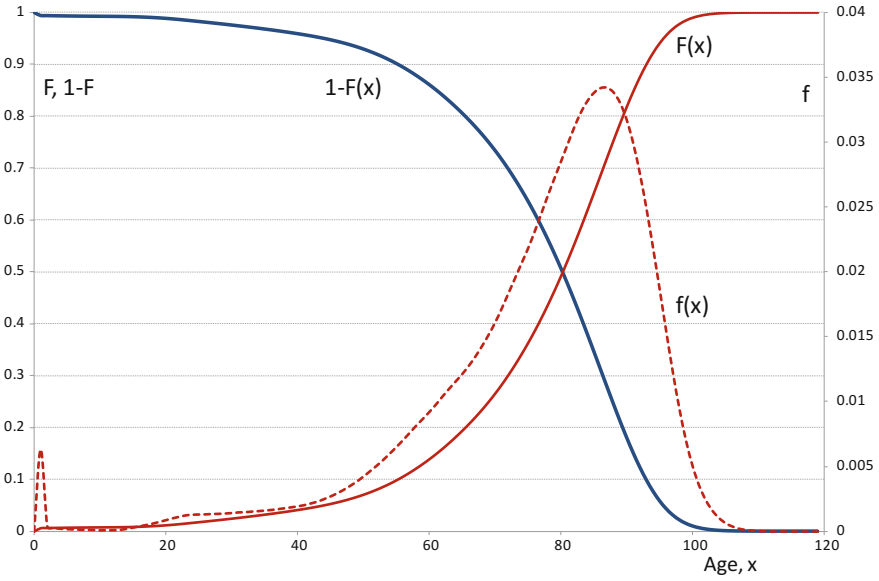


Fig. 4.6 Survival probabilities and complements for US males 2014

From the perspective of my current age x, the specified survival time in expression (4.4) is for exactly t years into the future. Weighting such terms over different values of t into the future gives my conditional life expectancy, relative to my current age of x. In continuous time, this would amount to

$$\frac{1}{(1-F(x)}\int_x^N Xf(X)dX - x,$$

with $X \sim x+t$.The resulting conditional life expectancy at age x can therefore be written as

$$\mu_s(x) = \mu_r(x) - x = E[X|X > x] - x.$$

Figure 4.7 plots this conditional life expectancy function as a function of current age x.

A complementary function to the conditional life expectancy can be defined as

$$\mu_m(x) = x - \mu_l(x) = \frac{1}{F(x)}\int_0^x tf(x-t)dt.$$

To interpret, first fix a nominal age x and consider all people who have died before that time. The denominator refers to the number of people who have died by

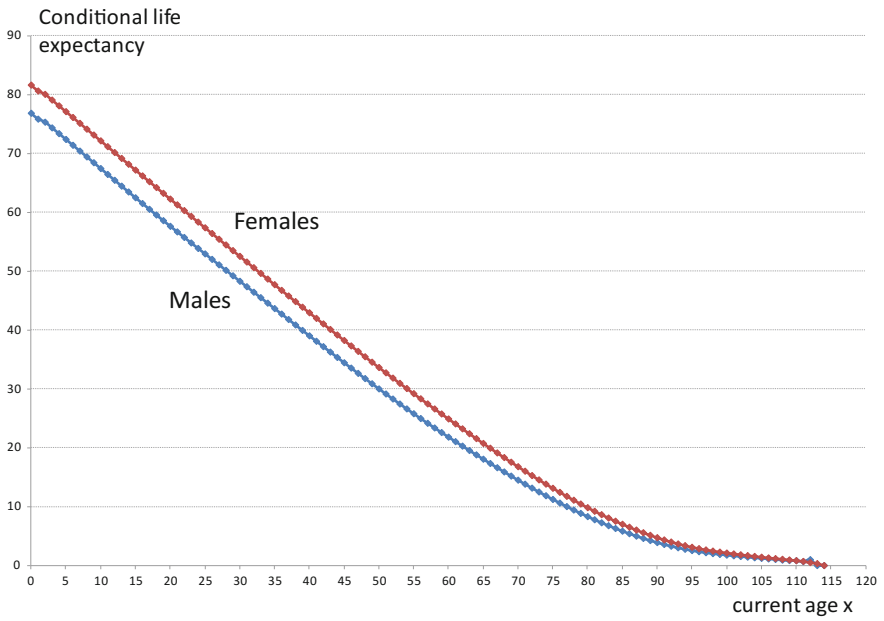


Fig. 4.7 Conditional life expectancy

the given age x. Then look back to the time since death, indicated here by t. The result can be termed a 'memorial function'. It is of less practical interest, except possibly to a monumental mason. However, in the context of the present US example it does have some points of interest, as it is not necessarily monotonic with nominal age x. Indeed, over low to middle age intervals the function for US males is below that for females, though this is reversed for the higher nominal age zone, as one would expect.

One can use the above formulations to derive a single metric that compares (in this case) male and female longevities. From Chap. 3, $\mu_R = E_x[\mu_r(x)]$. Averaged over all values of x, the average life expectancy or survival time is given by

$$\mu_s = E_x[\mu_r(x) - x] = \mu_R - \mu.$$

The right shifted mean μ_R can be evaluated as in Chap. 3, with reference to the unit right shifted distribution of $F_R(x)$. The result gives $\mu_s = 12.149$ for males and $\mu_s = 10.852$ for females. At first sight this result looks paradoxical. It arises because the average is taken with respect to the mortality density $f(x)$, so it is in essence a forfeiture function. The smaller figure for US females arises because they tend to live longer with the same absolute maximum date at death. Hence heavier weight is allocated to shorter conditional life expectancies at the longer end. The effect is apparent in Fig. 4.8. Relative to males, the density of ages at death for females is bunched more to the right. Chapter 5 develops more systematic comparative measures applicable to such contexts.

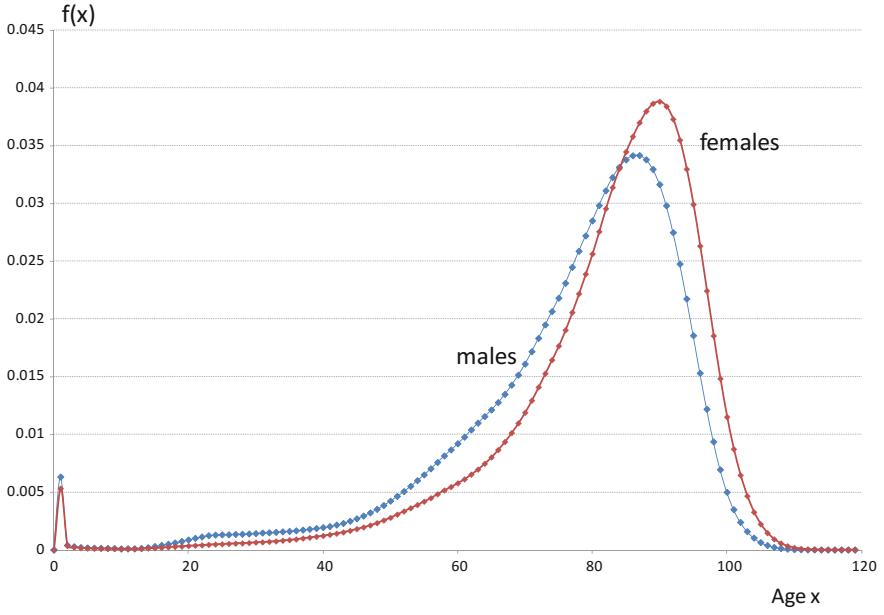


Fig. 4.8 Age mortality densities, US 2014

4.6 Literature Notes

The assistance of Daniel Ullmann and Peter Posch with the material of Sect. 4.3 is gratefully acknowledged.

There is an enormous literature in economics covering both the conceptual foundations and special wrinkles in computing Gini and other proposed welfare indexes. A general review of the whole area can be found in Jenkins and Van Kerm (2009). At the most abstract level, Arrow (1951), Goodman and Arrow (Goodman 1953) showed that the possibility of a consensus based social welfare function as earlier envisaged by Bergson (1938) and Samuelson (1947), was limited at best. Kraemer (1987) reviews the general axiomatic basis. As an extension, Sen (1970) proposed a metric that combines in itself both the mean income and the complement (1-G) of the Gini index.

Alternative or supplementary indexes compare distributions over time or location in terms of a user assigned inequality aversion parameter. Formulations of different kinds can be found in Atkinson (1970), Donaldson and Weymark (1980) and Yitzhaki (1983), Greselin and Zitikis (2015). The Atkinson inequality aversion coefficient is defined by

$$I = 1 - \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^{1-\varepsilon}\right)^{1/(1-\varepsilon)} \quad ;\ 0<\varepsilon<1$$

$$= 1 - \frac{1}{\mu}\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad ;\ \varepsilon = 1$$

Increasing aversion to inequality corresponds to $\varepsilon \to 1$. The Atkinson index does satisfy a number of axioms seen as desirable for such purposes, notably the Robin Hood transfer and subgroup decomposability (see Sect. 4.2).

Of the other indexes, Theil (1965) proposed Shannon entropy as a metric. For a population of size N with mean income μ, the Theil index is defined by

$$\mathrm{I} = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i}{\mu} \ln\left(\frac{x_i}{\mu}\right).$$

The Theil index is a point of departure for number of other contributions utilising Shannon entropy; thus Shorrocks (1980), also Foster et al. (1984). Finally, on practical aspects of all the above, including sampling theory, see e.g. Deltas (2003), Giles (2004).

The application of the *v, d* metrics to income distribution was first proposed in Bowden (2016a, b).

A comprehensive discussion of the income distribution changes in Europe that took place over the GFC can be found in Bowden et al. (2018), from which Figs. 4.3 and 4.4 are sourced. Some quite radical differences exist as between their respective v-d phase planes over the period, with some countries experiencing reduced asymmetry and spread, with others going quite the opposite way.

Turning to Sect. 4.3 on measuring stock market performance, standard measures are typically based on the Sharpe—Lintner CAPM model of market equilibrium (Sharpe 1964; Lintner 1965); thus Sharpe (1966). The CAPM model in the standard version is not well supported by empirical research, but remains highly influential if only as a starting point. Thus extensions have sought to add one or more general factors to the equilibrium pricing model. However they do not explicitly address the observed asymmetry of returns in many cases, even in relatively normal times (e.g. Fama 1965; Chunhachinda et al. 1997). If returns are asymmetric over any given interval, then extraordinary exposures to loss, or else opportunities for gain, can arise. Kraus and Litzenberger (1976) proposed skewness as a second factor in the traditional CAPM formula. Ang and Chua (1979) use this modification in order to define an excess return performance measure. For other approaches incorporating skewness, see Eling and Schuhmacher (2007), Farinelli et al. (2008).

A more extensive treatment of the empirical performance of the W metric relative to that of Sharpe, from which Fig. 4.6 is sourced, may be found in Bowden et al. (2017).

With reference to Sect. 4.4, actuarial theory and practice evolved somewhat independently of the general body of statistics and stochastic processes. As a result, it can be hard for generalists to read. This is not helped by some of the notation: as an instance, ${}_{t|k}q_x$ meaning the probability that someone aged x will survive for t more years, then die within the following k years. Further details and methodology can be found in number of actuarial textbooks, including those sponsored by the professional actuarial societies. Examples are Bowers (1997), Hickman et al. (1997), and Promislow (2011).

References

Ang, J. S., & Chua, J. H. (1979). Composite measures for the evaluation of investment performance. *Journal of Financial and Quantitative Analysis., 14,* 361–384.

Arrow, K. J. (1951). *Social choice and individual values*. Yale: University Press.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory, 2,* 244–263.

Bergson, A. (1938). A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics*, *52*, 310–334.

Bowden, R. J. (2016a). Giving Gini direction: An asymmetry metric for economic disadvantage. *Economics Letters, 138,* 96–99.

Bowden, R. J. (2016b) Dual spread and asymmetry distribution metrics based in partition entropy. Kiwicap Research. Retrieved from http://www.wellesley.org.nz/papers_public.asp.

Bowden, R. J., Posch, P. N., & Ullmann, D. (2017). Asymmetry and performance metrics for equity returns. Technical University of Dortmund, Faculty of Business, Economics and Social Science Discussion Paper SFB 44227.

Bowden, R. J., Posch, P. N., & Ullmann, D. (2018). Income distribution in troubled times: Disadvantage and dispersion dynamics in Europe 2005–2013. *Finance Research Letters, 25,* 36–40.

Bowers, N. (1997). *Actuarial mathematics* (2nd ed.). Schaumberg Illinois: Society of Actuaries.

Chunhachinda, P., Dandapani, K., Hamid, S., & Arun Prakash, P. (1997). Portfolio selection and skewness: Evidence from intemational stock markets. *Journal of Banking & Finance, 21,* 143–167.
Deltas, G. (2003). The small sample bias of the Gini coefficient: Results and implications for empirical research. *Review of Economics and Statistics, 85,* 226–234.
Donaldson, D., & Weymark, J. A. (1980). A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory, 22,* 67–86.
Eling, M., & Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking & Finance, 31,* 2632–2647.
Fama, E. F. (1965). Portfolio analysis in a stable Paretian market. *Management Science, 11,* 404–419.
Farinelli, S., Ferreira M., Rossello D., Thoeny M. & Tibiletti L. (2008). Beyond the Sharpe ratio: Optimal asset allocation using different performance ratios. *Journal of Banking and Finance*, *32*, 2057–2063.
Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica, 52,* 761–766.
Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: Some further results. *Oxford Bulletin of Economics and Statistics, 66,* 1468–1484.
Goodman, L. A. (1953). Social choice and individual values by Kenneth Arrow. *American Sociological Review*, *18*, 116–117.
Greselin, F., & Zitikis, R. (2015). Measuring economic inequality and risk: A unifying approach based on personal gambles, societal preferences and references. arXiv:1508.00127.
Hickman, J. C., Gerber, H. U., Nesbitt, C. J., Jones, D. A., & Newton, L. (1997). *Actuarial Mathematics, 2*. Schaumberg Illinois: Society of Actuaries.
Jenkins, S. P., & Van Kerm, P. (2009). *The measurement of economic inequality*. Oxford: Oxford University Press.
Kraus, A., & Litzenberger, R. (1976). Skewness preference and the valuation of risky assets. *Journal of Finance, 31,* 1085–1100.
Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics, 47,* 13–37.
Promislow, S. D. (2011). *Fundamentals of actuarial mathematics* (2nd edn.) New York: Wiley.
Samuelson, P. A. (1947). *Foundations of economic analysis*. Harvard Economic Studies. Harvard: University Press.
Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden Day.
Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19,* 425–442.
Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica* 48, 613–25.
Theil, H. (1965). The information approach to demand analysis. *Econometrica, 33,* 67–87.
Yitzhaki, S. (1983). On an extension of the Gini inequality index. *International Economic Review, 24,* 617–628.

Chapter 5
Binary Perspectives for Spread and Asymmetry

5.1 Introduction

The spread and asymmetry metrics developed in foregoing chapters could be regarded a parametric in nature, expressed as they are in terms of first moments. The agenda of this chapter is a nonparametric approach that seeks a conceptual condensation of the partition function, with two agendas in mind. The first continues on from previous themes of finding minimally descriptive measures for location, asymmetry and spread. The second associated agenda seeks a more far reaching decomposition such that the total entropic spread could be regarded as generated by just two polar outcomes.

An early approach to non parametric metrics for spread and asymmetry invoked the relationship between the orginal distribution function and its entropic shifts, the latter summarised in terms of the centred shift. The relationship between the corresponding partition entropy functions leads to complementary upper and lower intersection points associated with invariant distribution function values. The metrics for spread and asymmetry could then be defined in such terms. Section 5.1 is an exposition of this approach.

A subsequent approach developed in Sect. 5.2 is more far reaching, and has been developed with decision theoretic applications in mind. Following a well trodden path in the theory of finance in particular, it takes the form of a reduction of the original density to one using an equivalent probability measure with respect to which decision making is less complex. In the more promising of these, the condensation takes a bipolar form, so that the original distribution is entropically equivalent to an equally weighted combination of just two Dirac delta distributions: e.g. the 'bad' and the 'good' states. One can then balance up the two poles in terms of risk and reward. A consequence of this development is that the distance between these binary outcomes, or polar points, can be used to construct spread and asymmetry metrics that are non parametric in form.

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_5

Section 5.3 enlarges on the idea of polar perspectives. Illustrative applications are to income distribution (poor versus affluent) and financial investment (success versus failure). In the investment context it would be as though the investor acts as though there are now just the two polar outcomes, good and bad. An accompanying indication of entropic dominance, in this case of one investment over another, leads to a further point of contact with stochastic dominance in general.

Some applications follow. Section 5.4 revisits the issue of fund performance in finance. A debate here has concerned whether hedge funds, which charge expensive management fees, have in reality done better than merely investing in an exchange traded fund (ETF) that simply tracks the general index such as the S&P500. A surprising conclusion emerges: hedge funds in general can be regarded as more defensive than the general market index, not at all the aggressive outperformers commonly claimed in their publicity.

Section 5.5 revisits actuarial uncertainty. The polar entropic decomposition and associated asymmetry metric indicates that US males as a demographic group exhibit a longer right hand tail to their times of death than do females. More precisely, the mortality function for females is shifted bodily to the right relative to that for males. It is also more condensed; the polar asymmetry metric is smaller than it is for males.

The chapter concludes with the literature review as Sect. 5.6.

5.2 A First Approach Based on Spanning

The metrics *v, d* as proposed in Chap. 4 are not the only measures of asymmetry and spread that could be devised in the entropic context. In particular, since they are technically constructed in terms of means, they could be regarded a semiparametric in nature. It might therefore be useful to have a non parametric version, where such a requirement is the current focus of interest.

An early suggestion was to identify points such that spanned equal values of the partition entropy. So if $h(x_b) = h(x_a)$ where x_a, x_b were respectively above and below the median (denoted here as m), one could use the difference $x_a - x_b$ as a spread measure and the quantity $\frac{1}{2}(x_a + x_b) - m$ as a measure of asymmetry. Because there are an infinite number of point pairs x_a, x_b that would satisfy such a condition, it would become necessary to choose a particular pair that would tell us a little more about the nature of the asymmetry.

One such point of departure is to use as a starting point the centred shift of Chap. 2, which averages out the left and right unit entropic shifts. From the definition of the centred density, any point x^* where the natural $f(x)$ and centred $f_c(x)$ densities intersect must satisfy the condition

$$\xi_c(x^*) = -0.5 \ln[F(x^*)(1 - F(x^*))] = 1,$$

with $\xi_c(x)$ as the Radon-Nikodym derivative that generates the centred shift (Sect. 2.1). This yields two solutions $x_a*,\ x_b*$ to $F(x^*) = 0.5(1 \pm \sqrt{1 - 4e^{-2}})$ with common value for partition entropy at $h^* = 0.441948$.

Figure 5.1 illustrates with a unit scale Gumbel distribution. The two x-values $x_b* = -0.601,\ x_a* = 1.735$ correspond to invariant probabilities of 16.138% for the left and right hand distribution tails, invariant in the sense that they apply to any distribution. It is of passing interest to note that they are very close to the one-sigma tail of the standard normal distribution, which is 15.866%.

To further motivate the connections with spread and asymmetry, one could imagine some perturbation of the original distribution function F that lengthened the right hand tail at point $A = a$, while leaving the left hand tail probability of point $B = b$ unchanged. This amounts to an increase in partition entropy at B relative to A. To preserve equality, point B must move to the right. An increased value of $a - b$ signals the higher spread; likewise the degree of positive asymmetry would also rise.

The resulting metrics for spread and asymmetry do indeed satisfy some intuitively useful properties. Thus for any scalable distribution (standardisable in terms of central location and scale), the spread $a - b$ in terms of the unstandardised x will be $\Delta = \beta\tilde{\Delta}$, and hence proportional to the scale parameter β. The approach in general could be regarded as supplementary to classical metrics such as the inter-quartile range.

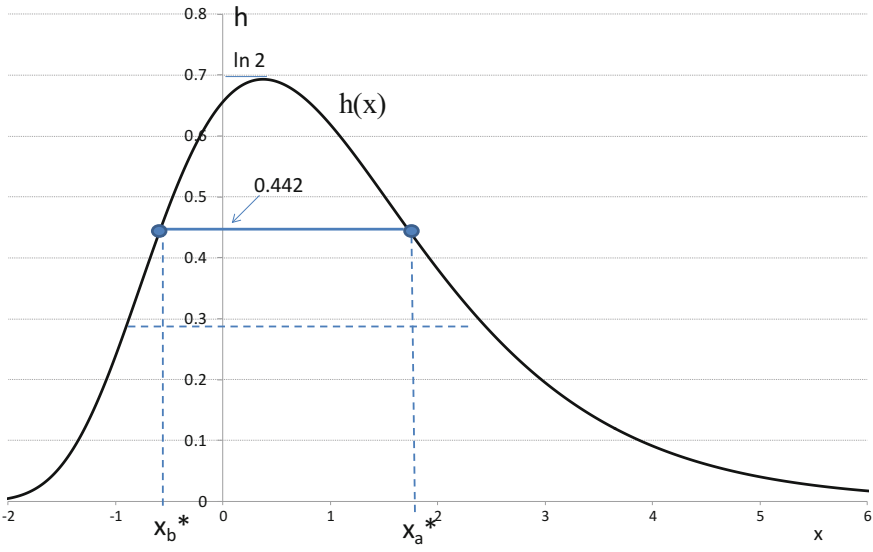


Fig. 5.1 Centred metric spread

On other hand, metrics of this kind do ignore important information, such as the remaining lengths of the distribution tails beyond the limit a on the upside, and of b on the downside. Perhaps more importantly, they lack any decision theoretic content: it is not immediately apparent just how one would use the resulting metrics. Measures of more substantive content are developed in what follows, that in turn call upon the entropic asymmetry and spread metrics v and d, viewed as descriptive measures.

5.3 Bipolarity: The Entropic Centre and Equivalent Width

The methodology that follows establishes points of equivalent entropic concentration such that the total partition entropy of the given distribution can be conceived of as equivalent to a mixture of two point densities located at the respective upper and lower width boundaries. This can have interpretive benefits as location points for summary binary distinctions (e.g. 'conservative' or 'liberal', or 'poor' versus 'affluent'). Distinctions of this kind are rarely absolute, but instead are comparative in nature.

A starting point is to consider what happens with densities that have been constructed in mixture distribution fashion by combining two elementary delta type densities. An approach to the latter is depicted in Fig. 5.2. Two normal densities are respectively centred at means $\mu = -1,\ 1$ with identical standard deviations of $\sigma = 0.05$. In the limit as $\sigma \to 0$, the respective densities become Dirac delta densities, with their partition entropy functions becoming a pair of vertical lines, each preserving the maximum of *ln2* at their respective medians.

Now consider what happens when two such distributions are combined as mixture distributions with equal weightings. Figure 5.3 illustrates with such a mixture, constructed as a combination of the two independent normal densities of Fig. 5.2, with polar means at (-1,1) but with different standard deviations for the components, respectively 0.05 for mixture A and 0.5 for mixture B. In other words,

$$f_A(x) = \frac{1}{2}n(x;\ 1,\ 0.05) + \frac{1}{2}n(x;\ -1,\ 0.05);$$
$$f_B(x) = \frac{1}{2}n(x;\ 1,\ 0.5) + \frac{1}{2}n(x;\ -1,\ 0.5)$$

If the standard deviations in such mixtures are allowed to tend to zero, this effectively generates bipolar Dirac type delta densities with all mass concentrated at just two points. The mixture partition entropy function $h(x)$ becomes asymptotically a rectangular box, of width in the present case as $w = 2$, and invariant height *ln2*. If w is the width of the box then its area would be $w \ln 2$. As the area underneath $h(x)$, its entropic spread (as in Sect. 3.4) is therefore $d = w \ln 2$.

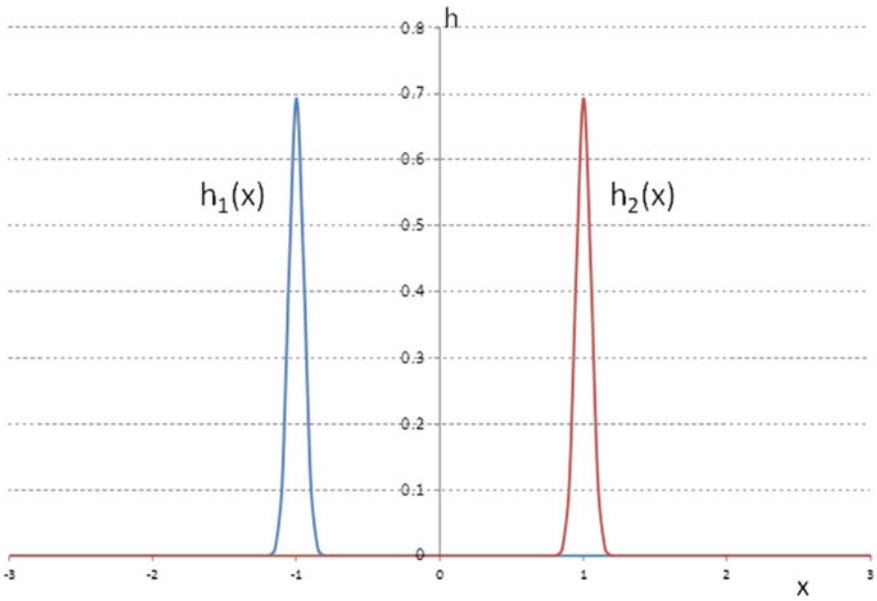


Fig. 5.2 Delta approach densities

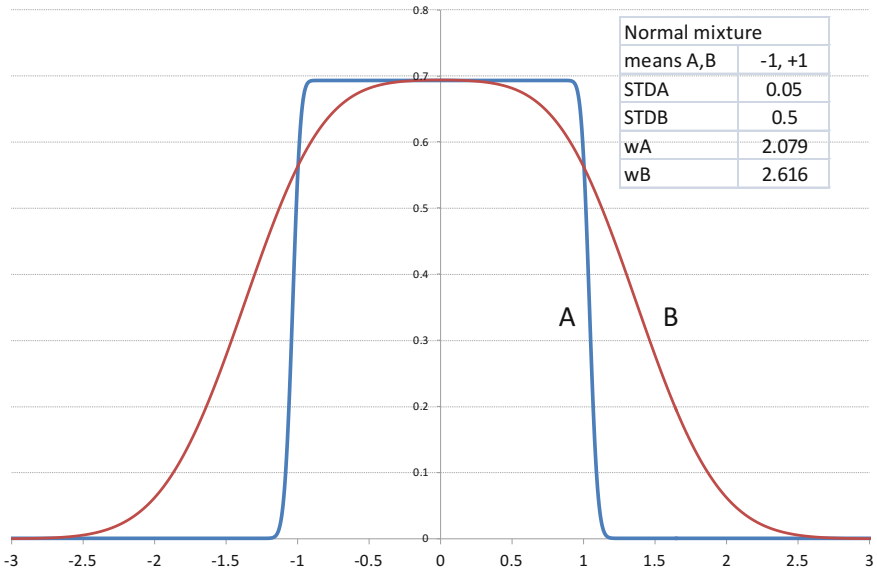


Fig. 5.3 Partition entropy functions, normal mixture

Turning to the general case with an arbitrary distribution function $F(x)$, suppose the value of the entropic spread metric is d as in Sect. 3.4, which is also equal to the total area underneath its partition entropy function $h(x)$. An equivalent width metric can be defined as $w = d/\ln 2$. This in turn will suffice to establish unique bounds for an interval of that width, such that the partition entropy function takes equal values at those bounds.

More formally, suppose $F(x)$ is any continuous distribution for which the partition entropy function $h(x)$ is integrable over the given domain. Given the value d as the entropic spread metric, let $w = d/\ln 2$. Then there is a unique interval $(x_L,\ x_U)$ with $x_U = x_L + w$ and $F(x_U) = 1 - F(x_L)$. The quantity w satisfies the interval dimension requirement for a width metric.

This result can be proved in different ways, but a reasonably concise demonstration is via the mean value theorem of integral calculus. For any $x < m$, the median, let $h = h(x)$. Then there exists $\tilde{x} > m$ such that $h(\tilde{x}) = h$. Conversely, any number $0 < h < \ln 2$ corresponds to a unique width dimension $W(x) = \tilde{x} - x$ such that $h(x) = h$. Thus for notational brevity, set $W(h) = W(x)$. The area underneath the partition entropy function can then be expressed as

$$d = \int_0^{\ln 2} W(h)dh = \int_{-\infty}^{m} W(x)h'(x)dx.$$

As $W(x), h(x)$ are both continuous, then using the second mean value theorem of integral calculus, together with $h(m) = \ln 2$, we must have

$$d = W(x_*) \int_{-\infty}^{m} h'(x)dx = W(x_*) \ln 2,$$

for some value $x_* < m$. Setting $w = W(x_*)$ then from the definition of W$(\cdot)$, the limits $x_L = x_*, x_U = x_* + w$ will be such that $x_L < m$, $x_U > m$ with $F(x_U) = 1 - F(x_L)$.

The resulting width w is not the same as that portrayed as the distance $a - b$ in Fig. 5.1. It is in fact wider than the latter, indicated on that diagram as the horizontal hatched line. Thus the two metrics, namely the entropic width and the spanning measure of Sect. 5.1, are not equivalent.

The average or midpoint of the two extreme points $x_c = 0.5 * (x_L + x_U)$ partitions into half the effective total entropy. This point could accordingly be called the centre of entropy of the given distribution. The degree to which it differs from the median, as $x_c - m$, is a measure of the degree to which uncertainty is generated more to one side of the range than the other; in effect, a metric for entropic asymmetry. For any symmetric distribution it is zero, so the common mean and median is also the centre of entropy. However this need not be true for asymmetric distributions. For the unit Gumbel distribution of Fig. 5.4, the median $m = 0.367$ but $x_c = 0.745$, which together with the single mode indicates more entropic uncertainty on the right hand side of the median.

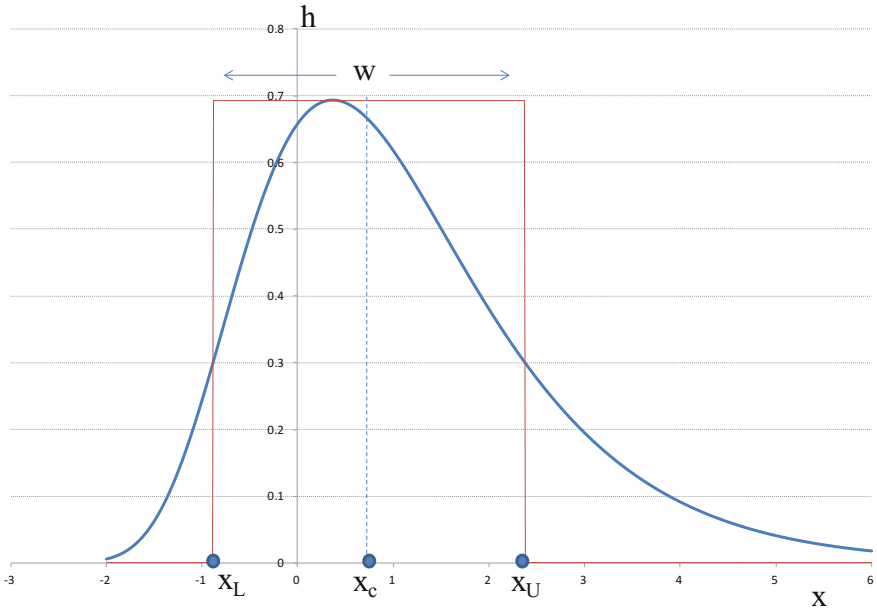


Fig. 5.4 Spread and polarities for the unit Gumbel distribution

Turning to operational matters, a two stage process is involved: (a) to find d and hence $w = d/ln2$; and (b) to solve for the limits $x_L,\ x_U$. Comments follow on both aspects.

(a) In some cases an analytical solution is available for d. Thus for a uniform distribution over the range $[0, N]$, $d = N/2$. A useful general relationship in other contexts is $d = \mathrm{cov}(x, \lambda(x))$, where $\lambda(x) = \ln\left(\frac{F(x)}{1-F(x)}\right)$ is the log odds function that $X \leq x$ versus $X > x$. Thus for a unit scale logistic distribution, the log odds function is linear, resulting in $d = \pi^2/3 \approx 3.289$. If an analytical solution is not available, a numerical integration of the function $h(x)$ can be used to find d and hence the entropic spread w.

(b) If the density $f(x)$ is symmetric, then the width interval can immediately be established as $x_L = m - w/2;\ x_U = m + w/2$. If the density is asymmetric, a convenient computational algorithm for x_L, x_U is to find the value $x < m$ that minimises $abs\,(h(x+w) - h(x))$, then set $x_L = x;\ x_U = x + w$. This converges quickly in Excel using *Solver* or *Goalseek*, provided the initial value for the iteration is a reasonable guess.

Of the common scalable two parameter distributions, the normal distribution has for unit scale the smallest spread (w), consistent with its differential entropy minimising properties among this class. The logistic has a wider entropic spread,

reflecting the longer tail at each end. The Gumbel shows up as positively skewed according to the entropic metric $x_c - m$, consistent with the density shape.

The required steps can be summarised as follows:

1. Given $F(x)$ find its left and right entropically shifted means μ_L , μ_R and hence or otherwise its entropic spread $d = \mu_R - \mu_L$;
2. Calculate its width as $w = d/\ln 2$;
3. Find $x = x_L$ to minimise $abs[h(x+w) - h(x)]$;
4. Set $x_U = x_L + w$.

5.4 Polar Asymmetry and Spread Metrics

With reference to Fig. 5.4, the rectangle with base as the entropic width w and height as *ln2* can be viewed as the partition entropy function for an equally weighted mixture of the two delta distributions respectively centred at $x_L,\ x_U$. In this sense, the variation inherent in the original distribution function (here, the unit Gumbel) can be viewed as homologous with just two polar outcomes at $x = x_L,\ x_U$ embodying the respective Dirac binary equivalents $\delta(x - x_L)$ and $\delta(x - x_U)$.

Thus in the social or economic domains, the two points x_L, x_U could be regarded as summary polar attitudes: 'conservative' or 'liberal', or 'poor' versus 'affluent'. Such references change in their numerical magnitudes over time, much as definitions of 'poor' are relative to those better off and not absolute in themselves. In such terms, social aversion to an increasing spread of an income distribution might be less if the lower point x_L increased along with the upper x_U; in this sense, the gains are more equally shared. The end points of the entropic width measurement can therefore convey information additional to the dispersion figure.

Figure 5.5 illustrates within the context of income distribution for the US, in this repect supplementing the income distribution discussion of Sect. 4.2. The lower entropic limit $x_L \approx \$14,500$, while the upper is given by $x_U \approx \$117,500$. An observer would certainly be inclined to characterise the former as 'poor' and the latter as 'affluent'.

Similar remarks might apply in the context of financial performance and portfolio selection, as in funds management or equity investment. A premise would be that the investor acts as though he or she replaces the true probability distribution of returns with its partition entropy function as a decision basis. A first possible justification arises from behavioural economics, according to which investors tend to overweight small probabilities of extreme outcomes. Thus a lottery is an unfair gamble—it has to be, to pay for the costs of running it—but the consequences of winning are life changing. Correspondingly, the partition entropy function overemphasises longer tails, relative to the parent distribution function. A second possible justification arises from a more formal investor utility function of the Friedman-Savage type, which is concave downward in the negative zone and

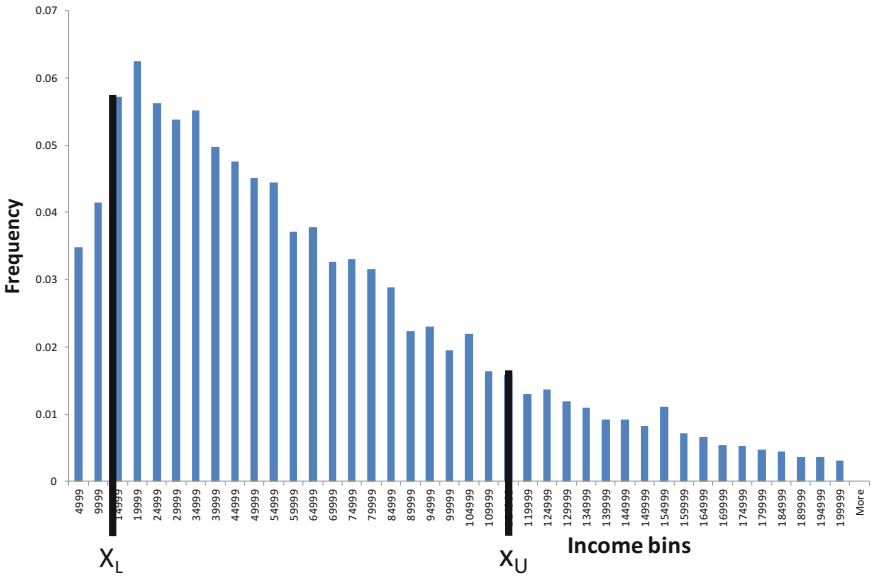


Fig. 5.5 Poor versus affluent: US income distribution 2013

convex upwards for positive outcomes. Again, this would cause investors to overweight smaller probabilities over a longer distribution tail.

As a decision aid, one could then replace the parent partition entropy function, with just the two polar outcomes at x_L, x_U. Analogies exist in the theory of derivatives pricing, where the original probabilities are replaced by an equivalent risk neutral measure. The investor could then act as though he or she was risk neutral with regard to this simplified set of outcomes. The expected outcome is then just

$$x_c = \frac{1}{2}(x_L + x_U).$$

Portfolio A would then be preferred to B if $x_{c,A} > x_{c,B}$.

In placing more weight on a longer right hand tail, one can expect the entropic centre to be greater than the mean for such distributions. Thus for a right handed unit scale Gumbel as in Fig. 5.4, the mean $\mu = 0.5772$ but the entropic centre is $x_c = 0.7450$. The two polar extremes are $x_L = -0.8836$ and $x_U = 2.3735$.

In a financial decision context, the two extremes have a supplementary role as representative 'bad' and 'good' outcomes. If the median $x_m = m$ an associated asymmetry metric is

$$x_c - m = \frac{1}{2}(x_L + x_m) - m. \tag{5.1}$$

A negative value for this indicates negative skewness. Possible applications also exist to decision theory. In such terms, one could imagine an investor with current wealth x_0 condensing the entropic uncertainty into just the two representative outcomes x_L for unfavourable and x_U for favourable, with equal probabilities. A proposed investment with outcome x for wealth would be judged worthwhile if $x_c > x_0$. Such an investor could be considered as risk neutral with respect to the equivalent entropic measure.

A further analogy exists with conventional stochastic dominance. Given two distributions A and B, if both $x_{A,L} \geq x_{B,L}$ and $x_{A,U} \geq x_{B,U}$ then one could say that A is entropically displaced relative to B, in the sense that uncertainty is focussed on higher outcome values along the common range. A correspondence with conventional stochastic dominance flows via the polar distribution outcomes. Given the two polar densities $\delta(x - x_L),\ \delta(x - x_U)$ their equally weighted combination leads to a distribution function $\tilde{F}(x;\ x_L,\ x_U)$ that has just two steps at $x = x_L,\ x = x_U$. In such terms, entropic dominance of A over B would be equivalent to first order stochastic dominance such that $\tilde{F}_A$ is FSD over $\tilde{F}_B$.

5.5 Fund Performance Measures

The performance of hedge funds and other activist investment vehicles has recently been the subject of adverse media commentary and falling investor confidence, with a switch to index tracking exchange traded funds (ETF's) and similar more passive vehicles. However, a proper comparison, given the objectives of activist vehicles, needs to take into account non standard distributional shapes, specifically those with extended or asymmetric tails.

In this respect two considerations have been overlooked in the comparison between activist and ETF investment vehicles. The first is the need to assess over a time scale sufficient to encompass bad as well as good times, while the second is the issue of just what measure should be used for the purpose. A simple historical mean return is often used, yet is arguably inappropriate for high risk- high reward vehicles, with their high net worth clientele better able to bear losses. A similar objection can apply to alternatives such as the Sharpe ratio or Jensen's alpha, if only because hedge funds, and the assets they invest in, are often not publicly traded. More structured alternatives for using classical third or fourth order moments for fat or asymmetric tails have been proposed in the literature.

The illustration that follows applies the bipolar equivalent methodology of Sect. 5.2 to a comparison of hedge fund returns with those on the S&P500 index. Hedge funds are not in general publicly traded (although one very recent 'fund of funds' that invests in hedge funds has achieved listing on a traded basis). They are instead unit trusts, that in bad times may have a withdrawal penalty, or even a hold, on investor withdrawal. They are intended to cater for high net worth clients, who are judged more capable of absorbing possible losses, though in recent times the proliferation of competing offerings has de facto lowered this investment hurdle.

But especially given their high management fees, a central question has always figured, namely whether such investment vehicles are really worthwhile, compared with the much simpler alternative of investing in an exchange traded fund that captures the returns on a general market index such as the S&P500 index.

In what follows, the monthly hedge fund return index takes the form of the arithmetic mean of funds reporting to the Barclay hedge fund database. It should not be construed as a single fund of funds in its own right. The data span is Jan 1997 to Dec 2016. Figure 5.6a, b illustrate with distribution and partition entropy functions for the Barclay HF index (F_B) and the S&P500 index (F_{SP}) monthly returns, measured as fractions (so 0.05 = 5%). The monthly medians are quite close at 0.078 and 0.098%, respectively. The longer tails of the S&P500 are evident in the corresponding h-functions. In this sense, the returns on the S&P500 are more uncertain than those of the Barclay HF index, even when the extreme negative outlier of the former (February 2009) is excluded.

In the investment context, uncertainty need not have negative connotations, as it can generated by probability mass in the right hand tail as well as the left. It is more a matter of balancing up one versus the other. The corresponding lower and upper entropic concentration limits x_L, x_U are marked in Fig. 5.6b, together with the respective widths w_{SP}, w_B. For the respective partition entropy diagrams, the rectangle with base x_L, x_U corresponds to the partition entropy function for the equally weighted Dirac mixture, and its entropic area is the same as that of the parent distribution. Table 5.1 summarises relevant measures.

The considerations implicit in Sect. 5.2 apply to the current comparison. Making use of partition entropy in an investment context could be regarded as invoking a behavioural model that replaces the density $f(x)$ as an object of concern with the value of the partition entropy $h(x)$. Thus suppose the point x is on the right hand tail of the return distribution, so a favourable value. If the distribution in question has a long right hand tail, then even though $f(x)$ might be small at the given point, the value of $h(x)$ would be enhanced by the values still appreciably further to the right than the given x. The same consideration applies with a longer left hand tail. In this case the uncertainty-driven investor will mentally inflate the prospective losses, relative to the natural distribution.

Figure 5.7 illustrates. Here the natural distribution function $F(x)$ is compared with the cumulated partition entropy function $H(x) = \int_*^x h(x)dx$ normalised to sum to unity as a distribution analogue. The longer left hand tail for the S&P500 inflates the partition entropy values in this region, relative to the natural probability.

In terms of the entropic certainty equivalent approach as in Sect. 5.2, the Barclay HF index is a clear winner, with a value of 0.76% for x_c (9.12% per annum) versus 0.08% (0.72% p.a.) for the S&P500. If the average US CPI inflation rate of 4.26% over the period is taken as benchmark, the hedge fund index outperforms on real returns. The S&P500 is clearly more exposed to adverse events, notably the full impact in October 2009 of the subprime crisis (later developing into the GFC). Apparent hedge fund returns may well have been buffered at the time by non traded or publicly listed assets in their balance sheets.

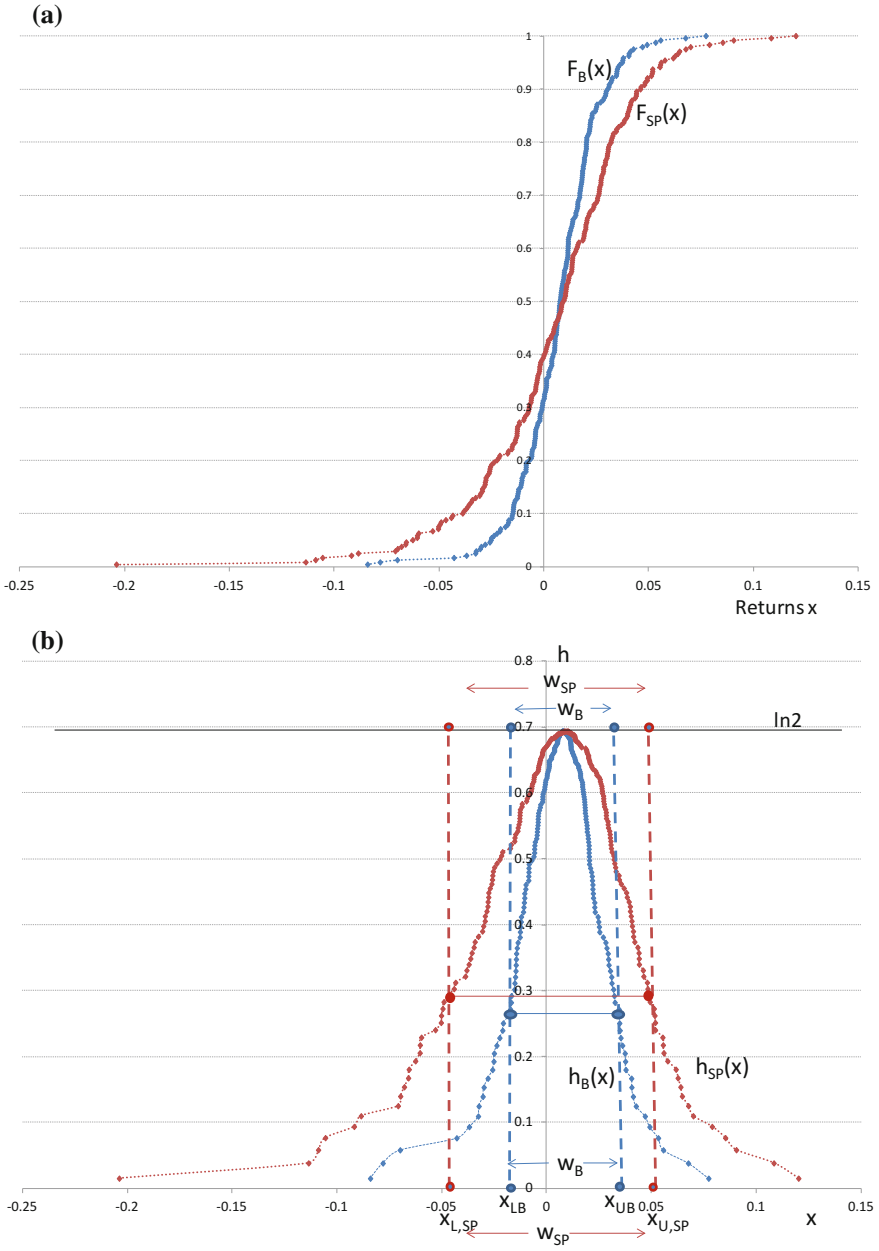


Fig. 5.6 **a** Distribution functions compared, **b** Polar points compared

Table 5.1 Key metrics for the Barclay Hedge Fund versus S&P500 index returns

	Barclay HF	S&P500
xL	−0.0173	−0.0460
xU	0.0326	0.0476
F(xL)	0.0792	0.0840
F(xU)	0.9208	0.9160
xc	0.0076	0.0008
median	0.0078	0.0096
xc-median	−0.0001	−0.0088
mean	0.0070	0.0053

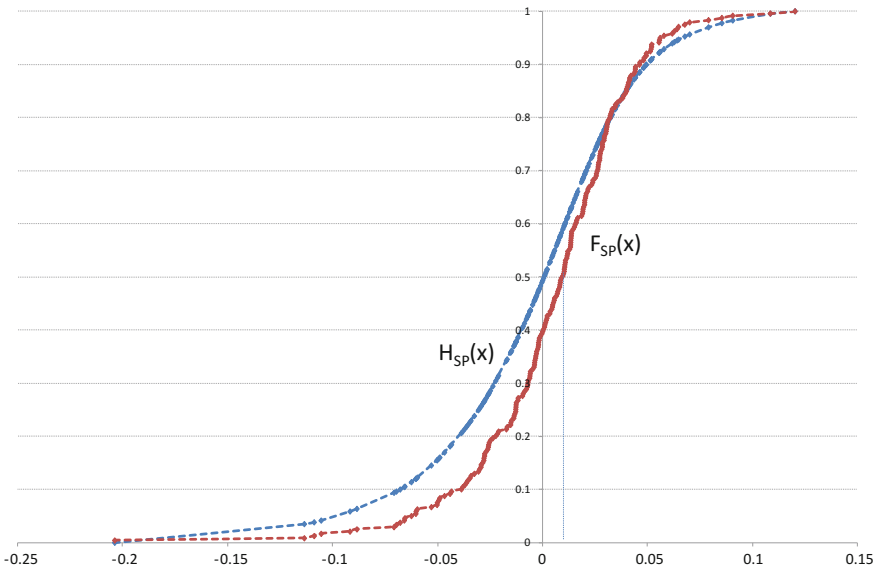


Fig. 5.7 Distribution function and partition entropy comparisons

The median return difference for the Barclay HF versus the S&P500 is in favour of the latter, with value of $0.0078 - 0.0096 = -0.0018$ per month. But the relativities are quite different in term of the entropic centre, at $0.0068 \sim 0.07\%$ per month in favour of the Barclay HF. The corresponding comparison difference for the mean is 0.0017.

In general, such differences may be influenced by the numerical sensitivity to extreme values, as well as distribution shape. The median and entropic centre are computed in terms of the distribution function, with the entropic centre more influenced by the differential tail behaviour on either side of any given value for the median. The arithmetic mean is further influenced by the actual numerical values at the extremes, especially for smaller sample sizes. So far as a predictive use is concerned, the issue remains one of uncertainty attached to any extreme values on

the return scale, whatever the precise magnitude might turn out to be. On such grounds, the use of the entropic centre would be a better guide to the risks or rewards to come.

It will be observed that the binary poles x_L, x_U for the S&P500 returns lie outside those of the Barclay HF index. Thus entropic dominance, in the sense of Sect. 5.2, does not exist. However, the respective lower and upper poles do indicate that the former distribution has longer tails at each end, suggesting that the Barclay HF is second order stochastically dominant with respect to the S&P500. A strict comparison of this kind does require a common mean or median, but even if the S&P500 is translated to the right by the difference in means, Fig. 5.6a would continue to indicate second order stochastic dominance. The problem is that the relative downside penalty $XL_{S\&P500} - XL_{BHF} = -0.127$ is not sufficiently compensated by the upside gain $XU_{S\&P500} - XU_{BHF} = 0.015$. This is consistent with the relative shape of the distributions. The respective values of $x_c - m$, where m is the median, indicate that whereas the S&P500 is almost symmetric, the S&P500 distribution is negatively skewed. An investor with Friedman-Savage type utility would not find the upside potential of the latter sufficient to counter the negative downside.

In summary, the differential responsiveness of the entropic centre to tail probabilities carries a potential for independent behaviour relative to classic measures of central tendency or dispersion. Thus one could imagine a switch in the distribution from negative to positive asymmetry that would leave the median, mean and standard deviation as before. But a significant change in the entropic centre x_c would signal the change in shape and invoke differential Friedman-Savage type preferences.

The point is relevant in considering the signalling content of the VIX index, which is based on backing out the implied volatility from current options prices. Published by the Chicago Board of Exchange, the VIX index is itself traded as an indicator of current market uncertainty. Textbook Black-Scholes pricing models indicate that a rise in the backed out σ could well be theoretically neutral in its effect on underlying physical stock prices, where the latter are based on the future expected mean return. On the other hand, the Black Scholes options pricing model does not price out of the money strike prices very well, which suggests that the market might be very concerned about directional uncertainty. If that is the case, then a fall in the entropic centre x_c would be accompanied by a fall in the underlying stock price. The signalling content of the VIX for the underlying stock prices therefore has to be considered with reference to the framework of directional uncertainty.

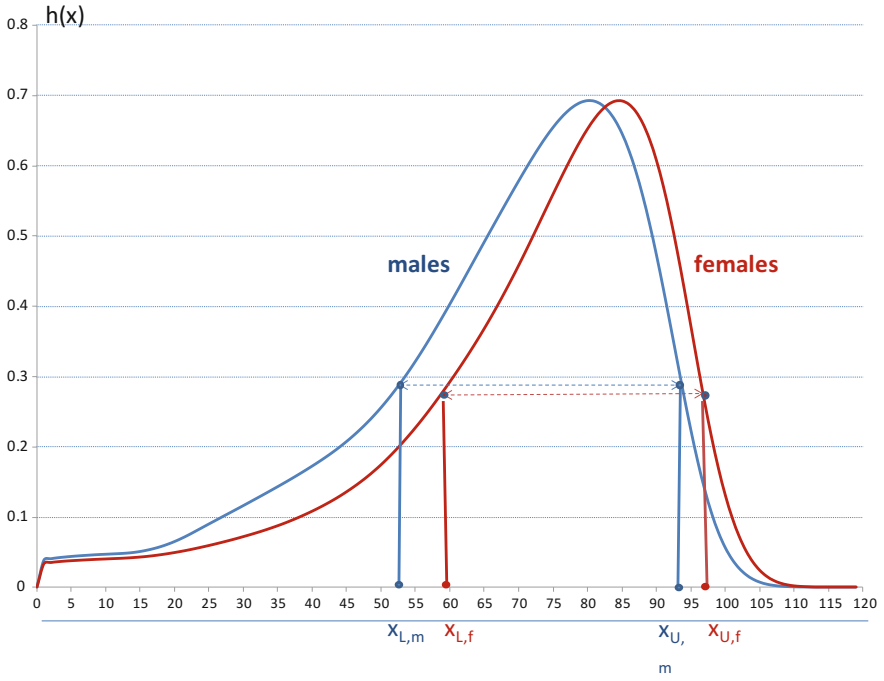


Fig. 5.8 US life table 2014, partition entropy function for mortality distributions

5.6 Actuarial Uncertainty Revisited

In an actuarial context, the respective partition entropy functions can be associated with the age-specific probabilities of surviving (U) or not (D). Continuing the discussion of Sect. 4.4, Fig. 5.8 depicts the partition entropy functions that correspond to a smoothing of the US probability distributions of ages at death, or males and females, respectively.

Also depicted are the polar entropic concentrations. For the males these are located at $x_{Lm} = 52.3$, $x_{Um} = 93.8$ years and for the females at $x_{Lf} = 59.0$, $x_{Uf} = 96.6$. The respective medians are 80.7 for males and 84.3 for females. So the respective polar entropic asymmetry metrics (5.1) are (−7.75) for males and (−6.5) for females. Both exhibit negative skewness, but more so for males. The latter have a more substantial left hand tail in their mortality distributions.

As the diagram suggests, the survival distribution for females is first order stochastically dominant over that for males, with the entropic concentration for the females shifted bodily to the right relative to that for the males. There is more uncertainty associated with the time of death for males, with an entropic width of 41.5, compared with 37.6 for females. Indeed, the time of death density for females resembles that for males but shifted entropically to the right. American men are evidently less attractive candidates for life insurance!

5.7 Literature Notes

The polar entropic framework of the present chapter was first proposed in Bowden (2017). The general concern with tail behaviour as a motivational influence has been reviewed in earlier chapters, as has the actuarial framework (Sect. 4.4). An earlier proposal as to the spanning metric of Sect. 5.1 was contained in Bowden (2012).

With respect to Sect. 5.3, more structured alternatives for fat or asymmetric tails can take the form of CAPM extensions (e.g. Kraus and Litzenberger 1976; Bawa and Lindenberg 1977) or as a portfolio selection criterion (Harvey and Siddique 2000; Ang et al 2006). Dittmar (2002) explicitly embeds such considerations into a pricing kernel or stochastic discount factor approach, in this case with possibly incomplete markets (one where arbitrages may not be possible).

The academic literature on hedge funds tends to be sceptical in nature, addressing issues such as performance and reporting bias. Stulz (2007) is a review, though it predates the GFC, at which time the hedge funds outperformed the general equity index. Performance measures in general have been addressed by Liang and Park (2010), while a number of authors consider the issue of how performance incentives of different kinds contribute to (the substantial) mortality risk in hedge funds.

References

Ang, A., Chen, J., & Xing, Y. (2006). Downside risk. *Review of Financial Studies, 19,* 1191–1239.

Bawa, V. S., & Lindenberg, E. B. (1977). Capital market equilibrium in a mean-lower partial moment framework. *Journal of Financial Economics, 5,* 189–200.

Bowden, R. J. (2012). Information, measure shifts and distribution metrics. *Statistics, A Journal of Theoretical and Applied Statistics, 46,* 249–262.

Bowden, R. J. (2017). Distribution spread and location metrics using entropic separation. *Statistics and Probability Letters, 124,* 148–153.

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *Journal of Finance, 57,* 369–404.

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *Journal of Finance, 55,* 1263–1295.

Kraus, A., & Litzenberger, R. (1976). Skewness preference and the valuation of risky assets. *Journal of Finance, 31,* 1085–1100.

Liang, B., & Park, H. (2010). Predicting hedge fund failure: A comparison of risk measures. *Journal of Financial and Quantitative Analysis, 45,* 199–222.

Stulz, R. (2007). Hedge funds: past, present, and future. *Journal of Economic Perspectives, 21,* 175–194.

Chapter 6
Higher Dimensions

6.1 Introduction

Many of the parent concepts of differential entropy generalise to two or more dimensions, with more or less natural extensions such as the entropy of the respective conditional distributions. Likewise, the mutual information between two random variables can be expressed in terms of the reduction in the entropy of the one that would follow by knowing the value of the other. The literature notes of the present chapter provide a short summary of relevant definitions and relationships.

While these concepts provide some useful background, the concern of the present chapter is with metric comparisons, for which a specific framework is more useful. In a single dimension, unit shifts to either left or right are unequivocal with reference to the axis x. But in higher dimensions the directions must now be defined with reference to a plane, such as (x_1, x_2) in two dimensions. However, elementary shifts can be defined with reference to each dimension separately, then combined with weights according to the desired direction in the (x_1, x_2) plane. The elementary shifts themselves can be obtained with Radon-Nikodym shift factors as the logs of the respective conditional distribution functions.

Left and right hand moments can now be defined both for marginal and conditional shifted distributions. However there is additional flexibility, for one can now consider smoothing a given variable with the benchmark set by a second. This is referred to as co-smoothing, and is introduced in Sect. 6.1.

Section 6.2 adapts the preceding development for contexts where a more or less natural weighting exists in the form of a welfare index or similar construct. In this case, one can explore the resulting index changes following distributional shifts in one or both of the constituent elements of the index.

In a bivariate or multivariate context, co-smoothing refers to the conditional expectation of a variable taken over a progressive range of a covariate, instead of its own. The ordered mean difference is of this generic form, exposited in Sect. 6.3. In the context of finance, this is a construction that characterises the risk- return profile

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_6

of a given security of return r against that of the market, denoted R, as a benchmark. In the terms of the core development of Sect. 6.1, the difference $r - R$ corresponds to the first variable x_1, and the benchmark return R correspond to the second, denoted x_2.

The empirical study confirms the findings of Sect. 5.4 that far from being investments of the advertised high risk-high reward profiles, the performance of hedge funds has on the average been defensive rather than aggressive.

The literature notes conclude.

6.2 Higher Dimensions: Directional Shifting Perspectives

In higher dimensional contexts there are two or more (depending on dimensionality) unit shift primary directions, which in turn can generate shifts in any desired direction in the form of linear combinations of the primary shifts. Themes of this sort are the subject of the present section. For expositional convenience the case of just two dimensions is the point of departure.

Thus consider a bivariate distribution function $F(x_1, x_2)$ with density $f(x_1, x_2)$. For expositional brevity, the range of each variate is sometimes indicated with an asterisk standing in for the range as e.g. $(-\infty, \infty)$ or the half axis $[0, \infty)$. To avoid an unwieldy proliferation of notation, the symbol x_1 will be often be taken to indicate either the parent random variable (hitherto X_1) or a specific value $X_1 = x_1$; similarly for x_2.

With this notation, the marginal means of each variable are given by

$$\mu_1 = E[x_1] = \iint_* x_1 f(x_1, x_2) dx_1 dx_2;\ \mu_2 = E[x_2] = \iint_* x_2 f(x_1, x_2) dx_1 dx_2,$$

with the integration taken over the entire double range of x_1 , x_2. Conditional distribution functions are denoted as $F(x_1|x_2) = P(X_1 \leq x_1|X_2 = x_2)$, with density $f(x_1|x_2)$; similarly for $F(x_2|x_1)$ and $f(x_2|x_1)$.

The factor $\xi_{L1}(x_1, x_2) = -\ln F(x_1|x_2)$ is nonnegative and has unit expected value with respect to both the joint density $f(x_1, x_2)$ and the conditional density $f(x_1|x_2)$. With respect to the joint density, it amounts to a Radon-Nikodym derivative that accomplishes a unit shifting along direction 1, i.e. the direction of the x_1 axis. The resulting density can be denoted by $f_{L,1}(x_1, x_2; 1)$, where the subscript indicates the direction of the shifting (leftward L, along the axis for x_1), and the index argument 1 indicates that this is a unit shift.

With the above conventions,

$$f_{L,1}(x_1, x_2; 1) = (-\ln F(x_1|x_2)) f(x_1, x_2) = f_L(x_1|x_2) f(x_2), \tag{6.1a}$$

where $f_L(x_1|x_2)$ refers to a unit left shift of the conditional density. The corresponding distribution function can be written as

$$F_{L1}(x_1, x_2; 1) = \int_*^{x_2} F_L(x_1|X_2) f(X_2) dX_2.$$

The notational convention indicates that this is the result of a unit left shift along the direction of x_1.

Similarly, one can derive a unit leftward shift, but in the direction of variable x_2.

$$f_{L,2}(x_1, x_2; 1) = (-\ln F(x_2|x_1)) f(x_1, x_2) = f_L(x_2|x_1) f(x_1), \tag{6.1b}$$

together with the distribution function as

$$F_{L2}(x_1, x_2; 1) = \int_*^{x_1} F_L(x_2|X_1) f(X_1) dX_1.$$

The extension to unit rightward shifts is homologous with the above. Thus for the densities,

$$f_{R,1}(x_1, x_2; 1) = -\ln(1 - F(x_1|x_2)) f(x_1, x_2) = f_R(x_1|x_2) f(x_2), \tag{6.2a}$$

$$f_{R,2}(x_1, x_2; 1) = -\ln(1 - F(x_2|x_1)) f(x_1, x_2) = f_R(x_2|x_1) f(x_1), \tag{6.2b}$$

with corresponding expressions for the distribution functions.

Shifts can be compound, potentially in any direction relative to the x_1, x_2 plane. Those of potential interest in applications are where x_1, x_2 shift in the same direction, either to the left or right together, though possibly to a different extent. The corresponding R-N derivatives are given by:

$$\xi_L(x_1, x_2; \theta_L) = -(\theta_L \ln F(x_1|x_2) + (1 - \theta_L) \ln F(x_2|x_1)); \; 0 \le \theta_L \le 1;$$

$$\xi_R(x_1, x_2; \theta_R) = -(\theta_R \ln(1 - F(x_1|x_2)) + (1 - \theta_R) \ln(1 - F(x_2|x_1))); \; 0 \le \theta_R \le 1.$$

In turn, these generate densities as

$$\begin{aligned} f_{L,\theta}(x_1, x_2; \theta_L) &= \xi_L(x_1, x_2; \theta_L) f(x_1, x_2) \\ &= \theta_L f_L(x_1|x_2) f(x_2) + (1 - \theta_L) f_L(x_2|x_1) f(x_1). \end{aligned}$$

$$\begin{aligned} f_{R,\theta}(x_1, x_2; \theta_R) &= \xi_R(x_1, x_2; \theta_R) f(x_1, x_2) \\ &= \theta_R f_R(x_1|x_2) f(x_2) + (1 - \theta_R) f_R(x_2|x_1) f(x_1) \end{aligned}$$

The special cases $\theta_L, \; \theta_R = 1$ correspond to the unit shifted joint densities (6.1a, 6.1b, 6.2a, 6.2b), hence the notation of the latter.

Figure 6.1 illustrates a rightward shift of a parent bivariate normal density, with $\theta_R = 0.5$. The distribution is shifted to the right, but the partial nature of the shift distributes its mass a little wider.

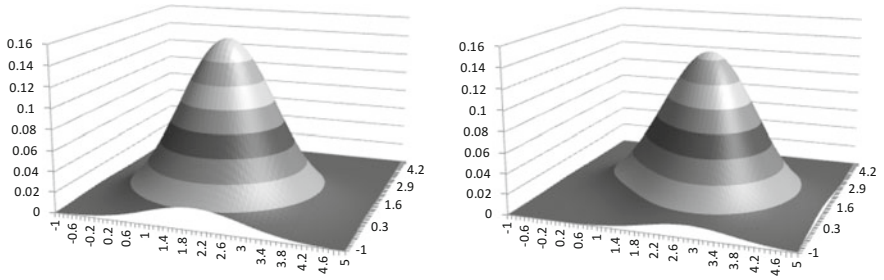


Fig. 6.1 Partial right shift of a bivariate normal density

Distributional moments can be defined with respect to the shifted distributions. Thus the marginal unit left shifted means from the parent joint density are:

$$\mu_{1L} = \iint_{*} x_1 f_{L,1}(x_1, x_2; 1)dx_1dx_2; \quad \mu_{2L} = \iint_{*} x_2 f_{L,2}(x_1, x_2; 1)dx_1dx_2, \tag{6.3}$$

with corresponding conventions for the unit right shifted means μ_{1R}, μ_{2R}. The definition extends naturally to the marginal means for the partial component shifts. So

$$\mu_{1L}; \theta_L = \iint x_1 \xi_L(x_1, x_2; \theta_L) f(x_1, x_2)dx_1dx_2$$
$$\mu_{1R}; \theta_R = \iint x_1 \xi_R(x_1, x_2; \theta_R) f(x_1, x_2)dx_1dx_2.$$

Similarly for $\mu_{1R}; \theta_R$, $\mu_{2R}; \theta_R$.

6.2.1 Left and Right Hand Smoothing Moments

As earlier noted, the same R-N derivatives can also be used with respect to the conditional measures $F(x_1|x_2)$, $F(x_2|x_1)$ to derive unit left or right shifted conditional densities of the form

$$f_L(x_1|x_2) = -\ln F(x_1|x_2) f(x_1|x_2) \tag{6.4a}$$

$$f_L(x_2|x_1) = -\ln F(x_2|x_1) f(x_2|x_1). \tag{6.4b}$$

Similarly for $f_R(x_1|x_2)$,$f_R(x_2|x_1)$ in terms of the complementary R-N derivative factors $-\ln(1 - F(x_1|x_2))$, $-\ln(1 - F(x_2|x_1))$.

Shifts of this kind are essentially those of a univariate distribution, as the respective conditionals. In this respect they contrast with shifts defined in terms of

the parent joint distribution function. Thus the conditional mean of x_1, given x_2, but computed with respect to the left shifted conditional distribution with density (6.4a) can be written as

$$\mu_{1L}(x_2) = \int_* x_1 f_L(x_1|x_2)dx_1. \tag{6.5a}$$

Likewise, the shifted conditional mean of x_2, given x_1, is obtained as

$$\mu_{2L}(x_1) = \int_* x_2 f_L(x_2|x_1)dx_2, \tag{6.5b}$$

with corresponding expressions for shifts to the right.

As in the univariate case, a double smoothing connotation exists. In connection with expression (6.5a), one can write

$$\mu_1(x_1|x_2) = E[X_1|X_1 \leq x_1, x_2].$$

The progressive conditional expectation is with respect to the first variable, while the second remains fixed throughout, essentially only incidental. Then

$$\mu_{1L}(x_2) = E[\mu_1(x_1|x_2)],$$

where again, the expectation is taken with respect to x_1, given x_2.

Similarly for the second variable:

$$\mu_2(x_2|x_1) = E[X_2|x_1,\ X_2 \leq x_2],$$

leading to

$$\mu_{2L}(x_1) = E[\mu_2(x_2|x_1)].$$

Carrying the smoothing one stage further, there is a connection between the marginal means as defined in expression (6.3) and those defined with respect to the shifted univariate conditional distributions. Combining expressions (6.5a, 6.5b) with (6.3) results in

$$\mu_{1L} = E[\mu_{1L}(x_2)];\ \mu_{2L} = E[\mu_{2L}(x_1)].$$

More generally,

$$\mu_{1L}; \theta_L = E_L[x_1; \theta_L] = \mu_1 + \theta_L(E[\mu_{1L}(x_2)] - \mu_1).$$

$$\mu_{1R}; \theta_R = E_R[x_1; \theta_R] = \mu_2 + \theta_R(E[\mu_{1R}(x_2)] - \mu_2).$$

A demonstration of this result uses expressions (6.5a, 6.5b) together with $f(x_1, x_2) = f(x_1|x_2)f(x_2) = f(x_2|x_1)f(x_1)$ to show that

$$\mu_{1L}; \theta_L = \theta \iint x_1 f_L(x_1|x_2) f(x_2) dx_1 dx_2 + (1-\theta) \iint x_1 f_L(x_2|x_1) f(x_1) dx_1 dx_2,$$

Corresponding expressions hold for the right shifted marginal means $\mu_{1R}; \theta$ and $\mu_{2R}; \theta$.

6.2.2 Co-smoothing

Smoothing, or the process of taking progressive conditional expectations, has to this point been cast within the frame of self reflexivity, so that one smoothes a given variable up to its current point of focus. In this process the value of the other variable or variables, remain fixed. For example,

$$\mu_l(x_1|x_2) = E[X_1|X_1 \leq x_1, x_2],$$

so that X_1 is being smoothed based on its own past, while X_2 remains fixed at $X_2 = x_2$.

However with two (or more) variables under consideration, the facility arises of smoothing the first variable over a designated range of the second. This can be referred to as co-smoothing. In this framework, an expression such as $E[X_1|X_2 \leq x_2]$ is in effect a marginal expectation with respect to variable X_1, but confined to the space where X_2 is less than or equal to a given value x_2.

In this case,

$$\begin{aligned} E[X_1|X_2 \leq x_2] &= \frac{1}{F(x_2)} \int_{-\infty}^{\infty} \int_{-\infty}^{x_2} X_1 f(X_1, X_2) dX_1 dX_2 \\ &= \frac{1}{F(x_2)} \int_{-\infty}^{x_2} \int_{-\infty}^{\infty} X_1 f(X_1|X_2) f(X_2) dX_1 dX_2. \end{aligned}$$

This reduces to

$$E[X_1|X_2 \leq x_2] = \frac{1}{F(x_2)} \int_{-\infty}^{x_2} \mu_1(X_2) f(X_2) dX_2 \ . \tag{6.6}$$

Here $\mu_1(X_2) = E[X_1|X_2]$ refers to the conditional expectation of the first variable given a specified value of the second variable. Expression (6.6) then says that one smoothes this up to a specified value $X_2 = x_2$.

Smoothing algorithms of this kind find application in Finance. Section 6.3 is an extended discussion and illustration in the context of the ordered mean difference benchmarking of security returns.

6.3 Index Combinations

In some applications it is useful to recombine two variables into a single index. For instance, income and wealth constitute a natural pairing in the context of economic distribution. If one is prepared to assign relative welfare weights, then a generalised inequality metric can be attached to the resulting index.

Such an index could be interpreted as either a generalised income or a generalised 'economic wellbeing' index, to indicate command over both current and future purchasing power. Indeed, one could interpret wealth in terms of its command over future purchasing power. Similarly, income is the command over current purchasing power, the capitalised value of which over current and future time is equivalent to the purchasing power interpretation of wealth. Older people in western societies may technically be income poor, but on the other hand, asset rich. Accordingly, the weights could reflect the terms on which wealth could be transferred into current income.

Thus if x_1 denotes income and x_2 denotes wealth, such an index for economic welfare might be $y = w_1x_1 + w_2x_2\ ; w > 0,\ w_1 + w_2 = 1$. In the context of income and wealth, the component v_1 would represent the income envy (as in Sect. 4.2) for any given level of wealth, averaged over all wealth levels. Likewise, the component v_2 is the wealth envy for any given level of income, averaged over all incomes. Measures for asymmetry (here, inequality) and for spread are then to be assigned to the welfare index, either as the progressive conditional expectation functions $v(y), d(y)$, or as their metric averages v_y, d_y as in Chap. 3.

If the two variables x_1, x_2 are statistically independent, then the envy metric for y reduces to the weighted sum of the two marginal envy metrics. However, for contexts such as income and wealth the net envy for any given level of income will likely depend upon the subject's wealth. Richer retirees are less likely to be bothered by the excesses of executive remuneration! So the income envy metric is first formulated with respect to a specific level of wealth, followed by an averaging process over both.

Such a process can be conceived of in two stages. First fix x_2 and consider the distribution of X_1 conditional on $X_2 = x_2$. For any given value x_1 the conditional relative is

$$v(x_1|x_2) = \mu_r(x_1|x_2) + \mu_l(x_1|x_2) - 2x_1.$$

In terms of the notation of Sect. 6.1, the conditional expected value, given x_2, is

$$E[v(x_1|x_2)] = \mu_{1R}(x_2) + \mu_{1L}(x_2) - 2E[x_1|x_2]. \tag{6.7a}$$

Similarly,

$$v(x_2|x_1) = \mu_r(x_2|x_1) + \mu_l(x_2|x_1) - 2x_2$$

is the conditional relative for x_2. Its conditional expected value, given x_1, is

$$E[v(x_2|x_1)] = \mu_{1R}(x_1) + \mu_{1L}(x_1) - 2E[x_2|x_1]. \tag{6.7b}$$

The proposed directional relative is the weighted average, with respect to the joint distribution $F(x_1, x_2)$:

$$E_F[w_1 v(x_1|x_2) + w_2 v(x_2|x_1)].$$

Utilising expressions (6.6) in conjunction with the double smoothing property of Sect. 3.2, this reduces to the weighted average

$$v(w) = w_1 v_1 + w_2 v_2,$$

where

$$v_1 = (\mu_{1L} + \mu_{1R}) - 2\mu_1.$$

$$v_2 = (\mu_{2L} + \mu_{2R}) - 2\mu_2.$$

If desired, the metrics v_1 , v_2 can be normalised by dividing by the weighted sum of the marginal means $w_1\mu_1 + w_2\mu_2$ and expressing the result as a percentage. Thus with respect to Fig. 6.1, the partial rightward shift has increased an equally weighted v from -0.70 to 1.84%.

As a final observation, it could be argued that simply averaging the income metric over x_2 constitutes an implicit applied welfare measure with respect to wealth, but otherwise a lack of formal interaction between the two. An extension might be to have income envy explicitly modified by societal welfare weightings of wealth, of the form $\psi(x_2)v(x_1|x_2)$ with $\psi'(x_2) < 0$.

6.4 Co-smoothing and the Ordered Mean Difference

The process of co-smoothing, where a second variable provides the benchmark for the first, has been introduced in Sect. 6.1. The application that follows will illustrate the process and the uses to which the construction can be put. The specific context is a construction for security benchmarking in Finance, referred to as the ordered mean difference. In turn, this is an empirical procedure that attempts to put numbers to a source of comparative gain in adding a given security to a standard benchmark.

More specifically, the equivalent margin is the name given to a constructive relationship between two variables, one regarded as benchmark and the other as a

prospect that may or may not add value to the given benchmark. In Finance, the benchmark is often taken as the return on a market index such as the S&P500, usually regarded as a standard of value or of performance returns. The problem is then how to assess whether a given security adds value to the benchmark, perhaps in conjunction. The resulting ordered mean difference (OMD) calculation, as the embodiment of the equivalent margin, is first exposited on a general level, followed by an application in context to financial returns and their benchmarking.

The OMD construction proceeds as follows. Given a set of observations $R_i,\ r_i\ ; i = 1, 2, \ldots, N$, on a benchmark R and a subject comparison variable r, reorder and tabulate these in ascending order of the benchmark. Then take the progressive moving averages of the differences $r - R$, collectively forming the OMD schedule.

Table 6.1 continues with the hedge fund performance example of Sect. 5.4. In column B, the benchmark has been designated as the S&P500 returns, and has first been re-ordered by increasing values, with the index i reflecting this new order. The comparison variable, denoted r, is the Barclay hedge fund index return and appears in column C. Columns D and E are the differences $r - R$ and the progressive sums thereof, with the last column F as the progressive moving averages, indicated here as $t(P)$ for future reference.

Given a benchmark value $R = P$, the ordered mean difference is then obtained as the last column of Table 6.1, amounting to

$$OMD(P) = t(P) = E[r - R | R \leq P]. \tag{6.8}$$

This can be interpreted in the bivariate framework of Sect. 6.1 by setting $x_1 = r - R;\ x_2 = R$, so x_2 is the benchmark and x_1 is the proposed enhancement

Table 6.1 The OMD construction

A	B	C	D	E	F
Observation number	S&P500	Barclay HF			t(P)
i	R	r	r-R	Progressive sums	Progressive moving average
1	−0.1694	−0.0841	0.0853	0.0853	0.0853
2	−0.1458	−0.0781	0.0677	0.1530	0.0765
3	−0.1100	−0.0124	0.0976	0.2506	0.0835
4	−0.1099	−0.0146	0.0953	0.3460	0.0865
5	−0.0923	−0.0144	0.0779	0.4239	0.0848
6	−0.0908	−0.0699	0.0209	0.4448	0.0741
7	−0.0860	−0.0173	0.0687	0.5134	0.0733
8	−0.0857	−0.0014	0.0843	0.5977	0.0747
9	−0.0820	−0.0323	0.0497	0.6474	0.0719
10	−0.0817	−0.0258	0.0559	0.7033	0.0703

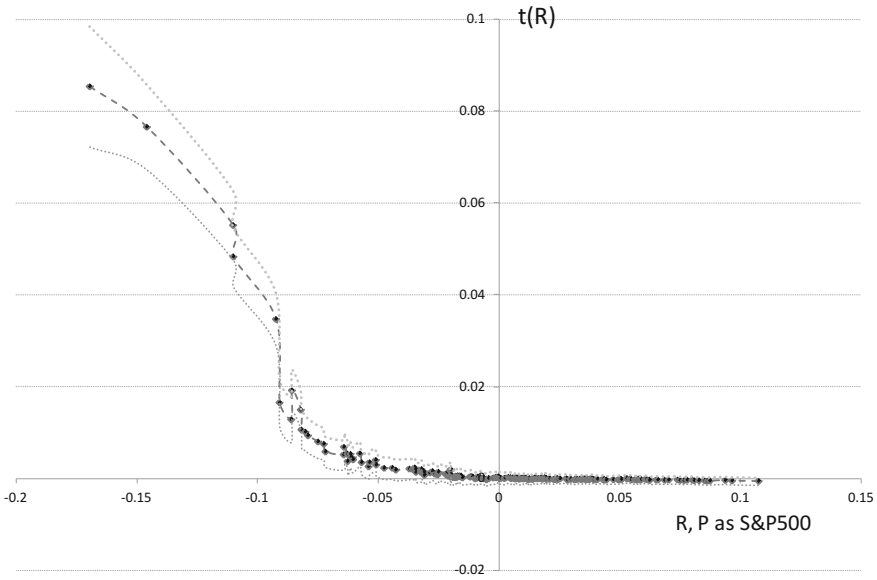


Fig. 6.2 The OMD for Barclay hedge funds index versus the S&P500

difference. Then the OMD of r with respect to R amounts to the smoothed conditional expectation as expression (6.6).

Approximate confidence bands can obtained by estimating a linear regression of regression of r on R, i.e. provisionally assuming $e(R) = \beta_0 + \beta_1 R$, to estimate the standard deviation of the residual: $\hat{\sigma} = STD(\varepsilon)$. One sigma confidence bands can then be added as $t(P_i) \pm \sigma/\sqrt{n_i};\ i = 1, 2, \ldots, N$. Figure 6.2 in the text that follows is an illustration which should be taken in conjunction with the exposition that follows.

6.4.1 *The OMD in Context: Some Financial Decision Theory*

In the context of finance and investments, let the benchmark R take the form of the return on a widely agreed standard of value, such as the US S&P500 index or a world equity market index such as the MSCI-World. Also let r denote the return on a security that may or may not add value to the benchmark R.

More specifically, suppose I have a dollar to invest, and I explore a combined portfolio with x units of r, and $(1 - x)$ units of the benchmark R. The proportion x could be negative (a short position in r) as well as positive (long in r). One way to assess just how much value that security r adds is to imagine a specific tax at rate t, levied on it, though not on the benchmark. In this case the net return would be

$x(r-t)+(1-x)R$. Naturally this will lessen the incentive to hold security r, provided of course that the original intention was to hold it long. One measure of its value is to increase the tax rate t until the investor just decides *not* to hold it, i.e. $x=0$.

Let $U(.)$ be the representative investor's utility function for returns. For any given tax rate t, the investor will choose the proportion x to maximise $U(x(r-t)+(1-x)R)$.The first order condition for this can be written as

$$E[(r-R-t)U'(x(r-t)+(1-x)R)]=0.$$

Setting the tax rate t to just drive the holding x to zero solves as

$$t_u=\frac{E_{r,R}[(r-R)U'(R)]}{E_R[U'(R)]}.$$

The quantity t_u is referred to as the equivalent margin. It is a measure of how much the investor would have to be compensated before he or she gives up the opportunity to invest in security r. It can be written in the form

$$t_u=E_{r,R}[\pi(R)(r-R)],$$

with nonnegative weights $\pi(R)=\frac{U'(R)}{E[U'(R)]}$ that sum to unity. States of the world with greater marginal utility receive more weight.

An equivalent formulation utilises the theoretical regression of r on R:

$$r=e(R)+\varepsilon;\quad E[\varepsilon|R]=0.$$

In this case,

$$t_u=\frac{E_R[(e(R)-R)U'(R)]}{E_R[u'(R)]}.$$

The value t_u is referred to as the equivalent margin for security r over the benchmark standard R.

One particular form of utility function is of special importance because it turns out to generate the value t_u for any arbitrary utility function as a linear combination of more elementary t values. For fixed number P, and a payoff y, define

$$U_P(y)=-\max(0,P-y);\quad -\infty<y<\infty.$$

Although an artificial construct, the function $U_P(y)$, for any given value of the marker P, does qualify as a non decreasing and concave risk averse utility function in its own right, with a further connection to finance concepts such as second order stochastic dominance.

The equivalent margin for such a utility function can be written as

$$t(P) = \frac{1}{F(P)} \int_{-\infty}^{P} (e(R) - R) dF(R),$$

with $e(R)$ as the regression of r on R, and $\lim_{P \to \infty} t(P) = \mu_r - \mu_R$. The importance of this version is that given an arbitrary risk averse utility function U, its own equivalent margin t_u can be expressed as a linear combination of the $t(P)$:

$$t_u = const + \int_{-\infty}^{\infty} w(P) t(P) dP.$$

The nonnegative weights $w(P)$ can be written as

$$w(P) = -\frac{U''(P)}{E[U'(R)]} F(P).$$

The weights depend on the underlying utility function as well as the distribution of the benchmark return R. A specific interest is with the second derivative of the underlying utility function at each marker point P. A proposed security r that has high value for $t(P)$ at a given point P, will be particularly attractive to an investor whose own benchmark utility function has a high value $w(P)$ at that point. Thus P could index a bad state of the world for the benchmark return R. This might be especially damaging to an investor whose utility function is characterised by high risk aversion in that zone, measured as $-U''(P)$. An enhancement r that adds value as $t(P)$ in that zone will be particularly attractive to such an investor.

The equivalent margin schedule $t(P)$ can take any shape when plotted against the benchmark $P = R$. If the schedule slopes negatively from a positive starting value, this is regarded as a defensive security relative to the market. The opposite profile would correspond to an aggressive enhancement. One can show that if a textbook capital market equilibrium (CAPM) exists, with R as the market index return, all the OMD schedules, should intersect at just the same benchmark return.

6.4.2 Do Hedge Funds Really Add Value?

The OMD construction as the equivalent margin can be used to elaborate on the discussion of Chap. 5 concerning the investment profile of hedge funds (with their expensive management fees) versus the S&P 500 as a publicly available benchmark. Figure 6.2 depicts the OMD schedule for the Barclay hedge funds index returns versus those of the S&P500. Also added are the approximate one sigma confidence bands as $t(P_i) \pm \sigma/\sqrt{n_i};\ i = 1, 2, \ldots, N$. These have been obtained by estimating a linear regression $e(R) = \beta_0 + \beta_1 R$, to estimate the standard deviation of the residual: $\hat{\sigma} = STD(\varepsilon)$.

The general import of Fig. 6.2 is that the hedge funds add their particular value in the zone where the benchmark S&P00 return is negative. Investors with a high risk aversion would find it more useful to supplement their portfolio with a hedge fund, or portfolio of such, whose returns mirror those of the Barclay index. In this respect, the results reinforce the earlier findings of Sect. 5.4.

This is a paradoxical finding, for one normally thinks of hedge funds as purporting to add value that in good times that exceeds that of an exchange traded fund (ETF) whose returns that mirror those on the broad market index. A conclusion might be that hedge funds have been much better at defensive offloading of risky assets at the right time to do so. They have been defensive, not aggressive.

6.5 Literature Notes

The information theory of bivariate and multivariate distributions was highlighted by authors such as Gelfand and Yaglom (1959), but also appears in standard reference works such as Pinsker (1964). The conditional entropy of Y given X can be written as

$$\kappa_{y|x} = -\int_y \int_x f(y|x) \ln f(y|x) dxdy,$$

with a corresponding definition for the conditional entropy of X given Y.

The mutual information of the one with the other is measured as the amount by which specifying X reduces the uncertainty (entropy) of Y. Thus $\kappa_y - \kappa_{y|x} = \kappa_x - \kappa_{x|y} = \kappa_x + \kappa_y - \kappa_{x,y}$, from which the mutual information can be expressed as

$$\iint_{y,x} f(x,y) \ln \frac{f(x,y)}{f(y)f(x)} dxdy.$$

If Y, X are jointly normal, this reduces to $-\frac{1}{2}\ln(1-r^2)$; or if vector valued, to the corresponding sum in terms of their canonical correlations. All this a useful way of demonstrating mutual information benefits, but it does not as such address the exigencies of comparison in a bivariate context.

Otherwise, much of the material of the present chapter is new. An exception is the ordered mean difference framework of Sect. 6.3, which was proposed in several earlier papers by the current author, e.g. Bowden (2000, 2005). An analysis of this kind can also be used in finance to check whether a CAPM holds in respect to any given security or a collection of securities at any one time. If it does, then the respective OMD schedules, with the market return as benchmark, should all intersect at the one point. This does not turn out to be the case, suggesting that the CAPM model has at best only illustrative applicability to the real world. A further

connection is with the computation of second order stochastic dominant portfolios, which can be generated using the OMD structure.

References

Bowden, R. J. (2000). The ordered mean difference as a portfolio performance measure. *Journal of Empirical Finance, 7,* 195–233.

Bowden, R. J. (2005). Ordered mean difference benchmarking, utility generators, and capital market equilibrium. *Journal of Business, 78,* 441–467.

Gel'fand, I. M., & Yaglom, A. M. (1959). Calculation of the amount of information about a random function contained in another such function. *American Mathematical Society Translations Series 2*, *12*, 199–246.

Pinsker, M. S. (1964). *Information and information stability of random variables and processes*. San Francisco: Holden Day.

Chapter 7
Entropy, Risk and Comparability

7.1 Introduction

The social sciences abound in loose ends, and so it is with any theory of comparisons. This chapter elaborates on some of the issues that arise. These range from the nature and scope of the probability judgements involved to the precise nature of the preference functions that are inevitably involved.

One of the recurring themes in much of the preceding development is that tail probabilities in themselves are not a sufficient guide to decision making. It is the length of the tail that matters: 'mine is a long and sad tale', as the dormouse says to Alice in the Lewis Carroll classic pun. In this respect, tail entropy is an important complexity indicator. The objective of Sect. 7.1 is to propose a standardisation that reduces the tail entropy for an arbitrary distribution to a standard based on the logistic distribution.

The logistic in itself might be viewed as a more or less standard distribution for financial returns, symmetric as between upwards and downwards. But it has the further useful property that the regime entropies (tail versus the rest) are simply proportional to their respective probabilities. Given a specified tail probability for the subject distribution, one can then establish what the tail probability would be for a logistic that has exactly the same tail entropy as the distribution under consideration. Thus a fund manager might be perturbed to be told that a nominal 5% left hand tail for his or her returns is equivalent to a 20% left hand probability for a logistic with the same amount of entropy in the tail. In a prudential management context, the effective value at risk (VaR), in terms of the potential range of outcomes, might be lot more than the nominal 5%.

In Sect. 7.2 the same idea can be adapted to deal with the conditional value at risk (CVaR), which refers to the expected monetary or value loss given that a value at risk critical probability limit has been violated. A related prudential metric in the insurance industry is the expected shortfall, which as the name suggests, refers to the loss given that a specified critical point has been breached.

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3_7

Section 7.3 turns to rather basic sort of question that concerns the logical nature of comparisons, for this is a book about the information theory of comparisons. From the point of view of economic theory, the theory of comparisons makes most sense when there exists a preference ranking of some kind between the objects of comparison. But in referring to alternative distribution outcomes, what exactly is meant when one says that outcome A is preferred to B; and is it possible to ascribe a numerical value to the difference? This introduces the issue of whether any given underlying preference function is cardinal or ordinal in nature. Section 7.3 exposits and explores issues of this kind.

In recent years some debate has taken place about how agents really react to uncertain outcomes. Do they follow the precepts of more or less objective probability theory; or do they adhere to a more subjectivist probability, and do so more or less in common? Or at one remove, do human agents follow decision procedures that meld both probabilities and preference functions into the one decision criterion? A notable example is the cumulative prospect theory associated with Tversky and Kahnemann. In Sect. 7.4 it is shown that such rules amount to the use of an equivalent probability distribution obtained via a suitably specified Radon-Nikodym derivative. With this substitution, decision rules based in entropic complexity can be carried out in terms of a biased partition entropy function where the two components, respectively involving F and $1 - F$, are weighted differently.

Section 7.5 broadens the range of entropic comparisons to organisation theory, one of the tangential areas raised in Sect. 1.8. The suggestion is advanced that discussion of an organisation's efficiency could be approached via its equivalent entropic complexity. An efficient organisation is one that filters more efficiently the incoming information complexity by the time it reaches senior management.

The chapter concludes with Sect. 7.6 as the literature review.

7.2 Tail Probabilities and Informational Measures

In many contexts of economic or social importance, specific interest lies in one of the tails of a given distribution. A leading instance arises in financial risk management, where prudential policy for banks or insurance companies addresses the possibility of a loss of capital beyond a pre-specified safety zone. The latter is usually determined as less than or equal to a given marker value for capital. Value at risk (VaR), which is concerned with the probability of loss beyond this prudential marker value, has been introduced in Sect. 2.6. A related concept, the conditional value at risk (CVaR), refers to the expected value of the loss given that the VaR critical marker value has been breached. There is a natural connection here with the left conditional expectation function of Chap. 3. Section 7.2 takes up this particular application.

In addition, there are other contexts involving specific concerns with the risk management of distribution tails, such as failure rates in mechanical or electrical equipment, or mortality in medical or demographic contexts. In what follows, some

of the groundwork is established for measuring and managing the risk associated with long tails by making use of an information theoretic framework.

The log odds function is a good starting point, as

$$\lambda(x) = \ln \frac{F(x)}{1 - F(x)} \tag{7.1}$$

With reference to a given marker value x, the event $\{X \leq x\}$ will be denoted as $R_L(x)$ or just R_L where the marker value is understood; and the complementary event $\{X > x\}$ as $R_U(x)$ or R_U. The log odds function in the form (7.1) specifies the log probability of an outcome in $R_L(x)$—the bad or unacceptable zone—versus the good zone. Risk management is not a cheerful business.

The derivative of the log odds function can be written in the form

$$\lambda'(x) = \frac{f(x)}{F(x)(1 - F(x))}.$$

In this form, it is analogous to a signal to noise ratio. The denominator can be viewed in the light of fuzzy logic as the product of fuzzy membership indexes, incorporating the degree of doubt as to which of R_L, R_U the next drawing of X will belong. The numerator $f(x)$ can be interpreted as the amount of information contained in the interval $x,\ x + dx$.

The logistic distribution with distribution function $F(x; \mu, \beta) = 1/(1 + \exp(-\frac{x-\mu}{\beta}))$ is a useful benchmark. In this case the log odds function is linear in x, with $\lambda'(x) = 1/\beta$, the inverse of the scale factor β. The logistic belongs to the extreme value family of distributions, obtained as the limiting maxima of a collection of identically distributed random variables. This may be of relevance in the analysis of rare events, where an extreme outcome may follow a cluster or sequence of good or bad outcomes.

Figure 7.1 illustrates its log odds function as the negative, i.e. as $-\lambda(x)$, along with those for the Cauchy and Normal, both also used for financial data. The comparisons are normalised on a Shannon entropy of 1.25. This is consistent in the case of the normal distribution with a standard deviation of 0.7% over the 10 day period used in the Basle guidelines for commercial bank value at risk. The Cauchy in particular has a sharp density peak at the median, resulting in a definite point of inflection in this zone, with a much slower decay in the tails.

The derivative of the log odds function has a useful connection with Shannon or total entropy, the latter defined in the usual form as

$$\kappa = -\int_{-\infty}^{\infty} f(x) \ln f(x) dx = -E[\ln(f(x))].$$

Utilising expression (7.1) together with $E[\ln F(x)] = E[\ln(1 - F(x))] = -1$, it follows that

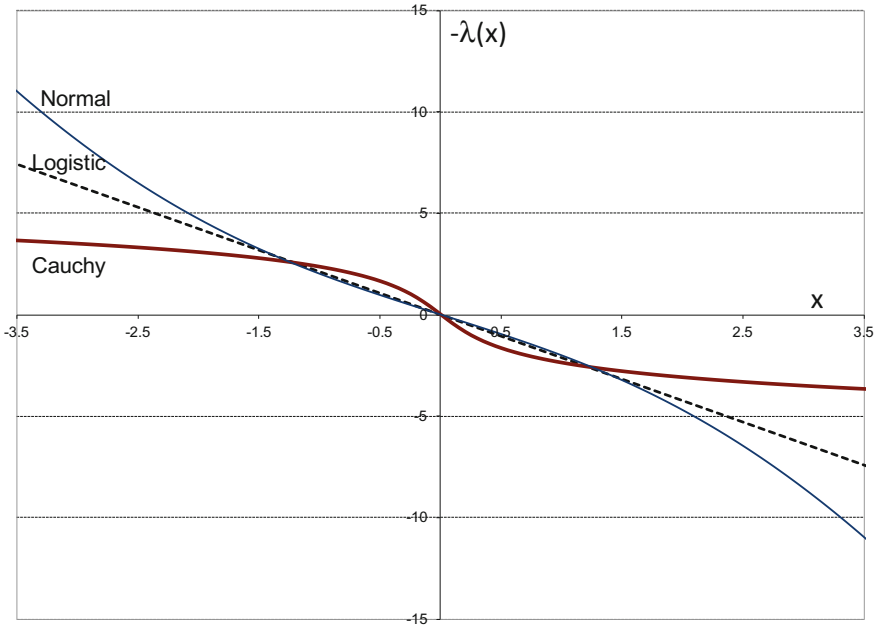


Fig. 7.1 The log odds functions for some standard distributions

$$\kappa = E[2 - \ln \lambda'(x)] \tag{7.2a}$$

Equivalently if one defines a function

$$\xi_d(x) = \frac{1}{\kappa}[2 - \ln \lambda'(x)] \tag{7.2b}$$

then $E[\xi_d(x)] = 1$. The function defined by expression (7.2b) will be referred to as a scaling function. For a logistic distribution the scaling function $\xi(x) \equiv 1$ i.e. has a constant value of unity. Otherwise it can differ quite radically in form as between distributions; thus for a Normal it has a maximum at the median, while for the Cauchy it has a minimum there.

For some leptokurtic distributions where the density falls away rapidly towards one or both tails, the function $\xi_d(x)$ can become negative in this region. To preserve an interpretation as a Radon Nikodym scaling factor, it may be desirable to truncate the distribution range to ensure that $\xi_d(x) \geq 0$.

Figure 7.2 depicts directional scaling functions for three common distributions, normalised to have the same differential entropy κ. It will be apparent that the shape varies considerably with the respective tail properties.

Given a designated marker value x, the scaling function can itself be used to attribute Shannon entropy into upper (U) and lower contributions (L) relative to x. The starting point is to isolate scaling functions specific to each regime as

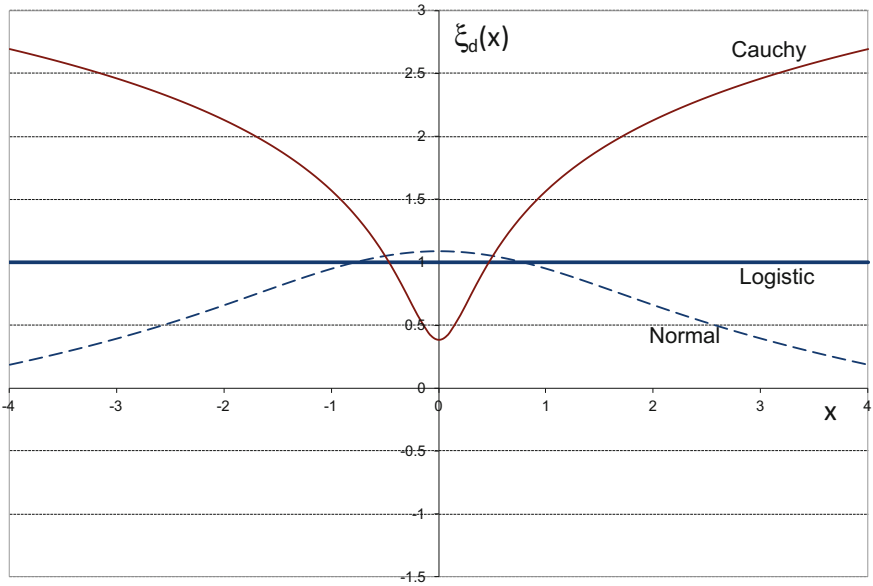


Fig. 7.2 Directional scaling functions

$$\begin{aligned} \xi_{d,L}(x) &= E[\xi_d(X)|X \leq x] \ ; \\ \xi_{d,U}(x) &= E[\xi_d(X)|X > x] \ . \end{aligned} \tag{7.3}$$

With the same marker value x, define

$$\kappa_L(x) = \kappa\xi_L(x)F(x) \ ;$$

$$\kappa_U(x) = \kappa\xi_U(x)(1 - F(x)).$$

Then total Shannon entropy divides into lower and upper contributions as

$$\kappa = \kappa_L(x) + \kappa_U(x)$$

The functions $\kappa_L(x)$, $\kappa_U(x)$ can be referred to as the lower and upper directional entropies, with the basic context as Shannon entropy.

The logistic distribution continues to provide a benchmark for comparisons, for in this case $\xi_d(x) = 1$ for all values x, meaning that $\xi_{d,L} = \xi_{d,U} = 1$ as well. Hence for any given marker value, the directional entropy contributions are proportional to the regime probabilities:

$$\kappa_L(x) = \kappa F(x); \ \kappa_U(x) = \kappa(1 - F(x)).$$

For other distributions, the directional entropies differ from the partition probabilities in more intrinsic fashion as the marker value x varies. Nonetheless, the

logistic distribution continues to have benchmark value in relation to these other distributions.

Given a distribution function $F(x)$, the two scaling functions of expression (7.3) can be employed as Radon Nikodym derivatives to obtain a change of measure, leading to a transformed distribution function defined as

$$F_q(x) = \xi_{d,L}(x)F(x), \tag{7.4}$$

with the complement as $1 - F_q(x) = \xi_{d,U}(x)F(x)$. A proviso is that the scaling functions $\xi(x)$ are nonnegative over their range, which is true of the distributions used in the present discussion. For a parent logistic distribution the scaling functions are identically unity, which means that $F_q(x) = F(x)$.

However, this is not true of more fat tailed distributions, where the general effect is to inflate the tails of the transformed distributions. Figure 7.3 illustrates for a parent Cauchy distribution. In a VaR context, an allowable loss probability fixed at 10%, would imply a critical marker point at $x = -0.855$. However this pays little, if any, attention to the possibility of much higher losses consistent with the fat tailed property of the Cauchy distribution. Under the equivalent $F_q(x)$ distribution, this would evidently inflate to a probability at the chosen VaR critical point of 22.7%, much higher. The tail probabilities of F_q could then be used to construct a corrected VaR for the natural distribution of gains or losses.

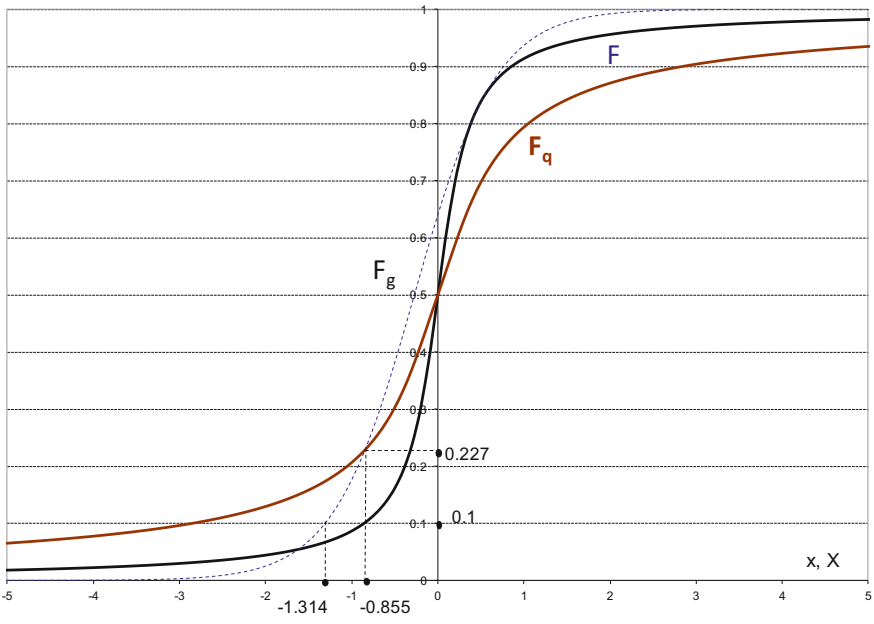


Fig. 7.3 Equivalent scaled distributions

Judgements of this kind can be validated by drawing on the invariance property of the logistic distribution, for which $F_q(x) = F(x)$. Let x be a predetermined marker value, in this context the VaR critical point. Then one can always find a locally equivalent logistic distribution $F_g(x)$ by choosing its location and scale parameters such that $F_g(x; \mu_X,\ \beta_X) = F_q(x)$. Specifically,

$$\beta_x = e^{\kappa - 1}\ ;\ \mu_x = x + \beta_x \ln(-\lambda_q(x),$$

where $\lambda_q(x)$ is the log odds function for the rescaled distribution function $F_q(x)$.

The equivalent logistic distribution function associated with the given 10% critical point $x_{0.1} = -0.855$ is labelled as F_g in Fig. 7.3. At the given critical point of 0.855, it shares with F_q the revised tail risk of 22.7% instead of the nominal 10% of the original F. To arrive at a 10% probability loss, the critical point would have to be relaxed to an effective value of -1.314.

In drawing conclusions of this kind, the logistic becomes an effective standard for assessing the probabilities of expected gains and losses, one which at the same time makes allowance for very long tails. In terms of the above example, prudential managers unwilling to accept a revised critical point would have to scale back investments that have the possibility of large adverse tail outcomes.

Finally, in the case of distributions that have shorter tails than the logistic benchmark, the adjustment relative to the logistic would have to go the other way. This is true for a normal distribution. Relative to its equivalent logistic benchmark, the effective VaR point moves to the right, with the consequences of a loss exceeding the VaR critical point viewed as less serious.

7.3 The Conditional Value at Risk

The conditional value at risk (CVaR) is an extension of VaR that has independently surfaced in a number of different contexts. In the insurance industry it is referred to as the expected shortfall, referring to the ability to meet larger claims stemming from exceptional but nonetheless very adverse insured events. In the finance industry it is employed as an additional prudential measure designed to capture and allow for the numerical values of losses that could lead to bankruptcy, or in the case of banks necessitate a rescue by the central bank.

Whatever the nomenclature, the idea is the same. The CVaR refers to the expected loss given that the VaR critical point has been reached. Thus if x is a preassigned VaR critical point, for a lower limit of outcomes x taken for expositional purposes as $-\infty$, then

$$CVaR(x) = E[X|X \leq x] = \frac{1}{F(x)} \int_{-\infty}^{x} Xf(X)dX.$$

An alternative expression is

$$CVaR(x) = x - \frac{\Phi(x)}{F(x)},$$

where $\Phi(x) = \int_{-\infty}^{x} F(X)dX$ is the further progressive accumulation of $F(x)$.

The associated quantity

$$x - CVar(x) = \Phi(x)/F(x)$$

is referred to in the insurance risk management literature as the 'expected shortfall'. It is a measure of how much the company can expect to lose once the critical point has been triggered. In this respect, literature references to the 'expected shortfall' can refer to either the difference $E_x[(X - x)_-]$, or else the difference $x - CVar(x) = \Phi(x)/F(x)$ as above.

As a distribution function accumulator, the function $\Phi(x)$ features in the theory of stochastic dominance. Distribution A is second order stochastic dominant over B if $\Phi_A(x) \leq \Phi_B(x)$ for all x. If the random variable x refers to returns on a security or portfolio, this means that every risk averse investor would choose asset A in preference to asset B. Second order stochastic dominance (SSD) is a weaker condition than first order stochastic dominance (FSD), which would be $F_A(x) \leq F_B(x)$.

Once again, the logistic distribution provides a useful benchmark for the expected shortfall. If $F(x)$ is a logistic with scale parameter β, then

$$CVaR(x) = x + \beta \frac{\ln(1 - F(x))}{F(x)}.$$

Then for any arbitrary distribution, one can form the equivalent distribution $F_q(x)$ as in expression (7.4) and obtain the logistic equivalent CVaR as

$$CVaR_q(x) = x + \frac{e^{\kappa - 2} \ln(1 - F_q(x))}{F_q(x)}.$$

In general, however, there is less motivation to use the logistic proxy once the actual expected values of losses that exceed the VaR marker $X = x$ are explicitly introduced as in CVaR. A possible exception is where the left hand tail is so long that the actual CVaR does not exist. The logistic equivalent may also be useful if a degree of agnosticism exists as to welfare criteria, specifically as to whether expected losses are a sufficient in themselves, as distinct from the welfare effects of higher order conditional moments.

In terms of the notation of the present study, the conditional value at risk can be recognised as the left conditional expectation $\mu_l(x)$, with the expected shortfall as $x - \mu_l(x)$ This is to be assessed at the designated critical point x. If the latter is legislative in nature, as in Basle style bank capital adequacy, then it is a stand alone provision. No matter how much upside there might be, the bank has to ensure that

its portfolio meets such a requirement, with the designated probability. In particular, this becomes a side constraint on the institution's profit maximising behaviour, and as such has its own shadow price as a binding constraint, in terms of foregone expected profit.

A less rigorous regime might recognise the potential value of the upside, assessed in this case as the right conditional expectation $\mu_r(x)$. For a given critical value the welfare function might be of a form such as

$$d_w(x) = w(x)(\mu_l(x) - \mu_r(x)),$$

with nonnegative weights $\int_* w(x)dx = 1$. Portfolios could be assessed by comparing their alternative net payoff profiles in such terms.

7.4 Cardinal Versus Ordinal Comparisons

If the substance of the present work has to do with the theory and practice of comparisons, a consequential issue arises as to their precise extent or scope. Just how far can one push things? If two countries are compared on their income distributions and the one yields half the value of v, or some other measure, does this means twice the social welfare, or merely an ordering as better or worse?

From the point of view of economic theory, the theory of comparisons makes most sense when there exists a preference ranking of some kind between the objects to be compared, such as consumption bundles. But the question arose early in the economics literature as to just how far welfare comparisons could go. Historical development took place along two major lines, summarised in what follows.

7.4.1 Utility Theory: A Short Review

The older of the two strands or welfare comparison was within the context of consumer preference theory. The ensuing utility function was ordinal in nature. It may not be possible to say that oranges and applies can be reduced to a common scale of like or dislike, such that one orange is worth 1.75 apples. But it may be possible to rank different bundles of apples and oranges, and to derive indifference sets of combinations such that the consumer would like equally well any bundle along such a surface. In turn, different levels of such indifference sets could be ranked with a numerical utility number; but as a function, the latter is determined only up to a positive monotone transform. Thus if x is a bundle, with given utility function $u(x)$ then so is $W(x) = w(u(x))$, where $w'(u) \geq 0$. Utility rankings of this kind are said to be 'ordinal'.

The (relatively) more recent line of development originated with the work of Oscar Morgenstern and John Von Neumann in game theory. In this case the

bundles are such that in principle it is possible to assign welfare numbers to gambles between one bundle and another. The resulting utility function is unique up to an affine linear transform i.e. choice of location and scale parameters. Such 'cardinal' utility functions underpin much of modern finance theory as well as the theory of games.

A more or less common set of basic axioms is shared by both cardinal and original approaches, with the two differing only in the overlay of the further axiomatic assumptions. There is a common starting point of a set of objects or events such as $\{A, B, C \ldots\}$, with which a preference ordering $\succ$ exists. An implied indifference relationship, denoted $\sim$, is such that $A \sim B \Leftrightarrow A \succ B \text{ and } B \succ A$. As well as being transitive in nature, this is assumed to be a complete preordering such that for any two objects or outcomes it must be that the one is either preferred or indifferent to the other.

From here, the two strands diverge. Underpinning the cardinal approach is the possibility of combining the events into compound events referred to as gambles. The continuity axiom then says that if $A \succ C \succ B$, then there is a gamble that gives A with probability λ and B with probability $1 - \lambda$, and this gamble is a compound event that is indifferent to C. From here it follows that a utility function u must exist that is unique up to an affine positive linear scaling: $w = \alpha + \beta u \,; \beta > 0$. This is a cardinal utility function, utilised in the theory of choice under risk. The work on the ordered mean difference in Chap. 6 requires a utility theory of this kind, as does the VaR and CVaR concepts of Sect. 7.2. One has to be able to rank differences in monetary outcomes, which in turn becomes a cardinal proxy for the investor's utility.

The second strand of utility theory, the ordinal approach, does not attempt to rank different outcomes with a common or universal standard of value. The equivalent relationship takes centre stage, embodied as indifference sets. For any given object (C, say), the space of objects is partitioned into those preferred to C and those preferred over C. Their intersection gives those objects indifferent to C, and it is assumed that both sets (upper and lower) contain all their boundary points. The entire space of outcomes can then be partitioned into indifference surfaces, as in the textbook 'indifference curves' of consumer economics. A utility function $u(x)$ exists that has the same value along the indifference curve containing x, and has higher value for points along an indifference curves containing points strictly preferred to x. In this sense indifference curves are level surfaces for the function u. However, any positive monotone transform $w(u)$ will do equally well for such a purpose. So the given utility function u empowers only an ordinal ranking of the underlying objects.

7.4.2 *Entropic Asymmetry and Social Utility*

With the above summary review as background, one can re-examine inequality measures established in Chaps. 3 and 4, specifically the asymmetry function

$v(x) = \mu_l(x) + \mu_r(x) - 2\mu$; or in the present context $-v(x)$, which is an increasing function of income or wealth x, the 'advantage function'. Could a function of this type quality as a utility function and is the resulting scale cardinal or ordinal?

Discussion divides according to whether the underlying income distribution function $F(x)$ is fixed in its scope and coverage and throughout, or on the other hand whether the comparison is intended to span a complete set of possible distributions as in inter-country comparisons. In this respect, there is a difference between a utility of function of the form $v(x : F_0)$ with the distribution function fixed, and one of the form $v(x, F)$ where the distribution function F is itself an argument.

Consider the former case, and imagine the following experiment, intended to adapt the social justice argument of John Rawls to the current context. I know the underlying distribution of incomes $F(x)$. But I do not yet know what income x or wealth I will be endowed with. Consider two possible income values x_a , x_b with $x_a > x_b$, and suppose I prefer the former to the latter. Now consider an intermediate value such that $x_a \succ x_c \succ x_b$. These may be considered as outcome events. Is there a probability λ such that I will be indifferent between getting x_c or the outcome of the gamble? If so, then it must follows that a cardinal utility function exists such that $u(x_c) = \lambda u(x_a) + (1 - \lambda)u(x_b)$. For any such individual i, the resulting utility function is $u_i(x)$. In the particular case where this is linear in x, the same choices would be made with

$$u_i(x) = \alpha_i + \beta_i x \,; with\ \beta_i > 0.$$

A collective or social utility function could then be obtained by aggregating over individuals i according to some given rule. Still in the spirit of the Rawls experiment, each such individual could have attitudes as to how his or her income x_i (whatever it turns out to be) will compares with that of others.

One such comparison could be calibrated in terms of comparison of any given x with $\mu_l(x_i)$, $\mu_r(x_i)$, as

$$u_i(x) = (x - \mu_l(x_i)) - (\mu_r(x_i) - x).$$

A personal utility function of this form would indeed be cardinal in nature, for if my income shifts from x to $x' > x$, all other individuals staying the same, then my personal welfare has increased linearly. If we now aggregate over individuals i with the same weights as their relative frequency of incomes, a collective or social utility would result, of the form

$$u_c(x) = 2x - (\mu_R + \mu_L).$$

All the above is predicated on just the one underlying distribution of incomes. However, a common agenda is to compare outcomes across a given set of distribution functions, as in comparisons between countries. The objects to be calibrated

and compared in such a case are now of the compound form (x, F), with the distribution function F a member of a given set of such: $\Im = \{F_1, F_2, \ldots, F_N\}$.

Thus suppose three objects, $A = (x_a, F_a)$,$B = (x_b, F_b)$ and $C = (x_c, F_c)$ are such that $A \succ C \succ B$. Under the continuity axiom, a cardinal comparison would then require that C be indifferent to a notional gamble with A as outcome with probability λ, and B with probability *1-*λ.

But it is unclear just how to define such a gamble. One could imagine a two stage process. For any given F a first stage gamble would be between (x_a, F), (x_b, F), (x_c, F) with respective probabilities π_a, π_b, π_c, resulting in $(\pi_a x_a + \pi_b x_b + \pi_c x_c,\ F)$. This could then be followed by second stage gamble over the choices of F, the combined effect viewed as a compound gamble. But a problem is then how to define a unique object C such that C is indifferent to A and B; there may well be a set of such.

Ordinal utilities are a possible recourse. Thus for any given F, one might consider the function $v_F(x)$ as an ordinal welfare outcome over different values of income x. A given indifference curve, if it exists, would then consist of order preserving combinations x, F such that $v_F(x) = v_0$ for a given value $v = v_o$. The key test would then be whether for higher or lower values of *v,* the indifference curves ever cross over one another. If a complete first order (FSD) stochastic dominance relationship exists as between $F \in \Im$ then a proper set of indifference curves does exist. However for most practical purposes this is an exceptional scenario. Thus in the context of cross country comparisons it might well be that there is one particular country that stochastically dominates all the others. But it unlikely that this applies in sequence to the remainder.

Whether the income advantage function qualifies as either a cardinal or ordinal social utility function therefore depends upon the scope of the implied comparison. For most potential applications it is best treated as ordinal in nature, sufficing to rank differences e.g. as between countries or different social groups; but not to say that one such ever going to be worth exactly twice the other.

7.5 Information Based Rescaling in Subjectivist Probability

The distinction between objective and subjective probability has a venerable history in economics and statistics, originating with influential authors such as Ramsay, de Finetti, and Savage. But while the epistemological foundations might differ, the general supposition was that subjectivist probability was nevertheless internally consistent in the sense of the game theoretic axioms of Von Neumann-Morgenstern.

In more recent times, however, Tversky and Kahneman have suggested that agents do not necessarily think and act in terms of the classic rules of probability. Their principal postulates may be summarised as: (a) Framing is important (the way choices are presented); (b) people overestimate probabilities of rare events; and

(c) in thinking about compound gambles, people do not recombine probabilities correctly. In particular, their 'cumulative prospect theory' could be applied to investment activity, in terms of which investors would systematically overweight one or both of the tails, relative to the true or natural distribution. Subsequently, the approach was utilised in the theory of finance to explain apparent empirical anomalies in asset pricing, such as the equity premium puzzle, size and value anomalies, or the momentum effect.

Cumulative prospect theory substitutes a 'weighting function' of wealth or return outcomes for the original or natural probability distribution. The result, as it applies to investment, is analogous to a subjective probability measure, although the resulting decision rules do necessarily adhere to those that might be derived from the Von Neumann-Morgenstern continuity and independence axioms. Likewise, the classical utility function is replaced by a 'value function', resulting in a decision criterion that is nevertheless analogous to expected utility. Prospect theory shares with the more traditional subjectivist approach a basis that even where objective or frequentist probabilities might exist, economic agents effectively map such probabilities into their own more subjective measures.

The choice of transformation might itself be guided by psychological heuristics or biases. Subjectivist authors drew on the widespread understanding and use of fractional odds in betting in horse races and the like. The odds at which a subject is willing to bet can be taken as a proxy for the probability that he or she has assigned to the outcome. Taking the log of the odds ratio enables the subject to rank events in an intuitively appealing way. However, at any given support point, two distributions might have the same tail probability but could differ in their tail length and therefore the prospect of large gains or losses. It seems reasonable to suppose that investors or gamblers are influenced by the odds over the entire tail area, rather than just the odds at any particular point, and it is this perspective that provides a link to partition entropy.

7.5.1 Formalising as a Change of Measure

The process could be formalised by supposing that outcomes such as share returns or lottery rewards have a natural (objective) distribution F. But investors do not act as though this is their decision basis. Instead they think and act in terms of a subjective distribution function W (or the 'cumulative weighting function' in prospect theory). In line with the prospect theory findings as to rare events, W might have longer tails than F, to a degree that could differ as between the upper and lower tails. Thus at the upper end, an investment in a dotcom IPO might turn into a life changing outcome for the lucky investor. At the lower end, a proposed portfolio could decimate an investor's carefully accumulated retirement capital. In each case the true probability might be small, but the investor reweights it to become larger, and acts accordingly.

The basic version of the measure transformation can be represented as

$$W = \Psi(F);\ W(x) = \Psi(F(x)).$$

The function $\psi(F)$ is specified as non decreasing (the monotonicity assumption) and is normalised so that $W(0) = 0,\ W(1) = 1$. The Radon-Nikodym (R-N) derivative, if it exists, is written as $\xi_w^0(F) = \frac{d\psi}{dF}$ so that $dW = \xi_w^0(F)dF$. Correspondingly, let

$$\xi_w(x) = \xi_w^0(F(x)) = \Psi'(F(x)).$$

Then $dW(x) = \xi_w(x)dF(x)$, and $w(x) = \xi_w(x)f(x)$ if the densities exist.

The specifications imposed on ψ imply that the function ξ is nonnegative, and $E_F[\xi_w(x)] = 1$. The outcome reweighting function $\xi_w(x)$ therefore plays the role of a R-N derivative for a change in distribution from F to Ψ, i.e. for a change from the natural to the subjective distribution (weighting function) of the investment outcome x. The general import is that the function rescales to a greater degree at one, or possibly both, tails of the original, making compensating re-weightings elsewhere. Thus if investors act as though events of rare good fortune have an inflated probability, they will act as though $1\text{-}W(F) > 1 - F$ for outcomes x such that $F(x)$ is near unity; equivalently, $\xi_w(x) > > 1$ in this zone.

Operationally, one might start with the suggestion that a given nonnegative function $s(F)$ could serve as a possible rescaling function. Normalising as $\xi_w^0(F) = s(F)/\int_0^1 s(F)dF$ and setting $\xi_w(x) = \xi_w^0(F(x))$ will assure that $E_f[\xi_w(x)] = 1$, providing the integral $\int_0^1 s(F)dF$ exists.

In cumulative prospect theory the zone of integration may be split into positive and negative values of x to reflect distinct re-weightings of the left and right hand tails:

$$\psi(F) = \begin{cases} \psi^-(F); & x \le 0 \\ \psi^+(\bar{F}); & x > 0,\ \bar{F} = 1 - F. \end{cases}$$

In this case, some adaptations are necessary with differentials of the form

$$\xi_w(x) = \frac{d}{dF}\psi^-(F(x)); \quad x \le 0$$

$$\xi_w(x) = -\frac{d}{dF}\psi^+(1 - F(x)); \quad x > 0.$$

Thus the investor weighting function chosen by Tversky-Kahneman can be cast as:

$$\psi(F) = \begin{cases} k\frac{F^{\gamma^-}}{(F^{\gamma^-}+(1-F)^{\gamma^-})^{1/\gamma^-}}; F \le F(0) \\ 1 - \frac{F^{\gamma^+}}{(\bar{F}^{\gamma^+}+(1-\bar{F})^{\gamma^+})^{1/\gamma^+}}; \bar{F}=1-F,\ F > F(0) \end{cases}$$

where γ^-, γ^+ are positive constants.

As it stands, the resulting cumulative weighting function is not continuous at zero, and may not even be monotonic. However, the constant k can be chosen so that the two halves splice together with no break at *F(0)*. For example, if $\gamma^- = \gamma^+ = \gamma$, then

$$k = \frac{\Delta}{F_0^\gamma} - \mu_0^\gamma;\ \textit{where}\ \Delta_0 = (F_0^\gamma + (1-F_0)^\gamma)^{1/\gamma},\ \mu_0 = \frac{1-F_0}{F_0}.$$

It is possible to generalise to allow for critical zones of x, where the value function adopts a shaper curvature upwards as in a life changing lottery outcome, so that $W(x) = W(F(x), x)$. In such terms,

$$\xi_w(x) = \frac{\partial W}{\partial F} + \frac{1}{f(x)}\frac{\partial W}{\partial x}.$$

In prospect theory discussions, the utilities of classical Von Neumann-Morgenstern choice theory are recast as 'value functions', and the 'weighting function' replaces probabilities, to arrive at a decision criterion that corresponds to expected utility. As earlier pointed out, however, much the same rescaling process can apply to either. In what follows, the more familiar expected utility framework is utilised for exposition.

The above framework can then be applied to utility comparisons and the consequent decision rules. Expected utility under the natural measure is

$$E_f[u(x)] = \int_{-\infty}^{\infty} u(x)dF(x)\ . \tag{7.5a}$$

Under the subjective measure it is

$$E_w[u(x)] = \int_{-\infty}^{\infty} u(x)dW(x). \tag{7.5b}$$

Note that expression (7.5b) could alternatively be written as

$$E_w[u(x)] = \int_{-\infty}^{\infty} \tilde{u}(x)dF(x);\ \tilde{u}(x) = \xi_w(x)u(x),$$

in other words, as a rescaling of utility instead of probability. For any given application (choice of F, ψ) there may be observational equivalence between transformed probability and transformed utility, although the equivalent utility function will not remain invariant to different specifications of $F(x)$.

Most subjective probability measures entail overweighting the tails of the natural distribution. This means that the original utility function, evaluated at subjective probabilities, is expected utility equivalent under the natural measure to a transformed utility function $\tilde{u}(x)$ that overemphasises extreme values of the domain, perhaps to the point of a radical change in shape or behaviour.

7.5.2 Information Based Rescaling

The foregoing exposition can be utilised to develop an information based rescaling. A suitable vantage point is the log odds function. Indeed, the longstanding use of fractional odds by bookmakers implies that people are comfortable in thinking and acting in such terms, as a matter of the psychology of gambling. In this respect, the use of the log odds facilitates judgments involving transitivity. Thus if event A is viewed as twice as likely to happen than B, and B three times as C, then the subject should intuitively consider event A as 6 times more likely than C (rather than 5).

More generally, a predisposition ('Gibrat's law') to think in terms of logs is familiar from everyday life, which is why % changes in salary or wage determinations mean more than dollar changes. In the present context, this could mean that in their subjective probability rescaling, people react to the log of tail areas: 'noticeability' is based on log F and log $(1 - F)$. In commercial life, investors are often preoccupied with tail length, proxied in terms of the average log odds before, or after, the chosen point. It is possible to formalise such behaviour in terms of the log odds function and its cumulative, the partition entropy function.

To avoid carrying negative signs in what follows, it is convenient to measure the log odds function in the negative of the usual, as referring to the event that the random variable $X > x$ versus $X \leq x$:

$$\lambda(x) = \ln\frac{1 - F(x)}{F(x)};\quad \Lambda(F) = \ln\frac{1 - F}{F}\ .$$

Reflecting subjectivist perspectives, the log odds function can be generalised so that the respective tails F, $1 - F$ are weighted unequally. The new distribution function is given by

$W_\theta(x)$ such that $W_\theta = \Lambda_\theta(F)$, where

$$\Lambda_\theta(F) = \ln[\frac{(e(1-F))^{1-\theta}}{(eF)^{\theta}}] \; , 0<\theta<1; \; \lambda_\theta(x) = \Lambda_\theta(F(x)) \; . \tag{7.6}$$

The cumulative gives a biased partition entropy corresponding to W_θ as

$$h_\theta(x) = -2[\theta F(x)\ln(F(x)) + (1-\theta)(1-F(x))\ln(1-F(x))] \; ; 0<\theta<1 \; ,$$

$$H_\theta(F) = -2[\theta F \ln F + (1-\theta)(1-F)(\ln(1-F)] \; ; 0<\theta<1 \; .$$

The normalising factor $e^{1-2\theta}$ implicit in expression (7.6) ensures that $E[\lambda_\theta(x)] = 0$ while $h_\theta(x)$ remains positive, with $\lim_{x\to\pm\infty} h_\theta(x) = 0$. The special case $\theta = \frac{1}{2}$ reverts to the unbiased version.

To reverse the logic, partition entropy can provide a potential measure for subjective over- or under-valuation of the tails, which are regions where information is highest. A suitable rescaling to be applied at any given point incorporates the average log odds up to that point; or reversing sign, after that point. Thus in a lottery type situation, one could visualise the subject, in reweighting any given point x, as weighing up the probabilities of the entire range of outcomes $X \geq x$. As x increases, and the prospective prizes become greater, the subject becomes more predisposed to do this. The density rescaling at any given point x is linked to the cumulative log odds of points thereafter, and hence to the information at that point.

In turn, this is reflected in the scaling function that gives the new subjective distribution. The rescaling functions incorporate an affine normalisation of the inverse information functions, to ensure that $\Psi(0) = 0; \Psi(1) = 1$ and $E_f[\xi_w(x)] = 1$. A suitable specification to accomplish this takes the form:

$$\xi_w^0(F;\beta,\theta) = 1 + \beta(0.5 - h_\theta(F)) \; , \quad \xi_w(x;\beta,\theta) = \xi_w(F(x);\beta,\theta) \; .$$

This leads to the subjective distribution function

$$\Psi(F;\beta,\theta) = -0.5 * \beta(1-\theta) + (1+0.5*\beta)F + \beta[\theta F^2(\ln F - 0.5) - (1-\theta)(1-F)^2(\ln(1-F) - 0.5)].$$

The parameter $\beta > 0$ can be chosen to reflect the slope or sensitivity of the scaling function. Limits have to be placed on β to ensure that ξ > 0. For θ = ½, the limit is $\beta < \frac{1}{\ln 2 - 0.5} = 5.177399$.

Note also that $\xi_{w\prime}^0(F;\beta,\theta) = -\beta\Lambda_\theta(F)$. As the log odds function is such that $\Lambda\theta'(F) < 0$, this means that the proposed rescaling function $\xi_w^0(F;\beta,\theta)$ is convex in F.

Figures 7.4a, b illustrate. As a general observation, information based rescaling weights the tail areas by more than does the Tversky-Kahneman formula, drawing more mass away from the median area.

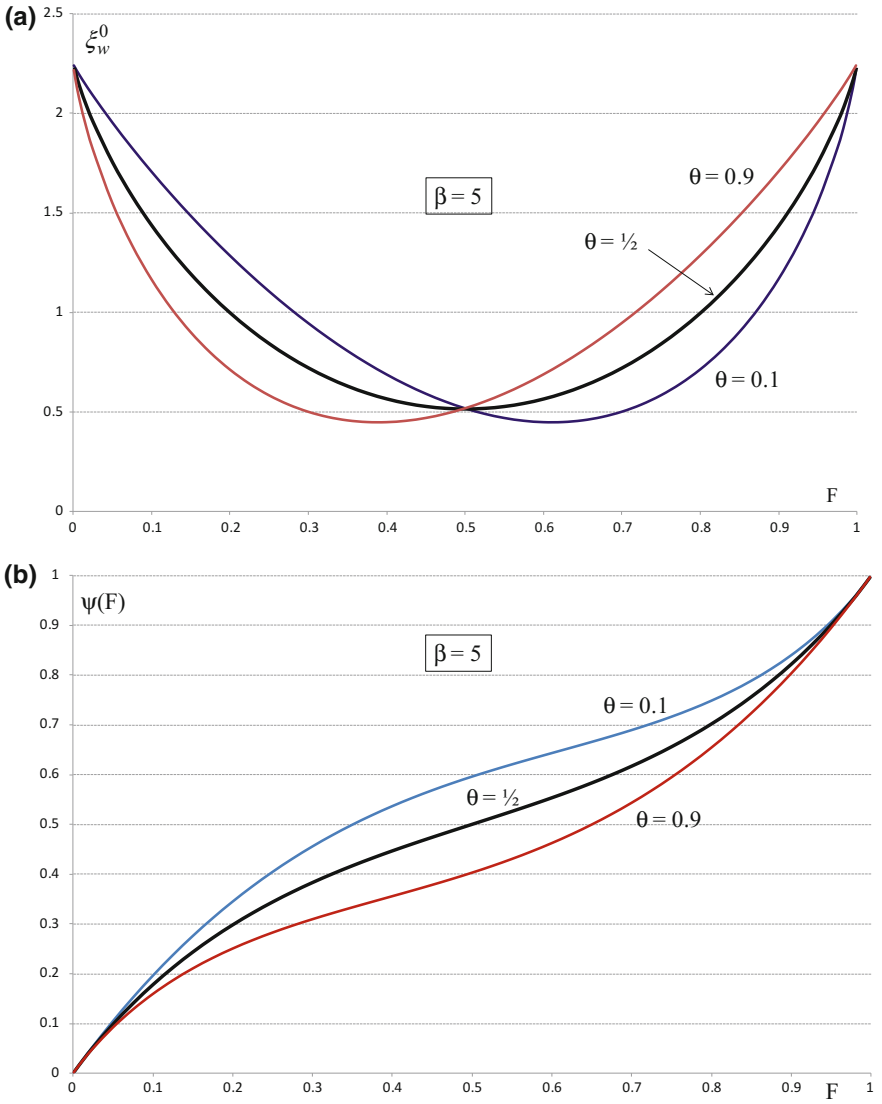


Fig. 7.4 **a** Density scaling functions, **b** Distribution function scaling

7.6 Concluding Remarks: The Scope of Informative Comparisons

The scope of the present contribution has been within the frame of comparisons between more or less well defined frequency distributions: incomes, ages, investment returns, opinion polling, grading, and other fields of application. But as

pointed out in Sect. 1.8, the general notion of information has a familial relationship with that of complexity, even where this is implicit or not even well defined.

One such area is organisation theory in the management sciences. The issue in this case has been how to characterise organisational complexity. What does it mean, precisely, when we claim that organisation A is more complex than organisation B?

One possible framework is to consider an organisation such as a corporation as an ordered network of reporting, command and control. At the top is a CEO (technically a shareholder elected Board, but let us be realistic here). Distributed along the bottom tier are the 'coalface' units that have direct contact with customers (bank tellers, sales staff and the like). Intermediary managers link the two, with their own reporting hierarchies, the whole constituting a multistage network of command and control.

Information flows in two directions; from the coalface upward and from command and control downward. Starting with the very bottom tier, one role of the first layer of management, as the next tier up, as is to filter out happenings of more or less everyday occurrence. The constructive progression from bottom to resembles that of wavelet theory. The counter staff are charged with dealing with the very short fluctuations, passing the result up to their supervisors as the smoothed second wavelet, and so on from there. Senior management, and from there the Board, deal with the wavelet of longest periodicity.

In coding terms, only the longer code lengths, representing more unusual occurrences, would be passed upward along the chain, transformed at this level into shorter coded messages. By the time the information flow gets to senior management, the everyday has been filtered out. What remains takes the form of shorter recoded messages as to just what does demand attention at their level, and in turn, further upward reporting. In such terms, an organisation is less complex if the information flow is so well managed that the CEO needs be preoccupied only with a limited range of matters of systemic importance. Thus from the bottom up, this is a story in coding terms of a reduction in information complexity. Likewise, the command and control flow in the reverse direction should be one with minimal required coding from the top.

Organisation A would therefore be judged as less complex than B if the expected code length is more economical for messages flowing between the top and the bottom layers. In theory it might possible to construct a distribution of message code lengths among units of the organisation and draw inferences as to comparative informational efficiency from its position and shape. A construction of this kind would draw together the distributional comparisons and metrics of the present book with those of organisational theory at large. However, this and related formalisations of the theory of social organisation remain an agenda for further research.

7.7 Literature Notes

The material in Sect. 7.1 on tail probabilities and risk management is based on Bowden (2011). Suarez and Menendez (2005) enlarge on the long tail problem for financial returns data. Many statistics textbooks carry some coverage on the extreme value family; an example is Johnson et al. (1994) as an authoritative source.

For Sect. 7.3, utility functions, and the cardinal-ordinal distinction, are covered in most textbooks of microeconomics, or in mathematical economics. However, books on game theory sometimes have a more apposite coverage. A concise and very readable example is Owen (1968), Chap. 6.

Turning to Sect. 7.4 on subjectivist probability and utility, this has from time to time resurfaced in statistics, and in the economics literature as behavioural economics. Notable contributions are Ramsay (1926), de Finetti (1937, 1970), Savage (1954), and Tversky and Kahneman (1974, 1979, 1992). Pfanzagl (1967) is useful for axiomatic foundations.

On fractional odds, see de Finetti (1937, 1970). Gibrat's law (Gibrat 1931) originally referred to growth in firm size as a rationale for the lognormal distribution, with no specific reference to subjectivism, i.e. the way people assess outcomes. Influences in the area of finance extend to the equity premium puzzle (Benartzi and Thaler 1995); size and value anomalies (Giorgi and Hens 2006); or the momentum effect (Grinblatt and Han 2005; Menkhoff and Schmeling 2006).

The mention of wavelets in Sect. 7.5 is intended only as an analogy, but nevertheless has some points of potential development in the general context of complexity. Daubechies (1992) remains a very readable introduction for non specialists, but there are many others.

References

Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, *100*, 73–92.

Bowden, R. J. (2011). Directional entropy and tail uncertainty, with applications to financial hazard. *Quantitative Finance, 11,* 437–446.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM: CBMS-NSF Regional Conferences Series in Applied Mathematics.

De Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. H. E. Kyburg, Jr. translation; reproduced in *Breakthroughs in statistics*. In S. Kotz and N. J. Johnson (Eds.), *1*, 134–174, New York: Springer Verlag, 1993.

De Finetti, B. (1970). *Theory of probability* (Vols. 1, 2). New York: Wiley. Also trans. A. Machi & A. F. M. Smith. New York: Wiley, 1974–5.

Gibrat, R. (1931). *Les inégalités économiques*. Paris: Librairie du Receuil Sirey.

Giorgi, E. D., & Hens, T. (2006). Making prospect theory fit for finance. *Financial Markets and Portfolio Management*, *20*, 339–360.

Grinblatt, M., & Han, B. (2005). Prospect theory, mental accounting and momentum. *Journal of Financial Economics*, 78, 311–339.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continous univariate statistics* (Vol. 1, 2nd ed). New York: Wiley.

Menkhoff, L., & Schmeling, M. (2006). A prospect theoretical interpretation of momentum returns. *Economics Letters*, *93*, 360–366.

Owen, G. (1968). *Game theory*. Philadelphia: W.R. Saunders.

Pfanzagl, J. (1967). Subjective probability derived from the Morgenstern—von Neumann utility concept. In M. Shubik (ed.) *Essays in mathematical economics in honor of Oscar Morgenstern*. Princeton: University Press.

Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and Other Logical Essays* London: Routledge. Also in D. H. Mellor (ed.) *Foundations*: *Essays in Probability, Logic, Mathematics and Economics*, 1978. London: Routledge.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Suarez, S., & Menendez, S. C. (2005). Computational tools for the analysis of market risk. In S. C. Menendez, A. S. Calle & L. Seco (Eds.), *Risk management in finance: Proceedings of the 1st Risk Lab international conference*. Bilbao: Fundacion B.B.V.A.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*. New Series (Vol. 185, pp. 1124–1131).

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–292.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation under uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Appendix

Excel VBA Code for the Compressed Smoothing Kernel

The code below can be copied into the Excel Developer studio to generate the complete smoothed series, including the end correction, as YSTAR. In the code below the output appears as a column of the same dimension (nobs) as the input series to the smoothed. As given in the code, the input series is read in as a range, but it can also be entered as a named vector. Note that the matrix-vector 'enter' must be used in either case. The first few elements of YSTAR are just the original elements of the input series until the first complete window is available.

The code below can also be used to output Shannon entropy, also the entropic bandwidth as in Chap. 5. A choice of kernels is available. The default utilised below is for the partition entropy kernel (denoted BiK). However for the function 'kernel' that appears in the main program code below, one can if preferred substitute the Epanechnikov or any other kernel, in a format that is consistent with the substantive code.

The code that follows is a based on the work of Tom Blaesche (utilised in Blaesche et al. 2016) and has been made freely available. The present author has made some amendments in order to ensure the code is fully consistent with the text. In the listing that follows, continuations of some longer commands or non executable notes are reverse indented.

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3

```
Option Explicit     ' Forces explicit declaration of all variables.
Option Base 1       ' Declares the default lower bound for array subscripts.
```

```
Public Function kernel(x As Integer, m As Integer) As Double
'The code below is for the BiK kernel. If the Epanechnikov kernel is used, it should span (-m-1
to m+1 values when edges are included)
'kernel = (1 / (m + 1)) * (3 / 4) * (1 - ((x - m) / (m + 1)) ^ 2)
'x ranges from 0 to 2m
Dim W As Double
Dim sum As Double
Dim n As Integer
sum = 0
For n = 0 To 2 * m Step 1
W = 0.5 * (1 + n) / (1 + m)
sum = sum - (W * Log(W) + (1 - W) * Log(1 - W))
Next n
W = 0.5 * (1 + x) / (1 + m)
kernel = -(W * Log(W) + (1 - W) * Log(1 - W)) / sum
End Function


Public Function YSTAR (data As Range, m As Integer) As Double()
   Const BandwidthLowerBound As Double = 0.161378  'Entropy Effective Bandwidth lower
bound
   Const BandwidthUpperBound As Double = 0.838622  'Entropy Effective Bandwidth upper
bound
   Dim window As Integer
   Dim nObs As Integer
   Dim lambda As Double
   Dim increment As Double
   Dim expect As Double
```

```
    Dim temp As Double
    Dim min As Integer, max As Integer
    Dim i As Integer, j As Integer
    Dim n As Integer, s As Integer, t As Integer
    Dim EffectiveBandwidth() As Integer
    Dim ShannonEntropy() As Integer
    Dim ShiftedKernel() As Double
    Dim ShiftedKernelDist() As Double
    Dim yOut() As Double
    window = 2 * m          ' 0 to 2m contains 2m+1 values
    nObs = data.Rows.Count  ' Number of observations
    lambda = 0              ' Starting parameter for partial kernel shifts
    increment = 0.00001     ' Increment of changing lambda in solving for best lambda according
to expectation value
     ReDim EffectiveBandwidth(0 To m)
    ReDim ShannonEntropy(0 To m)
    ReDim ShiftedKernel(0 To m, 0 To window)
    ReDim ShiftedKernelDist(0 To m, 0 To window)
    ReDim yOut(1 To nObs, 1)
    ' Calculate Shifted Kernels for all shifts from 1 To m
    For i = 0 To m Step 1
        If (i = 0) Then
            ' Define Base Kernel
        For j = 0 To window Step 1
           ShiftedKernel(0, j) = kernel(j, m)
        Next j
                ' Calculate current expectation value
        expect = 0
        For j = 0 To window Step 1
           expect = expect + j * ShiftedKernel(i, j)
        Next j
           Else
        Do
           ' Shift Kernels
```

```
            For j = 0 To window Step 1
               ShiftedKernel(i, j) = ShiftedKernel(i - 1, j) * (1 - lambda * (1 +
Log(ShiftedKernelDist(i - 1, j))))
            Next j
             ' Calculate Normalizing Coefficient
            temp = 0
            For j = 0 To window Step 1
               temp = temp + ShiftedKernel(i, j)
            Next j
               ' Normalize Shifted Kernel
            For j = 0 To window Step 1
               ShiftedKernel(i, j) = ShiftedKernel(i, j) / temp
            Next j
            ' Calculate current expectation value
            expect = 0
            For j = 0 To window Step 1
               expect = expect + j * ShiftedKernel(i, j)
            Next j
               ' Redefine lambda
            lambda = lambda + increment
                     ' Error handler
            If lambda >= 1 Then
               Exit Do
            End If
            ' Re-calculate shifted Kernel if numerical expectation value is not close enough to
analytical value
         Loop Until (expect - (m - i) < 0)
               End If
      ' Calculate Distribution Functions for base Kernel and all shifts from 1 to m
      min = 0
      max = 0
      For j = 0 To window Step 1
         If (j = 0) Then
            ShiftedKernelDist(i, 0) = ShiftedKernel(i, 0)
```

```
            Else
               ShiftedKernelDist(i, j) = ShiftedKernelDist(i, j - 1) + ShiftedKernel(i, j)
            End If
                  If (ShiftedKernelDist(i, j) >= BandwidthLowerBound) And (min = 0) Then
               min = j
            End If
                  If (ShiftedKernelDist(i, j) >= BandwidthUpperBound) And (max = 0) Then
               max = j
            End If
            Next j
         ' Calculate Effective Bandwidth
         EffectiveBandwidth(i) = max - min
         ' Calculate Shannon Entropy
         ShannonEntropy(i) = 0
         For j = 0 To window Step 1
         ShannonEntropy(i) = ShannonEntropy(i) - ShiftedKernel(i, j) *
WorksheetFunction.Ln(ShiftedKernel(i, j))
         Next j
      Next i
      For t = 1 To m Step 1
         yOut(t, 1) = data(t, 1)
      Next t
      For t = m + 1 To nObs Step 1
               If t <= nObs - m Then
            s = 0
         Else
            s = t - (nObs - m)
         End If
               For n = 0 To window Step 1
            yOut(t, 1) = yOut(t, 1) + data(t - s + n - m, 1) * ShiftedKernel(s, window - n)
         Next n
            Next t
        YSTAR = yOut

End Function
```

Subject Glossary

Activist fund An investment fund whose managers try to beat the market. The opposite is a passive fund: an exchange traded fund (ETF) is a passive variety whose portfolio weights simply mirror those of a major index such as the S&P500

Actuarial survival function The proportion of people who survive to a given age. Often tabulated in terms of people born in a specific year (a cohort life table), or age groups at one particular time (a period life table)

Asymmetry metric A signed number that measures the extent to which a given density is mirrored about its median. To be accepted as such it must satisfy a number of qualifying conditions

Bandwagon effect Copycat or social comfort driven behaviour in attitudes or investment behavior. Can lead to rational expectations where it is optimal to jump on the bandwagon

Binary code A way of representing symbols in common use such that they can be recognized and acted upon by a digital computer. Hence elements as 0, 1 for 'on' or 'off'

Cardinal ordinal utility Where the subject is able to rank differences in his or her satisfaction. The resulting utility function is unique up to a linear affine transformation. As distinct from an ordinal utility function, for which any positive monotone transformation will do

Conditional value at risk The expected value of an outcome variable, given that it is less than an assigned prudential benchmark value

Co–smoothing A process of taking the expected value of one variable, conditional upon values of another being less than some assigned value or benchmark

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3

Cumulative prospect theory A behavioural formalisation of the idea that people do not behave according to the strict laws of probability. Proposes to substitute the notion of a weighting function for the latter

Differential entropy The expected value of the log density, as the integral over a continuous range of the given random variable. The version of Shannon entropy for such a variable

Dirac delta function A spike of technically infinite height at a given value along the range, so all the mass is concentrated at this point. As such, a technical device used in control theory. In the case of a random variable, becomes the Dirac delta density

Double smoothing property The overall expected value of a variable or random function that is itself formed as a conditional expectation up to any assigned point

End correction Techniques used in the context of kernel based time series smoothing. Required when the time of current focus approaches the last observation, so that the full width of the kernel is not available

Entropic asymmetry metric A measure for distribution asymmetry that is based on the algebraic difference between the average above and the average below any given point along the range. The expected value of such differences turns out to be equal to the average of means of the right and left entropic shifts less than the mean of the original distribution

Entropic spread metric As for the asymmetry metric except that the difference is of the absolute values of the progressive conditional expectations. The result is the difference between the means of left and right entropic shifts. The greater this difference, the wider the spread

Equivalent width A non parametric spread measure as the distance between the left and right entropic delta concentrators. Proportional to the area under the partition entropy function

Expected shortfall In risk management, refers to the expected value of the difference between an accepted prudential benchmark and a prospective outcome, conditional upon violation of the benchmark, i.e. the bad scenario

Friedman Savage utility function Is concave downwards for losses, and convex upwards for gains. Investors are assumed to be risk averse in the loss region but risk loving in the gain region

Gibrat's law People think or respond more to proportional (or log) changes than to arithmetic ones

Gini coefficient Measures the lack of concordance between the progressive proportion of people and the progressive proportion of the total income they enjoy

Gini mean absolute difference Measures the average weighted distance between any two randomly drawn observations from the same distribution function, with the weights provided by the density function

Internal observer A construction that averages how subjects might perceive themselves in relation to others

Lake Woebegone (or Wobegone) effect Where everybody is better than the average. A humorous sign off in the US radio show of Garrison Keillor

Locational entropy Synonym for partition entropy (q.v.)

Lorenz curve In income distribution, the graph of the proportion of income earned by the progressive proportion of people. Leads to the Gini income distribution metric

Message set An input sequence of symbols, that is to be coded into binary form

Mixture distribution Combines two or more distributions defined on the same range, where the constituent distributions are of the same general type. Combining a continuous distribution with a discrete one would be referred to as a mixed distribution

Ordered mean difference A schedule that graphs the conditional expected values of one variable against values of another benchmark variable, the latter as the conditioning factor. See also 'co-smoothing'

Ordinal utility function A weaker preference ordering than cardinal (q.v.). Objects or outcomes can be preferred one to the other, but not their respective differences (how much more preferred)

Partition entropy At any given point along its range of a random variable, is the entropy of a binary variable that takes value 1 if values exceed the given point, or 0 if less. The collective over all such given points, is the partition entropy function

Rawls social justice criterion Says that the only fair social orderings should be established by ex ante agreements based on advance ignorance of who or what (e.g. how wealthy) I will turn out to be

Retracement point Trading recommendation that says to buy when the price has fallen to a designated proportion off its previous peak. So the price recovers

Scalable kernel A smoothing kernel specification that adjusts automatically for alterations in the window length

Shannon entropy The expected value of the log of the density, technically with the binary 2 as the base for the log, but in practice usually as the natural log. Captures the expected length of a message code

Subjective probability Postulated existence of systematic biases in peoples' assessment of future events. Commonly to overweight extremely favourable outcomes, as in lotteries

Tversky-Kahneman postulates Stress the importance of the way that prospective events are presented to a subject, also bias in recombining probabilities of compound events, and overestimating rare outcomes

Value at risk An assigned critical point for value outcomes in the lower tail, commonly 1 or 5%, such that the chosen portfolio should not fall short with a greater probability, over a designated time interval

Index

R. Bowden, *The Information Theory of Comparisons*,
https://doi.org/10.1007/978-981-13-1550-3