Every student of probability knows that $E(X + Y) = E(X) + E(Y)$. In words, given two random variables, the expected value of their sum is the sum of their expected values.

Based on the logarithm, this book introduces the fundamental concept of expected information $E_I(X)$, which has the remarkable property $E_I(X \cdot Y) = E_I(X) + E_I(Y)$.

Consistently, the corresponding theory is developed, which tackles

- higher logarithmic moments of random variables & their distributions, and the 'information variance' in particular,
- modified information parameters and associated logarithmic functions,
- the discrete and the multidimensional cases,
- general patterns — most notably the algebra of random variables, and the stochastic Askey scheme, i.e., a logical map of all distributions,
- the Signal-to-Noise paradigm.

Since the book extends basic stochastic ideas and touches upon combinatorics and special functions, it should be of interest to all those dealing with probability and its ilk — most statisticians, many mathematicians, quite a few scientists, and some philosophers.

The book of nature is written in the language of mathematics, and information ranks among its key concepts.

About the Author

Uwe Saint-Mont is a mathematician and statistician whose research focuses on the foundations of probability theory, information theory, and the philosophy of statistics. He has published extensively on these topics, including works on the history and philosophy of statistics. This book represents a synthesis of ideas that began in 2018, when he sought a unified mathematical framework connecting information, probability distributions, and the signal-to-noise paradigm.

Information is Key

Uwe Saint-Mont

Information is Key

Parametric Information Theory, The Signal-to-Noise Paradigm, Stochastic Distributions and Logarithmic Functions

Uwe Saint-Mont

Information is Key

Parametric Information Theory, The Signal-to-Noise Paradigm, Stochastic Distributions and Logarithmic Functions

By Uwe Saint-Mont

First Edition, 2026

ISBN 979-8-249695-18-7

This book was designed and typeset using LaTeX,
based on the LODE template
https://www.lode.de/template.

This document uses Libertinus, STIX Two Math, Source Sans Pro, and Source Code Pro fonts, which are licensed under the SIL Open Font License, Version 1.1.

Contents

Dedication

This book is dedicated to the memory of my much admired teacher
Ester Samuel-Cahn (1933-2015)

Preface

It is always a pleasure to thank all those who helped with such a large project.

First and foremost, I would like to thank my wife who has endured extended phases of mental absence and shifting moods from her husband. You know that I did not really want to write this book, but somehow the subject wanted to be developed. Now it is finished and I am back. Mea culpa for all the inconvenience. At least our children, students at the time of working on this project, learned first-hand that research has much to do with inspiration and transpiration. (Some say that the proportion is 20:80 or even 10:90.) Since you are all very bright, your input and support was very welcome!

Second, several colleagues listened to what I had to say, and their reaction helped me in proceeding. Back in 2018, Thomas Augustin and his colleagues in Munich gave me initial feedback. Later, my former colleague and dear friend Georg Baumbach read a draft of the manuscript. When I presented the subject matter in Bochum at the IMS World Congress on Probability and Statistics 2024, and the following year in Berlin at the DaGStat-conference, the response was encouraging, especially that of the younger generation.

Thirdly, I would also like to thank my former publisher, who did not know what to do with the manuscript. Thus, I had to search for an alternative and found that direct online publishing was a contemporary and attractive option. If you do most of the work, why shouldn't you be the one to profit? All you need is a competent partner, and I was very happy to find that Clemens Lode (www.lode.de) could provide all the services I needed.

Last but not least, since I am not a native speaker of English, I (still) need some assistance with respect to language. (Well, who does not?) I would have loved to rely more on trusted native speakers such as Conna Craig (Lode Publishing) and my former in-company teacher and English language coach Mike Seymour, Leeds. The latter has been instructing me on the subtleties of the English language for more than twenty years. However, due to monetary constraints, I had to rely mainly on Writeful, Grammarly, and DeepL, all of which were helpful.

Finally, I would like to thank my employer, who gave me sufficient time and resources to develop the ideas that I am now able to present. In particular, our library was able to procure anything I needed, our IT services department kept the machines running, and our Research Advisory Board did not hesitate to grant me a research sabbatical.

Of course, any remaining errors - may there be few - are mine alone.

P.S.: This list would be incomplete without mentioning the countless and mainly anonymous people who have built such wonderful tools as Mathematica, LaTeX (WinEdt, Overleaf), Google Scholar and their siblings. In a nutshell, this work could not have been done without their assistance, and the information they provided proved to be crucial.

Uwe Saint-Mont

Nordhausen, Germany, February 2026

“

The truth always turns out to be simpler than you thought.

—Richard Feynman

1 Information in an observation

1.1 Motivation and introduction

Shannon (1948) defined the information in a **probability** p to be $I(p) = \ln(1/p) = -\ln p \geq 0$. The standard interpretation is that the smaller $p \in [0,1]$, the larger the surprise (information) in p. Thus, $I(p)$ is a decreasing function in p. If $p = 1$ (the sure event), one does not learn anything new, and $I(1) = 0$. On the other hand, an impossible event A with $p = p(A) = 0$ is truly remarkable and exceeds one's frame of reference, therefore $I(0) = \infty$. Moreover, the logarithm is a natural choice if information should be additive, and the information in a probability is always non-negative.[1]

The crucial idea of almost everything that follows is to extend the latter idea to an **observation** x. That is, given $x \geq 0$, one defines $I(x) = \ln(1/x) = -\ln x$. More generally, $I(x) = -\ln|x|$ if x is any real number. The interpretation of this straightforward definition is that the information is located in the distance of the point x from the origin, where the logarithm has a singularity. The farther away from the origin, the less information in x about zero. Like before, $I(x) = -\ln x$ is a decreasing function for all $x \in \mathbb{R}^+$.

This basic idea occurred to me in early 2018, when I was looking for a general and elegant mathematical model that could be able to describe convergent and divergent phenomena. I thought in terms of signal S versus noise N, and my intuition at that time was that if $S > N$, information should accrue and lead to some kind of convergence. However, if $S < N$, some given structure would rather dissolve or disassemble. Working with several distinguished scientists on a book on the history and philosophy of statistics at that time, Reinhard Viertl from Vienna Technical University always emphasised that observations are Janus-faced. Being an outstanding engineer, he insisted that although statisticians like to emphasise the information in their observations, they also like to downplay their dark side. Observations are plagued with errors and vagueness, and can even be straightforwardly misleading.

Moreover, ever since I had written my account (Saint-Mont 2011) of statistics' philosophy, I have had the impression that classical information theory was incomplete. Somehow, the latter theory started very simple, general, and promising. However, it has not really lived up to expectations; in other words, one should be able to do more with the idea of information. Extending the classic information function $I(x)$ beyond the unit interval seemed to be a promising first step in formalising this set of ideas. The tour de horizon that follows is its natural consequence.

[1] It may be mentioned that the logarithm also connects Kolmogorov complexity and universal probability, and since the logarithm is a 1:1 function, these two concepts are, in this sense, equivalent. For details, see Cover and Thomas (2006), Section 14.11.

Rather than just describing that information, probability and statistics are closely related, and proposing that these fields should be combined into a single data science, this work fuses two basic concepts of probability and information theory — i.e., probability distributions and information — into a unified conceptual framework. On the one hand, the latter framework has deep mathematical roots in the theory of (analytical and special) functions and geometry; on the other hand, it may serve as the centrepiece of a general paradigm, emphasising signal and noise.

In more detail, this book is about

1. distributions and the transformations that are tying them together, thus producing a large (hierarchical) net or system of distributions
2. random variables and their algebra, in particular products and ratios
3. the logarithm and related functions
4. logarithmic moments that supplement ordinary moments and integral transforms such as the Laplace transform
5. a parametric theory of information, based on the logarithmic distance relative to a centre
6. unifying the basic concepts of information and probability.
7. the extraordinarily tight connections between the fields of stochastics, geometry and the theory of analytic functions
8. the signal-to-noise paradigm, distinguishing between 'good' and 'bad' information, which - implicitly - has been the leading 'statistical philosophy'. In particular, statistics is applied information theory.

So what have we got? Given some distinguished point x_0 (which may be interpreted as the "true value"; for convenience, very often, $x_0 = 0$), all observations x close to x_0 possess positive information about x_0; observations that are far away are misleading and thus possess negative information. A "direct hit" ($x = x_0$) means an infinite amount of information. $-\ln|x - x_0| = 0 \Leftrightarrow |x - x_0| = 1$ may be interpreted as a kind of crucial distance, separating points close to the centre x_0 (i.e., central points) from those far from the centre (peripheral points). Or, thinking in terms of statistical distributions with expected value μ and standard deviation σ, the crucial distance is defined by $-\ln|(x - x_0)/\sigma| = 0 \Leftrightarrow |x - x_0| = \sigma$.

If, without real loss of generality, $x_0 = 0$, one distinguishes, qualitatively speaking, three kinds of observation or set:

1. "good" observations, $I(x) > 0 \Leftrightarrow |x| < 1$, defining the centre $\mathcal{C} = \{x|I(x) > 0\}$,
2. "neutral" observations, $I(x) = 0 \Leftrightarrow |x| = 1$, i.e., the margin $\mathcal{M} = \{x|I(x) = 0\}$,

3. "bad" observations, $I(x) < 0 \Leftrightarrow |x| > 1$, that is, the periphery $\mathcal{P} = \{x|I(x) < 0\}$.

Good observations add to the 'signal'. That is, they tell something about truth and, if taken together, they may come closer to truth than any single one of them. Yet, bad observations constitute 'noise' that dilutes our state of knowledge. In the worst case, observations are treacherous and lead us down the primrose path.[2]

Note that the transformation $y = -x$ interchanges left and right. Symmetric probability densities f respect that transformation, i.e. $f(-x) = f(x)$. In the same vein, the transformation $y = 1/x$ exchanges the centre and the periphery. The corresponding symmetry is $p(x) = p(1/x)$ for all $x \neq 0$ in the case of a discrete distribution. Due to the singularity of the logarithm at the origin, the mass in zero requires special treatment in the case of discrete probability distributions.

A more abstract point of view would be to start with the positive semi-axis $\mathbb{R}^+$ or the reals $\mathbb{R}$. One has the additive group $(\mathbb{R}, +)$ and the multiplicative group $(\mathbb{R}, \cdot)$. Their neutral elements, that is, the distance $|1 - 0|$, defines an (absolute) scale. Thus $\mathcal{C} = (-1, 1), \mathcal{M} = \{-1, 1\}$, and $\mathcal{P} = \{x|1 < |x|\}$.

The transformation $y = -x$ maps the left semi-axis to the right semi-axis and vice versa. In other words, it exchanges the sign. The transformation $y = 1/x$ exchanges big and small (the centre and the periphery). That is, a number that is close to the origin is mapped to a remote value with a large absolute value. Restricted to the positive semi-axis, $1/x$ is an injective mapping that interchanges the origin and the point ∞. Given $\mathbb{R}$, the points $\pm\infty$ are both mapped to 0. Note that the mass located to the right of the origin remains on the positive semi-axis, since $1/x$ is a point reflection at the point 1. The same is true for the mass to the left of the origin. There, -1 is the fixed point of the transformation $1/x$.

Given a symmetric probability density f about the origin, it is tantamount to study $2f(x)$ on the positive semi-axis. Let S be a random variable such that $p(S = -1) = p(S = 1) = 1/2$ (sometimes named after Rademacher), and suppose $X \geq 0$ has pdf h(x) on $\mathbb{R}^+$. If S and X are independent, the random variable $S \cdot X$ has pdf $h(x)/2$ on $\mathbb{R}$, since the mass in each $x > 0$ is distributed equally among the points $x, -x$.

Moreover, given any function g with $g(x) = g(1/x)$ on $\mathbb{R}^+$, then the parameter integral $K(s) = \int_0^\infty \frac{g(x)}{(x^s+1)x}\,\mathrm{d}x$ does not depend on s (see Boros and Moll (2004), pp. 252-253). Here is their short proof: "Split the integral into two pieces on $[0, 1]$ and $[1, \infty)$ and make the substitution $x \mapsto 1/x$ in the second. Then

$$K(s) = \int_0^1 \frac{g(x)}{x^s + 1}\frac{\mathrm{d}x}{x} + \int_0^1 \frac{x^s g(1/x)}{x^s + 1}\frac{\mathrm{d}x}{x} = \int_0^1 g(x)\frac{\mathrm{d}x}{x}$$

and the last expression is independent of s." This result is pivotal for their so-called "Master Formula."

[2]More generally and neutrally speaking, there is always signal and noise, see Section 6. Depending on the circumstances, upon accumulating events (observations), some structure may be static, consolidate or erode (remain as it is, converge or diverge).

1.2 Distance relative to a distinguished point

The idea of measuring distance with respect to a pole or some other interesting point is by no means extraordinary. In fact, it is used in many areas of mathematics.

For instance, there is only a single norm on $\mathbb{R}$ that does not depend on a distinguished point, which is the ordinary norm $|x|$. Consistently, Euclidean distance between two real numbers $d(x, y)$ does not depend on some origin or a 'third number'. In a sense, this is an exception, since in spherical and hyperbolic geometry (Poincaré's disk model), there is a distinguished point: the centre of the circle considered. Likewise, using polar coordinates, the origin stands out, since if $z = r\cos\varphi \in \mathbb{C}$, r is just the distance from zero.

Given $\mathbb{Q}$, we have the same phenomenon, i.e., all other kinds of distance on this set are based on a prime p, and p-adic distance or norm (cf. Gouvêa (1997)), is defined by $|a|_p = 1/p^m$, where $a = b/c \in \mathbb{Q}$ and p^m is the largest power of p in the prime number decomposition of a, i.e., the largest natural number m in the equation $a = p^m b'/c'$, where neither b' nor c' are multiples of p. With respect to this norm, a number is small if it 'contains' a large power of p.

In graph theory, trees are particularly important. Trees are connected graphs with a minimum number of edges, and they possess a root, i.e., a vertex that may serve as the point of origin of the graph. A well-known example to mathematicians is the Erdős number.[3] In general, the (non-parametric) distance of some vertex from the root is the number of vertices between the root and the vertex in question. Of course, this idea can be modified in a variety of ways.[4]

Kullback-Leibler divergence $D(f||g)$, see Section 1.10, is also a concept of a distance, *relative* to a distinguished (and fixed) function g. Since g is a function and not a single number, there is some reason to name it 'non-parametric'. In a very similar vein, $-\ln|x| = -\ln|x-0|$ measures the distance of x from zero. Therefore, it makes sense to think of it as a parametric (and logarithmic) kind of distance, see Section 3.22, in particular.

From a philosophical point of view, the use of the logarithm is just a minor technical detail. However, it will turn out that from a mathematical point of view, the logarithm is fundamental. Therefore, we may have used the subtitle "a study of form based on the logarithm". Moreover, since the gamma function appears on every other page, there

[3] Paul Erdős (1913-1996) published an incredible number of papers. He is associated with the number 0. If a mathematician wrote a paper with him, this person has Erdős number 1. People with Erdős number two did not publish with Erdős but with somebody with Erdős number 1, etc. Thus one gets a graph with Paul Erdős at its centre. (The graph of all mathematicians is not connected, since there are researchers who have never published with anybody else. Consistently, their Erdős number is ∞.)

[4] A parametric kind of distance can be found in a weighted graph, i.e., every edge has a certain length to it. Considering two airports or stations, their degree of linkage could also be defined by the number of planes or trains connecting them in a day, say. In other words, two airports are close if there are many flights connecting them. (Note that $d(A, B)$ is not a distance in the mathematical sense: If B is a hub and A, C are regional airports, $d(A, C)$ may be larger than $d(A, B) + d(B, C)$.) The same idea could also be applied to the brain: if A and B are brain regions, they are close, if $d(A, B) = 1/n$, or $d(A, B) = 1/(1 + \ln n)$ is small, where n is the number of neuronal links between these areas, and $d(A, A) = 0$.

was also some justification to mention that special function in the title, or 'information functions' which generalise the logarithm, for instance $\ln(\Gamma(x))$.

1.3 The Cauchy distribution

Considering pdfs on $\mathbb{R}$, a probability density that observes both kinds of symmetry (about 0 and 1, respectively) demands special attention. As already mentioned, any symmetric density about the origin observes $f(x) = f(-x)$. For the second symmetry about one, the following lemma is needed:

Lemma 1.3.1. *(Density of the inverse random variable). Suppose X is a real-valued r.v. with density $f(x)$ and distribution function $F(x)$. Then we have for the density $g(y)$ of the r.v. $Y = 1/X = X^{-1}$*

$$g(y) = \frac{f(1/y)}{y^2} \qquad \text{if} \qquad y \neq 0.$$

Proof. Suppose $y < 0$ and thus also $x = 1/y < 0$. This yields for the distribution function G of Y on the negative semi-axis

$$G(y) = p(Y \leq y) = p(0 > X \geq 1/y) = F(0) - F(1/y)$$

Differentiating with respect to y, one gets $g(y) = -f(1/y)(-1/y^2) = f(1/y)/y^2$.

(Obviously, we also have $p(Y \leq 0) = G(0) = F(0) = p(X \leq 0)$.) If $y > 0$ and thus also $x = 1/y > 0$, we obtain

$$\begin{aligned} G(y) &= p(Y \leq y) = p(Y \leq 0) + p(0 < Y \leq y) \\ &= F(0) + p(X \geq 1/y) = F(0) + (1 - F(1/y)) \end{aligned}$$

By the same token, we obtain on the positive semi-axis $g(y) = -f(1/y)(-1/y^2) = f(1/y)/y^2$, which concludes the proof. □

Note that f is *not* symmetric about 1 (or -1), i.e., $f(x) \neq f(1/x) = f(y)$, and also $g(y) \neq g(1/y)$. It is also *not* true that $f(x)$ and $g(y)$ coincide, i.e., in general $f(x) \neq g(y)$. Rather, the last lemma says that $f(x) = f(1/y) = y^2 g(y)$.

The intuition is that the mass in a small environment $U_\delta(x)$ of $0 < x < 1$ is stretched to a large environment $U_{\varepsilon(\delta)}(1/x) = U_\varepsilon(y)$ of y. A small $\delta > 0$ implies a large $\varepsilon = \varepsilon(\delta)$. Thus, the density $g(y)$ at some peripheral point $y = 1/x$ has to be multiplied by a huge factor (y^2) to get $f(x) = f(1/y)$ back.

Of course, this consideration does not apply if X has a discrete distribution, then $p(X = x) = p(Y = y)$ where $y = 1/x$.

Definition 1.3.2. *(Density of the Standard Cauchy). If $X \sim C(0,1)$, the density of this r.v. is $c(x) = \frac{1}{\pi(1+x^2)}$.*

Note that

$$c(x) = \frac{1}{\pi(1+x^2)} = \frac{1}{\pi(1+(-x)^2)} = c(-x)$$

and

$$\frac{c(1/x)}{x^2} = \frac{1}{x^2\pi(1+(1/x)^2)} = \frac{1}{\pi(x^2+1)} = c(x).$$

That is, the Cauchy distribution observes both symmetries, $X \sim C(0,1)$ is a fixed point with respect to the transformation $y = 1/x$. In other words, $X \overset{d}{=} 1/X$, and we also have

$$\int_{-\infty}^{-1} c(x)\,\mathrm{d}x = \int_{-1}^{0} c(x)\,\mathrm{d}x = \int_{0}^{1} c(x)\,\mathrm{d}x = \int_{1}^{\infty} c(x)\,\mathrm{d}x = \frac{1}{4}. \tag{1.1}$$

Of course, there are other densities with property (1.1), for example $h(x) = 1/(4x^2)$ if $x \in \mathcal{P}$, and $h(x) = 1/4$ otherwise. Due to its peculiar shape, $h(x)$ might be named "the stump", see Fig. 1.1.

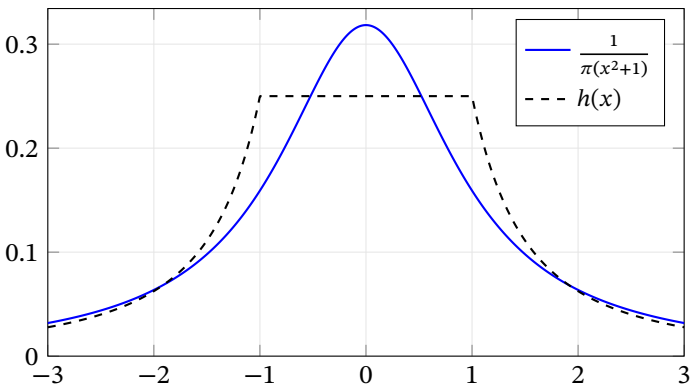

Figure 1.1: Density of the Standard Cauchy, and *h(x)*, see text.

Koosis (1988, 2009) has written two books about the "logarithmic integral"

$$\int_{-\infty}^{\infty} \frac{\log M(t)}{1+t^2}\,\mathrm{d}t$$

with quite diverse functions $M(t)$, since this integral is important in (functional) analysis. Given the perspective of this book, it is straightforward to extend his analysis by studying

1. various pdfs $g(t)$ of important statistical distributions,

2. functions of the form $I(t) = (\pm)\ln f(t)$, named information-extracting functions (iefs) in what follows.

Of particular interest are the integrals $\int I(t)g(t)\,\mathrm{d}t$.

1.4 Logarithmic expected value

The information at some point x is $I(x) = -\ln x$. Given a random variable X with distribution $\mathcal{D}(X)$, it is of course straightforward to ask for the mean or expected information of the random variable $E_I(X)$. This value can also be interpreted as the typical amount of information that an observation x from the parent $\mathcal{D}(X)$ contains. It turns out that this particular number is like the 'signature' of the distribution in question, and helps tremendously in describing the 'landscape' of distributions.

Definition 1.4.1. *(Expected Information). Suppose $X \sim \mathcal{D}(X)$, and $p(X = 0) = 0$. If X is discrete, define*

$$E_I(X) = E_I(\mathcal{D}(X)) = \sum_i (-\ln|x_i|)p(X = x_i). \tag{1.2}$$

If X has a density $f(x)$, let

$$E_I(X) = E_I(\mathcal{D}(X)) = E_I(f) = \int (-\ln|x|)f(x)\,\mathrm{d}x\,. \tag{1.3}$$

Of course, $f(0)$ is irrelevant in the latter case, since $p(X = 0)$ is zero.

Examples with densities:

1. **Standard Cauchy** $C(0,1)$:

$$\begin{aligned} E_I(C(0,1)) &= \int_{-\infty}^{\infty} (-\ln|x|)c(x)\,\mathrm{d}x = 2\int_0^{\infty} (-\ln x)\frac{1}{\pi(1+x^2)}\,\mathrm{d}x \\ &= 2\left(\int_0^1 \frac{-\ln(x)}{\pi(1+x^2)}\,\mathrm{d}x + \int_1^{\infty} \frac{-\ln(x)}{\pi(1+x^2)}\,\mathrm{d}x\right) \\ &= \frac{2}{\pi}(\beta_2 - \beta_2) = 0, \end{aligned}$$

where $\beta_2 = -\int_0^1 \ln(x)/(1+x^2)dx \approx 0.915965$ is Catalan's constant (see Section 7.1.5). In the second equation, we use the axial symmetry with respect to the

y-axis. The transformation $y = 1/x$ (and thus $x = 1/y$ and $\mathrm{d}x = d(1/y) = (-1/y^2)\,\mathrm{d}y$) yields for the last integral

$$\int_1^\infty \frac{-\ln x}{\pi(1+x^2)}\,\mathrm{d}x = \int_1^0 \frac{-\ln(1/y)}{\pi(1+1/y^2)}\left(\frac{-1}{y^2}\right)\mathrm{d}y = \frac{1}{\pi}\int_0^1 \frac{\ln y}{y^2+1}\,\mathrm{d}y = -\frac{\beta_2}{\pi}$$

and is well-known in the theory of functions. Figuratively speaking, centre and periphery cancel each other out. Physically speaking, some 'centripetal force' is exactly as strong as its 'centrifugal' counterpart.

2. **Standard Normal** $N(0,1)$: The density here is $\varphi(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$. Using the transformation $y = x^2/2$ and thus $x = \sqrt{2y}$ and $\mathrm{d}x = \mathrm{d}y/\sqrt{2y}$, the integral becomes

$$\begin{aligned}
E_I(N(0,1)) &= \int_{-\infty}^{\infty}(-\ln|x|)\varphi(x)\,\mathrm{d}x = 2\int_0^\infty \frac{(-\ln x)\exp(-x^2/2)}{\sqrt{2\pi}}\,\mathrm{d}x \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\int_0^\infty \frac{(-\ln(\sqrt{2y}))e^{-y}}{\sqrt{2y}}\,\mathrm{d}y \\
&= \frac{1}{\sqrt{\pi}}\int_0^\infty \left(-\frac{1}{2}\right)(\ln 2 + \ln y)\frac{e^{-y}}{\sqrt{y}}\,\mathrm{d}y \\
&= -\frac{1}{2\sqrt{\pi}}\left((\ln 2)\int_0^\infty \frac{e^{-y}}{\sqrt{y}}\,\mathrm{d}y + \int_0^\infty (\ln y)\frac{e^{-y}}{\sqrt{y}}\,\mathrm{d}y\right) \\
&= \frac{-\left((\ln 2)\sqrt{\pi} - \sqrt{\pi}(\gamma + 2\ln 2)\right)}{2\sqrt{\pi}} = \frac{\gamma + \ln 2}{2} \approx 0.6352 > 0,
\end{aligned} \tag{1.4}$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant, and $\ln 2 \approx 0.6931$. Both constants will appear very often throughout this text. Moreover, we used the integrals $\int_0^\infty e^{-y}/\sqrt{y}\,\mathrm{d}y = \sqrt{\pi}$ and $\int_0^\infty (\ln y)e^{-y}/\sqrt{y}\,\mathrm{d}y = -\sqrt{\pi}(\gamma + 2\ln 2)$. For a more general result, see Moll (2014), p. 71.

The last illustration shows that the Cauchy has heavier tails than the Normal. However, apart from this rather coarse received point of view, one also observes that the term $I(x)\cdot f(x)$ is larger for the Normal in the central area $\mathcal{C}$. In the interval $[1, \sqrt{-2W_{-1}\left(-\frac{1}{\sqrt{2e\pi}}\right) - 1}] \approx [1, 1.85123]$, where W_{-1} is the lower branch of the Lambert W function, the Cauchy is better than the Normal. Overall, a typical observation from a Normal produces a positive amount of information, which is the deeper reason why iid $X_i \sim N(0,1)$ yield that $\bar{X}_n = \sum_{i=1}^n X_i/n$ converges toward a number (LLN), whereas iid $X_i \sim C(0,1)$ produce $\bar{X}_n \sim C(0,1)$, i.e., a r.v. with a non-trivial distribution.

3. **Standard Exponential** Exp(1). The density of this distribution is $f(x) = e^{-x}$ for $x \geq 0$. Notice that $-\ln x$ is just the inverse function of e^{-x}. The expected

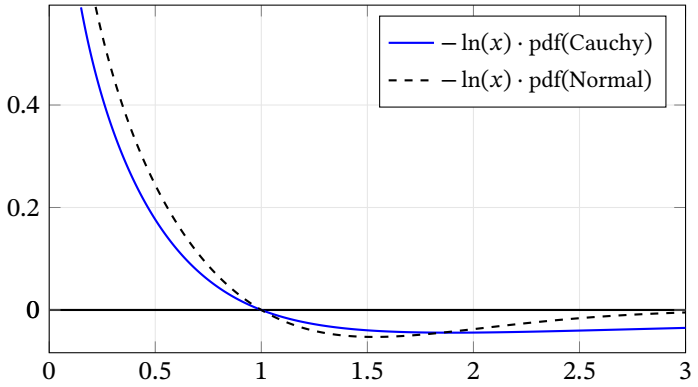

Figure 1.2: Comparison of the information in a Cauchy and in a Normal.

information $\int_0^\infty (-\ln x)e^{-x}\,dx = \gamma > 0$ is sometimes used as the definition of the Euler-Mascheroni constant γ.

4. **Standard Lévy** Lévy(0, 1). If $X \sim N(0,1)$, then $Y = 1/X^2$ has a Lévy distribution with density $f(x) = e^{-\frac{1}{2x}} x^{-3/2}/\sqrt{2\pi}$ for $x > 0$. It will be shown later that $E_I(Y) = -\gamma - \ln 2 < 0$, see Corollary 1.5.3.

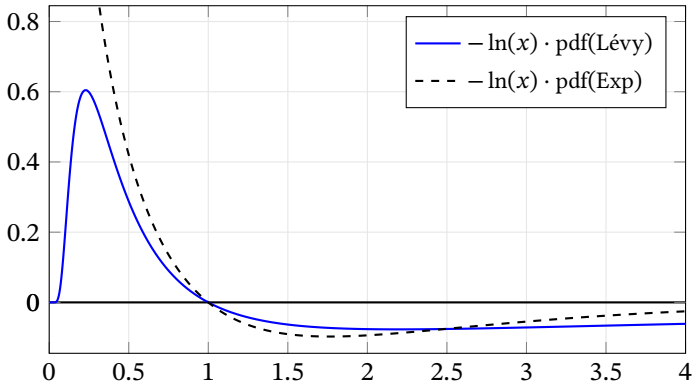

Figure 1.3: Comparison of the information in an Exponential and in a Lévy.

Fig. 1.3 is quite similar to Fig. 1.2: The Lévy has a considerably heavier tail than the Exponential, and the Exponential generates more information in $\mathcal{C}$. Again, there is an interval $(1, x_0)$, where $I(x) \cdot f(x)$ is larger for the Lévy (having smaller overall E_I). Straightforward calculations show that $x_0 \approx 2.486$ is the unique point in $[1, \infty)$, where $2x - 3\ln x = \ln(2\pi) + 1/x$.

5. **Uniform** distribution $U(0, c)$ on $[0, c]$, i.e. $f(x) = 1/c$ for $0 \le x \le c$. We obtain $E_I(U(0,c)) = \int_0^c (-\ln x)/c\,dx = 1 - \ln c$. Note that $E_I(U(0,1)) = 1$ for the Standard Uniform ($c = 1$), and $E_I(U(0,c^*)) = 0 \Leftrightarrow c^* = e$.

6. If $X \sim U(0,1)$, then $Y = 1/X$ has a Standard Type-1 **Pareto** distribution Pareto$(1,1)$ with density $f(x) = 1/x^2$ on the interval $[1,\infty)$. (In general, a Pareto(c,a) is defined on the interval (c,∞) and has pdf $f(x) = ac^a x^{-a-1}$ there.) For the Standard Pareto we find

$$\begin{aligned} E_I(Y) &= \int_1^\infty \frac{-\ln y}{y^2}\,\mathrm{d}y = \int_1^0 \frac{-\ln(1/x)}{1/x^2}\left(\frac{-1}{x^2}\right)\mathrm{d}x = \int_0^1 \ln x\,\mathrm{d}x \\ &= [x\ln x - x]_0^1 = -1 = -E_I(U(0,1)), \end{aligned}$$

where we put $0\cdot(\ln 0) = 0$, as usual. Note that $I(x)$ is positive for every x, since all $x \in \mathcal{C}$, and $I(y)$ is negative for every y, since all $y \in \mathcal{P}$.

Many rather elementary transformations will appear in the sequel. Therefore, we collect the corresponding densities in the following technical

Lemma 1.4.2. *(Densities of transformed random variables). Suppose X has density $f(x)$ and distribution function F. Corresponding, let $Y = t(X)$ have pdf $g(y)$ and cdf $G(y)$. Then one obtains:*

	Domain of x	$t(x)$	*Domain of y*	*pdf $g(y)$*
1.	$\mathbb{R}\backslash\{0\}$	$1/x$	$\mathbb{R}\backslash\{0\}$	$f(1/y)/y^2$
2.	$\mathbb{R}^+$	$\ln x$	$\mathbb{R}$	$e^y f(e^y)$
3.	$\mathbb{R}^+$	$-\ln x$	$\mathbb{R}$	$e^{-y} f(e^{-y})$
4.	$\mathbb{R}$	$\exp(x)$	$\mathbb{R}^+$	$f(\ln y)/y$
5.	$\mathbb{R}$	$\exp(-x)$	$\mathbb{R}^+$	$f(-\ln y)/y$
6.	$\mathbb{R}^+$	$\sqrt{x}$	$\mathbb{R}^+$	$2y f(y^2)$
7.	$\mathbb{R}$	x^2	$\mathbb{R}^+$	$(f(\sqrt{y}) + f(-\sqrt{y}))/(2\sqrt{y})$
8.	$\mathbb{R}^+$	x^n	$\mathbb{R}^+$	$\frac{y^{1/n-1}}{n}\cdot f(y^{1/n})$
9.	$(-\pi/2,\pi/2)$	$\tan(x)$	$\mathbb{R}$	$f(\arctan(y))/(1+y^2)$
10.	$\mathbb{R}$	$\arctan(x)$	$(-\pi/2,\pi/2)$	$f(\tan(y))\sec^2(y)$

1. *Proof.* See Lemma 1.3.1 or consider $G(y) = p(1/X \le y) = p(1/y < X) = 1 - p(X \le 1/y)$. Thus $g(y) = -f(1/y)\cdot(-1/y^2) = f(1/y)/y^2$
2. Since $G(y) = p(\ln(X) \le y) = p(X \le e^y) = F(e^y)$, one has $g(y) = e^y f(e^y)$
3. $G(y) = p(-\ln(X) \le y) = p(\ln(X) \ge -y) = p(X \ge e^{-y}) = 1 - p(X \le e^{-y}) = 1 - F(e^{-y})$, thus $g(y) = e^{-y} f(e^{-y})$

4. $G(y) = p(e^X \le y) = p(X \le \ln y) = F(\ln y)$, thus $g(y) = f(\ln y)/y$
5. $G(y) = p(e^{-X} \le y) = p(-X \le \ln y) = p(X \ge -\ln y) = 1 - p(X \le -\ln y) = 1 - F(-\ln y)$, thus $g(y) = f(-\ln y)/y$
6. $G(y) = p(\sqrt{X} \le y) = p(X \le y^2) = F(y^2)$, thus $g(y) = 2yf(y^2)$
7. $G(y) = p(X^2 \le y) = p(-\sqrt{y} \le X \le \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y})$; therefore $g(y) = (f(\sqrt{y}) + f(-\sqrt{y}))/(2\sqrt{y})$. Moreover, $g(y) = 2f(\sqrt{y})/\sqrt{y}$ if X is symmetric about the y-axis.
8. $G(y) = p(X^n \le y) = p(X \le y^{1/n}) = F(y^{1/n})$, thus $g(y) = \frac{y^{1/n-1}}{n} \cdot f(y^{1/n})$
9. $G(y) = p(\tan(X) \le y) = p(X \le \arctan(y)) = F(\arctan(y))$, thus $g(y) = \frac{f(\arctan(y))}{1+y^2}$
10. $G(y) = p(\arctan(X) \le y) = p(X \le \tan(y)) = F(\tan(y))$, thus $g(y) = \frac{f(\tan(y))}{1+\tan^2(y)} = \frac{f(\tan(y))}{\cos^2(y)} = f(\tan(y)) \cdot \sec^2(y)$

□

The following transform is particularly important:

Lemma 1.4.3. *(Exponentially transformed random variable I). Suppose Y has density $g(y)$, and $X = e^{-Y}$ has density $f(x)$.*

Then we have $f(x) = g(-\ln x)/x$ for $x > 0$, and $E_I(X) = EY$.

Proof. Due to the general transformation theorem (e.g. Feller (1959), Theorem 1 for expected values, p. 208), we have for any transformation $t(x)$ that $E(t(X)) = \int t(x)f(x)\,dx$.

The mapping $y = t(x) = -\ln x$ is continuously differentiable, strictly decreasing and convex. In particular, it is a bijection from $\mathbb{R}^+$ to $\mathbb{R}$ with the inverse mapping $x = t^{-1}(y) = e^{-y}$. $x = 0$ is mapped to $y = \infty$, and $x = \infty$ is mapped to $y = -\infty$. Thus

$$E_I(X) = E_I(e^{-Y}) = \int_{-\infty}^{\infty} (-\ln e^{-y}) g(y)\,dy = \int_{-\infty}^{\infty} y \cdot g(y)\,dy = EY$$

which proves the second claim. For the first claim, see the last lemma (5.), or note that $dy = -dx/x$, which implies

$$\begin{aligned} \int_0^{\infty} (-\ln x) f(x)\,dx &= E_I(X) = \int_{-\infty}^{\infty} (-\ln e^{-y}) g(y)\,dy \\ &= \int_{\infty}^{0} (-\ln x) \cdot g(-\ln x) \cdot (-1)\frac{dx}{x}, \end{aligned}$$

and $f(x) = g(-\ln x)/x$ follows. □

Lemma 1.4.3 demonstrates that $\mu_I = E_I(X)$ is similar to the traditional expected value $\mu = E(X)$. Moreover, it can be interpreted in several ways:

- The usual expected value is computed relative to an 'exponentially transformed' random variable, namely $X = \exp(-Y)$.
- Likewise, one may say that X is evaluated with respect to a logarithmic scale, i.e., $-\ln|x|$ 'counterbalances' the exponential transform.

These considerations imply that μ_I exists for most well-known statistical distributions. To this end, consider a density on $\mathbb{R}^+$. Since μ_I is the larger, the more mass is concentrated in the neighbourhood of zero, it suffices to study the uniform $U(0,c)$. Although the expected information $E_I(U(0,c)) = 1 - \ln c$ goes to infinity if $c \to 0$, growth is "logarithmically slow." In the limit, the uniform distribution degenerates to ϵ_0 (or $X \equiv 0$), and $E_I(X) = \infty$. However, as long as there is a density or at least not an atom at the origin,[5] this cannot happen.

Examples:

1. Due to $E_I(X) = EY$, where $X = e^{-Y}$, the existence of EY is equivalent to the existence of $E_I(X)$.
2. Starting with $Y \sim N(0,1)$, we get $E_I(e^{-Y}) = E_I(e^Y) = EY = 0$, where $X = \exp(-Y)$ is the 'Lognormal'. Actually, it should be named InvLogNormal or ExpNormal, having the well-known density $f(x) = \varphi(-\ln x)/x = \frac{\exp(-(\ln x)^2/2)}{x\sqrt{2\pi}}$ for $x > 0$. It follows that
$$\int_0^1 (-\ln x)\frac{\exp(-(\ln x)^2/2)}{x\sqrt{2\pi}}\,\mathrm{d}x = -\int_1^\infty (-\ln x)\frac{\exp(-(\ln x)^2/2)}{x\sqrt{2\pi}}\,\mathrm{d}x\,.$$
3. For the Cauchy, $Y \sim C(0,1)$ we know that $E_I(e^{-Y}) = EY$ is not defined. Nevertheless, it is easy to calculate the density $f(x) = \frac{1}{\pi x(1+(\ln x)^2)}$ for all $x > 0$ if $X = \exp(-Y)$. Fig. 1.4 displays the pdf, its derivative $f'(x)$, and $(-\ln x)f(x)$. We will see in a while that
$$\int_0^1 \ln(|\ln(x)|)/(\pi x(1+(\ln x)^2))\,\mathrm{d}x$$
$$= \int_1^\infty \ln(|\ln(x)|)/(\pi x(1+(\ln x)^2))\,\mathrm{d}x = 0$$

[5] We will come to that later, see, in particular, Section 4.5.

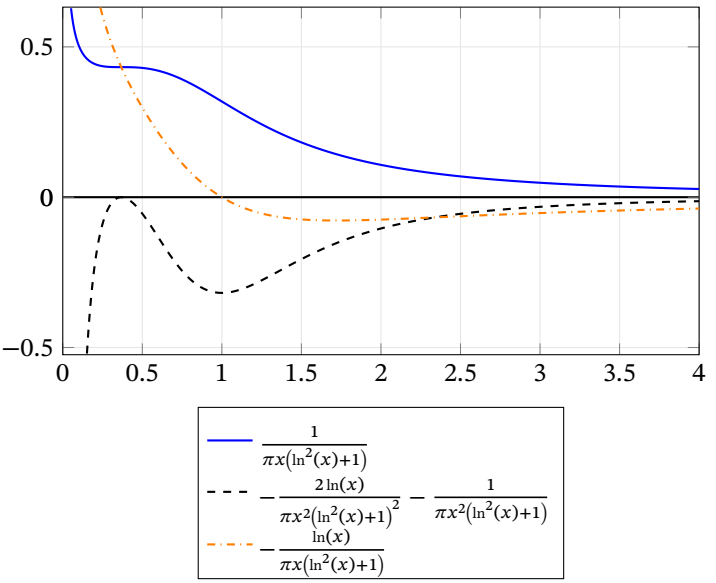

Figure 1.4: The ExpCauchy, for details see text.

4. For the Standard Lévy $Y \sim \text{Lévy}(0,1)$, $E_I(X) = E_I(e^{-Y})$ is also not defined. Here, $h(y) = \frac{\left(\frac{1}{y}\right)^{3/2} e^{-\frac{1}{2y}}}{\sqrt{2\pi}}$ implies $f(x) = \frac{e^{\frac{1}{2\ln(x)}}\left(-\frac{1}{\ln(x)}\right)^{3/2}}{\sqrt{2\pi}x}$ for $0 < x < 1$, see Fig. 1.5. Note that the density is increasing on the interval $(1/e, 1/\sqrt{e})$.

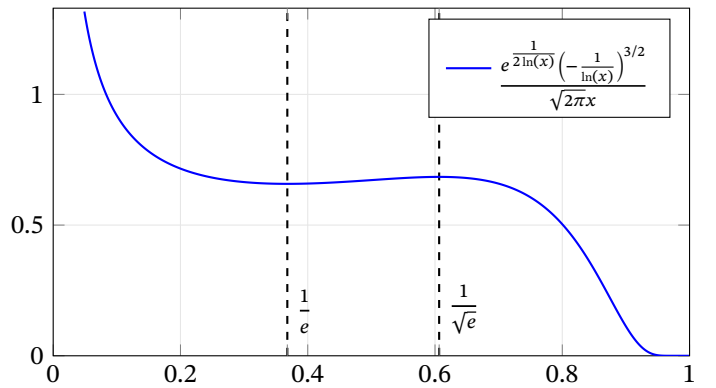

Figure 1.5: Density of ExpLévy.

Note that the domain of Y, i.e., $\mathbb{R}^+$, is mapped to the unit interval. Since the Lévy has heavy tails, the density of ExpLévy near the origin becomes large.

The density of $Z = -\ln Y$ can also be given explicitly. For that r.v., the pdf is $h(z) = \frac{\exp((z-e^z)/2)}{\sqrt{2\pi}}$, and $E_I(Z) \approx 0.2$.

1.5 Algebra of random variables

In classical theory, $E(X + Y) = EX + EY$ holds if the expected values involved exist. In the case of independence, one also has $E(XY) = EX \cdot EY$ and $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$. The analogue here is the fundamental

Theorem 1.5.1. *(Additivity of logarithmic expected values). Given random variables X, Y such that $p(X = 0) = p(Y = 0) = 0$, we have*

$$E_I(X \cdot Y) = E_I(X) + E_I(Y).$$

Proof. The proof of Theorem 2 in Feller (1959), p. 208, can be transferred almost literally to our case:

$$\begin{aligned} E_I(X) + E_I(Y) &= \sum_j (-\ln|x_j|)p(x_j) + \sum_k (-\ln|y_k|)p(y_k) \\ &= \sum_j \sum_k (-\ln|x_j|)p(x_j, y_k) + \sum_j \sum_k (-\ln|y_k|)p(x_j, y_k) \\ &= -\sum_j \sum_k (\ln|x_j| + \ln|y_k|)p(x_j, y_k) \\ &= \sum_j \sum_k (-\ln(|x_j \cdot y_k|))p(x_j, y_k) = E_I(XY) \end{aligned}$$

The same holds for convergent series and densities. □

The crucial relationship $\ln(xy) = \ln x + \ln y$ is valid for all positive real numbers x, y, and can be extended to complex numbers with $Re(x), Re(y) > 0$, see Abramowitz and Stegun (1964), p. 67.[6]

Note that just as in the case of the classical expected value, dependence is *irrelevant*. One may "always" calculate logarithmic and "ordinary" expected values. The case of perfect dependence, i.e. $X(\omega) = Y(\omega)$, demonstrates that $E_I(X^2) = 2E_I(X)$. More generally, we have

Corrolary 1.5.2. *(Calculation rule for general dispersion). Suppose X is a r.v. such that $p(X = 0) = 0$ and $Y = s \cdot X$, where $s > 0$. Then $E_I(Y) = E_I(X) - \ln s$.*

Proof. Since the constant s may be treated as a constant random variable $S^* \equiv s$, we immediately obtain $E_I(Y) = E_I(sX) = E_I(s) + E_I(X) = E_I(X) - \ln s$. □

The latter result may be interpreted as a decomposition: The expected information in any (ordinary) density may be decomposed into the information of its basic form plus the information due to its dispersion. That is, the dispersion parameter always creates an additive term, and it is interesting to compare the shapes of the densities, i.e., $E_I(\mathcal{D})$ for various distributions $\mathcal{D}$.

[6] It might be added that in number theory additive functions f on $\mathbb{N}_1$ are defined by $f(mn) = f(m) + f(n)$, and multiplicative functions g satisfy $g(mn) = g(m)g(n)$, see for instance Iwaniec and Kowalski (2004), pp. 9f.

Corrolary 1.5.3. *(Calculation rules for logarithmic expected values). Given random variables X_i such that $p(X_i = 0) = 0$ and $E_I(X_i) = \nu$. Then $E_I(\prod_{i=1}^n X_i)^p = pn\nu$, where p is an arbitrary real number.*

Proof. Because of the last theorem, $E_I(\prod_{i=1}^n X_i) = n\nu$; moreover,

$$\begin{aligned} E_I(X^p) &= \int (-\ln|x^p|) f(x)\,\mathrm{d}x = \int (-\ln|x|^p) f(x)\,\mathrm{d}x \\ &= p\int (-\ln|x|) f(x)\,\mathrm{d}x = pE_I(X). \end{aligned}$$

□

If $p = 0$, note that $X^0 \equiv 1$ (in other words, the distribution degenerates to δ_1). Therefore, $E_I(1) = -\ln 1 = 0$.

Odd powers obey the reflections at 1 and -1, i.e., for any $x \in (-\infty, -1]$, $x \in [-1, 0]$, $x \in [0, 1]$ or $x \in [1, \infty)$, x^{2n+1} lies in the same interval.

Even powers $2n$ map the centre, i.e., the interval $(-1, 1)$ to the unit interval $(0, 1)$ and the periphery to the interval $(1, \infty)$. Thus we have in general:

Corrolary 1.5.4. *(Logarithmic expected value of the power of a random variable). Suppose X is a real-valued random variable and $p(X = 0) = 0$. Then we obtain for X^p:*

$$E_I(X^p) = pE_I(X)$$

Proof. Let $n = 1$ in Corollary 1.5.3. □

Corrolary 1.5.5. *(Logarithmic expected value of the inverse random variable). Suppose X is a real-valued r.v. and $p(X = 0) = 0$. Then we obtain for $Y = 1/X$:*

$$E_I(Y) = -E_I(X).$$

Proof. Let $p = -1$ in Corollary 1.5.3. □

Corrolary 1.5.6. *(Logarithmic expected value of the ratio of two random variables). Suppose X, Y are r.v. such that and $p(X = 0) = p(Y = 0) = 0$. Then we have for $Z = X/Y$:*

$$E_I(Z) = E_I(X) - E_I(Y)$$

Proof. Let $X/Y = X \cdot Y^{-1}$ in Corollary 1.5.3. □

More generally, if X and Y have the same logarithmic expected value, this implies $E_I(X/Y) = 0$.

Corrolary 1.5.7. *(Logarithmic expected value of the geometric mean). Suppose X_i are random variables such that $p(X_i = 0) = 0$ and $E_I(X_i) = \nu$, $i = 1, \dots, n$. Then we have*

$$E_I\left(\left(\prod_{i=1}^{n} |X_i|\right)^{1/n}\right) = \nu.$$

Proof. Let $p = 1/n$ in Corollary 1.5.3. □

Note that Theorem 1.5.1 and its consequences hold for any kind of random variable, as long as there is no probability mass at the origin.

1.6 Logarithmic moments

In received probability theory, moments $EX^p = \int_{-\infty}^{\infty} x^p f(x)\,\mathrm{d}x$ are important, in particular if $p = 1, 2$. Looking at the integrand, $f(x)$ is multiplied by x^p. On $\mathbb{R}^+$, $\ln x$ grows slower, and $\exp(x)$ grows faster than any power of x. Therefore it is straightforward to consider $\int_0^\infty (-\ln x)^p f(x)\,\mathrm{d}x$ and $\int_0^\infty (\exp(-x))^p f(x)\,\mathrm{d}x$.

Restricting attention to the positive semi-axis ($X \geq 0$ having pdf f), and thinking of transformations, $\mathcal{M}_s(X) = \mathcal{M}_f(s) = \int_0^\infty x^{s-1} f(x)\,\mathrm{d}x$ ist the Mellin transformation of f (see Boros and Moll (2004), p. 196, Nolan (2020), pp. 270,272, J. Bertrand, P. Bertrand and Ovarlez (2010), and the monograph Paris and Kaminski (2001)). For independent $X, Y \geq 0$ with respective densities f and g,

$$\begin{aligned} \mathcal{M}_s(XY) &= \int_0^\infty \int_0^\infty (xy)^{s-1} f(x) g(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \mathcal{M}_f(s) \cdot \mathcal{M}_g(s) = \mathcal{M}_s(X) \cdot \mathcal{M}_s(Y), \end{aligned}$$

which is a generalisation of $E(XY) = EX \cdot EY$.

For $X \geq 0$, $\pi_X(s) = \mathcal{L}_f(s) = E(e^{-sX}) = \int_0^\infty e^{-sx} f(x)\,\mathrm{d}x$ is the Laplace transform of f. In particular, $\int e^{-x} f(x)\,\mathrm{d}x = \mathcal{L}_f(1)$.

Due to symmetry, the logarithmic transform

$$\mathcal{T}(s) = \int_0^\infty (-\ln x)^s f(x)\,\mathrm{d}x$$

should have similar properties. In a table:[7]

Name	Product	Integral transform
Laplace	$e^{-sx} f(x)$	$\mathcal{L}_f(s) = \int_0^\infty e^{-sx} f(x)\,\mathrm{d}x$
Mellin	$x^{s-1} f(x)$	$\mathcal{M}_f(s) = \int_0^\infty x^{s-1} f(x)\,\mathrm{d}x$
Logarithmic	$(-\ln x)^s f(x)$	$\mathcal{T}_f(s) = \int_0^\infty (-\ln x)^s f(x)\,\mathrm{d}x$

[7] For a generalisation see Section 1.10. Although for fixed s, the product contains more detailed information than the integral transform, we rather consider products $I(x)f(x)$ with interesting information-extracting functions $I(x)$.

In particular, $\mathcal{I}_f(1) = E_I(X)$ is the first logarithmic moment of X. Consistently, if X is real-valued, one may define higher logarithmic moments:

$$M_I^p(X) = M_I^p(f) = \int_{-\infty}^{\infty} (-\ln|x|)^p f(x)dx,$$

and absolute logarithmic moments

$$M_I^{|p|}(X) = M_I^{|p|}(f) = \int_{-\infty}^{\infty} |(-\ln|x|)^p| f(x)dx.$$

Since the logarithm is also defined for complex-valued random variables, it makes sense to consider signed moments, i.e.

$$M_S^p(X) = M_S^p(f) = \int_{-\infty}^{\infty} (-\ln x)^p f(x)dx.$$

Remark: Ordinary moments consider the (lack of) symmetry with respect to the y-axis, i.e., the origin. Logarithmic moments consider (the lack of) symmetry with respect to a reflection in 1 (and -1, if X is real-valued). Correspondingly, Figures 1.6, 1.7 and 1.8 demonstrate that the functions involved in the higher logarithmic moments become small in the interval $\mathcal{C}^* = (1/e, e)$ and large outside of this interval.

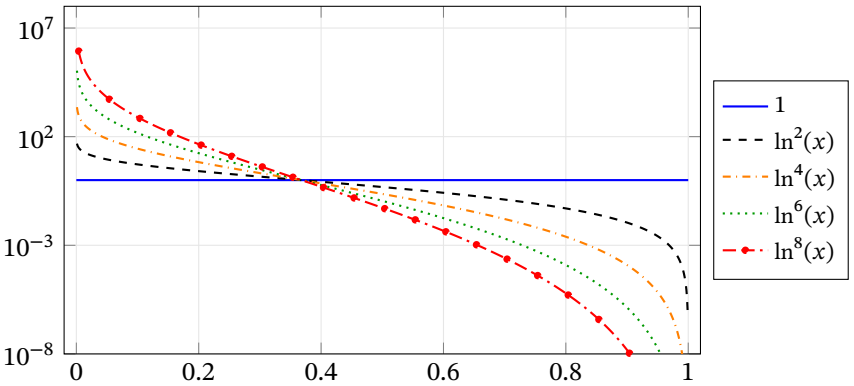

Figure 1.6: Functions involved in the logarithmic moments of the Standard Uniform (logarithmic scale).

Given a symmetric density about the y-axis, it suffices to consider the positive semi-axis. In particular, the expected information in a "folded Normal" $|X|$, where $X \sim N(0,1)$, is equal to $E_I(X)$.

Lemma 1.6.1. *(Logarithmic moments of symmetric distributions). Suppose X is a real-valued r.v. with pdf $f(x)$ such that $f(-x) = f(x)$. Then we have*

$$M_I^p(|X|) = M_I^p(X) \qquad \text{and} \qquad M_I^{|p|}(|X|) = M_I^{|p|}(X)$$

Proof. Owing to the definition of logarithmic moments, all that matters is the distance from the origin, and not the direction. Therefore,

$$\begin{aligned} M_I^p(X) &= \int_{-\infty}^{\infty} (-\ln|x|)^p f(x)\,\mathrm{d}x = 2\int_0^{\infty} (-\ln x)^p f(x)\,\mathrm{d}x \\ &= \int_0^{\infty} (-\ln x)^p (2f(x))\,\mathrm{d}x = M_I^p(|X|) \end{aligned}$$

and an analogous result holds for the p-th absolute logarithmic moment. □

Note that the symmetry assumption in the last lemma can be dropped, since if $g(x)$ denotes the pdf of $|X|$ on $\mathbb{R}^+$, we have $f(-x) + f(x) = g(x)$ for all $x > 0$. Therefore, $\int_{-y}^{y} (-\ln|x|)^p f(x)\,\mathrm{d}x = \int_0^y (-\ln x)^p g(x)\,\mathrm{d}x$ for all $y > 0$, which implies the equality $\int_{-\infty}^{\infty} (-\ln|x|)^p f(x)\,\mathrm{d}x = \int_0^{\infty} (-\ln x)^p g(x)\,\mathrm{d}x$ in the limit.

In other words, without loss of generality, we could assume $X \geq 0$ in what follows.

Lemma 1.6.2. *(Logarithmic moments of powers). Suppose X is a real-valued r.v. Then we have*

$$M_I^p(X^r) = r^p M_I^p(X) \qquad \textit{and} \qquad M_I^{|p|}(X^r) = |r^p| M_I^{|p|}(X).$$

Proof.

$$\begin{aligned} M_I^p(X^r) &= \int_{-\infty}^{\infty} (-\ln|x^r|)^p f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} (-\ln|x|^r)^p f(x)\,\mathrm{d}x \\ &= r^p \int_{-\infty}^{\infty} (-\ln|x|)^p f(x)\,\mathrm{d}x = r^p M_I^p(X), \end{aligned}$$

and an analogous result holds for the p absolute logarithmic moment. □

Lemma 1.6.3. *(Logarithmic moments of mixtures 1). Given densities f, g on $\mathbb{R}$, let $m(x) = \lambda f(x) + (1-\lambda)g(x)$ be their mixture $(0 \leq \lambda \leq 1)$. Then*

$$M_I^p(m) = \lambda M_I^p(f) + (1-\lambda) M_I^p(g)$$

and in particular,

$$E_I(m) = \lambda E_I(f) + (1-\lambda) E_I(g).$$

Proof.

$$\begin{aligned} M_I^p(m) &= \int (-\ln|x|)^p (\lambda f(x) + (1-\lambda)g(x))\,dx \\ &= \lambda \int (-\ln|x|)^p f(x)\,dx + (1-\lambda)\int (-\ln|x|)^p g(x)\,dx \\ &= \lambda M_I^p(f) + (1-\lambda) M_I^p(g) \end{aligned}$$

Obviously, an analogous result holds for finite convex combinations of densities (f_i, λ_i) where $0 \leq \lambda_i \leq 1$ and $\sum_{i=1}^n \lambda_i = 1$. □

The extension to a larger family of parameters is straightforward:

Lemma 1.6.4. *(Logarithmic moments of mixtures 2). Suppose there is a parametric family of densities $f_\vartheta(x) = f(x|\vartheta)$. If ϑ is a r.v. such that $E(\vartheta) = \bar{\vartheta}$, we obtain for the expected information*

$$E_I = \int_{-\infty}^{\infty}\int_{\vartheta}(-\ln x)f(x|\vartheta)\,\mathrm{d}x\,d\vartheta = \int_{-\infty}^{\infty}(-\ln x)f_{\bar{\vartheta}}(x)\,\mathrm{d}x = E_I(f_{\bar{\vartheta}})$$

and

$$E_I = \int_{\vartheta}\int_{-\infty}^{\infty}(-\ln x)f(x|\vartheta)\,\mathrm{d}x\,d\vartheta = \int_{\vartheta}E_I(f_\vartheta)d\vartheta$$

if the exchange of integrations is justified, that is, if the situation is sufficiently smooth. (For particular conditions, see, e.g, Bronstein and Semendjajew (1987), Section 3.1.9.) Obviously, an analogous result holds for higher logarithmic moments.

The last lemma states that $E_I(X)$, just like EX or some ordinary moment in general, may be computed in two ways: Either one first determines the mean density $f_{\bar{\vartheta}}$ and then calculates the expected information with respect to that density; or one first calculates the expected information conditional on ϑ, and combines these expected values in a second step. Since the integration of expected values is typically easier than the combination of densities, the first approach is preferable.

Examples:

1. The logarithmic moments of the Standard Exponential are (see Srivastava and J. Choi (2012), p. 373):

$$\int_0^\infty(-\ln y)^n e^{-y}\,\mathrm{d}y = (-1)^n\int_0^1(\ln(-\ln x))^n\,\mathrm{d}x = (-1)^n\Gamma^{(n)}(1) \qquad (1.5)$$

   The authors above add that, completely analogously,

$$\int_0^\infty(\ln y)^n e^{-y}/\sqrt{y}\,\mathrm{d}y = \int_0^1(\ln(-\ln x))^n/\sqrt{-\ln x}\,\mathrm{d}x = \Gamma^{(n)}(1/2).$$

2. Since the Cauchy is symmetric about 1 and -1 (see Definition 1.3.2), we have $M_I^{|1|}(C(0,1)) = \frac{2}{\pi}(\beta_2+\beta_2) = 4\beta_2/\pi$, and $M_I^{2n+1}(C(0,1)) = 0$ for all $n \in \mathbb{N}_0$.

   If n is an even number, the logarithmic moments M_I^n are given by the well-known integral $\int_0^\infty \frac{(\ln x)^n}{1+x^2}\,\mathrm{d}x = |E_n|(\pi/2)^{n+1}$, where E_n are the Euler numbers, see Section 7.1.7. Altogether we obtain:

$$M_I^n(C(0,1)) = \int_{-\infty}^{\infty} \frac{(-\ln\ |x|)^n}{\pi(1+x^2)}\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\infty} \frac{(-\ln\ x)^n}{1+x^2}\,\mathrm{d}x = |E_n|\left(\frac{\pi}{2}\right)^n \quad (1.6)$$

It may be mentioned that Mathematica gives for even n and $Re(n) > -1$:

$$\int_0^{\infty} \frac{\ln^n(x)}{x^2+1}\,\mathrm{d}x = \frac{2\cdot\Gamma(n+1)}{4^{n+1}}\left(\zeta\left(n+1,\frac{1}{4}\right) - \zeta\left(n+1,\frac{3}{4}\right)\right)$$

where $\zeta(n,z)$ is the Hurwitz zeta function (see Section 7.5.2). The latter expression may be simplified with the help of equation (7.64), leading to (1.6).

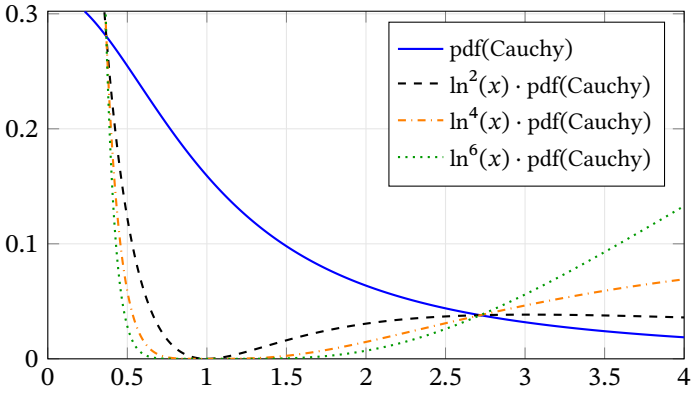

Figure 1.7: Functions involved in the logarithmic moments of the Standard Cauchy.

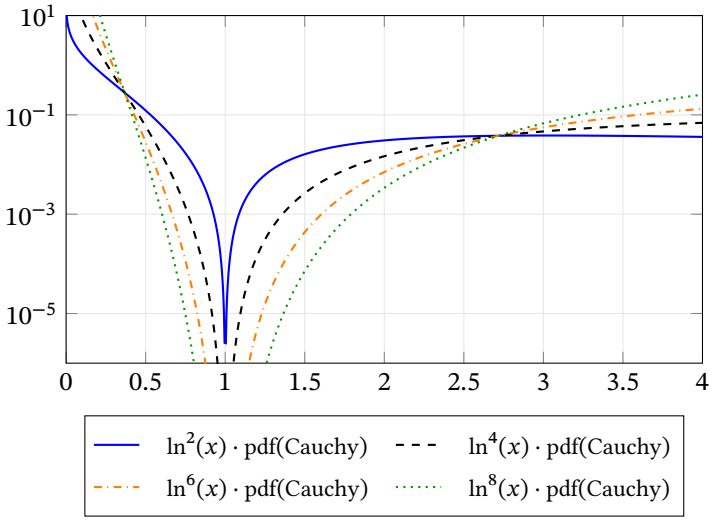

Figure 1.8: The same functions as in Fig. 1.7 (logarithmic scale).

3. For the Standard Normal, Mathematica finds

$$\begin{aligned} M_I^2(N(0,1)) &= \int_{-\infty}^{\infty} (-\ln|x|)^2 \varphi(x)\,\mathrm{d}x = \frac{\pi^2 + 2\ln^2(2) + \gamma(2\gamma + \ln(16))}{8} \\ &= \frac{\pi^2}{8} + \frac{(\ln 2)^2}{4} + \frac{\gamma^2}{4} + \frac{\gamma \ln 2}{2} = \frac{1}{2} \cdot \left(\frac{\pi}{2}\right)^2 + \left(\frac{\gamma + \ln 2}{2}\right)^2 \end{aligned} \tag{1.7}$$

4. Let X be a real-valued r.v. with density f and cdf F. Using $\ln(-x) = \ln x + i\pi$ for $x > 0$, one may calculate the first signed logarithmic moment:

$$\begin{aligned} M_S^1(X) &= \int_{-\infty}^{\infty} (-\ln x) f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{0} (-\ln(-x) - i\pi) f(x)\,\mathrm{d}x + \int_{0}^{\infty} (-\ln x) f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} (-\ln|x|) f(x)\,\mathrm{d}x - i\pi \int_{-\infty}^{0} f(x)\,\mathrm{d}x = E_I(X) - i\pi F(0) \end{aligned}$$

For the Normal, one thus has $M_S^1(N(0,1)) = (\gamma + \ln 2 - i\pi)/2$. For the Cauchy, $M_S^1(C(0,1)) = -i\pi/2$. Moreover, $M_S^2(C(0,1)) = -\pi^2/4$, $M_S^3(C(0,1)) = i\pi^3/8$, $M_S^4(C(0,1)) = \pi^4/16$, $M_S^5(C(0,1)) = -i\pi^5/32$, etc.

1.7 Iterated moments and scales

Since $(\log x)^2 \neq \log(x^2)$, one might think that Lemma 1.4.3 cannot be generalised to higher logarithmic moments. However, a closer look reveals:

Lemma 1.7.1. *(Exponentially transformed random variable II). Suppose Y has pdf $g(y)$, and $X = e^{-Y} \geq 0$ has pdf $f(x)$. Then $M_I^p(X) = EY^p$ holds.*

Proof. The proof of Lemma 1.4.3 may be extended in a straightforward manner:

$$\begin{aligned} M_I^p(X) &= \int_0^{\infty} (-\ln x)^p f(x)\,\mathrm{d}x = \int_{\infty}^{0} (-\ln x)^p \cdot g(-\ln x) \cdot (-1/x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} (-\ln e^{-y})^p g(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} y^p \cdot g(y)\,\mathrm{d}y = EY^p. \quad \square \end{aligned}$$

This means that any logarithmic moment may be interpreted as an ordinary moment and vice versa. More specifically: If $X = \exp(-Y)$, or, equivalently, if $Y = -\ln X$, where $X \geq 0$, the logarithmic moments M_I^n of X are the ordinary moments M^n of Y.

Examples:

1. The logarithmic moments of $U(0,1)$ coincide with the ordinary moments of Exp(1), since

$$\int_0^1 (-\ln x)^n \, \mathrm{d}x = \int_0^\infty y^n e^{-y} \, \mathrm{d}y = \Gamma(n+1) = n! \tag{1.8}$$

   where $\Gamma(\cdot)$ is the gamma function (see Section 7.4.1 for many properties of this basic function). One could also say that the gamma function agrees with the ordinary moments of the Standard Exponential and the logarithmic moments of the Standard Uniform, respectively.

   There is a further way to interpret this equation: $f(x) = \ln(1/x) = -\ln x$ is a density on the unit interval. Its logarithmic moments are $\int_0^1 (-\ln x)^n f(x) \, \mathrm{d}y = \int_0^1 (-\ln x)^{n+1} \, \mathrm{d}x = (n+1)!$ It is also no coincidence that the corresponding r.v. $X = U_1 \cdot U_2 \sim$ ProdUniform is the product of two independent $U_i \sim U(0,1)$, see Proposition 2.0.1, p. 35.

2. Williams (1973), p. 26, differentiates equation (1.8) with respect to n and obtains $\Gamma'(n+1) = \int_0^\infty y^n (\ln y) e^{-y} \, \mathrm{d}y$. For $n = 0$ he gets with the substitution $y = st$,

$$\begin{aligned} \Gamma'(1) &= \int_0^\infty e^{-y} (\ln y) \, \mathrm{d}y = s \int_0^\infty e^{-st} (\ln st) \, \mathrm{d}t \\ &= s(\ln s) \int_0^\infty e^{-st} \, \mathrm{d}t + s \int_0^\infty e^{-st} (\ln t) \, \mathrm{d}t = (\ln s) + s\mathcal{L}(\ln t) \end{aligned}$$

   Therefore, $-\mathcal{L}(\ln t) = \frac{\ln s - \Gamma'(1)}{s} = \frac{\gamma + \ln s}{s}$. For $s = 1$, this is the expected information in a Standard Exponential, for $s = 2$ it is the expected information in a Standard Normal.

3. If $Y \sim N(\mu, \sigma)$, $X = \exp(Y) \sim$ ExpNormal$(\mu, \sigma) =$ 'Lognormal'(μ, σ). Thus $E_I(X) = -EY = -\mu$ and $M_I^2(X) = \sigma^2(Y) + \mu^2 = \sigma^2 + \mu^2$.

Equation (1.8) may be extended tremendously. First, we are going to study transformations of the Uniform (i.e., generalisations of the left-hand side). Second, we are going to study the gamma distribution (the right-hand side) and observe that it takes centre stage in transformation theory and beyond.

The next logical step is to study extensions of the gamma function, in particular $\Gamma(x) \cdot \zeta(x,s)$ where $\zeta(x,s)$ is the Hurwitz zeta function on the right-hand side. Thus, one gets distributions associated with Gamma. More specifically, on the left-hand side, this leads to simultaneous (i.e., simultaneous ordinary and logarithmic) moments of probability distributions. "In between" there are interesting, irresistible and (almost

impossible) integrals or integral representations, see Boros and Moll (2004), Nahin (2015) and Vălean (2019). In a sense, much of what follows is an elaboration of this sketchy scheme.

Corrolary 1.7.2. *(Iterated logarithmic expected value). Given the assumptions and the notation of Lemma 1.7.1, let* $E_{II}(X) = \int_0^\infty -\ln(|-\ln x|)\ f(x)\ \mathrm{d}x$.

Then we have $E_{II}(X) = E_I(Y)$, *and for the higher moments*

$$\int_0^\infty (-\ln(|-\ln x|))^p\ f(x)\ \mathrm{d}x = \int_{-\infty}^\infty (-\ln|y|)^p\ g(y)\ \mathrm{d}y$$

Proof.

$$\begin{aligned} E_{II}(X) &= \int_0^\infty (-\ln(|-\ln x|))\ f(x)\ \mathrm{d}x = \int_{-\infty}^\infty (-\ln|-\ln(e^{-y})|)\ g(y)\ \mathrm{d}y \\ &= \int_{-\infty}^\infty (-\ln|y|)\ g(y)\ \mathrm{d}y = E_I(Y) \end{aligned}$$

and also

$$\int_0^\infty (-\ln(|-\ln x|))^p\ f(x)\ \mathrm{d}x = \int_{-\infty}^\infty (-\ln|y|)^p\ g(y)\ \mathrm{d}y\,. \qquad \square$$

Examples:

i) The last equation has been given for the Exponential, see (1.5), p. 19. Note that the domain of $Y \sim \mathrm{Exp}(1)$ is $\mathbb{R}^+$, and thus the domain of $X \sim U(0,1)$ is the unit interval. In particular, $E_{II}(X) = \gamma$.

ii) Let $Y \sim C(0,1)$ with pdf $g(y) = 1/(\pi(1+y^2))$, $y \in \mathbb{R}$. Thus $X = \exp(-Y) \sim$ ExpCauchy$(0,1)$ with pdf $f(x) = 1/(\pi x(1+\ln^2(x)))$, $x \in \mathbb{R}^+$.

Then we have $E_{II}(X) = -\int_0^\infty \ln(|\ln(x)|) \Big/ \left(\pi x\left(\ln^2(x)+1\right)\right) \mathrm{d}x = E_I(Y) = 0$. Moreover, in the "double-logarithmic periphery" $\mathcal{P}^* = \{x|0 \le x < 1/e\} \cup \{x|x > e\}$ we have

$$\int_0^{1/e} \frac{-\ln(-\ln x)}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = \int_e^\infty \frac{-\ln(\ln(x))}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = \int_1^\infty \frac{-\ln y}{\pi(1+y^2)}\,\mathrm{d}y = -\frac{\beta_2}{\pi},$$

and in the "double-logarithmic centre" $\mathcal{C}^* = \{x|1/e < x < e\}$ we find

$$\int_{1/e}^1 \frac{-\ln(-\ln(x))}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = \int_1^e \frac{-\ln(\ln(x))}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = \int_0^1 \frac{-\ln y}{\pi(1+y^2)}\,\mathrm{d}y = \frac{\beta_2}{\pi}.$$

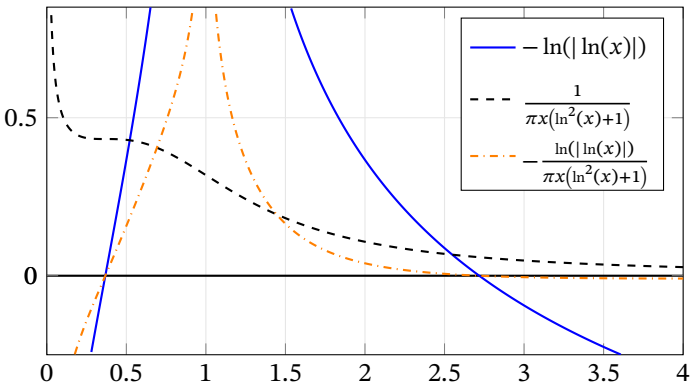

Figure 1.9: The double-logarithmic information function and ExpCauchy.

The higher logarithmic moments are also those of the Cauchy, in particular

$$\int_0^\infty \frac{\ln^2|(-\ln x)|}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = 4\int_0^{1/e} \frac{\ln^2(-\ln x)}{\pi x\left(\ln^2(x)+1\right)}\,\mathrm{d}x = \frac{\pi^2}{4} = M_I^2(C(0,1))$$

In other words, with respect to logarithmic moments, the double-logarithmic information function $I_p(x) = (-\ln(\ln|x|))^p$ deletes the effect of the exponential transformation $\exp(-Y)$.

Another exponentiation yields:

	Cauchy	exp(Cauchy)	exp(exp(Cauchy))	Sech
Domain of definition	$\mathbb{R}$	$[0,\infty)$	$[1,\infty)$	$\mathbb{R}$
Density	$\frac{1}{\pi(1+x^2)}$	$\frac{1}{\pi x(1+\ln^2(x))}$	$\frac{1}{\pi x\ln(x)(1+\ln^2(\ln x))}$	$\frac{1}{e^{\pi x/2}+e^{-\pi x/2}}$
Info. extracting function	$-\ln\lvert x\rvert$	$-\ln\lvert\ln x\rvert$	$-\ln\lvert\ln(\ln x)\rvert$	x
Expected Information	0	0	0	$EX = 0$

In the last column, we took the logarithm. That is, $Z = \frac{2}{\pi}\ln|Y|$ has a standard hyperbolic secant distribution Sech(0, 1), which is symmetric about the y-axis. Thus, in a sense, $I(x) = x$ is also an 'information-extracting function' that returns the location of some pdf $g(x)$. Moreover, $I(x) = x^2$ is crucial for the variance.

The notation exp(Cauchy) emphasises the transformation of the Cauchy, ExpCauchy the resulting distribution. In the same vein, the density of the double-exponential Standard Normal (shorthand notation ExpExpNormal) on the interval $(1,\infty)$ is $\frac{e^{-\frac{1}{2}\ln^2(\ln x)}}{\sqrt{2\pi}x\ln(x)}$. Since $E_{II}(\text{ExpExpNormal}) = E_I(\text{ExpNormal}) = E(N(0,1)) = 0$, we have

$$\int_1^\infty \frac{\ln(\ln x)e^{-\frac{1}{2}\ln^2(\ln x)}}{\sqrt{2\pi}x\ln(x)}\,\mathrm{d}x = 0.$$

Analogously, the expected value of Log|Normal| is

$$EX = \int_{-\infty}^{\infty} x\sqrt{\frac{2}{\pi}}e^{x-\frac{e^{2x}}{2}}\,\mathrm{d}x = -\frac{\gamma+\ln 2}{2},$$

i.e., $X \sim \ln(|Y|)$, where $Y \sim N(0,1)$. It may be mentioned that the corresponding cdf is just $F(x) = \Phi(e^x/\sqrt{2})$.

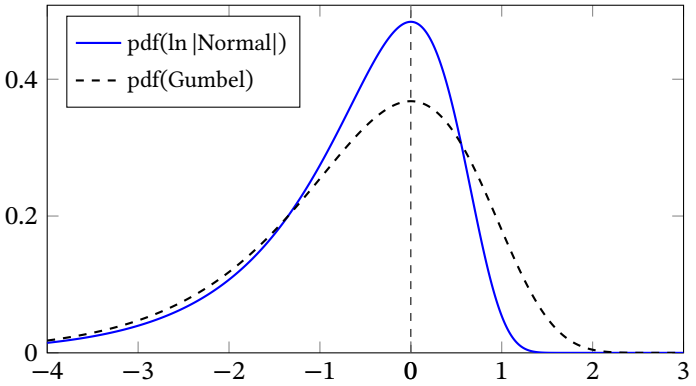

Figure 1.10: Densities of Log|Normal| and the Standard Gumbel.

Since Gumbel$(0,1)$ = Log(Exponential), the density of a Standard Gumbel is e^{z-e^z} for $z \in \mathbb{R}$. Thus we have $E_{II}(U(0,1)) = E_I(\mathrm{Exp}(1)) = \gamma = -E(\mathrm{Gumbel}(0,1))$, i.e.,

$$\int_0^1 -\ln(-\ln x)\,\mathrm{d}x = \int_0^\infty (-\ln y)e^{-y}\,\mathrm{d}y = \gamma = -\int_{-\infty}^{\infty} ze^{z-e^z}\,\mathrm{d}z\,.$$

(Iterated) logarithmic moments make the corresponding integrals small, i.e. those moments typically exist. With ordinary moments or transformations involving the exponential function, this may be quite different. If they exist, logarithmic moments, ordinary moments, the Mellin- and the Laplace transformations are closely related (cf. J. Bertrand, P. Bertrand and Ovarlez (2010) and Poularikas (2010)).

From a mathematical point of view, one "just" works with several similar scales ($x, \ln x, e^x$, etc.) With respect to interpretation, however, they are different. The expected value can be interpreted as the barycentre of the distribution, yet E_I is the typical amount of information in an observation from that distribution. All these scales measure different aspects of a distribution. It will be very instructive to combine them (see Section 5.5, in particular).

Another point of view would be that each logarithmic or exponential transformation bends the real axis. Thus, geometrically speaking, the straight line becomes a family of convex curves, derived from powers of e, and a family of concave curves, defined by iterating logarithms. These scales are very closely related, since the inverse functions of $e^x, \exp(e^x), \ldots$ are $\ln x, \ln(\ln x) \ldots$ (see Fig. 1.11).

In one sense, the operations of addition and multiplication are very similar. Thus,

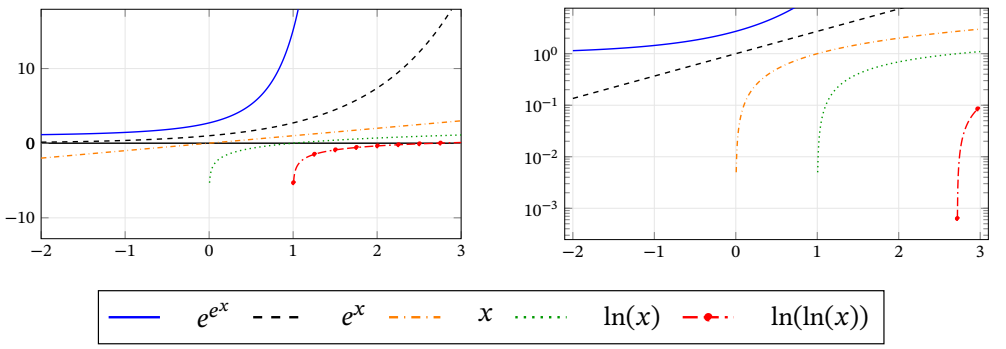

Figure 1.11: Different scales on the real axis. (Left: standard plot. Right: logarithmic plot)

they can be treated in an analogous way. For instance, $E(X+Y) = EX + EY$ and $E(XY) = EX \cdot EY$ (or the Mellin transform, in general).

However, the neutral element of addition, zero, is an isolated singularity with respect to multiplication. Thus, in another sense, there is a vast difference. In particular, just like the ordinary logarithm, $E_I(XY) = E_I(X) + E_I(Y)$ connects the additive and multiplicative structures.

Analogously, if $G(z) = \sum_{j=0}^{\infty} p_j z^j = E(z^X)$ is the probability generating function (pgf) of a discrete random variable X, where $p_j = p(X = j)$ for $j = 0, 1, \ldots$, then $G(z) = G_1(z) \cdot G_2(z)$, where G_1, G_2 are the pgf's of independent r.v.'s X_1, X_2 respectively, and $X = X_1 + X_1$, see Johnson, Kemp and Kotz (2005), p. 60.

1.8 Information variance

Given classical moments, EY and $\sigma^2(Y) = E(Y - EY)^2 = EY^2 - (EY)^2$ are of paramount importance. Due to Lemma 1.4.3, $E_I(X)$ can be interpreted as the expected value of Y, where $X = \exp(-Y)$, and Lemma 1.7.1 extends this to general (raw) moments. Therefore, it is straightforward to consider

$$\sigma_I^2(X) = M_I^2(X) - E_I^2(X).$$

We name the latter expression *information variance* since it measures the fluctuations in the information provided by the distribution $\mathcal{D}(X)$. Note that the straightforward definition $\sigma_I^2(X) = \int((-\ln|x|) - E_I(X))^2 f(x)\,\mathrm{d}x$ would yield the same result, since the latter expression is equal to

$$\begin{aligned} & \int \left((-\ln(|x|)^2 - 2(-\ln|x|)E_I(X) + (E_I(X))^2\right) f(x)\,\mathrm{d}x \\ = & \int (\ln|x|)^2 f(x)\,\mathrm{d}x - 2E_I(X)\int(-\ln|x|)f(x)\,\mathrm{d}x + (E_I(X))^2 \cdot \int f(x)\,\mathrm{d}x \\ = & \; M_I^2(X) - 2E_I^2(X) + (E_I(X))^2 = M_I^2(X) - E_I^2(X). \end{aligned}$$

The ordinary variance has the well-known properties $\sigma^2(X+b) = \sigma^2(X)$ and $\sigma^2(rX) = r^2\sigma^2(X)$. Analogously, we get here:

Lemma 1.8.1. *(Scaling does not affect information variance). Suppose $Y = bX$, where $b \neq 0$. Then $\sigma_I^2(Y) = \sigma_I^2(X)$,*

and

Lemma 1.8.2. *(Information variance of powers). Let $Z = X^r$. Then $\sigma_I^2(Z) = r^2\sigma_I^2(X)$,*

which implies

Corrolary 1.8.3. *The transformation $1/x$ does not change the information variance. That is, $\sigma_I^2(X^{-1}) = \sigma_I^2(X)$.*

Proof. Lemma 1.8.1 is straightforward:

$$\begin{aligned}\sigma_I^2(bX) &= \int((-\ln|bx|) - E_I(bX))^2 f(x)\,\mathrm{d}x \\ &= \int(-\ln|b| - \ln|x| - (E_I(X) - \ln|b|))^2 f(x)\,\mathrm{d}x \\ &= \int(-\ln|x| - E_I(X))^2 f(x)\,\mathrm{d}x = \sigma_I^2(X)\end{aligned}$$

To prove Lemma 1.8.2, one may evaluate $\sigma_I^2(X^r) = \int((-\ln|x^r|) - E_I(X^r))^2 f(x)\,\mathrm{d}x$ and find that the exponent r becomes the factor r^2 in front of the integral, or consult Lemma 1.4.3, which gives

$$\sigma_I^2(Z) = \sigma_I^2(X^r) = M_I^2(X^r) - E_I^2(X^r) = r^2M_I^2(X) - (rE_I(X))^2 = r^2\sigma_I^2(X)$$

For the corollary, either set $r = -1$ in Lemma 1.8.2 or use Lemma 1.6.2, which says that $M_I^p(X^{-1}) = (-1)^pM_I^p(X)$. Thus,

$$\begin{aligned}\sigma_I^2(X^{-1}) &= M_I^2(X^{-1}) - (E_I(X^{-1}))^2 = (-1)^2M_I^2(X) - (-E_I(X))^2 \\ &= M_I^2(X) - E_I^2(X) = \sigma_I^2(X). \quad \square\end{aligned} \tag{1.9}$$

EX is the typical (mean) observation, and $\sigma(X)$ indicates how much observations differ from one other. Analogously, $E_I(X)$ gives the amount of information that is to be expected, and $\sigma_I(X)$ indicates how much a single observation may deviate from the latter expectation. If $E_I(X) > 0$, information accrues, and the smaller $\sigma_I(X)$, the less volatile the learning process.

Starting with entropy, which is a kind of nonparametric analogue to expected information (see the Section 1.10), Tribus (1979) says that this concept is also used "to determine the degree of uncertainty about a proposition" and "as a measure of the rate of information acquisition". Here, $E_I(X)$ measures the amount of knowledge about the origin, and the parameters E_I and σ_I give detailed information about sampling from a population $X \sim \mathcal{D}(X)$.

Examples:

1. Since for the Cauchy $E_I(X) = 0$, we have, according to equation (1.6), $\sigma_I^2(X) = M_I^2(X) = (\pi/2)^2 = \pi^2/4$, and thus $\sigma_I = \pi/2$.

2. For the Normal, equations (1.4) and (1.7) yield

$$\sigma_I^2 \quad = \quad \frac{1}{2} \cdot \left(\frac{\pi}{2}\right)^2 + \left(\frac{\gamma + \ln 2}{2}\right)^2 - \left(\frac{\gamma + \ln 2}{2}\right)^2 = \frac{\pi^2}{8}$$

   and thus $\sigma_I = \pi/(2\sqrt{2})$.

3. For the Exponential, we find $M_I^2 = \gamma^2 + \pi^2/6$ and thus $\sigma_I^2 = \pi^2/6$ and $\sigma_I = \pi/\sqrt{6}$.

4. For the Lévy L, $\sigma_I^2 = \gamma^2 + \pi^2/2 + \ln^2(2) + \gamma \ln(4) - (-\gamma - \ln 2)^2 = \pi^2/2$, i.e., $\sigma_I = \pi/\sqrt{2}$. We could also use the fact that $L = N^{-2}$, where N is Standard Normal, and Lemma 1.8.2 with $r = -2$ to get $\sigma_I^2(L) = (-2)^2 \sigma_I^2(N) = 4\pi^2/8 = \pi^2/2$.

5. Since $\int_0^1 \ln^2(x)\,\mathrm{d}x = 2$, we have $\sigma_I^2(U(0,1)) = 2 - 1^2 = 1 = \sigma_I$.

6. For the Standard Pareto, we get $\sigma_I^2 = \int_1^\infty \ln^2(x)/x^2\,\mathrm{d}x - (-1)^2 = 2 - 1 = 1 = \sigma_I$. Alternatively, apply Corollary 1.8.3.

Note that the Normal is quite exceptional: the expected rate of information acquisition is rather high ($E_I(X)$ is large), and the flow of information is rather steady, since $\sigma_I(X)$ is the smallest compared to other important distributions.

Ordinary variance and information variance are connected in the following way:

Lemma 1.8.4. *(Exponentially transformed random variable: variance). Suppose $X = e^{-Y}$. Then $\sigma_I^2(X) = \sigma^2(Y)$ holds.*

Proof. Lemma 1.7.1 yields

$$\sigma_I^2(X) = M_I^2(X) - E_I^2(X) = E(-\ln X)^2 - (E(-\ln X))^2 = EY^2 - (EY)^2 = \sigma^2(Y) \square$$

Moreover, there is an important corollary to the last lemma:

Corrolary 1.8.5. *(Exponentially transformed random variables: decomposition of information variance). Suppose $X_i = e^{-Y_i} \Leftrightarrow Y_i = -\ln X_i$. If the $X_1, \ldots, X_n$, or, equivalently, $Y_1, \ldots, Y_n$ are independent, then $\sigma_I^2(\prod_{i=1}^n X_i) = \sum \sigma^2(Y_i) = \sum \sigma_I^2(X_i)$.*

Proof.

$$\begin{aligned}
\sigma_I^2\left(\prod_{i=1}^n X_i\right) &= E\left((-\ln\prod_{i=1}^n X_i)\right)^2 - \left(E(-\ln(\prod_{i=1}^n X_i))\right)^2 \\
&= E\left(\sum_{i=1}^n(-\ln X_i)\right)^2 - \left(E\left(\sum_{i=1}^n(-\ln X_i)\right)\right)^2 \\
&= E\left(\sum Y_i\right)^2 - \left(E(\sum Y_i)\right)^2 = \sigma^2\left(\sum Y_i\right) \\
&= \sum_{i=1}^n \sigma^2(Y_i) = \sum_{i=1}^n \sigma_I^2(X_i) \quad \square
\end{aligned}$$

Note that owing to Lemma 1.6.1 and the remark thereafter, the last corollary holds for any real-valued random variable. In the proof, one just has to use $|X_i|$ instead of X_i.

Examples:

1. If $X \sim \text{ExpNormal}(\mu, \sigma)$, then $E_I(X) = -\mu$ and $\sigma_I^2(X) = \sigma^2$, see p. 22.
2. Geometrically speaking, the Standard Cauchy C has much more mass in the periphery. Therefore, due to Corollaries 1.8.3 and 1.8.5, its logarithmic torque is twice as large as that of the Standard Normal: Suppose N_1, N_2 are independent and Standard Normal, then $\sigma_I^2(C) = \sigma_I^2(N_1 N_2^{-1}) = \sigma_I^2(N_1) + \sigma_I^2(N_2^{-1}) = \pi/8 + \pi/8 = \pi/4$.

At the heart of a classical ANOVA, one finds the formulas $E(\sum X_i) = \sum EX_i$ and $\sigma^2(\sum X_i) = \sum \sigma^2(X_i)$. Here, $E_I(\prod X_i) = \sum E_I(X_i)$ and $\sigma_I^2(\prod X_i) = \sum \sigma_I^2(X_i)$ may serve as the starting point of an ANOIVA, i.e., an Analysis of the Information Variances.

1.9 Closing a theoretical gap

Looking at this work from the angle of integrals, we are dealing with integral transforms, see Section 1.6, in particular. The Laplace transform is 'made for' non-negative random variables, which may be extended to the Fourier transform for real-valued r.v.'s. Both transforms use the exponential function, whereas the Mellin transform stems from powers of x.[8] In probabilistic terms, one is thus dealing with moments of a r.v.

Independent of jargon, it is consistent to extend these considerations to logarithmic moments, i.e., to products $(-\ln x)^s f(x)$, where f is a pdf or some other interesting function. Of course, this transform cannot be inverted without further effort, and that

[8] For the close connection between the latter transform and Dirichlet series, see e.g. Kowalski (2021), appendix A.4.

may be a reason why it has been overlooked. Nevertheless, it should have become clear that it has useful properties, $E_I(XY) = E_I(X) + E_I(Y)$, in particular. Moreover, σ_I^2 and other higher logarithmic moments are also straightforward parameters, embedding the new ansatz seamlessly in the received theory.

Actually, the parameters $E(X), \sigma^2(X)$; $E_I(X), \sigma_I^2(X)$ and entropy characterise a certain distribution quite nicely, and we are going to deal with the latter in the next section.

1.10 Received information theory

Classic information theory has a focus on probability distributions. Concrete realisations x are not relevant, just the probability $p(x)$ or density $f(x)$ that is associated with them. In this sense, the classical theory is not metric. It is also nonparametric, since the whole pdf and not just a few parameters are taken into consideration.

However, at least formally, the similarities to the account given above are striking. Starting with the integral

$$S_I(X) = \int I(x) f(x)\,\mathrm{d}x,$$

we have

$I(x)$	$S_I(X)$	Concept
x^n	$M^n(X)$	classical moments
x	EX	expected value
$(-\ln \lvert x\rvert)^n$	M_I^n (X)	logarithmic moments
$-\ln \lvert x\rvert$	$E_I(X)$	expected information
$-\ln(f(x)) = \ln(1/f(x))$	$E_I(f)$	entropy

In particular, the entropy $E_I(f) = \int(-\ln f(x)) f(x)\,\mathrm{d}x$ is the expected information of X with respect to the 'nonparametric' information-extracting function $-\ln(f(x))$. Since the latter expression should be well-defined, we assume tacitly that the integral extends over the carrier C_f of $f(x)$, i.e., $C_f = \{x | f(x) > 0\}$. For a probability, $\ln(1/p)$ is the information of (in) that probability. Consistently, for a discrete distribution with X assuming the values x_i with probability p_i, $E_I = \sum_i (-\ln p_i) p_i$ is the entropy of that distribution.

More generally: Given the pdf f, the ratio $g(x)/f(x)$ measures the deviation of g from f at the point x. This leads to the conditional information $I_{g|f}(x) = \ln(g(x)/f(x)) = \ln g(x) - \ln f(x)$ at x. The corresponding expected information is

$$E_I(\mathrm{g}||f) = \int \ln\left(\frac{g(x)}{f(x)}\right) f(x)\,\mathrm{d}x = -\int \ln\left(\frac{f(x)}{g(x)}\right) f(x)\,\mathrm{d}x = -D(f||\mathrm{g}),$$

where $D(f||g)$ is the Kullback-Leibler divergence (see p. 292, Kullback (1959), p. 5, and Cover and Thomas (2006)). Note that $E_I(f||f) = 0$, and it is well-known that (up

to a set of measure zero) $D(f||g) \geq 0$, i.e., $E_I(g||f) \leq 0$. Given this perspective, the entropy $E_I(f)$ stems from a comparison of f with the constant function 1. That is,

$$E_I(1||f) = \int \ln\left(\frac{1}{f(x)}\right) f(x)\,\mathrm{d}x = \int (-\ln f(x)) f(x)\,\mathrm{d}x = E_I(f)$$

Defining $E_I(g|f) = \int (\ln g(x)) f(x)\,\mathrm{d}x$, one obtains

$$\begin{aligned} E_I(g|f) &= \int \ln\left(\frac{g(x)}{f(x)}\frac{f(x)}{1}\right) f(x)\,\mathrm{d}x \\ &= \int \ln\left(\frac{g(x)}{f(x)}\right) f(x)\,\mathrm{d}x - \int \ln\left(\frac{1}{f(x)}\right) f(x)\,\mathrm{d}x = E_I(g||f) - E_I(f) \end{aligned}$$

The notation just introduced is not confusing upon observing that the sign "|" indicates how often the function behind that sign appears (once or twice). Note that $E_I(g||f) = E_I(g|f) + E_I(f)$, $E_I(f|f) = -E_I(f)$ and that $E_I(1|f) = 0$.

H. Jeffreys (1961), p. 179 (first ed. 1946), introduced the symmetrised KL divergence

$$\begin{aligned} J(f,g) &= D(f||g) + D(g||f) = -E_I(g||f) - E_I(f||g) \\ &= -E_I(g|f) - E_I(f|g) - E_I(f) - E_I(g) \end{aligned} \tag{1.10}$$

and showed that much of classical statistics can be treated elegantly (ibid., pp. 180-195) with the help of this crucial concept.

Kullback (1959) introduces KL divergence and J immediately afterwards (p. 6). His examples include entropy (p. 7) and the information that is given in the deviation from independence (p.8), that is

$$\begin{aligned} D(h||f\cdot g) &= \int\int \ln\left(\frac{h(x,y)}{f(x)g(y)}\right) h(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= -\int\int \left(\ln\frac{f(x)g(y)}{h(x,y)}\right) h(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= -E_I(f\cdot g||h) = -E_I(fg|h) - E_I(h) \geq 0. \end{aligned} \tag{1.11}$$

In other words, $D(h||fg)$ or equivalently $E_I(fg||h)$, compare h with the independent case fg. Given h, a large value of $D(h||fg)$ indicates a considerable amount of dependence.

Lindley (1956) employs a random parameter Θ with density $g(\theta)$, and thus obtains the information that an experiment is supposed to provide about the parameter (see Kullback (1959), p. 10):

$$D(h||f,g) = \int\int h(x,\theta) \ln\left(\frac{h(x,\theta)}{f(x)g(\theta)}\right) \mathrm{d}x\,d\theta = -E_I(fg||h) \geq 0.$$

An observation x says much about θ, if the common distribution $h(x,\theta)$ of x and θ deviates considerably from the independent case, i.e. $f(x)\cdot g(\theta)$. That is, if there is (in that sense) a strong dependence between x and θ.

Information may also be contained in the change of a parameter's value. Given a parameterized density $f(x,\theta)$, this leads to Fisher information (s. Kullback (1959), Kap.

6, pp. 26-28). Given two parameter values θ_0, θ_1, let $E_I(\theta_i|\theta_j) = \int(\ln f_{\theta_i}(x))f_{\theta_j}(x)\,\mathrm{d}x$, where $i, j = 0, 1$. Then equation (1.10) becomes

$$J(\theta_0, \theta_1) = -E_I(\theta_0, \theta_1) - E_I(\theta_1|\theta_0) - E_I(\theta_0|\theta_0) - E_I(\theta_1|\theta_1).$$

Above, we compared two densities f, g. Now, we compare two parameter values θ, $\theta + \Delta\theta$ or, more precisely, their corresponding values $f_\theta(x)$ and $f_{\theta+\Delta\theta}(x)$, and integrate out x. Thus it becomes apparent how much of f's change is due to a change in θ. If the dependence is strong, a slight change in θ leads to largely different values $f_\theta(x)$ and $f_{\theta+\Delta\theta}(x)$. Given certain regularity conditions, guaranteeing smoothness, $\ln(1 + \Delta f(x,\theta)/f(x,\theta)) \approx \Delta f(x,\theta)/f(x,\theta)$ where $\Delta f(x,\theta) = f(x,\theta+\Delta\theta) - f(x,\theta)$. Thus it turns out that

$$J(\theta, \theta + \Delta\theta) \approx \int \left(\frac{\Delta f(x,\theta)}{f(x,\theta)}\right)^2 f(x,\theta)\,\mathrm{d}x$$

More generally: If a function $g_\theta(x)$ provides information about a parameter θ, one should study

$$E_I(\theta) = \int(\ln g_\theta(x))f(x)\,\mathrm{d}x = E_I(g_\theta|f).$$

Suppose, in addition, that the parameter has to be estimated from the data $\mathbf{x} = (x_1, \dots, x_m; x_{m+1}, \dots, x_n) = (\tilde{\mathbf{x}}, \mathbf{x}^*)$. Thus one obtains straightforwardly

$$E_I(\hat{\theta}) = \int_{\mathbf{x}^*}\int_{\tilde{\mathbf{x}}}(\ln g_{\hat{\theta}(\tilde{\mathbf{x}})}(\mathbf{x}^*))f(\mathbf{x}^*)d\tilde{\mathbf{x}}d\mathbf{x}^*$$

Hirotugu Akaike's information criterion (AIC, see Akaike (1998)) approximates $-E_I(\hat{\theta})$. Not quite surprisingly, the estimator $AIC = -2\ln(\hat{\theta}_{MLE}) + 2k$ is the sum of the logarithm of the maximum likelihood estimator and the "complexity" k of the model considered. (The factor 2 is not important.) Gideon Schwarz' Bayesian Information Criterion (BIC, see Schwarz (1978)) has the similar form $BIC = -2\ln(\hat{\theta}_{MLE}) + 2k\ln(n)$. In other words, when selecting a model, AIC and BIC consider the information in the data and model complexity. A "good model" gives a reasonable estimate and has a rather low complexity.[9]

[9] Thus, these criteria avoid underfitting (the model being too primitive, i.e, the estimate is far off), and overfitting (the model being too close to the data). Such a model would not be robust - a similar sample would produce a completely different estimate.

1.11 The importance of logarithmic expected information

Logarithmic expected values are also at the heart of classical information theory. In general, there is an information-extracting function $I(x) = -\ln g(x)$ or $I_\vartheta(x) = -\ln g_\vartheta(x)$ and a pdf $f(x)$. Further interesting iefs will be studied throughout this text, and in Section 3.23, in particular. The information in some observation is always given by the product $(-\ln g(x)) \cdot f(x)$, and the corresponding expected information is

$$E_I(g_\vartheta(X)) = \int_x (-\ln g_\vartheta(x)) f(x)\,\mathrm{d}x$$

In classical information theory, higher moments do not seem to be important, but it is pivotal that the KL divergence and J are sums of expected information terms. Moreover, the concept of relative information is important.

Note also that at least in the Bayesian statistical tradition, data x and parameter θ both have a probability distribution. It is no coincidence that this fits in nicely here. Our basic idea is that of a (logarithmic) "distance" from a distinguished point (a singularity). Qualitatively speaking, there is a centre and a periphery. However, the formulas do not distinguish between

- the dispersion of measured values (stemming from an error-prone measurement process or unknown nuisance factors),
- a certain lack of knowledge (prior knowledge being vague or focused, resulting in a large or small deviation of that parameter's distribution)
- or the natural variation in a population.

Within our paradigm, (positive) information always means to be close to "truth", i.e., that the spread of the distribution in question is small. This also means that we treat location and variance (deviation, dispersion) widely differently, which is close to R.A. Fisher's view that (at least an important part of) statistic may be regarded as the study of variation (see Fisher (1925)).[10]

Going into the details, we thus arrive at distributions and their particular shape. They are the natural objects that come to mind if one wants to elaborate on the basic idea in a simple and nuanced (i.e., elegant) way. Moreover, it is straightforward to start with unimodal densities. The limiting case of perfect or complete information is thus represented by a one-point distribution δ_a or, equivalently, by a constant random variable $X \equiv a$. Probabilistic convergence toward δ_a has the consequence that the information in the corresponding distributions goes to infinity with logarithmic speed. This slow kind of divergence also allows for a detailed analysis that we begin in the next chapter.

[10] In contrast, classical statistical theory emphasises that location and variance are both parameters, and was successful in establishing a general theory of parameter estimation.

2 Transformation theory

Essentially, this chapter is a study of shape based on the transformation $1/x$ and related functions, the logarithm, in particular.

There are vast collections of statistical distributions, and a huge number of facts and relationships are known (see Johnson, Kemp and Kotz (2005), Johnson, Kotz and Balakrishnan (1994) and Kotz, Kozubowski and Podgórski (2001) but also Mathematica and Wikipedia, as first references). Alas, a general system like the Periodic Table of the Elements seems to be missing. In this section, many (if not most) continuous distributions of classical statistics will be put into a natural framework, and the concept of expected information will be crucial in classifying them.

R.A. Fisher's thoughts revolved around the normal distribution. χ^2, F, t and their ilk are all close relatives of the Normal. His congenial colleague H. Jeffreys (1961) went well beyond the Normal and used nonlinear transformations in a systematic way. Gnedenko and Kolmogorov (1959) also extended the classical theory (CLT, see H. Fischer (2011)) to stable distributions that are still unimodal. In the 1960s, Mandelbrot started to apply them to real-world problems, and today "power laws" are in vogue.

Due to $E_I(XY) = E_I(X) + E_I(Y)$, we are now able to extend these approaches to products and ratios of random variables. With the help of $E_I(\mathcal{D})$ and its domain of definition, each distribution $\mathcal{D}$ can be given its proper place. Moreover, just a few nonlinear transformations suffice to relate not just the Cauchy, the Exponential, the Normal, the Uniform and the Pareto, but also their entourage.

If not otherwise indicated, we will work with the common information function, the "standard scale" $I(x) = -\ln|x|$. In order to calculate the distributions explicitly, one also has to assume independence.

The following result of Curtiss (1941) is very helpful, and will be used very often:

Proposition 2.0.1. *Suppose X, Y are independent random variables with pdfs $f(x), g(x)$, respectively.*

1. *Then the pdf of $Z = XY$ may be computed with the help of the formula*

$$h(z) = \int_{-\infty}^{\infty} \frac{f(x)g(z/x)}{|x|}\,\mathrm{d}x$$

2. *Equivalently, the pdf of $Z = X/Y$ is*

$$f_Z(z) = \int_{-\infty}^{\infty} |y| f(yz)g(y)\,\mathrm{d}y\,. \tag{2.1}$$

2.1 The Uniform and its neighbourhood

Every pdf is associated with its domain of definition. So, upon transforming densities, one is also transforming sets. It is customary to identify some set A with its indicator function 1_A. If A is a bounded set, the uniform distribution there is this set's natural probabilistic representative. If $vol(A)$ is the volume of A, the corresponding density would be $f(x) = 1_A(x)/vol(A)$. In particular, if $X \sim U(a,b)$, we have $f(x) = 1/(b-a)$ if $x \in (a,b)$, and $f(x) = 0$ for all other $x \in \mathbb{R}$.

The most natural set to start some theory is the open $(0,1)$ or the closed $[0,1]$ unit interval. Upon dealing with densities, the slight difference between these sets does not really matter (although, at times, the boundary points demand special consideration), and thus we write $U(0,1)$ for the uniform distribution there. Moreover, suppose $U, U_i \sim U(0,1)$.

The "immediate neighbourhood" of $U(0,1)$ may be defined with the help of the functions $1/x, x^2, \ln x$ and $\exp x$, i.e, the operations of inverting, squaring, taking the logarithm and the exponent, respectively.

Transformation	Image Distribution	pdf	Domain	E_I
none	$U \sim U(0,1)$	1	$(0,1)$	1
$1/U$	Pareto(1, 1)	$1/x^2$	$x > 1$	-1
U^2	Beta(1/2, 1)	$1/(2\sqrt{x})$	$0 \le x \le 1$	2
$1/U^2$		$1/(2x^{3/2})$	$x > 1$	-2
$\ln U$	$-$Exp(1)	e^x	$x < 0$	γ
$-\ln U$	Exp(1)	e^{-x}	$x > 0$	γ
$S \cdot (\ln U^{-1})$	Laplace(1)	$e^{-\lvert x\rvert}/2$	$\mathbb{R}$	γ
$\ln(U^{-2})$	Exp(1/2), $\chi^2(2)$	$e^{-x/2}/2$	$x > 0$	$\gamma - \ln 2$

The pdf of $X_\lambda \sim \text{Exp}(\lambda)$ is $\lambda e^{-\lambda x}$ for $x > 0$. Since the parameter λ signifies the 'intensity' of some process and $Exp(\lambda)$ is the waiting time until the next random event, $X_\lambda = X_1/\lambda$ (the larger the intensity, the smaller the waiting time), and $E_I(X_\lambda) = E_I(X_1/\lambda) = E_I(1/\lambda) + E_I(X_1) = \gamma + \ln\lambda$ becomes larger. Note that since the distribution function of Exp(1) is $F(x) = y = 1 - e^{-x}$, its quantile function is $x = F^{-1}(y) = -\ln(1-y)$.

Moreover, a Laplace or 'double exponential' is just a symmetrised Exponential with $\mu = 1/\lambda$. Thus $Z \sim$ Laplace(ν) has pdf $f(z) = exp(-|z|/\nu)/(2\nu)$ and $\nu > 0$. This means that $|Z| \sim \text{Exp}(1/\nu) = \text{Exp}(\lambda)$, and

$$E_I(\text{Laplace}(\nu)) = E_I(\text{Exp}(\lambda)) = \gamma + \ln\lambda = \gamma - \ln\nu.$$

On the other hand, if $V_1, V_2 \sim \text{Exp}(1/\lambda)$ are independent, $V_1 - V_2 \sim \text{Laplace}(\lambda)$ and also if $U_1, U_2 \sim U(0,b)$ are independent, $\ln(U_1/U_2) = \ln U_1 - \ln U_2 \sim$ Laplace(1).

Since the logarithm favours small values, information should increase upon taking logarithms, and decrease in the case of the exponential function. This idea is correct for classical statistical distributions. (See, however, Theorem 2.10.2.)

With respect to the Uniform, on the one hand, $\ln(1/U) \sim \text{Exp}(1)$. On the other hand, $Z \sim \exp(U(0,1))$ has pdf $f(z) = 1/z$ on the interval $(1, e)$ with $E_I(Z) = -1/2$, and $Z^* \sim \exp(U(-1,1))$ has pdf $g(z) = 1/(2z)$ on the interval $(1/e, e)$ with $E_I(Z^*) = 0$.

More on $\exp(U)$ and its ilk can be found in Sections 4.2 and 4.9. The transformation 10^U leading to the distribution function $f(x) = \log_{10}(x)$ on the interval $(1, 10)$ will be discussed in Section 3.12. We continue with the discussion of distributions derived from the Uniform in Section 2.10.

2.2 The Cauchy, the Normal and the Chi-squared distribution

Suppose X, X_i are independent Standard Normal random variables. Then it is well-known that $X_1^2 + X_2^2 \sim \chi^2(2)$ (also see the last table) and that $Y = X^2 \sim \chi^2(1)$ with pdf$f(y) = \frac{1}{\sqrt{2\pi}} \frac{e^{-y/2}}{\sqrt{y}}$ for $y > 0$.

Because of this particular structure, one should be able to construct the Normal from more elementary distributions. Since $C = X_1/X_2$ is Standard Cauchy (for a derivation see equation 2.4), one might think of C^2. Now, $E_I(C^2) = 2 \cdot E_I(C) = 0$, but Mathematica easily calculates the pdf of $1 + C^2$ to be $g(t) = 1/(\pi t\sqrt{t-1})$ if $t > 1$, and zero otherwise. (The author borrowed the crucial idea of looking at the 'Shifted (1) Squared Cauchy' $1 + C^2$ from the page "Normally Distributed and uncorrelated does not imply independent" of the English Wikipedia.) Thus, to demonstrate that

$$S \cdot \sqrt{\frac{Y_2}{1 + C^2}} \sim N(0,1),$$

where $Y_2 \sim \chi^2(2)$, all one has to show is that $Y_1 = Y_2/(1 + C^2) \sim \chi^2(1)$. Fortunately the second part of Proposition 2.0.1 readily gives

$$\begin{aligned} f_Z(z) &= \int_1^\infty y \frac{1}{2} e^{-yz/2} \frac{1}{\pi y \sqrt{y-1}} \, \mathrm{d}y = \frac{1}{2\pi} \int_1^\infty \frac{e^{-yz/2}}{\sqrt{y-1}} \, \mathrm{d}y \\ &= \frac{1}{2\pi} \frac{e^{-z/2}\sqrt{2\pi}}{\sqrt{z}} = \frac{1}{\sqrt{2\pi}} \frac{e^{-z/2}}{\sqrt{z}} \end{aligned}$$

which is the pdf of a $\chi^2(1)$ distributed r.v. Moreover, summarising the successive transformations, i.e.,

Distribution / r.v.	$U(0,1)$	U^2	$-\ln U^2$
Domain	$[0,1]$	$[0,1]$	$\mathbb{R}^+$
Expected Information	1	2	$\gamma - \ln 2$
Distribution / r.v.	$C(0,1)$	C^2	$C^2 + 1$
Domain	$\mathbb{R}$	$\mathbb{R}^+$	$[1, \infty)$
Expected Information	0	0	$-2 \ln 2$

yields

$$E_I(Y_1) = E_I\left(\frac{-\ln U^2}{C^2+1}\right) = \gamma - \ln 2 - (-2\ln 2) = \gamma + \ln 2. \tag{2.2}$$

It may be mentioned that an elementary calculation, using the substitution $x = u^2$ and $E_I(X) = (\gamma + \ln 2)/2$ leads to the same result:

$$\begin{aligned} E_I(Y_1) &= \int_0^\infty (-\ln x)\frac{1}{\sqrt{2\pi}}\frac{e^{-x/2}}{\sqrt{x}}\,dx = \int_0^\infty (-\ln u^2)\frac{1}{\sqrt{2\pi}}\frac{e^{-u^2/2}}{u}2u\;du \\ &= 2\int_0^\infty (-\ln u)\frac{2e^{-u^2/2}}{\sqrt{2\pi}}\;du = 2\frac{\gamma+\ln 2}{2} = \gamma + \ln 2. \end{aligned}$$

Suppose $V \sim \text{Exp}(1), Y \sim \chi^2(1)$ are independent and $Q = Y/V$. Thus $E_I(Q) = E_I(Y/V) = E_I(Y) - E_I(V) = \gamma + \ln 2 - \gamma = \ln 2$. With the help of Proposition 2.0.1, one gets the pdf

$$g(x) = \frac{1}{\sqrt{\frac{x}{x+2}}(x+2)^2} = \frac{(x+2)^{-3/2}}{\sqrt{x}} \tag{2.3}$$

of Q on $\mathbb{R}^+$. Moreover, the next figure (Fig. 2.1) shows that the density of Y is sandwiched by the densities of Q and V, respectively.

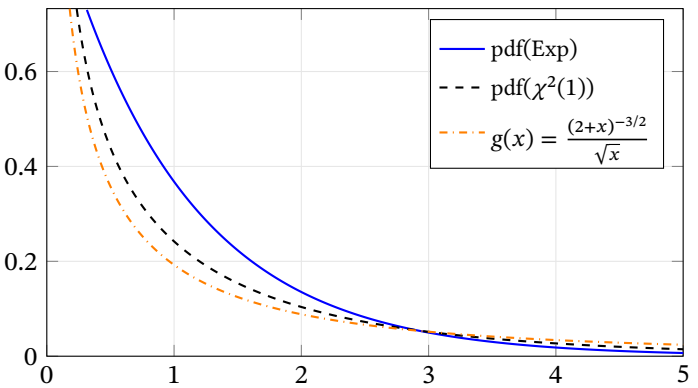

Figure 2.1: The densities of V, Y, and their ratio Q=Y/V.

Suppose the r.v. Q^* has the pdf of equation (2.3). Then the expected information in the geometric mean of Q^* and V coincides with the Standard Normal's, since

$$E_I(\sqrt{Q^*V}) = (E_I(Q^*) + E_I(V))/2 = (\ln 2 + \gamma)/2 = E_I(N(0,1)) = E_I(\sqrt{QV})$$

However, given the additional assumption that Q^* is independent of Y, the pdf of Q^*V is $h(x) = \frac{\sqrt{\pi}e^{x/2}(x+1)\Phi\left(\sqrt{x/2}\right)}{2^{3/2}\sqrt{x}} - 1/2$, and therefore the pdf of $\sqrt{Q^*V}$ is $k(x) = 2xh(x^2) = \sqrt{\frac{\pi}{2}}e^{\frac{x^2}{2}}\left(x^2+1\right)\Phi\left(x/\sqrt{2}\right) - x$, which does not coincide with the pdf of a HalfNormal, i.e., the pdf of $|X|$, where $X \sim N(0,1)$, see Fig. 2.2.

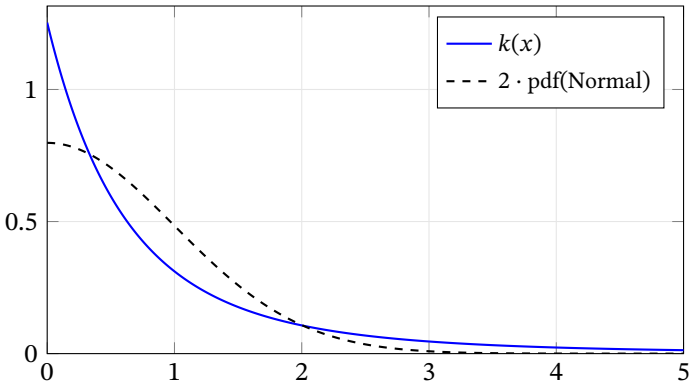

Figure 2.2: The densities of $\sqrt{Q^*V}$ (see text) and the HalfNormal.

The distributions just studied are very closely related. In particular, if $X, X_1, \ldots X_4 \sim N(0,1)$ and $Y_2 \sim \chi^2(2)$ are independent, then $X_1X_2 + X_3X_4$, $X_1X_2 - X_3X_4$, and $\sqrt{Y_2}X$ are all $Laplace(1)$, see, in particular, Kotz, Kozubowski and Podgórski (2001), p. 26.[1]

2.3 Exponential Power distributions

The exponential power distribution ExpoPower(a,b), where $a, b > 0$, interpolates between the Laplace and the Normal. On $\mathbb{R}$ it has the pdf

$$f(z) = \frac{a^{-1/a}e^{-\frac{\left(\frac{|z|}{b}\right)^a}{a}}}{2b\Gamma\left(1+\frac{1}{a}\right)}$$

which is symmetric about the origin (see Fig. 2.3, where $a \in \{1/2, 1, 2, 8\}$). If $a = 1$, ExpoPower$(1,b)$ = Laplace$(0,b)$, if $a = 2$, ExpoPower$(2,b)$ = Normal$(0,b)$.

The moments of $Z_{a,b} \sim$ ExpoPower(a,b) are:

EZ	$\sigma^2(Z)$	$E_I(Z)$	$\sigma_I^2(Z)$
0	$\frac{a^{2/a}b^2\Gamma\left(\frac{3}{a}\right)}{\Gamma\left(\frac{1}{a}\right)}$	$\frac{-a\ln(b)-\ln(a)-\psi^{(0)}\left(\frac{1}{a}\right)}{a}$	$\frac{\psi^{(1)}\left(\frac{1}{a}\right)}{a^2}$

It may be noted that $E_I(Z)$ is a monotonically increasing and concave function in a, and $\sigma_I^2(Z)$ is a monotonically decreasing and convex function in a. For both functions,

[1] $\sqrt{Y_2}$ has a $\chi(2)$ or Rayleigh(1) distribution, see p. 85.

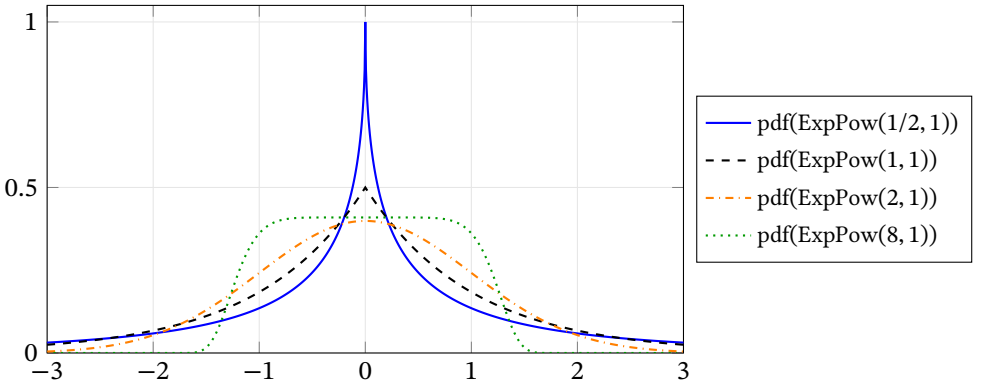

Figure 2.3: Densities of ExpoPower*(a,1)*, see text.

we have $\lim_{a\to\infty} E_I(Z) = \lim_{a\to\infty} \sigma_I^2(Z) = 1$.

The pdfs of the inverse random variables $1/Z_{a,1}$ also have interesting bimodal shapes (see Fig. 2.4, where $a \in \{1/4, 1/2, 1, 2, 4, 8\}$).[2] Straightforward analysis shows that the maxima occur at the points $\pm 2^{-1/a}$. The smallest maximum occurs if $\psi(1+1/a)+\ln a = 1/2$, i.e. if $a \approx 1.225$, i.e., 'between' the inverse Standard Laplace and the inverse Standard Normal.

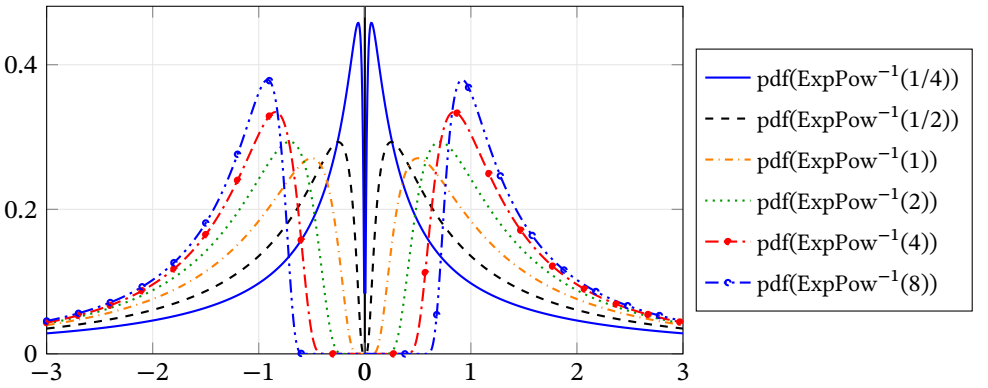

Figure 2.4: Inverses of Exponential Power distributions (densities), see text.

In the classical theory of moments, addition is trivial, since always $E(X \pm Y) = EX \pm EY$. Here, it is easy to deal with $E_I(XY)$ and $E_I(X/Y)$. However, just as in the case of the Mellin transform, the result of a shift is not trivial (cf. Nolan (2020), p. 270).

In order to find $E_I(X+Y)$, it is inevitable to study the distribution of $X+Y$, and to this end, independence is needed. Moreover, the result of a shift to the right, i.e. from $\int_0^\infty (-\ln x) f(x)\,\mathrm{d}x$ to $\int_c^\infty (-\ln x) f(x-c)\,\mathrm{d}x$ depends crucially on the concrete distribution at hand.

[2] More on these later, in particular in Section 3.20.

Already the example of $U(n, n+1)$, $n \in \mathbb{N}$, and $E_I(U(n, n+1)) = 1 - n\ln(1+1/n) - \ln(1+n)$ demonstrates that the result of such a shift is not trivial. For the "Shifted (n) Squared Cauchy", i.e., the random variable $X_n = n + C^2$, where $C \sim C(0,1)$, one obtains the pdf $f(x) = \frac{1}{\pi(x-n+1)\sqrt{x-n}}$ for $x > n$ and $f(x) = 0$ if $x \le n$.

2.4 The Algebra of distributions

If X and Y are independent r.v.'s, the distributions of $X+Y$ and $X-Y$ may be computed using classical methods (Fourier transforms, characteristic functions). For specific distributions $\mathcal{D}(X)$, $\mathcal{D}(Y)$, the product and the ratio distributions, i.e., $\mathcal{D}(XY)$ and $\mathcal{D}(X/Y)$, respectively, may also be accessed. Because of the end of the last section, it is tempting to write $\mathcal{D}(X) = \mathcal{D}(Y)$ instead of $X \overset{d}{=} Y$ and to calculate with distributions directly. For instance,

$$\frac{\chi^2(2)}{\mathcal{D}(1+C^2)} = \chi^2(1) \qquad \text{and} \qquad \pm\sqrt{\chi^2(1)} = N(0,1).$$

Note, however, that $Z = X/Y$ implies $Z \overset{d}{=} X/Y$, but not vice versa. Equality in distribution is weaker than pointwise equality $Z(\omega) = X(\omega)/Y(\omega)$. At times, this distinction is important. Moreover, calculating with distributions is associated with a numerical calculation. For instance, $E_I(\chi^2(1)) = E_I(\chi^2(2)) - E_I(\mathcal{D}(1+C^2))$, see equation (2.2).

Rational numbers z have a non-trivial decomposition x/y. Analogously, one could think of $\chi^2(2)$ as a "rational distribution". By the same token, not every distribution $\mathcal{D}(Z)$ can be decomposed into (non-trivial) distributions $\mathcal{D}(X), \mathcal{D}(Y)$, such that $\mathcal{D}(Z) = \mathcal{D}(X)/\mathcal{D}(Y)$. For instance, since the variance of $Z \equiv \pi$ is zero, X and Y also would have to be constants (almost surely). However, since π is irrational, δ_π cannot be decomposed into a fraction of non-trivial distributions.

For real numbers, the ratio a/b is well defined (if $b \neq 0$), and $b/b = 1$ is almost trivial. This is not so for random variables, even if they are iid. It is extremely important and *not* trivial that $C \sim C(0,1)$ can be written as a ratio of iid $X, Y \sim N(0,1)$. Here is a classical derivation using equation (2.1):

$$f_C(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} |y| e^{-(zy)^2/2} e^{-y^2/2}\,\mathrm{d}y = \frac{1}{2\pi}\int_{-\infty}^{\infty} |y| e^{-y^2(z^2+1)/2}\,\mathrm{d}y = \frac{1}{\pi(z^2+1)}, \tag{2.4}$$

since $\int_0^\infty y e^{-cy^2}\,\mathrm{d}y = 1/(2c)$, and, due to symmetry, $\int_{-\infty}^{\infty} |y| e^{-cy^2}\,\mathrm{d}y = 1/c$. In other words, $C \overset{d}{=} X/Y$ or $C(0,1) = N(0,1)/N(0,1)$, implicitly assuming independent random variables. (Of course, in the case of perfect dependence, $X(\omega) = Y(\omega) \neq 0$ always implies $X/Y = 1$.)

In general, decompositions $Z = X/Y$ with $E_I(Z) = 0$ demand special attention, since they define a "dual pair" of distributions. In more detail: If $X_i \sim \mathcal{D}(X) = \mathcal{D}$,

$\bar{X}_i \sim \mathcal{D}(1/X) = \bar{\mathcal{D}}$, we have $E_I(\mathcal{D}) = -E_I(\mathcal{D})$. Moreover, if all random variables are independent,

$$Z := X_1/X_2 \overset{d}{=} \bar{X}_1/\bar{X}_2 \overset{d}{=} X_1 \cdot \bar{X}_1, \tag{2.5}$$

since a r.v. distributed according to $\mathcal{D}$ in the numerator is stochastically indistinguishable from a r.v. distributed according to $\bar{\mathcal{D}}$ in the denominator. Tacitly assuming independence, $\mathcal{D}(Z) = \mathcal{D}/\mathcal{D} = \bar{\mathcal{D}}/\bar{\mathcal{D}} = \mathcal{D} \cdot \bar{\mathcal{D}}$.

Thus, in a sense, a distribution with expected information zero (a 'zero-information distribution') has been split into a positive and a negative part, and the variability in the numerator is perfectly counterbalanced by the same *kind* of variability in the denominator.[3] For instance, the dual distribution to a $U(0,1)$ is $P(1,1)$, and the dual distribution to $\chi^2(1)$ is Lévy$(0,1)$.

Of course, in every algebra $\mathcal{A}$, there is a unique zero, i.e., an element with the property $a + 0 = a$ for every $a \in \mathcal{A}$. Here, there is a large variety of distributions $\mathcal{D}$ with $E_I(\mathcal{D}) = 0$. Moreover, given real numbers, a/b is difficult to treat if $b = 0$, and $0/0$ is notorious. Here, it is possible to put a distribution $\mathcal{D}$ with vanishing expected information in the denominator, i.e., to divide by a zero-information distribution. For instance, if C, C_i are independent and Standard Cauchy, C_1/C_2 is straightforward. The reason is that $C(0,1)$ is *self-dual*, i.e., $\bar{\mathcal{D}}(C) = \mathcal{D}(1/C) = \mathcal{D}(C)$, see Definition 1.3.2. Therefore, because of equation (2.5), the distribution of the product and the ratio coincide, $\mathcal{D}(C_1C_2) = \mathcal{D}(C_1/C_2)$. More explicitly, the pdf of Z is $f(z) = \frac{2\ln|z|}{\pi^2(z^2-1)}$ on $\mathbb{R}$, and that of $W = \sqrt{|Z|}$ is $g(w) = \frac{16w\ln w}{\pi^2(w^4-1)}$ on $\mathbb{R}^+$. Thus the distribution of the geometric mean of two independent Cauchy r.v.'s is *not* Cauchy, but $\tilde{C} = 2/(1/C_1 - C_2) \sim C(0,1)$.

Just before Section 1.7, we mentioned that $E_S(C) = M_S^1(C) = -i\pi/2$. Since the "partial moment" $\int_0^\infty \frac{-\ln x}{\pi(1+x^2)}\,\mathrm{d}x$ vanishes, we know that $E_S(C) = 2\int_{-\infty}^{-1} \frac{-\ln x}{\pi(1+x^2)}\,\mathrm{d}x$. It turns out that

$$\int_{-\infty}^{-1} (-\ln x) f(x)\,\mathrm{d}x = -\frac{\beta(2)}{\pi} - \frac{i\pi}{4}$$

which yields

$$\int_{-1}^{0} (-\ln x) f(x)\,\mathrm{d}x = \frac{\beta(2)}{\pi} - \frac{i\pi}{4} \qquad \text{and} \qquad \int_0^1 \frac{-\ln(x)}{\pi(1+x^2)}\,\mathrm{d}x = \frac{\beta(2)}{\pi} \approx 0.2916.$$

Of course, $E_I(|C|) = E_I(C) = E_I(C_1 \cdot C_2) = E_I(\sqrt{C_1C_2}) = 0$. Therefore the integration

[3] Of course, the *amount* of variability does not vanish. For instance, $\sigma^2(X_1/X_2) \neq 0$ if $\sigma^2(X) \neq 0$. More precisely, it suffices to counterbalance $E_I(X) = c$ in the numerator with $E_I(Y) = c$ in the denominator, since $E_I(X/Y) = E_I(X) - E_I(Y) = c - c = 0$.

$$\int_0^1 (-\ln x)\frac{2\ln x}{\pi^2(x^2-1)}\,\mathrm{d}x = \frac{7\zeta(3)}{2\pi^2}$$

implies

$$\int_0^1 (-\ln x)\frac{8x\ln x}{\pi^2(x^4-1)}\,\mathrm{d}x = \frac{7\zeta(3)}{4\pi^2} \approx 0.2131,$$

where $\zeta(3) = \sum_{k=0}^{\infty} 1/k^3$ is Apéry's constant (cf. Boros and Moll (2004), Section 11.4). The constant will appear more often later, and the role of the Riemann ζ function will become more transparent later.

Since the density of the ExpNormal on the positive semi-axis is $f(x) = \frac{\exp(-(\ln x)^2/2)}{x\sqrt{2\pi}}$, ExpNormal is self-dual there. On the finite interval $(1/e, e)$, the density $f(x) = 1/(2x)$ is self-dual.

A different ansatz would be as follows: Let X be any distribution or corresponding r.v. with existing expected information, and X_i copies thereof (not necessarily independent). If $Q = X_1/X_2$, then $E_I(Q) = 0$, and assuming independence, Q is also self-dual since the distributions of $Q = X_1/X_2$ and $1/Q = X_2/X_1$ coincide. If $Q_1 = X_1/X_2$ and $Q_2 = X_3/X_4$, the same is true for $Q_1Q_2 = (X_1X_3)/(X_2X_4)$, which in the case of independence also has the remarkable property $Q_1Q_2 \stackrel{d}{=} Q_1/Q_2$.

In total generality, given $n \geq 2$ copies Q_j, *any* kind of ratio or product K of these random variables has expected information zero (Theorem 1.5.1). Moreover, in the case of independence, $\sigma_I^2(K) = n\sigma_I^2(Q) = 2n\sigma_I^2(X)$, since it then suffices to considers the product $P = \prod_{j=1}^{n} Q_j$, Corollary 1.8.5 gives the first equation, Lemma 1.8.2 with $r = -1$ the second.

Vice versa, we always have $E_I(X_1/X_2) = 0$ and $E_I(X_1X_2) = 2E_I(X)$. So the ratio and the product of two r.v.'s may only have the same distribution if $E_I(X) = 0$. A sufficient condition for $X_1X_2 \stackrel{d}{=} X_1/X_2$ is, of course, that X has a ratio distribution, $X_j = Y_{j,1}/Y_{j,2}$ say, such that in the case of independence $X_1X_2 \stackrel{d}{=} \frac{Y_{1,1}}{Y_{1,2}}\frac{Y_{2,1}}{Y_{2,2}} \stackrel{d}{=} X_1/X_2$.

The most important example is $X_j \sim N(\mu_j, \sigma_j)$. Given independence, we have $X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$. If, moreover, $\mu_j = 0$, we have $C_1C_2 \stackrel{d}{=} \frac{X_1X_3}{X_2X_4} \stackrel{d}{=} \frac{C_1}{C_2}$. In addition, for $Y_j = \exp(X_j) \sim$ExpNormal$(0, \sigma_j)$ one obtains

$$\begin{aligned} Y_1Y_2 &= e^{X_1}e^{X_2} = e^{X_1+X_2} \stackrel{d}{=} e^{X_1-X_2} = e^{X_1}/e^{X_2} = Y_1/Y_2 \\ &\sim \text{ExpNormal}\left(0, \sqrt{\sigma_1^2+\sigma_2^2}\right) \end{aligned}$$

It is also natural to study $\ln|Q| = \ln|X_1/X_2| = \ln|X_1| - \ln|X_2|$. If X is the Normal, this gives the Sech; if X is the Uniform, one gets the Laplace; and if X is the Exponential, $\ln Q$ is the Logistic, see Section 2.10.3 for details. Finally, it may be mentioned that the Normal essentially is the product of an Arcsin and an Exponential (see p. 56).

2.5 Kinds of error

A basic model of statistics is $X = \mu_X + \varepsilon$, where μ_X is interpreted as some "true" but unknown value, and $\varepsilon \sim N(0,1)$ adds a random amount of error to μ_X. More generally speaking, there is some latent deterministic structure *plus* random error. The error does not change the location, it does not introduce some kind of systematic bias; it just adds noise.

Analogously, here, one starts with some random variable Y and *multiplies* noise ε^*, i.e., a random variable with $E_I(\varepsilon^*) = 0$. This leads to $E_I(\varepsilon^* Y) = E_I(Y) + E_I(\varepsilon^*) = E_I(Y)$. Although the "contaminated distribution" $\mathcal{D}(\varepsilon^* Y)$ deviates from $\mathcal{D}(Y)$, the expected information remains the same.

For instance, if $C \sim C(0,1)$ and $N \sim N(0,1)$ are independent r.v.'s, one obtains with the incomplete gamma function $\Gamma(0,x) = \int_x^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$ (cf. Section 7.4.1), the following densities:

Random variable	Domain	pdf
$C \cdot N$	$\mathbb{R}$	$\frac{1}{\sqrt{2}\,\pi^{3/2}} \exp(x^2/2)\,\Gamma\left(0, x^2/2\right)$
$\sqrt{C \cdot N}$	$\mathbb{R}^+$	$\frac{2\sqrt{2}}{\pi^{3/2}} x \exp(x^4/2)\,\Gamma\left(0, x^4/2\right)$

We know that $E_I(CN) = E_I(N) = (\gamma + \ln 2)/2$ and $E_I(\sqrt{CN}) = (\gamma + \ln 2)/4$. However, $\frac{N_1}{N_2} \overset{d}{=} C$ does *not* imply $N_1 \overset{d}{=} N_2 \cdot C$. The product of a Normal and a Cauchy is *not* normally distributed. Also, the product of two independent Cauchys is *not* a Cauchy (see Springer and Thompson (1966)).

Equality only holds with respect to the corresponding logarithmic expected values, i.e., in a weaker sense. Algebraically speaking, there is not a single zero, but several distributions with $E_I(\mathcal{D}) = 0$.

Returning to the statistical issue, since $E_I(\varepsilon) > 0$, the statistician is able to learn the underlying deterministic structure in the long run. Typically, one assumes that the errors ε_i are independent and normally distributed. $\sum \varepsilon_i/n$ then has a smaller variance than ε_1. Extracting information from $\varepsilon^* Y$ is more difficult, even if $E_I(Y) > 0$. In the limit, it seems reasonable that Y can be reduced to $EY = \mu_Y$. However, standard methods won't be able to shrink a multiplicative error ε^*, since if $\varepsilon_i^* \sim C(0,1)$ are independent r.v.'s, say, $\sum \varepsilon_i^*/n \sim C(0,1)$. However, it is possible to divide by some constant larger than n, or to consider $\ln|(\prod_{i=1}^n \varepsilon_i^* Y)|$, see Section 6.8.

A slightly different perspective could hold that classical statistics focusses on ordinary moments of r.v.'s, the expected value and the variance in particular. Since $E(bX + c) = bEX + c$ and $\sigma^2(bX + c) = b^2\sigma^2(X)$, linear transformations and location-scale families are very important. Since $E_I(bX^c) = \ln b + cE_I(X)$, we are rather interested in scale-shape families here. Thus, powers and products of r.v.'s are rather straightforward to handle, and also exponential transformations. However, additive shifts pose a major challenge and lead to interesting new distributional forms, just think of the 'Shifted (1) Squared Cauchy', p. 37.

2.6 The Normal family

The "close neighbourhood" of the Normal may be defined with the help of the functions $1/x, e^x$, and x^2. Suppose $X, X_i \sim N(0,1)$ are independent r.v.'s, then[4]

Transformation	Name	Domain	$f(x)$	E_I
none	$N(0,1)$	$\mathbb{R}$	$\frac{\exp(-x^2/2)}{\sqrt{2\pi}}$	$(\gamma + \ln 2)/2$
$1/X$	$\bar{N}(0,1)$	$\mathbb{R}$	$\frac{\exp(-1/(2x^2))}{\sqrt{2\pi}x^2}$	$-(\gamma + \ln 2)/2$
X^{-2}	Lévy$(0,1)$	$\mathbb{R}^+$	$\frac{\exp(-1/(2x))}{\sqrt{2\pi}x^{3/2}}$	$-\gamma - \ln 2$
X^2	$\chi^2(1)$	$\mathbb{R}^+$	$\frac{\exp(-x/2)}{\sqrt{2\pi}\sqrt{x}}$	$\gamma + \ln 2$
$X_1 \cdot X_2$	VG$(1/2,1)$	$\mathbb{R}$	$K_0(\lvert x\rvert)/\pi$	$\gamma + \ln 2$
X_1/X_2	$C(0,1)$	$\mathbb{R}$	$1/(\pi(1+x^2))$	0
$\exp(X), \exp(-X)$	ExpNormal 'Lognormal'	$\mathbb{R}^+$	$\frac{\exp(-\ln^2(x)/2)}{\sqrt{2\pi}x}$	0

In the table, VG stands for 'Variance Gamma distribution' (also abbreviated Variance-Gamma in what follows), which will be discussed in Section 7.6.7. For an illustration, see Fig. 2.5. $K_0(x)$ is a modified Bessel function of the second kind (BesselK in Mathematica; see Section 7.6.1, and Abramowitz and Stegun (1964), p. 374). The Bessel functions $K_n(z) = K(n,z)$ satisfy the differential equation $z^2y'' + zy' - (z^2+n^2)y = 0$.

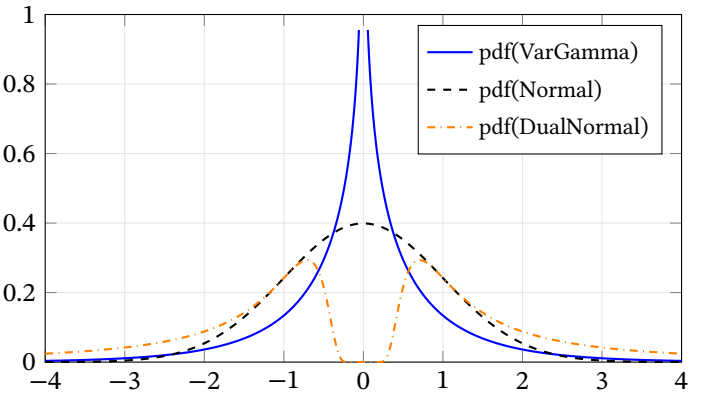

Figure 2.5: Densities of a VarianceGamma*(1/2,1)*, a Standard Normal, and its inverse, the Dual Normal (cf. the last table).

Note that the distributions of X^2 and $X_1 \cdot X_2$ differ. Nevertheless, the general theory teaches that $E_I(X^2) = E_I(X_1 \cdot X_2) = \gamma + \ln 2$. We will discuss this case in more detail on p. 56. A similar remark holds for the Cauchy and ExpNormal (see the illustrations).

[4]For the distribution of $\ln(X)$ see Fig. 1.10.

Using the specific connection between the Exponential and $\chi^2(i)$, $(i = 1, 2)$, see Table 2.1, p. 36 and Fig. 2.6, one obtains $\mathcal{D}(X_1X_4 - X_2X_3)$ ∼Laplace(1), cf. Kotz, Kozubowski and Podgórski (2001), p. 25, which can also be seen as an extension of equation (2.16).

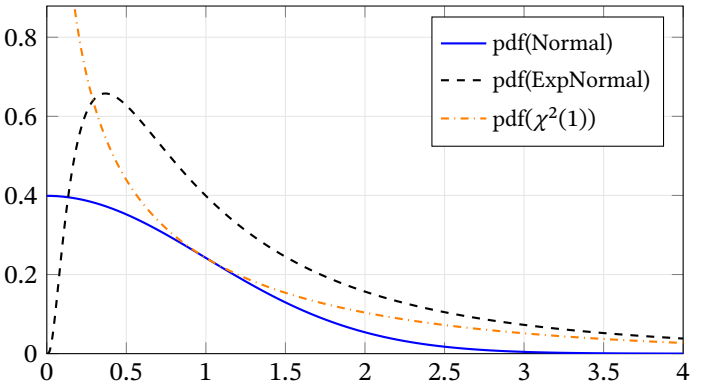

Figure 2.6: Densities of a Standard Normal, an ExpNormal (classically named "Lognormal"), and a Chi-squared distribution with one degree of freedom.

For the Standard Normal, $f(x) = \varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$. This implies for $\bar{N}(0, 1)$:

$$\begin{aligned} g(x) &= \frac{x^{-2}}{\sqrt{2\pi}} e^{-1/(2x^2)} \\ g'(x) &= \frac{x^{-5} - 2x^{-3}}{\sqrt{2\pi}} e^{-1/(2x^2)} \\ g''(x) &= \frac{6x^2 - 7 + x^{-2}}{\sqrt{2\pi}} e^{-1/(2x^2)} \end{aligned}$$

Proceeding from left to right, the Normal is convex, then concave and finally convex again. More precisely, it is convex in the periphery and concave in the centre, with one maximum at the origin and two reflection points at $x = \pm 1$. Its Dual has four inflexion points and two maxima. Straightforwardly, we get the inflection points $g''(x) = 0 \Leftrightarrow x = \pm 1/\sqrt{6}$ or $x = \pm 1$, and the maximum $g'(x) = 0 \Leftrightarrow x = \pm 1/\sqrt{2}$.

It seems to be no coincidence that, left to the origin, $\bar{N}(0, 1)$ switches from convex to concave to convex, and back again to the right of the origin, i.e., from convex to concave to convex. Moreover, g has two maxima. For classical pdfs, that is quite a typical behaviour. Heuristically, if there is much mass at the centre focusing in a single peak, the inverse has much mass in the periphery, and this mass may cumulate in two 'hills' at the points $-1 - \delta$ and $1 + \delta$, respectively. However, already the Cauchy demonstrates that the inverse of a unimodal distribution need not be bimodal.

At least, it is obvious that a symmetric pdf f always yields a symmetric pdf g, since the probability mass left (and right) to the origin remains 1/2 and is just reflected at the points 1 and -1, respectively (the transformation $1/x$ does not change the sign). Formally:

Lemma 2.6.1. *Suppose the pdf $f(x)$ of X has support $\mathbb{R}$ and is symmetric about the y-axis. Then $\mathcal{D}(1/X)$ is also symmetric about the origin.*

Moreover, g has the following properties:

$$\begin{aligned} g'(x) &= \frac{-f'(1/x) - 2xf(1/x)}{x^4} \\ g''(x) &= \frac{6f(1/x)}{x^4} + \frac{6f'(1/x)}{x^5} + \frac{f''(1/x)}{x^6}. \end{aligned} \tag{2.6}$$

W.l.o.g suppose $x > 0$. Any extremum has to satisfy

$$g'(x) = 0 \Leftrightarrow 2xf(1/x) = -f'(1/x) \Leftrightarrow x = -\frac{f'(1/x)}{2f(1/x)}, \tag{2.7}$$

and any inflexion point needs to abide by

$$g''(x) = 0 \Leftrightarrow 6x^2 f(1/x) + 6xf'(1/x) + f''(1/x) = 0. \tag{2.8}$$

Proof. $f(-x) = f(x)$ readily implies $g(-x) = f(-1/x)/(-x)^2 = f(1/x)/x^2 = g(x)$.

The further properties of g are straightforward. □

The last equation has some similarity with a homogeneous Euler differential equation,

$$a_2 x^2 y''(x) + a_1 x y'(x) + a_0 y(x) = 0,$$

where $a_2 \neq 0$ and $x > 0$. However, despite the formal similarity, we are not looking for a function y that satisfies some differential equation. Rather, f is given and we are trying to conclude something about $g(x) = f(1/x)/x^2$.

The behaviour of the normal and other "nice" densities may be understood with the help of equations (2.7, 2.8). Without loss of generality, suppose $x > 0$. If the density $f(1/x)$ were constant, the maximum would occur halfway between the inflexion points, i.e., we could calculate the inflexion points of the Normal with the help of the formula

$$\frac{1}{\sqrt{2}} \pm \frac{\sqrt{3\frac{2e^{-2}}{2\pi} - 2\frac{e^{-1}}{\sqrt{2\pi}}\frac{e^{-1}}{\sqrt{2\pi}}}}{\sqrt{12}\frac{e^{-1}}{\sqrt{2\pi}}} = \frac{1}{\sqrt{2}} \pm \frac{1}{\sqrt{3}}$$

Actually, these points $1/\sqrt{2} - 1/\sqrt{3} \approx 0.1298$ and $1/\sqrt{2} + 1/\sqrt{3} \approx 1.2845$ are close to $1/\sqrt{6} \approx 0.4082$ and 1. More generally speaking, if a density is symmetric, has much density at the centre and does not fluctuate much, its dual should be bimodal.

Without any assumption on the shape of f, however, a density (and thus also its dual) may behave "weird". For instance, on $\mathbb{R}^+$, the density of ExpCauchy is convex - concave - convex, but there is no extremum, since $f'(1/e) = f''(1/e) = 0$ (see Figs. 1.4, 1.9). On the same domain, ExpLévy is first convex and then concave, but it possesses a minimum *and* a maximum (see Fig. 1.5). A pdf may even have arbitrarily many maxima. For instance (see Stuart and Ord (1987), Exercise 6.21),

$$f_{k,s}(x) = \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{(\ln(x))^2}{2}\right)(1 + s\sin(2k\pi \ln(x)))$$

where $-1 \leq s \leq 1$ and $k \in \mathbb{Z}$, defines a familiy of pdfs. $f_{k,0}$ is the pdf of the Standard ExpNormal. However, if $k \neq 0$ and s is increasing, due to the sine, the number of maxima goes to infinity. Nevertheless, these densities all have the same moments. For the first logarithmic moment, we obtain

$$E_I(f_{k,s}) = -2sk\pi e^{-2k^2\pi^2}.$$

The last lemma can be extended:

Lemma 2.6.2. *(Symmetry is hereditary). Suppose X, Y are independent, real-valued r.v.'s with symmetric pdfs f, g about the y-axis. Then this is also true for $X + Y, X - Y, XY,$ and X/Y.*

Proof. $f(-x) = f(x)$ is equivalent to $X \overset{d}{=} -X$. Lemma 2.6.1 shows that $1/X$ is also symmetric.

For all $x, y \in \mathbb{R}$, $f(x) = f(-x)$ and $g(y) = g(-y)$. Thus, due to independence, the masses at the points (x, y), $(-x, y)$, $(x, -y)$, and $(-x, -y)$ have to be identical. Since the four algebraic operators have no preference with respect to direction, this proves the other claims. For example, the mass in $x + y$ is balanced by the mass in $-x - y$, and that in $y - x$ is compensated by the mass in $x - y = -(y - x)$.

□

2.7 Wald or "Inverse Gaussian" distribution

The Wald distribution Wald(a, b), $a, b > 0$ is defined on $\mathbb{R}^+$. Since it is related to the Normal, it is often called "Inverse Gaussian" (for instance, in Mathematica). Formally, the characteristic functions of the Wald and the Normal are related.

More vividly, suppose there is a Brownian motion that starts at the origin and has positive drift. Eventually, a particle that obeys this process will hit some point $m > 0$. Wald(a, b) describes the distribution of time until the first arrival at that point, i.e., it answers the question: How long does it take to get from the origin to m?

Equivalently, if one looks at the same process after some fixed amount of time t, the straightforward question is: Where is the particle? Since the position of the particle follows a Normal distribution, there is some justification to name the Wald distribution "inverse Normal". (In the same sense, the Exponential and the Poisson (to be dealt with soon) are equivalent. Considering random events that occur at a

constant rate (a radioactively decaying substance, say), one may proceed as follows: 1) Fix the number of decays and ask how long it takes to observe them. Equivalently, having just observed a decay, how long do I have to wait until I observe the next decay? This perspective leads to the (continuous) Exponential distribution. 2) Fix the observational time span and count the number of decays during this period of time. This perspective leads to the (discrete) Poisson. Since exactly the same process is described in two different ways, these descriptions have to be equivalent. Also see Section 4.9.2 on this.)

The pdf of $W \sim \text{Wald}(a, b)$ on $\mathbb{R}^+$ is (see, e.g., Jørgensen (1997), p. 261)

$$\begin{aligned} f(x) &= \frac{\sqrt{\frac{b}{x^3}}e^{-\frac{b(x-a)^2}{2a^2x}}}{\sqrt{2\pi}} = \sqrt{\frac{b}{2\pi x^3}}\exp\left(-\frac{b}{2}\left(\frac{x}{a^2}-\frac{2}{a}+\frac{1}{x}\right)\right) \\ &= \sqrt{\frac{b}{2\pi x^3}}e^{b/a}\exp\left(-\frac{b}{2}\left(\frac{x}{a^2}+\frac{1}{x}\right)\right) \end{aligned} \tag{2.9}$$

Thus $EW = a$, $\sigma^2(W) = a^3/b$, and

$$E_I(W) = -\ln a + \sqrt{\frac{2b}{\pi a}}e^{b/a}\left.\frac{\partial K(s, b/a)}{\partial s}\right|_{s=1/2}$$

which is monotonically decreasing in b. (For the definition of $K_n(z)$ see p. 45 and Section 7.6.1.) If $a = 1$, the general formula is

$$M_I^n(W) = \sqrt{\frac{2b}{\pi}}e^{b}\left.\frac{\partial^n K(s, b)}{\partial s^n}\right|_{s=1/2}$$

The next Figure 2.7 shows the pdf $f(x)$ of a Wald(1, 1), and the associated function $(-\ln x)f(x)$. Note that $E_I(Wald(1, 1)) \approx 0.361$, and $\sigma_I^2(W(1, 1)) \approx 0.475$.

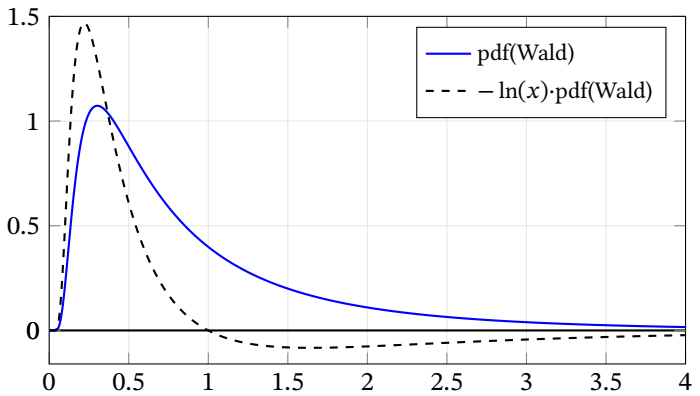

Figure 2.7: Density of a Wald*(1,1)*, and associated information.

2.8 Scales

2.8.1 Levels of measurement

Given that zero and one are fixed, we have an absolute scale. In other words, neither the choice of the origin nor that of the unit of measurement is arbitrary, but due to the basic structure of mathematics.

However, a scale may also be defined relative to the domain or the distribution of a random variable. Suppose the domain is a bounded interval (a, b). Then this interval's endpoints, but also the arithmetic mean $\bar{x} = (a + b)/2$ are natural choices for an origin o. A natural unit of measurement u would be the maximum distance from these points, i.e., the length of the interval if a or b is the origin, and $(b - a)/2$ in the case of the origin $\bar{x}$.

Since we have a choice, the scale on such a domain should not be named absolute. However, unlike classical measurement theory, having chosen the origin, the unit of measurement is also defined, whereas in classical measurement theory, a ratio scale has a non-arbitrary origin and an arbitrary unit of measurement, yet. However, these two kinds of scale have in common that the researcher (or his instrument) has exactly one choice to make. Moreover, since in both cases relative distances $|x - o|/u$ are well defined, it may be reasonable to speak of a 'ratio scale' here, too. See the next section for more on this.

An interval scale leaves the choice of origin *and* unit of measurement to the researcher (or, rather, his measurement device). Given a random variable X with mean μ and standard deviation σ, it is of course straightforward to measure the distance in multiples of σ. Since μ and σ depend on the concrete r.v. at hand, and there are exactly two choices to make, it may be reasonable to speak of an 'interval scale' here, too. In other words, if the origin and the unit of measurement are due to mathematics (or the phenomenon under study), we have an absolute scale. If they are local, crucially depending on the specific situation or object under study, we have an interval scale.

Given a unimodal pdf $f(x)$, where f is concave around the modus and f has convex tails, it is reasonable to put the origin in the modus m, define the centre as that part of $\mathbb{R}$ (or some other domain, in particular $\mathbb{R}^p$), where f is concave and the periphery as the set, where f is convex. Thus, the margin would be the set of all inflexion points of f, where the curvature of the function changes sign. Of course, in the case of the Normal, the latter consideration (curvature) coincides with the former (standard deviation).

2.8.2 Scales on continuous sets

$\ln|1/x|$ defines an absolute scale with origin 0 and unit of measurement $|1 - 0| = 1$ on $\mathbb{R}$. This might seem inappropriate for other sets. For instance, with respect to the interval $[-1, 1]$, the origin should be 0 and the unit of measurement 1/2, since $\mathcal{M} = \{-1/2, 1/2\}$ there. However, in the sense of transformations, we are always dealing with the same scale, decomposing a given set in a 'fair' way into centre and periphery. Equivalently, the decomposition may be described by the origin, the

upper and the lower margin points, and the upper and the lower limit of the set $(x_{inf}, x_{small}, x_0, x_{big}, x_{sup})$. Thus one obtains:

Set	Inf. x_{inf}	Lower margin x_{small}	Origin x_o	Upper margin x_{big}	Sup. x_{sup}	Transf.
$(0,1)$	0	$1/4$	$1/2$	$3/4$	1	$-$
$(0,\pi/2)$	0	$\pi/8$	$\pi/4$	$3\pi/8$	$\pi/2$	$\cdot\pi/2$
$(0,\pi)$	0	$\pi/4$	$\pi/2$	$3\pi/4$	π	$\cdot\pi$

For instance, given the first domain, the observation $x = 1/3$ yields the information $-\ln(4 \cdot |1/3 - 1/2|)$, since one should have $-\ln(4 \cdot |1/4 - 1/2|) = 0$, i.e., a point on the margin is associated with zero information. In general, the information function corresponding to the above intervals is

$$I(x) = -\ln\left(\frac{|x - x_o|}{x_{big} - x_o}\right).$$

Although the scales just introduced depend on the set on which they are defined, one could also think of them as a single scale adapted to the particular set at hand. This idea and the observation that $-\ln|x|$ does not depend on the sign of x lead to the equivalent scales

Set	Inf.	Lower margin	Origin	Upper margin	Sup.	Transf.
$(-1,1)$	-1	$-1/2$	0	$1/2$	1	$-$
$(-\pi/2,\pi/2)$	$-\pi/2$	$-\pi/4$	0	$\pi/4$	$\pi/2$	$\cdot\pi/2$
$(-\pi,\pi)$	$-\pi$	$-\pi/2$	0	$\pi/2$	π	$\cdot\pi$

Starting with the reals, the absolute value function $|\cdot|$ maps $\mathbb{R}$ to $\mathbb{R}^+$. By the same token, $-\infty \mapsto \infty$, $-1 \mapsto 1$, and $0 \mapsto 0$.

Set	Inf.	Lower margin	Origin	Upper margin	Sup.	Transformation
$\mathbb{R}$	$-\infty$	-1	0	1	∞	$-$
$\mathbb{R}^+$	0	0	0	1	∞	$\max(0,\cdot)$
$(0,\infty)$	0	$1/e$	1	e	∞	$\exp(\cdot)$
$(1,\infty)$	1	$e^{1/e}$	e	e^e	∞	$\exp(\exp(\cdot))$
(e,∞)	e	$e^{(e^{1/e})}$	e^e	$e^{(e^e)}$	∞	$\exp(\exp(\exp(\cdot)))$
			...			

2.8.3 Scales on discrete sets

The function $\{x\}$, called the fractional part of x, maps the reals to the interval $[0,1)$, e.g., $\{3.7\} = 0.7$. Obviously, any non-negative real number x may be decomposed into a natural number n and its fractional part: $x = n + \{x\}$, where $n = floor(x) = \lfloor x \rfloor \in \mathbb{N}$.

In other words, $x - n = \{x\}$ informs how much x exceeds $\lfloor x \rfloor$; and $n + \{x\}$ maps $\mathbb{R}$ to $[n, n+1)$. Moreover, $t(x) = \{x\} - 1/2$ takes values in $[-1/2, 1/2)$. This function gives the distance of x to $\mathbb{N} + 0.5 := \{0.5, 1.5, 2.5, ...\}$:

Set	Inf.	Lower margin	Origin	Upper margin	Sup.	Transf.
$[n, n+1)$	n	$n+1/4$	$n+1/2$	$n+3/4$	$n+1$	$-$
$[0,1)$	0	1/4	1/2	3/4	1	$\{\cdot\}$
$[-1/2, 1/2)$	$-1/2$	$-1/4$	0	1/4	1/2	$\{\cdot\} - 1/2$

In a sense, one thus studies "$\mathbb{R}$ modulo $\mathbb{N}$", with the (non-logarithmic) 'information-extracting functions' $I(x) = \{x\}$ and $\tilde{I}(x) = \{x\} - 1/2$ complementing each other. For an explanation of why non-logarithmic iefs are quite natural for discrete distributions, see Section 4.1.

For instance, the fractional part relative to the Standard Pareto is $\int_1^\infty \frac{\{x\}}{x^2}\,\mathrm{d}x = 1 - \gamma$. Since $\int_1^\infty \frac{1}{x}\,\mathrm{d}x$ does not converge, i.e. $1/x$ is not a pdf on $(1,\infty)$, it is reasonable to make the numerator smaller; in particular, $\int_1^\infty \frac{\{x\}-1/2}{x}\,\mathrm{d}x = \ln(\sqrt{2\pi}) - 1$, see Boros and Moll (2004), pp. 178, 93. Moreover, Coffey (2011) proves that

$$\int_0^1 \int_0^1 \cdots \int_0^1 \left\{\frac{1}{x_1 x_2 \cdots x_n}\right\} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n = 1 - \sum_{j=0}^{n-1} \frac{\gamma_j}{j!},$$

where γ_j are the Stieltjes constants. In particular, $\gamma_0 = \gamma$.

2.9 Geometric considerations

Probability distributions are often associated with certain interesting geometric structures. Thus, in this section, we study important one-dimensional sets and their transformations.

2.9.1 The Circle and the Interval (the sine and the Arcsin Distribution)

Looking at expected information, a reflection with respect to the $y-$axis does not make any difference, thus $E_I(U(0,1)) = E_I(U(-1,1)) = 1$. However, on the real line, dilating the unit interval decreases expected information, for instance, $E_I(U(-\pi,\pi)) = 1 - \ln(\pi)$, see p. 9.

However, one can also argue that $U(-\pi,\pi)$ represents the unit circle's circumfer-

ence, that is, $\delta B = \{z = (x, y) | x^2 + y^2 = 1\}$. The Arcsin distribution transparently connects both sets:

Distribution	Domain	$f(x)$	$y = F(x)$	$x = F^{-1}(y)$
$U(0,1)$	$[0,1]$	1	x	y
$U(-1,1)$	$[-1,1]$	$1/2$	$\frac{1}{2} + \frac{1}{2}x$	$2(y - 1/2)$
$U(-\pi,\pi)$	$[-\pi,\pi]$	$1/(2\pi)$	$\frac{1}{2} + \frac{1}{2\pi}x$	$2\pi(y - 1/2)$
$Y \sim \text{Arcsin}(-1,1)$	$(-1,1)$	$\frac{1}{\pi\sqrt{1-x^2}}$	$\frac{2}{\pi}\arcsin\left(\sqrt{\frac{1+x}{2}}\right)$	$2\sin^2(\pi y/2) - 1$
$X \sim \text{Arcsin}(0,1)$	$(0,1)$	$\frac{1}{\pi\sqrt{x(1-x)}}$	$\frac{2}{\pi}\arcsin(\sqrt{x})$	$\sin^2(\pi y/2)$

and

$$\text{Arcsin}(0,1) \underset{(\cdot)^{1/2}}{\overset{(\cdot)^2}{\leftrightarrows}} \text{Arcsin}(-1,1) \underset{\arcsin(\cdot)}{\overset{\sin(\cdot)}{\leftrightarrows}} U(-\pi,\pi) \underset{|\cdot\frac{1}{\pi}|}{\overset{\cdot\pi S}{\leftrightarrows}} U(0,1)$$

In other words, the chain of transformations defining the Standard Arcsine distribution starts on the unit interval, proceeds to the circumference of the unit circle δB, and returns to the unit interval. The middle part of the transformations defines an automorphism of the interval $(-1,1)$, i.e., if $U \sim U(-1,1)$, then $Y = \sin(\pi U) \sim$Arcsin$(-1,1)$; and $Y \sim$Arcsin$(-1,1) \Leftrightarrow Y^2 \sim$Arcsin$(0,1)$ connects the intervals $(-1,1)$ and $(0,1)$.

It may further be mentioned that given $U(-\pi,\pi)$, the sine can be replaced by the cosine, and that $\sin(T) \sim$Arcsin$(-1,1)$ if T has a symmetric triangular distribution on $(-\pi,\pi)$. Assuming independence, we have in a picture:

$$\begin{array}{ccc} & & U \sim U(-\pi,\pi) \\ & & \cos \downarrow \sin \\ U_i \sim U(-\pi/2,\pi) & \overset{\sin}{\rightarrow} & \text{Arcsin(-1,1)} \\ & & \uparrow \sin \\ U_1 + U_2 & \sim & \text{Triangular}(-\pi,\pi) \end{array}$$

By the same token, we get for the Raised Cosine Distribution $R_{\mu,\sigma} = R(\mu-\sigma, \mu+\sigma)$ with pdf $f(x) = \frac{1}{2\sigma}\left(1 + \cos\left(\frac{x-\mu}{\sigma}\pi\right)\right)$ on the interval $[\mu - \sigma, \mu + \sigma]$,

Distribution	Domain	$f(x)$	$y = F(x)$	$x = F^{-1}(y)$
$U(0,1)$	$[0,1]$	1	x	y
$U(-\pi,\pi)$	$[-\pi,\pi]$	$1/(2\pi)$	$\frac{1}{2} + \frac{1}{2\pi}x$	$2\pi(y - 1/2)$
$R(-\pi,\pi)$	$(-\pi,\pi)$	$\frac{1+\cos x}{2\pi}$	$\frac{1}{2} + \frac{x+\sin x}{2\pi}$	$s^{-1}(2\pi(y - 1/2))$
$R(0,1)$	$(0,1)$	$1 + \cos(\pi(2x-1))$	$x + \frac{\sin(\pi(2x+1))}{2\pi}$	$r^{-1}(2\pi y)$

where $s^{-1}(\cdot)$ is the inverse function of $s(x) = x + \sin x$ on the interval $(-\pi, \pi)$. s^{-1} exists since $s(x)$ is strictly increasing on $(-\pi, \pi)$, for $s'(x) = 1 + \cos(x) > 0$, and $r^{-1}(\cdot)$ is the inverse function of $2\pi x + \sin(\pi(2x+1))$ which is strictly increasing on the unit interval. Altogether,

$$R(0,1) \underset{\pi(2(\cdot)-1)}{\overset{\cdot\frac{1}{2\pi}+\frac{1}{2}}{\leftrightarrows}} R(-\pi,\pi) \underset{s(\cdot)}{\overset{s^{-1}(\cdot)}{\leftrightarrows}} U(-\pi,\pi) \underset{|(\cdot)/\pi|}{\overset{\cdot\pi S}{\leftrightarrows}} U(0,1)$$

2.9.2 The Circle and the Straight Line (the tangent and the Cauchy distribution)

In the same vein, one may also argue that $U(-\pi/2, \pi/2)$ represents the unit circle's circumference in the half plane, that is, $\delta B^+ = \{z = (x,y) | x^2 + y^2 = 1\} \cap \mathbb{C}^+$, and $\mathbb{C}^+ = \{(x,y) | x \geq 0\}$. In polar coordinates, $\delta B^+ = \{z = (r,\phi) | r = 1, -\pi/2 \leq \phi \leq \pi/2\}$. Owing to the last section, it is straightforward to define the origin of δB^+ to be the point $(1,0) \Leftrightarrow \phi = 0$, the centre would be $\delta B^+_{\mathcal{C}} = \{z = (r,\phi) | r = 1, -\pi/4 < \phi < \pi/4\}$, its margin $\delta B^+_{\mathcal{M}} = \{z = (r,\phi) | r = 1, |\phi| = \pi/4\}$, and $\delta B^+_{\mathcal{P}} = \{z = (r,\phi) | r = 1, \pi/4 < |\phi| \leq \pi/2\}$ the periphery. Moreover, it is appropriate to measure information with the help of the function $I(z) = I(\phi) = -\ln |4\phi/\pi|$, such that $I(\pm\pi/4) = 0$.

In the following, we will often need the set $\mathbb{R}_1 = \{(x,y) | x = 1, y \in \mathbb{R}\}$, i.e., the line $x = 1$, or a subset of it.

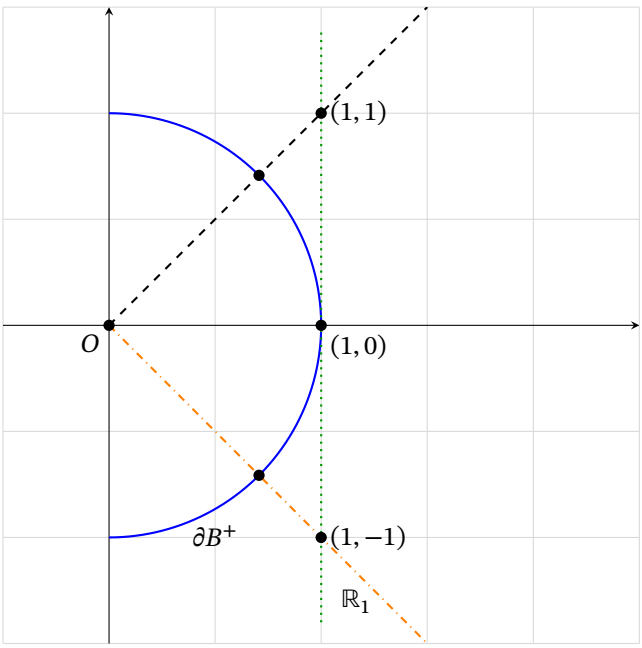

Figure 2.8: The circle, the lines $y = x, y = -x$, and $\mathbb{R}_1$ in the positive half plane.

The tangent maps $(-\pi/2, \pi/2)$ to the reals, in particular

$$\tan(\pm\pi/4) = \pm 1 \Leftrightarrow \arctan(\pm 1) = \pm\pi/4.$$

Given the interpretation of Fig. 2.8, $\tan(\delta B^+) = \mathbb{R}_1$, and "margin is mapped to margin". That is, since $\tan(\pi/4)/1 = \sin(\pi/4)/\cos(\pi/4)$, in the plane, the point $(1/\sqrt{2}, 1/\sqrt{2})$ is mapped to the point $(1,1)$. Of course, an analogous results holds for $-\pi/4 \in \delta B^+$ which is mapped to the point $(1,-1) \in \mathbb{R}_1$. Moreover, the probability mass in each of the four sections $(x_{inf}, x_{small}), (x_{small}, x_o), (x_o, x_{big}), (x_{big}, x_{sup})$ is mapped to the corresponding sections of the other set:

δB^+	$(-\pi/2, -\pi/4)$	$(-\pi/4, 0)$	$(0, \pi/4)$	$(\pi/4, \pi/2)$
$Mass$	p_1	p_2	p_3	p_4
$\mathbb{R}_1$	$(-\infty, -1)$	$(-1, 0)$	$(0, 1)$	$(1, \infty)$

Given the uniform $U(\delta B^+)$ on the first set, it is straightforward to determine the image distribution on $\mathbb{R}_1$. A variant of this construction is known as the 'witch of Agnesi', and can be found in various books (for instance, Johnson, Kotz and Balakrishnan (1994), p. 299). Formally, one has

Distribution	Domain	$f(x)$	$y = F(x)$	$x = F^{-1}(y)$
$U \sim U(0,1)$	$[0,1]$	1	x	y
$U(-1,1)$	$[-1,1]$	$1/2$	$\frac{1}{2} + \frac{1}{2}x$	$2(y - 1/2)$
$U(-\pi/2, \pi/2)$	$[-\pi/2, \pi/2]$	$1/\pi$	$\frac{1}{2} + \frac{1}{\pi}x$	$\pi(y - 1/2)$
$C \sim C(0,1)$	$\mathbb{R}_1(\mathbb{R})$	$\frac{1}{\pi(1+x^2)}$	$\frac{1}{2} + \frac{\arctan(x)}{\pi}$	$\tan(\pi(y - 1/2))$

or, equivalently,

$$C \underset{\arctan(\cdot)}{\overset{\tan(\cdot)}{\leftrightarrows}} U(-\pi/2, \pi/2) \underset{\cdot 2/\pi}{\overset{\cdot \pi/2}{\leftrightarrows}} U(-1,1) \underset{|\cdot|}{\overset{\cdot S}{\leftrightarrows}} U(0,1) \underset{F^{-1}(\cdot)}{\overset{F(\cdot)}{\leftrightarrows}} C(0,1).$$

In other words, we have

$$U \overset{d}{=} |\frac{2}{\pi} \arctan(C)| \iff C \overset{d}{=} \tan\left(\frac{\pi}{2} SU\right) \tag{2.10}$$

This means that the Cauchy is *not* completely different from the Uniform. Rather, it is just the projection of the uncurved Uniform (on the curved boundary of the circle δB^+) to the curved Cauchy (on the uncurved real line $\mathbb{R}_1$). Moreover, there is also a direct link between the Cauchy and the Arcsin: Starting with the r.v. $1 + C^2$ on the interval $(1, \infty)$, the pdf of its inverse $1/(1 + C^2)$ is $1/(\pi\sqrt{x(1-x)})$ on the unit interval, i.e., that of an Arcsin$(0,1)$.

Owing to the results on pp. 37, 53, it is also no coincidence that given the independent r.v.'s $A \sim \text{Arcsin}(0,1) \overset{d}{=} \sin^2(\pi U/2)$, $X \sim \text{Exp}(1/2) \overset{d}{=} \ln U^{-2}$, and Rademacher's S, the product $S(AX)^{1/2}$ is the Standard Normal N. In more detail:

	pdf	E_I	σ_I^2
A	$1/(\pi\sqrt{t(1-t)})$	$2\ln 2$	$\pi^2/3$
X	$\exp(-t/2)/2$	$\gamma - \ln 2$	$\pi^2/6$
AX	$\exp(-t/2)/\sqrt{2\pi t}$	$\gamma + \ln 2$	$\pi^2/2$
$\pm\sqrt{AX}$	$\exp(-t^2/2)/\sqrt{2\pi}$	$(\gamma + \ln 2)/2$	$\pi^2/8$

In other words, a Standard Normal is the geometric mean of an Exponential and an ArcSin. Thus, quite fittingly, its expected information is the arithmetic mean of γ and $\ln 2$, the expected information in $X^* \sim \text{Exp}(1)$, and $A^* \sim \text{ArcSin}(-1,1)$, respectively. Actually, we have that

$$\begin{array}{ccccc} N^2 \sim \chi^2(1) & \overset{d}{=} & A \sim \text{ArcSin}(0,1) & \cdot & X \sim \text{Exp}(1/2) \\ & & \uparrow (\cdot)^2 & & \uparrow \cdot 2 \\ N_1 N_2 \sim \text{VG}(1/2,1) & \overset{d}{=} & A^* \sim \text{ArcSin}(-1,1) & \cdot & X^* \sim \text{Exp}(1) \end{array}$$

where N_1, N_2 are independent Standard Normal r.v.'s, and VG is a Variance Gamma distribution, see p. 45. Note that $E_I(N^2) = E_I(N_1 N_2) = \gamma + \ln 2$. However, $\sigma_I^2(AX) = \pi^2/2$, whereas $\sigma_I^2(A^*X^*) = \pi^2/4$, see Corollary 1.8.5, p. 28, and p. 458.

Given the Raised cosine, the Arcsin and the Uniform on δB^+, the Uniform has no particular preference with respect to centre and periphery, yet the other two distributions have (see the next illustrations 2.9 and 2.10).

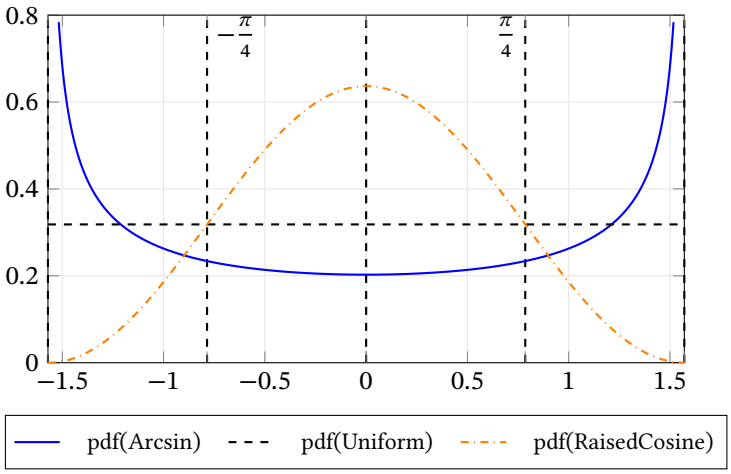

Figure 2.9: The pdfs of the Arcsin, Raised Cosine and Uniform Distributions on the interval $(-\pi/2, \pi/2)$.

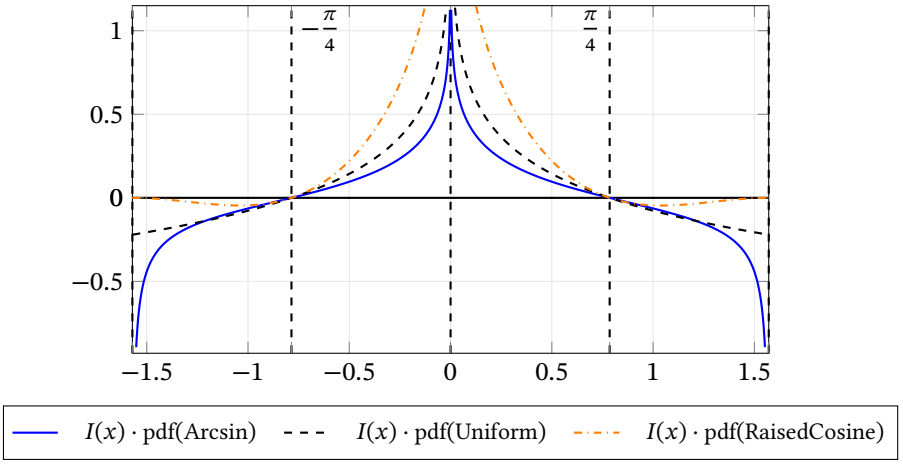

Figure 2.10: The pdfs of Fig. 2.9 multiplied by $I(x) = -\ln(4|x|/\pi)$

The Raised Cosine on $(-\pi/2, \pi/2)$ has pdf $f(x) = (1 + \cos(2x))/\pi$. Since

$$\int_0^{\pi/2} (-\ln(4x/\pi)(1 + \cos(2x))/\pi = \frac{1}{2} - \frac{\ln 2}{2} + \frac{\int_0^{\pi} \sin(t)/t \, dt}{2\pi}$$

we get the expected information

$$E_I(R(-\pi/2, \pi/2)) = 1 - \ln 2 + \frac{1}{\pi} \int_0^{\pi} \frac{\sin(t)}{t} \, dt \approx 0.8963$$

The function $Si(x) = \int_0^x (\sin(t)/t) \, dt$ is called sine integral (see Section 7.3.7), and the number $\int_0^{\pi} \sin(t)/t \, dt \approx 1,8519$ has been named the Wilbraham Gibbs constant (see, in particular, H. Jeffreys and B. Jeffreys (1988), Section 14.07, pp. 445-446 on the Gibbs-Wilbraham phenomenon). The Uniform on $(-\pi/2, \pi/2)$ is associated with the expected information $\int_{-\pi/2}^{\pi/2} (-\ln|4x/\pi|)/\pi \, dx = 1 - \ln 2 \approx 0.3069$. Finally, we have the following for the Arcsin distribution:

$$E_I(\text{Arcsin}(-\pi/2, \pi/2)) = \int_{-\pi/2}^{\pi/2} \frac{-\ln(4|x|/\pi)}{\pi\sqrt{\left(\frac{\pi}{2} - x\right)\left(x + \frac{\pi}{2}\right)}} \, dx = 0.$$

Moreover, letting $R \sim R(-\pi/2, \pi/2)$, $U \sim U(-\pi/2, \pi/2)$ and $X \sim \text{Arcsin}(-\pi/2, \pi/2)$, we know that on δB^+ and $\mathbb{R}_1$

		Mass on the interval			
δB^+	$\mathbb{R}_1(\mathbb{R})$	(x_{inf}, x_{small})	(x_{small}, x_o)	(x_o, x_{big})	(x_{big}, x_{sup})
R	$tan(R)$	$\frac{\pi-2}{4\pi}$	$\frac{\pi+2}{4\pi}$	$\frac{\pi+2}{4\pi}$	$\frac{\pi-2}{4\pi}$
U	$Cauchy$	1/4	1/4	1/4	1/4
X	$tan(X)$	1/3	1/6	1/6	1/3

since $F_R(-\pi/4) = 1/2 - (\pi/4)/\pi + \sin(-\pi/2)/(2\pi) = (\pi - 2)/(4\pi) \approx 0.0908$, and $F_X(-\pi/4) = 1/3$. It is also no coincidence that the pdf $f(x) = \frac{2}{\pi(x^2+1)^2}$ of $\tan(R)$ has a more pronounced maximum at the origin than the Normal, and that the pdf $g(x) = \frac{2}{\pi(x^2+1)\sqrt{\pi^2-4\arctan(x)^2}}$ of $\tan(X)$ has 'heavy tails' (see Fig. 2.11):

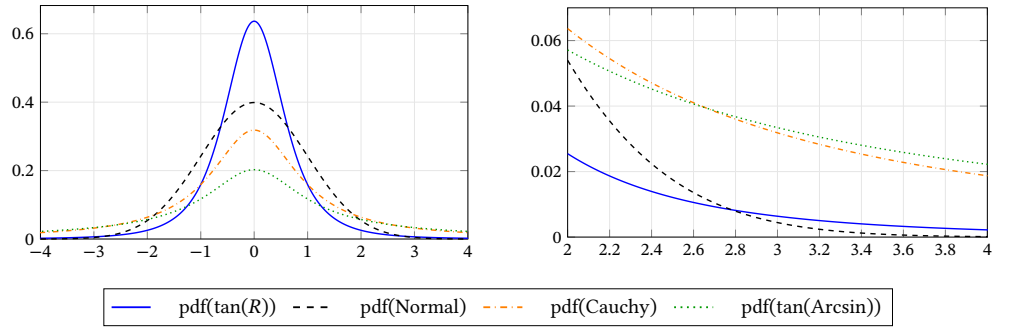

Figure 2.11: The pdfs of the Standard Normal, the tangent of the Arcsin, the Raised Cosine and the Uniform Distributions, respectively, on $\mathbb{R}$.

Since the tangent maps δB^+ to $\mathbb{R}_1$, it is also natural to start with the Uniform on some interval $(a, b) \in \mathbb{R}_1$ and to use the inverse mapping - Arctan - to get back to δB^+. For instance, $\arctan(U(-1, 1))$ leads to the pdf $\sec^2(x)/2$ and the cdf $(1 + \tan(x))/2$ on the interval $(2\arctan\left(1 - \sqrt{2}\right), -2\arctan\left(1 - \sqrt{2}\right)) = (-\pi/4, \pi/4)$, since the cdf of $\arctan(U(0, 1))$ is $\tan(x)$ on the interval $(0, \pi/4)$.

Knowing the link between the Cauchy and the circle, one should not be too surprised to learn that the following integral (see Nahin (2015), chap. 6.2. 'Ahmed's integral') can be expressed in terms of π:

$$\int_0^1 \frac{\arctan(\sqrt{2 + x^2})}{(1 + x^2)\sqrt{2 + x^2}} \, dx = \frac{5}{3 \cdot 2^5}\pi^2.$$

2.9.3 The Line and the Hyperbola (hyperbolic functions)

Although the notation $U(-\pi/2, \pi/2)$ and its ilk is (deliberately!) not unequivocal, it is very useful to think about sets and their transformations in geometric and probabilistic terms. On the one hand, a set has a certain shape (a curvature, in particular) and is located in some space. On the other hand, probabilistic distributions are always defined on specific sets, may represent these sets, and also have characteristic shapes. Transforming distributions is often tantamount to changing the shape of the underlying set. Moreover, particularly interesting sets and their transformations are associated with particularly important standard distributions.

In this vein, we are now going to extend the interval $(-1, 1) \subseteq \mathbb{R}_1$ to the hyperbola $H^+ = \{(x, y)|x^2 - y^2 = 1; x, y > 0\}$. Much of what follows can be derived directly from the basic geometric setting:

Every straight line $y = ax$ in $\mathbb{C}^+$ maps a point $t \in (-1, 1) \subseteq \mathbb{R}_1$ to a point y on

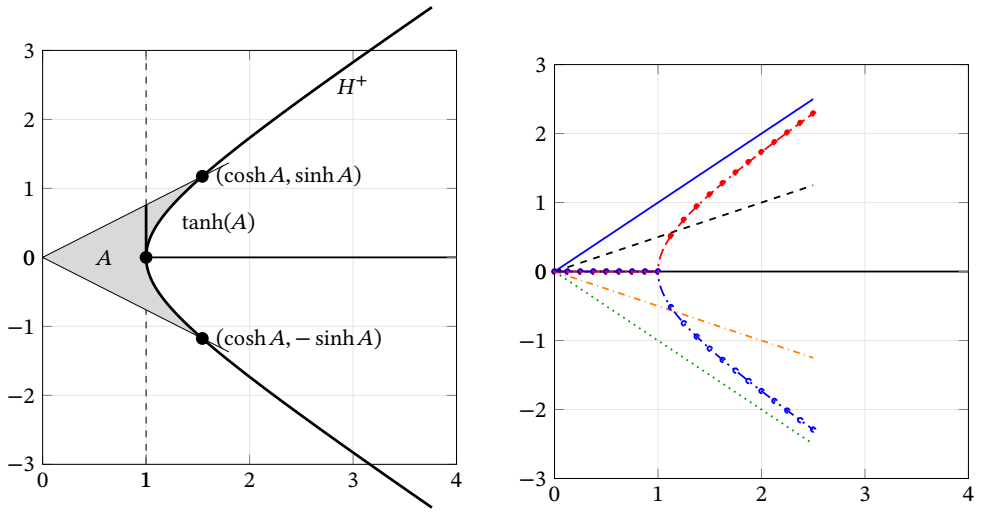

Figure 2.12: The hyperbola and associated functions in the positive half plane.

the hyperbola. In particular, the coordinates $(x, y) = (\cosh A, \sinh A)$ of the image of $1/2 \in \mathbb{R}_1$ are

$$\frac{1/2}{1} = \frac{\sqrt{x^2-1}}{x} = \sqrt{\frac{x^2-1}{x^2}} = \sqrt{1-1/x^2} \Leftrightarrow x = 2/\sqrt{3} \approx 1.155,$$

and $y = \sqrt{4/3-1} = 1/\sqrt{3}$. Thus we obtain the associations

$x \in \mathbb{R}_1$	x_{inf}	x_{small}	x_o	x_{big}	x_{sup}
	-1	$-1/2$	0	$1/2$	1
$y \in H^+$	$-\infty$	$(2/\sqrt{3}, -1/\sqrt{3})$	0	$(2/\sqrt{3}, 1/\sqrt{3})$	∞
	y_{inf}	y_{small}	y_o	y_{big}	y_{sup}

and the unit of measurement on the hyperbola becomes $y_{big} - y_o$ which may be calculated with the help of an Elliptic integral of the second kind (also see Section 7.6.1):

$$y_{big} - y_o = \int_0^{arcosh(2/\sqrt{3})} \sqrt{1 + 2\sinh^2(x)}\, dx = -iE\left(\frac{i}{2}\ln(3)\middle|\, 2\right) \approx 0.603.$$

The corresponding angle $\phi \in (-\pi/4, \pi/4)$ may be computed with the help of the point $(\cos\phi, \sin\phi)$, where the line $y = x/2$ and δB^+ intersect. $\cos^2(\phi)/4 = \sin^2(\phi) = 1 - \cos^2(\phi)$ implies $(\cos\phi, \sin\phi) = (2/\sqrt{5}, 1/\sqrt{5})$, and thus $\phi = \arcsin(1/\sqrt{5}) \approx 26.57°$.

Suppose $S \subseteq \mathbb{C}^+$ is the set delimited by $y = \pm x$ and H^+. The area functions $\cosh(A)$, $\sinh(A)$, $\tanh(A)$, *etc.* are defined on S. The yellow area $A \subseteq S$ in Fig. 2.12 is the area of the "hyperbolic triangle", defined by the vertices $(0,0)$, $(\cosh(A), -\sinh(A))$, and $(\cosh(A), \sinh(A))$. That is, the section between $(0,0)$ and $(\cosh(A), -\sinh(A))$

is straight, and so is the section between $(0,0)$ and $(\cosh(A), \sinh(A))$. However, the connection between the points $(\cosh(A), -\sinh(A))$ and $(\cosh(A), \sinh(A))$ is the hyperbola.

In the case above, $1/2 = \tanh(A) = \sinh(A)/\cosh(A)$ implies $\cosh(A) = 2\sinh(A)$ and thus $A = \ln 3/2 \approx 0.549$. Moreover, a natural measure with respect to the hyperbola may be defined with respect to the areas $B = A/2$ of the "hyperbolic triangles", i.e.,

B_{inf}	B_{small}	B_o	B_{big}	B_{sup}
$-\infty$	$-(\ln 3)/4$	0	$(\ln 3)/4$	∞

It is also well-known that the area defined by artanh can be expressed with the help of the logarithm:

$$B_x = 2\,\mathrm{artanh}(x) = \ln\left(\frac{1+x}{1-x}\right) = \ln(1+x) - \ln(1-x) = \int_{1-x}^{1+x} \frac{1}{t}\,\mathrm{d}t,$$

if $|x| < 1$. Actually, in the old days, the natural logarithm was called *hyperbolic* logarithm, which is no coincidence, since the latter area is just that of a "standard hyperbolic rectangle" with vertices $(1-x,0)$, $(1+x,0)$, $(1+x,1/(1+x))$, $(1-x,1/(1-x))$, and the upper side replaced by the hyperbola $y = 1/t$, see Fig. 2.13.

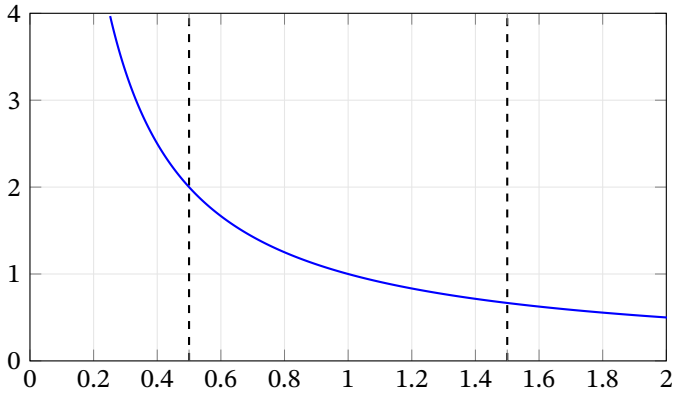

Figure 2.13: The hyperbola $y = 1/t$ with $x = 1/2$ (see text).

2.9.4 The Circle, the Line, and the Hyperbola (Gudermann's functions)

Next, we are now going to join δB^+, $\mathbb{R}$, and the hyperbola.[5] Much of what follows can be derived directly from a basic geometric figure (see Fig. 2.14).

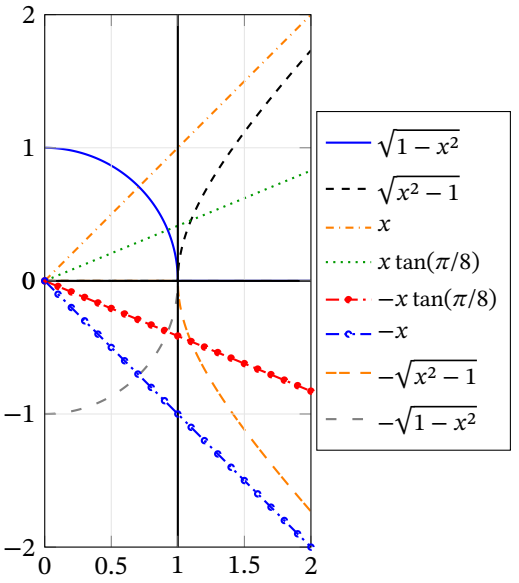

Figure 2.14: The circle $x^2 + y^2 = 1$, $\mathbb{R}_1$, the hyperbola $x^2 - y^2 = 1$, and the lines corresponding to angles 45°, 22.5°, −22.5°, −45° ($\pi/4, \pi/8, -\pi/8, -\pi/4$) in the positive half plane.

$U(-\pi/4, \pi/4)$ may represent the boundary of the unit circle $\delta B_{0,1}$ in the wedge $W = \{(x, y)|x \geq 0, -x \leq y \leq x\}$, that is, $\delta B_W = \{z = (x, y)|x^2 + y^2 = 1\} \cap W$. In polar coordinates, $\delta B_W = \{z = (r, \phi)|r = 1, -\pi/4 \leq \phi \leq \pi/4\}$. With respect to the latter set, it is straightforward to define the origin as the point $(1, 0) \Leftrightarrow \phi = 0$, the centre would be the set $\delta B_{\mathcal{C}}^W = \{z = (r, \phi)|r = 1, -\pi/8 < \phi < \pi/8\}$, its margin $\delta B_{\mathcal{M}}^W = \{z = (r, \phi)|r = 1, \phi = \pm\pi/8\}$, and $\delta B_{\mathcal{P}}^W = \{z = (r, \phi)|r = 1, \phi < -\pi/8 \vee \phi > \pi/8\}$ the periphery. Finally, it is appropriate to measure information with respect to this set, i.e. $I(z) = I(\phi) = -\ln(8|\phi|/\pi)$, such that $I(\pm\pi/8) = 0$.

Although, of course, $\mathbb{R}_1$ is not an axis of symmetry, there is a 1:1:1 correspondence between any point $z \in \delta B_W$, $t \in \mathbb{R}_1$, and $\tilde{z} \in H^+$. In particular, if $\phi = 22.5° = \pi/8$, the point $z = (\cos(\pi/8), \sin(\pi/8)) \in \delta B_W$ corresponds to $t = (1, \tan(\pi(8)) \in \mathbb{R}_1$ and $\tilde{z} = (\cosh(\pi/8), \sinh(\pi/8)) \in H^+$.

[5] Thus, in a sense, we are extending the 'witch of Agnesi'.

More explicitly, $\tan\left(\frac{\pi}{8}\right) \approx 0.4142 < 1/2$, and $\cosh^2\left(\frac{\pi}{8}\right) = \frac{1}{1-\tan^2(\pi/8)}$, $\sinh^2\left(\frac{\pi}{8}\right) = \frac{\tan^2(\pi/8)}{1-\tan^2(\pi/8)}$. We obtain the associations

$\phi \in \delta B_W$	ϕ_{inf}	ϕ_{small}	ϕ_o	ϕ_{big}	ϕ_{sup}
	$-\pi/4$	$-\pi/8$	0	$\pi/8$	$\pi/4$
$y \in H^+$	$-\infty$	$\left(\frac{1}{\sqrt{1-\tan^2(\frac{\pi}{8})}}, \frac{-\tan(\frac{\pi}{8})}{\sqrt{1-\tan^2(\frac{\pi}{8})}}\right)$	0	$\left(\frac{1}{\sqrt{1-\tan^2(\frac{\pi}{8})}}, \frac{\tan(\frac{\pi}{8})}{\sqrt{1-\tan^2(\frac{\pi}{8})}}\right)$	∞
	y_{inf}	y_{small}	y_o	y_{big}	y_{sup}

and the unit of measurement on the hyperbola becomes $y_{big} - y_o$, which may be calculated with the help of the following integral:

$$\begin{aligned} y_{big} - y_o &= \int_0^{arcosh(1/\sqrt{1-\tan^2(\pi/8)})} \sqrt{1 + 2\sinh^2(x)}\,dx \\ &= -iE\left(\frac{i}{2}\operatorname{arsinh}(1)\middle|\, 2\right) \approx 0.4687. \end{aligned}$$

Actually, the standard link between the three sets just considered is defined by

$$e^x = \tan\left(\frac{\pi}{4} + \phi\right) = \frac{1 + \tan\phi}{1 - \tan\phi} \tag{2.11}$$

where $-\pi/4 \le \phi \le \pi/4$. If $\alpha = 2\phi$, and thus $-\pi/2 \le \alpha \le \pi/2$, one obtains

$$\tan(\alpha/2) = \frac{e^x - 1}{e^x + 1} = \tanh(x/2), \tag{2.12}$$

and

$$\tan\alpha = \sinh x \qquad \text{or, equivalently,} \qquad \sin\alpha = \tanh x \tag{2.13}$$

There is much to be said about this.

First, when we introduced the hyperbola, we used the Cartesian parametrisation $\tilde{z} = (\cosh A, \sinh A)$ for every $\tilde{z} \in H^+$, where $A \in \mathbb{R}$. However, based on δB^+, one may also write $\tilde{z} = (1/\cos\alpha, \tan\alpha) = (\sec\alpha, \tan\alpha)$, where $-\pi/2 < \alpha < \pi/2$. In particular, $\tilde{z} = (1, 0)$ if $\alpha = 0$.

Second, note that $\varphi = \phi + \pi/4$ moves δB_W and thus the scale represented by

$$(-\pi/4, -\pi/8, 0, \pi/8, \pi/4)$$

to the first quadrant (rotating it by 45°). More explicitly, $U(0,\pi/2)$ may represent the boundary of the unit circle $\delta B_{0,1}$ in the set $Q^{++} = \{(x,y)|x,y \geq 0\}$, that is, $\delta B^{++} = \{z = (x,y)|x^2 + y^2 = 1\} \cap Q^{++}$.

In polar coordinates, $\delta B^{++} = \{z = (r,\varphi)|r = 1, 0 \leq \varphi \leq \pi/2\}$. The origin of this set is the point $(1/\sqrt{2}, 1/\sqrt{2}) \Leftrightarrow \varphi = \pi/4$, the centre would be the set $\delta B_{\mathcal{C}}^{++} = \{z = (r,\varphi)|r = 1, \frac{1}{8}\pi < \varphi < \frac{3}{8}\pi\}$, its margin $\delta B_{\mathcal{M}}^{++} = \{z = (r,\varphi)|r = 1, \varphi = \pi/8 \vee \varphi = 3\pi/8\}$, and $\delta B_{\mathcal{P}}^{++} = \{z = (r,\varphi)|r = 1, \varphi < \pi/8 \vee \varphi > \frac{3}{8}\pi\}$ the periphery. Finally, it is appropriate to measure information with respect to this set, i.e. $I(z) = I(\varphi) = -\ln(\frac{8}{\pi}|\varphi - \pi/4|)$, such that $I(\pi/8) = I(3\pi/8) = 0$. Equivalently, $I(\phi) = -\ln(\frac{8}{\pi}|\phi|)$ has the property that $I(-\pi/8) = I(\pi/8) = 0$ (see Figs. 2.15, 2.16).

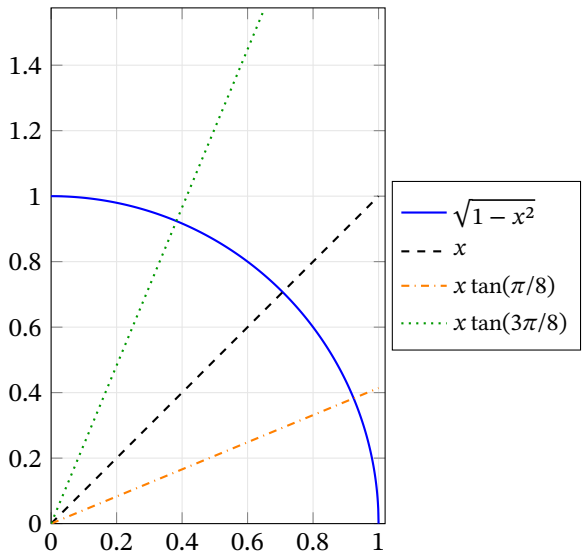

Figure 2.15: The situation in the first quadrant.

Third, due to the rotation, $y = \tan(\varphi) \in \mathbb{R}_1^+$. Thus it makes sense to write $y = e^x$, and the inverse $\ln y = x = \ln(\tan(\varphi))$ is well-defined, mapping an angle $\varphi \in (0,\pi/2)$ to an area $x \in [0,\infty)$, see Fig. 2.12, equation (2.14), and Sections 2.10, 7.3.2.

Fourth, $\tan(\cdot)$ maps the semi-circle δB^+ to $\mathbb{R}_1$, and δB_W to $(-1,1) \in \mathbb{R}_1$. Almost symmetrically, $\tanh(\cdot)$ maps the area S between the lines $y = \pm x$ and the hyperbola to $(-1,1) \in \mathbb{R}_1$. Figuratively speaking, tan and tanh "make ends - circle and hyperbola - meet" on the straight line $\mathbb{R}_1$. More precisely and prosaically, equations (2.11) and (2.12) say that if $t = \frac{y-1}{y+1} = \frac{e^x-1}{e^x+1} \in (-1,1)$,

δB_W	$\rightarrow$	$\mathbb{R}_1$	$\leftarrow$	S
$\phi = \alpha/2$	tan	t	tanh	$x/2$
$\phi = \alpha/2$	arctan	t	artanh	$x/2$
δB_W	$\leftarrow$	$\mathbb{R}_1$	$\rightarrow$	S

Fifth, going from right to left in the last table, the Gudermannian function gd(·), named after Christoph Gudermann (1798-1852), see Olver et al. (2010), p. 121,[6] maps S to δB_W. Equivalently, since $\alpha = 2\phi$, the area x is mapped to the corresponding angle $\alpha \in \delta B^+$,

$$\alpha = \mathrm{gd}(x) = 2\arctan(\tanh(x/2)) = \arctan(\sinh x) = \int_0^x \frac{\mathrm{d}t}{\cosh t} = \int_0^x \mathrm{sech}(t)\,\mathrm{d}t\,.$$

Therefore, sixth, the inverse Gudermannian function argd(·)

$$\begin{aligned} x &= \mathrm{argd}(\alpha) = \ln\tan\left(\frac{\pi}{4} + \frac{\alpha}{2}\right) = \ln(\sec\alpha + \tan\alpha) \\ &= \mathrm{arsinh}(\tan\alpha) = \mathrm{artanh}(\sin\alpha) = \int_0^\alpha \frac{\mathrm{d}t}{\cos t} = \int_0^\alpha \sec(t)\,\mathrm{d}t \end{aligned} \tag{2.14}$$

connects δB^+ and S, and one may also write $e^x = \sec\alpha + \tan\alpha$.

Finally, note that $\mathrm{argd}(\phi) = \ln\tan\left(\frac{\pi}{4} + \phi\right)$ has two poles in $-\pi/4$ and $\pi/4$, respectively. Thus, in a sense it is the "dual" to $I(\phi) = -\ln(\frac{8}{\pi}|\phi|)$. The former function indicates how close $\pi/4 + \phi$ is to the endpoints 0 and $\pi/2$ of the set δB^{++}, the latter has one pole at the origin.

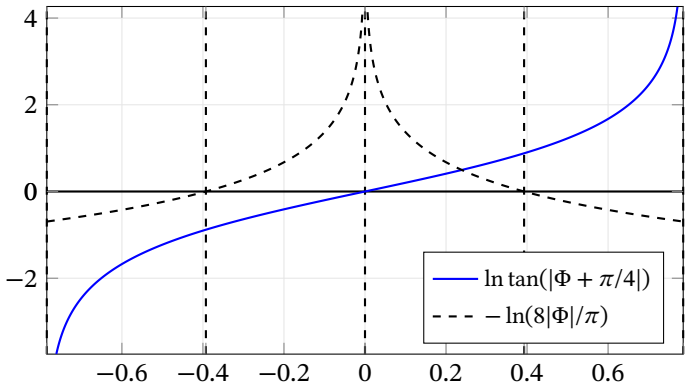

Figure 2.16: Information functions indicating the endpoints and the origin, respectively, of the finite interval $(-\pi/4 \le \phi \le \pi/4)$.

[6] There is also a printed edition of this book, Olver et al. (2010). In what follows, we are always going to refer to the online version.

2.9.5 The Circle, the Line, the Hyperbola, and the Lemniscate

Arguably, the circle, the line, the hyperbola, and the lemniscate (see Fig. 2.17) are the most prominent of sinusoidal spirals (cf. Zwikker (2005), pp. 225-227). A sinusoidal spiral is a curve in the plane whose points $z = (r, \varphi)$ in polar coordinates satisfy

$$r^n = c^n \cos(n\varphi) = \frac{(2p)^n}{2} \cos(n\varphi)$$

with a constant c, and n a rational number. In particular, $r^n = (-1)^n \cos(n\varphi)$ defines the following shapes:

n :	-2	-1	1	2
Sinusoidal Spiral:	$x^2 - y^2 = 1$ Hyperbola	$\mathbb{R}_1$ Line '$x = 1$'	$\delta B_{(1/2,0)}(1/2)$ Circle	$(x^2 + y^2)^2 = x^2 - y^2$ Lemniscate of Bernoulli

The length of a sinusoidal spiral is $l = p\Gamma^2(1/2n)/\Gamma(1/n)$. In particular, starting with $\Gamma(1/2) = \sqrt{\pi}$,

$$\pi = 2\int_0^1 \frac{\mathrm{d}t}{\sqrt{1-t^2}} \approx 3.146, \qquad \text{and} \qquad \varpi = 2\int_0^1 \frac{\mathrm{d}t}{\sqrt{1-t^4}} = \frac{(\Gamma(1/4))^2}{2\sqrt{2\pi}} \approx 2.622,$$

the arc length of the lemniscate of Bernoulli turns out to be 2ϖ, since $p = 1/\sqrt{2}$. (The focal points of a lemniscate are $(\pm p, 0)$.) If $n = 1$, we have $c = p = 1$, which corresponds to a circle of radius 1/2. Therefore, for the unit circle, $p = 2$, and thus that circle's perimeter is 2π.

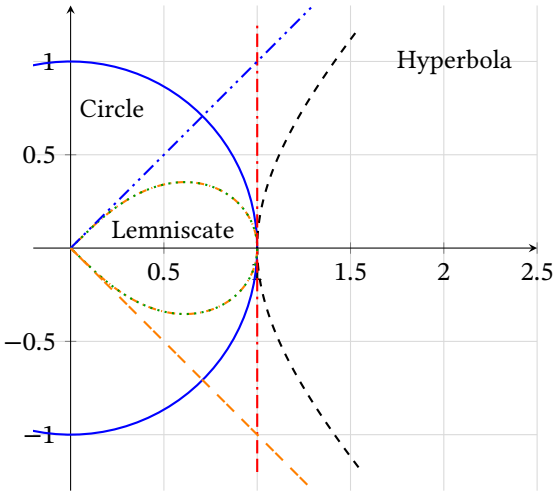

Figure 2.17: The circle $x^2 + y^2 = 1$, the hyperbola $x^2 - y^2 = 1$, Bernoulli's lemniscate $(x^2 + y^2)^2 = x^2 - y^2$, and the lines $y = x, y = -x, x = 1$ in the positive half plane.

Moreover, it is well-known that the inverse of a sinusoidal spiral with respect to the unit circle $\delta B_0(1)$ is another sinusoidal spiral whose value of n is the negative of the original curve's value of n. In particular, the inverse of the lemniscate of Bernoulli is the rectangular hyperbola.

Considering the wedge, every angle $\phi \in (-\pi/4, \pi/4)$ defines a line $y = mx = (\tan\phi)x$ that determines points of intersection with Bernoulli's lemniscate L_B, the boundary of the unit circle, the line '$x = 1$' and the rectangular hyperbola. If, due to symmetry, w.l.o.g. $\phi \in (0, \pi/4)$, these points are, respectively,

1. $(\cos(\phi)/r, \sin(\phi)/r)$, where $r = \sqrt{\frac{1+\tan^2(\phi)}{1-\tan^2(\phi)}}$ on L_B
2. $(\cos\phi, \sin\phi) = \left(\frac{1}{\sqrt{1+\tan^2(\phi)}}, \frac{\tan(\phi)}{\sqrt{1+\tan^2(\phi)}}\right)$ on δB_W
3. $(1, \tan\phi) = (1, \tanh\phi)$ on $\mathbb{R}_1$
4. $(\cosh\phi, \sinh\phi) = \left(\frac{1}{\sqrt{1-\tan^2(\phi)}}, \frac{\tan(\phi)}{\sqrt{1-\tan^2(\phi)}}\right)$ on H^+

More explicitly: $(\tan\phi)x = y = \sqrt{1-x^2}$ determines the intersection of the straight line through the origin and δB in the first quadrant. Elementary bookkeeping yields $x = 1/\sqrt{1+\tan^2(\phi)}$ and $y = \tan(\phi)/\sqrt{1+\tan^2(\phi)}$. In complete analogy, $(\tan\phi)x = y = \sqrt{x^2-1}$ determines the intersection of the same straight line and the hyperbola, which yields $x = 1/\sqrt{1-\tan^2(\phi)}$ and $y = \tan(\phi)/\sqrt{1-\tan^2(\phi)}$.

Since the lemniscate is the inverse of the hyperbola, $(x, y) = (r, \phi) \in H^+$ corresponds to $(x', y') = (1/r, \phi) \in L_B$. Now

$$r^2 = x^2 + y^2 = \frac{1+\tan^2(\phi)}{1-\tan^2(\phi)} \geq 1,$$

and thus

$$\left(\frac{1}{r}\right)^2 = (x')^2 + (y')^2 = \frac{1-\tan^2(\phi)}{1+\tan^2(\phi)} \leq 1$$

which has the consequence $\cos\phi = x'/(1/r) \Leftrightarrow x' = \cos(\phi)/r = \cos(\phi)\sqrt{\cos(2\phi)}$ and $y' = \sin(\phi)/r = \sin(\phi)\sqrt{\cos(2\phi)}$.

Moreover, the standard scale on δB_W is transformed to a scale on L_B:

Set	ϕ_{inf}	ϕ_{small}	ϕ_o	ϕ_{big}	ϕ_{sup}
δB_W	$-\pi/4$	$-\pi/8$	0	$\pi/8$	$\pi/4$
L_B	$(0,0)$	$\left(\frac{\cos\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}, \frac{-\sin\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}\right)$	$(1,0)$	$\left(\frac{\cos\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}, \frac{\sin\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}\right)$	$(0,0)$

More explicitly, the length of $L_B \cap \mathbb{C}^+$ is $\varpi \approx 2.6221$. In order to determine the position of $\left(\frac{\cos\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}, \frac{\sin\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}\right)$ on the lemniscate, one has to calculate $s(q) = \int_0^q \frac{dt}{\sqrt{1-t^4}} = F\left(\arcsin\left(\frac{1}{\sqrt[4]{2}}\right)\middle| -1\right) \approx 0.8956$, where $q = \sqrt{\frac{\sin^2\left(\frac{\pi}{8}\right)}{\sqrt{2}} + \frac{\cos^2\left(\frac{\pi}{8}\right)}{\sqrt{2}}} \approx 0.8409$.[7] Interpretation: $s = s(q)$ is the distance travelled on the lemniscate from the origin, and $q = sl(s)$ is the lemniscate sine (sinus lemniscatus) of s, i.e., the Euclidean distance of $\left(\frac{\cos\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}, \frac{\sin\left(\frac{\pi}{8}\right)}{\sqrt[4]{2}}\right)$ from the origin. Thus the above scale on L_B could also be written as $(-\varpi/2, -\varpi/2 + s, 0, \varpi/2 - s, \varpi/2)$, and the image of $U(-\pi/4, \pi/4)$ on δB_W is a distribution on L_B that puts half of its mass in the (relatively short) interval $(-\varpi/2 + s, \varpi/2 - s) \approx (-0.4154, 0.4154)$.

Of course, one could reverse the direction of the argument and start with the Uniform on the Lemniscate in the positive half plane $U(-\varpi/2, \varpi/2)$ and map this distribution to $H^+, S, (-1, 1) \subseteq \mathbb{R}_1$ and δB_W. It would also be natural to put half of the mass on the lemniscate to the right of its extrema in $\mathbb{C}^+$. The maximum and the minimum, respectively, occur at the points $(\sqrt{6}/4, \pm\sqrt{2}/4)$ that correspond to the angles $\phi = \pm 30° = \pm\pi/6$.

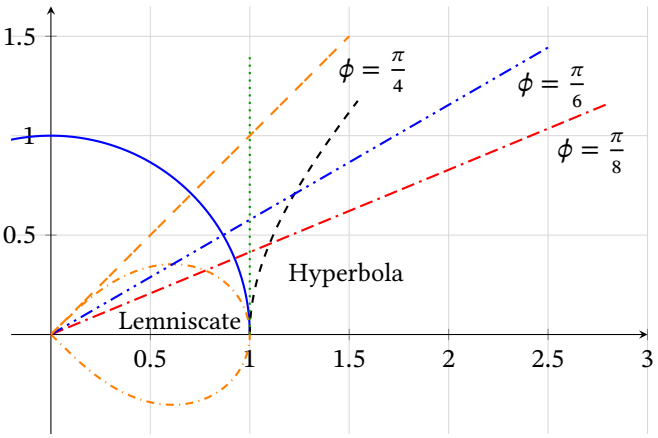

Figure 2.18: Bernoulli's lemniscate, the circle, the hyperbola, and the lines '$x = 1$', $y = \tan(\phi)x$, $(\phi = \pi/8, \pi/6, \pi/4)$, in the positive half plane.

In total generality, any distribution on a single sinusoidal spiral defines a family of distributions on the family of sinusoidal spirals (see Fig. 2.17). In particular, the most natural starting point seems to be the Uniform on δB_W or some other subset of δB. The mass is then transported along rays $y = ax$ that pass through the origin, and may also be reflected (e.g., from the hyperbola to the lemniscate, see Fig. 2.18).

[7] $F(\alpha|m) = \int_0^\alpha 1/\sqrt{1 - m\sin^2(\theta)}\ d\theta$ denotes an elliptic integral of the first kind, $-\pi/2 \le \alpha \le \pi/2$.

2.10 Logarithmic distributions

A whole book could be devoted to distributions generated by geometric arguments. In what follows, we will only highlight a few of them. To this end, it is very useful to consider the logarithm in more detail.

2.10.1 The logarithm's various roles

For a considerable number of reasons, the natural logarithm is omnipresent in transformation theory:

1. $\ln(x)$ is the antiderivative of $1/x$.
2. $\ln(\cdot)$ is the inverse mapping of $\exp(\cdot)$. In particular, e^{-x} maps $\mathbb{R}^+$ to the unit interval, and $-\ln(y)$ maps the unit interval on $[0, \infty)$.
3. The transformation $t(x) = -\ln(x) = \ln(1/x)$ maps $\mathbb{R}^+$ on $\mathbb{R}$.
4. The logarithm has a hyperbolic aspect to it. In particular $2\operatorname{artanh}(x) = \frac{\ln(1+x)}{\ln(1-x)}$ and
$$\begin{aligned}\ln x &= \operatorname{arcosh}\left(\frac{x^2+1}{2x}\right) = \operatorname{arsinh}\left(\frac{x^2-1}{2x}\right) = \operatorname{artanh}\left(\frac{x^2-1}{x^2+1}\right) \\ &= 2\operatorname{artanh}\left(\frac{x-1}{x+1}\right)\end{aligned}$$
5. The information $I(x) = -\ln|x|$ at the point x is the 'logarithmic distance' from the singularity at the origin.
6. Logarithmic moments are important, in particular $E_I(X) = \int(-\ln x)f(x)\,\mathrm{d}x$.

Moreover, logarithmically transformed random variables have nice properties:

2.10.2 Random variables with a ratio structure

Random variables having the form $Z = X/Y$ have particular properties:

Theorem 2.10.1. *Suppose $Z = X/Y$. Then we always have*

$$\begin{aligned}E_I(Z) &= E_I(X) - E_I(Y) \\ E(\ln|Z|) &= E(\ln|X|) - E(\ln|Y|)\end{aligned}$$

If $X \overset{d}{=} Y$, moreover $E_I(Z) = 0$, and $E(\ln|Z|) = 0$.

Proof. See Corollary 1.5.6 for $E_I(Z)$. For the expected value, note that $E(\ln|Z|) = E(\ln(|X/Y|)) = E(\ln(|X|/|Y|)) = E(\ln|X|) - E(\ln|Y|)$.

□

Theorem 2.10.2.

1. *Given iid r.v. $X_1, X_2 > 0$, then $\ln(X_1/X_2)$ is symmetric about the y-axis.*
2. *Given iid r.v. X, X_i such that $p(X = 0) = 0$, then $Z = \ln|X_1/X_2|$ is symmetric about the y-axis. Moreover,*

$$E_I(Z) \geq 1 - E[Z|Z > 0].$$

Proof. i) $Y_1 = \ln X_1$ and $Y_2 = -\ln X_2$ are real-valued random variables. Since $p(Y_1 = y) = p(Y_2 = -y)$, we have, for each pair $(y_1, -y_2)$

$$\begin{aligned} p(Y_1 = y_1, Y_2 = -y_2) &= p(Y_1 = y_1) \cdot p(Y_2 = -y_2) \\ &= p(Y_1 = y_2) \cdot p(Y_2 = -y_1) = p(Y_1 = y_2, Y_2 = -y_1). \end{aligned} \tag{2.15}$$

Let $Z = Y_1 + Y_2 = \ln X_1 - \ln X_2 = \ln(X_1/X_2)$, and thus, due to total probability, $p(Z = z)$ is the sum of all points $(y_1, -y_2)$, such that $y_1 + (-y_2) = y_1 - y_2 = z$. Owing to equation (2.15), this is perfectly counterbalanced by pairs $(y_2, -y_1)$ that add up to $p(Z = -z)$. Since $z \in \mathbb{R}$ is arbitrary, $\mathcal{D}(Z)$ is symmetric about the y-axis.

The same kind of reasoning holds for densities and integrals.

ii) The same argument as before may be applied to $Z = \ln|X_1/X_2| = \ln|X_1| - \ln|X_2|$.

For the inequality, note that

$$\begin{aligned} E_I(Z) &= \int_{-\infty}^{\infty} (-\ln|z|) f(z)\,\mathrm{d}z = 2\int_0^{\infty} (-\ln z) f(z)\,\mathrm{d}z \geq 2\int_0^{\infty} (1-z) f(z)\,\mathrm{d}z \\ &= 2\left(\int_0^{\infty} f(z)\,\mathrm{d}z - \int_0^{\infty} z f(z)\,\mathrm{d}z\right) = 1 - \int_0^{\infty} z \cdot (2f(z))\,\mathrm{d}z \\ &= 1 - E[Z|Z > 0]. \end{aligned}$$

□

Remark: If $X_i > 0$, we have $Z > 0 \Leftrightarrow X_1 > X_2$. This has the consequence that $E_I(Z) > 0$ if the ratio of iid random variables does not become too large.

If there is a classical statistical unimodal distribution with much mass at the centre, X_1 and X_2 are - typically - of about equal size. Thus X_1/X_2 does not deviate much from 1, and $|Z|$ is small. Therefore, $E[Z|Z > 0]$ is also close to zero. At least, it is far smaller than 1, and thus the inequality guarantees that $E_I(Z)$ is positive.

A bit more generally: although $E_I(X_1/X_2)$ is always precisely zero, if there is sufficient mass at the centre, the transform $\ln(X_1/X_2)$ creates information. In particular, the (Standard) Uniform, Normal and Exponential all yield information (see the next section). In the same vein, we should expect that $\exp(X)$ destroys information (see pp. 74 and 82 for examples).

The opposite are 'polarised' distributions with little mass at the centre. Consider, for instance, 'Bernoulli-like' X, X_i such that $p(X = 1) = p$, and $p(X = \varepsilon) = 1 - p$. Here,

$E[Z|Z > 0]$ may be arbitrarily large: If $X_1 = X_2 = 1$ or $X_1 = X_2 = \varepsilon$, $\ln(X_1/X_2) = 0$. However, $X_1 = 1$ and $X_2 = \varepsilon$ makes $Z = \ln(X_1/X_2) = \ln(1/\varepsilon)$ very large. Finally, if $X_1 = \varepsilon$ and $X_2 = 1$, we have $Z << 0$. This implies that $E[Z|Z > 0] >> 0$. In other words: Even if $X_1 = X_2$ occurs often, which implies that Z puts much mass at the origin, $|Z|$ assumes very large values, which has the consequence that $E_I(|Z|)$ is negative. The polarisation in X (on the unit interval) leads to a polarisation of $|Z|$ on $[0, \infty)$.

2.10.3 The Laplace, the Logistic, and the Sech distributions

Due to their common structure, $\ln(U_1/U_2)$, where $U_i \sim U(0,1)$, $\ln(Y_1/Y_2)$, where $Y_i \sim \text{Exp}(1)$, and $\ln(N_1/N_2)$, where $N_i \sim N(0,1)$ have similar properties. In particular, given independence, they all have symmetric distributions (see the next Figure 2.19).

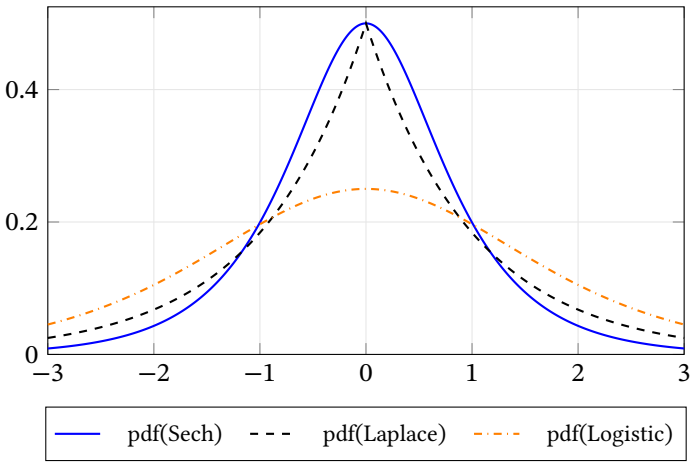

Figure 2.19: Standard Laplace, Logistic and Sech distributions.

If U_1, U_2 are independent, the distribution of U_1/U_2 is a 'mixture' of $U(0,1)$ and a Standard Pareto in the sense that $f(x) = 1/2$ for $x \leq 1$, and $f(x) = x^2/2$ if $x > 1$. We know from Corollary 1.5.6 that $E_I(U_1/U_2) = E_I(U_1) - E_I(U_2) = 0$, and Lemma 1.6.3 yields the same result.

In Section 2.1, Table 2.1, we also saw that a logarithmic transformation of the Uniform yielded the Standard classical Laplace distribution, Laplace(1), with pdf $e^{-|x|}/2$ for all $x \in \mathbb{R}$. It turns out (see, e.g., Kotz, Kozubowski and Podgórski (2001), pp. 23-24) that $\ln(U_1/U_2)$ also is Laplace(1), i.e., a symmetrised Exp(1) distribution. Thus $E_I(\ln(U_1/U_2)) = \gamma$.

Note, that the ratio of two independent Standard Pareto random variables P_i (for instance, $P_i = 1/U_i$) and the difference of two independent $Y_i \sim \text{Exp}(1)$ yields the same result, since

$$\ln(U_1/U_2) \stackrel{d}{=} (-(-\ln U_1) - \ln U_2) \stackrel{d}{=} \ln(P_2/P_1) \stackrel{d}{=} Y_2 - Y_1. \tag{2.16}$$

Starting with independent $Y, Y_i \sim \text{Exp}(1)$, first leads to a 'shifted (1) Pareto' (more commonly known as a (Standard) Lomax)[8], i.e., $Y_1/Y_2 \sim \text{Pareto}(1,1,0)$ with pdf $f(x) = 1/(1+x)^2$ for $x > 0$. Of course, $E_I(Y_1/Y_2) = 0$. Then, $X = \ln(Y_1/Y_2)$ has a Standard Logistic distribution $L(0,1)$:

Name	Domain	$f(x)$	$y = F(x)$	$x = F^{-1}(y)$
$U, U_i \sim U(0,1)$	$(0,1)$	1	x	y
$Y, Y_i \sim \text{Exp}(1)$	$\mathbb{R}^+$	e^{-x}	$1 - e^{-x}$	$-\ln(1-y)$
$L(0,1)$	$\mathbb{R}$	$\frac{e^{-x}}{(e^{-x}+1)^2}$	$\frac{1}{e^{-x}+1}$	$-\ln(1/y - 1)$

The Logistic can also be constructed with the help of a single uniformly or exponentially distributed random variable: Since $U/(1-U) \sim \text{Pareto}(1,1,0)$, consistently $\ln U - \ln(1-U) = \ln(U/(1-U)) \sim L(0,1)$, and thus also $\ln(e^{-Y}/(1-e^{-Y})) \sim L(0,1)$. For the expected information, one gets $E_I(L(0,1)) = \gamma + \ln 2 - \ln \pi \approx 0.1256$, and $\sigma_I^2(L(0,1)) \approx 1.382$ for the information variance.

Finally, $\ln|N_1/N_2| = \ln|C|$, where C is a Standard Cauchy, begets a hyperbolic secant distribution, $\text{Sech}(0, \pi/2)$. Therefore, since $\text{sech}(x) = 2/(e^x + e^{-x})$, we obtain

$$g(x) = \frac{\text{sech}\left(\frac{\pi x}{2s}\right)}{2s} = \frac{1}{2s\cosh\left(\frac{\pi x}{2s}\right)} = \frac{1}{s\left(\exp(\frac{\pi x}{2s}) + \exp(-\frac{\pi x}{2s})\right)}$$

as the density of $\text{Sech}(0, s)$. For $X = \frac{2}{\pi}\ln|C| \sim \text{Sech}(0,1)$ we get the expected information $E_I(X) = 4\ln\Gamma(1/4) - 3\ln 2 - 2\ln\pi \approx 0.7832$, and

Name	Domain	$f(x)$	$y = F(x)$	$x = F^{-1}(y)$
$C \sim C(0,1)$	$\boldsymbol{R}$	$\frac{1}{\pi(1+x^2)}$	$\frac{1}{2} + \frac{\arctan(x)}{\pi}$	$\tan(\pi(y - 1/2))$
$\text{Sech}(0,1)$	$\boldsymbol{R}$	$\frac{1}{2}\text{sech}\left(\frac{\pi}{2}x\right)$	$\frac{2}{\pi}\arctan\left(\exp(\frac{\pi}{2}x)\right)$	$\frac{2}{\pi}\ln(\tan(\frac{\pi}{2}y))$

Both distributions are closely related. First, since the pdf of $L(0,1)$ may be rewritten,

$$f(x) = \frac{e^{-x}}{(e^{-x}+1)^2} = \frac{1}{4}(\text{sech}(x/2))^2,$$

the Logistic is sometimes called 'Sech Squared Distribution'. Second, considering the density of $Y = X^2$, one obtains

$$h(y) = \frac{\text{sech}\left(\frac{\pi\sqrt{y}}{2}\right)}{2\sqrt{y}} \qquad \text{for all} \qquad y > 0.$$

[8] This line of thought will be extended later (see p. 101).

Third, with the Stieltjes constants $\gamma_1(1/4)$ und $\gamma_1(3/4)$ (see Appendix 7.1.4 and p. 462), compute

$$\begin{aligned} 2E_I(X) &= E_I(Y) = \frac{2\left(\gamma_1\left(\frac{1}{4}\right) - \gamma_1\left(\frac{3}{4}\right) + \gamma\pi + \pi\ln(2\pi)\right)}{\pi} \\ &= 8\ln\Gamma(1/4) - 6\ln 2 - 4\ln\pi \approx 1.5663776 \end{aligned}$$

Since $\Gamma(5/4) = \Gamma(1/4)/4$ and $(\Gamma(1/4))^2 = 2g\sqrt{2\pi^3}$, with the Gaussian constant

$$g = \frac{\varpi}{\pi} = \frac{2}{\pi}\int_0^1 \frac{\mathrm{d}x}{\sqrt{1-x^4}} = \frac{B(1/4,1/2)}{2\pi} = \int_0^\infty \frac{\mathrm{d}x}{\sqrt{\cosh(\pi x)}} \approx 0.8346 \approx \frac{2.622}{3.142},$$

we finally obtain

$$E_I(Y) = \ln\left(\frac{1024\left(\Gamma\left(\frac{5}{4}\right)\right)^8}{\pi^4}\right) = \ln(g^4\pi^2) = \ln(\varpi^4/\pi^2) = 4\ln\varpi - 2\ln\pi \quad (2.17)$$

and $E_I(X) = 2\ln g + \ln\pi = 2\ln\varpi - \ln\pi = -\ln(\pi/\varpi^2) \approx 0.783$.

Since sech$(\cdot)$ is a hyperbolic function, this result underlines that the hyperbola lies 'between' the circle and the lemniscate (see Section 2.9.5). The close connection between the circle, the line and the hyperbola also becomes obvious in various versions of Vardi's integral (Espinosa and Moll 2002; Vardi 1988; Weisstein 2000a). We refer to Medina and Moll (2009), p. 92, where a part of the Uniform, the Cauchy and the Sech are linked without much ado:

$$\begin{aligned} \frac{2}{\pi}\int_{\pi/4}^{\pi/2} \ln(\ln\tan x)\,\mathrm{d}x &= \frac{2}{\pi}\int_0^1 \frac{\ln(-\ln y)}{y^2+1}\,\mathrm{d}y \\ &= \frac{1}{\pi}\int_0^\infty \frac{\ln z}{\cosh z}\,\mathrm{d}z = \ln\left(\frac{\sqrt{2\pi}\Gamma(3/4)}{\Gamma(1/4)}\right) \end{aligned}$$

Note that the step from the circle to $\mathbb{R}_1$ erases the tangent, and the step from there to S (the hyperbola) deletes one of the logarithms.

Although the Logistic is considerably more popular than the Sech, using Stieltjes constants of higher order, the latter's higher logarithmic moments (and thus, in particular, the information variance) can be calculated explicitly. For instance,

$$\begin{aligned} M_I^2(X) &= \frac{\gamma_2\left(\frac{1}{4}\right) - \gamma_2\left(\frac{3}{4}\right)}{\pi} + \frac{2\left(\gamma_1\left(\frac{1}{4}\right) - \gamma_1\left(\frac{3}{4}\right)\right)(\gamma + \ln(2\pi))}{\pi} \\ &\quad + \gamma^2 + \frac{\pi^2}{6} + \ln^2(2) + 2\gamma\ln(2\pi) + \ln(\pi)\ln(4\pi) \approx 2.00 \end{aligned}$$

Due to the above interpretation, one should be able to simplify this expression considerably. Although hardly anything is known about Stieltjes constants of higher order, the most explicit expression seems to be due to Coffey (2011), namely

$$\begin{aligned}\gamma_k(a) &= -\frac{1}{k+1}\ln^k(a+1)+\frac{\ln^k a}{a}\\ &\quad+\sum_{j=1}^{\infty}\int_0^1\left(\frac{\ln^k(j+a)}{j+a}-\frac{\ln^k(x+j+a)}{x+j+a}\right)\mathrm{d}x,\end{aligned}$$

this guess turns out to be correct, since after some bookkeeping

$$\sigma_I^2(X)=M_I^2(X)-(E_I(X))^2=\frac{\pi^2}{6}+\frac{\gamma_2\left(\frac{1}{4}\right)-\gamma_2\left(\frac{3}{4}\right)}{\pi}-\left(\frac{\gamma_1\left(\frac{1}{4}\right)-\gamma_1\left(\frac{3}{4}\right)}{\pi}\right)^2\approx 1.388$$

Altogether, we thus have

Distribution	E_I	σ_I^2
L(0,1)	$\gamma+\ln(2/\pi)\approx 0.126$	1.382
Exp(1)	$\gamma\approx 0.577$	$\pi^2/6\approx 1.645$
N(0,1)	$(\gamma+\ln 2)/2\approx 0.635$	$\pi^2/8\approx 1.234$
Sech(0,1)	$\ln(\varpi^2/\pi)\approx 0.783$	1.388

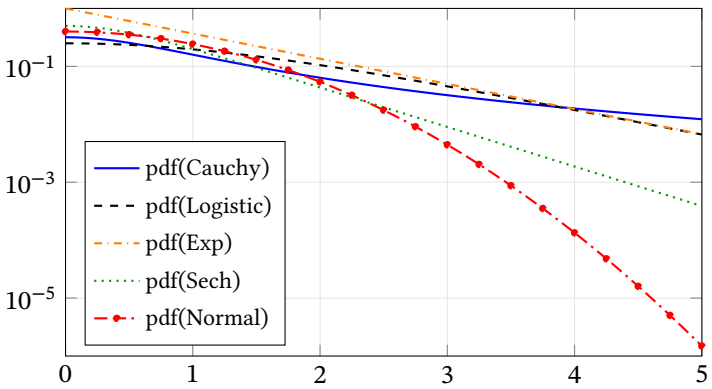

Figure 2.20: Tail behaviour of popular distributions.

Classically, one just looks at the tail behaviour, which becomes transparent in a logarithmic plot (see Fig. 2.20). Notice that in this respect, the Exponential and the Logistic are almost indistinguishable, and the Logistic is also rather similar to the Sech. Both have so-called 'semi-heavy' tails (cf. M. Fischer (2013), pp. 5f). However, E_I also considers the behaviour close to the origin, which is considerably different (see Fig. 2.21).

Finally, it may be mentioned that, although defined quite differently, an exponential

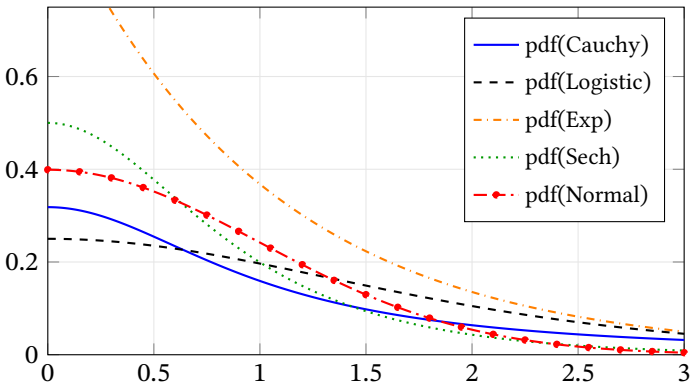

Figure 2.21: The distributions of Fig. 2.20, behaviour next to the origin.

transform of the Standard Normal, the Logistic and the Sech destroys the expected information. That is, owing to Lemma 1.7.1 and symmetry, the expected information of the following distributions, all defined on $\mathbb{R}^+$, is zero (see Fig. 2.22).

Name	Transformation	pdf
'Lognormal'	$\exp(N(0,1))$	$e^{-\frac{1}{2}\ln^2(x)}/(\sqrt{2\pi}x)$
'Loglogistic'	$\exp(\text{Logistic}(0,1/a)), a>0$	$ax^{a-1}/\left(x^a+1\right)^2$
'Logsech'	$\exp(\text{Sech}(0,1))$	$\text{sech}\left(\frac{\pi}{2}\ln x\right)/(2x)$

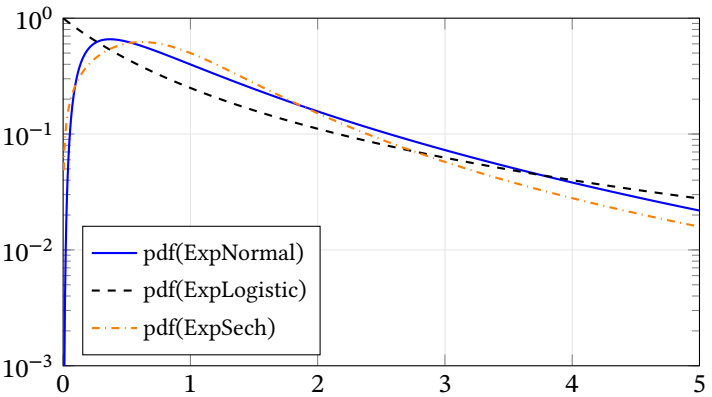

Figure 2.22: Exponentially transformed (Standard) Normal, Logistic and Sech distributions (logarithmic plot).

In Section 7.6.7 we will consider the Normal, the Sech and the Logistic one more time and show that they are indeed very closely related (p. 459).

2.11 The (closed) circle and π

A well known paper by Wigner (1960) starts with (his emphasis):

> THERE IS A story about two friends, who were classmates in high school, talking about their jobs. One of them became a statistician and was working on population trends. He showed a reprint to his former classmate. The reprint started, as usual, with the Gaussian distribution and the statistician explained to his former classmate the meaning of the symbols for the actual population, for the average population, and so on. His classmate was a bit incredulous and was not quite sure whether the statistician was pulling his leg. "How can you know that?" was his query. "And what is this symbol here?" "Oh," said the statistician, "this is π." "What is that?" "The ratio of the circumference of the circle to its diameter." "Well, now you are pushing your joke too far," said the classmate, "surely the **population has nothing to do with the circumference** of the circle."

Actually, it has. The crucial part of a Normal's pdf is the term $e^{-x^2/2}$, or just e^{-x^2}. In two dimensions, one thus gets

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,dr d\varphi = \pi\int_0^{\infty} e^{-u}\,du = \pi$$

Since $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y = \left(\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x\right)\cdot\left(\int_{-\infty}^{\infty} e^{-y^2}\,\mathrm{d}y\right) = \left(\int_{-\infty}^{\infty} e^{-t^2}\,\mathrm{d}t\right)^2$, we have

$$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi} \qquad \text{and} \qquad \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}.$$

However, there is nothing special about the Normal. That is, the number π is also closely related to the Cauchy and the Sech, since, on the one hand

$$\int_{-\infty}^{\infty} \operatorname{sech}(x)\;\mathrm{d}x = \int_{-\infty}^{\infty} 1/\cosh(x)\,\mathrm{d}x = \pi,$$

and on the other hand,

$$\int_{-\infty}^{\infty} 1/(1+x^2)\,\mathrm{d}x = \pi.$$

In a certain sense, the above transformation theory is based on the circle, and thus π and its representations. The latter are interpreted in a stochastic manner (distributions) and mapped to other geometrically interesting sets. Alternatively, it is straightforward to study trigonometric and linear transformations, but also transformations involving the exponential function and the logarithm. Those connections become transparent upon considering logarithmic moments.

Decomposing π into four equal parts, i.e.,

$$\pi = \int_{-\infty}^{\infty} \frac{1}{1+t^2}\,\mathrm{d}t = 4\int_{0}^{1} \frac{1}{1+t^2}\,\mathrm{d}t = 4\int_{1}^{\infty} \frac{1}{1+t^2}\,\mathrm{d}t$$

gives a basic property of the Cauchy, and it is straightforward to connect the latter integral to the series representation of the arctan:

If $|x| < 1$,

$$\arctan(x) = \int_0^x \frac{1}{1+t^2}\,\mathrm{d}t = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1}$$

The latter series also converges if $x = 1$, and one gets Leibniz' series: $\pi/4 = 1 - 1/3 + 1/5 - 1/7 + \ldots$ Here is an elementary argument to the same end: $\sin(45°) = \cos(45°)$, therefore $\tan(\pi/4) = 1$ and $\pi/4 = \arctan(1)$.

Of course, further representations of π could be investigated. For instance

$$\int_{-\infty}^{\infty} 1/(\sqrt{x(1-x)})\; dx = \pi,$$

which is related to sinh, the Arcsin distribution, (see p. 53), and the sine cardinalis, i.e., $si(x) = \frac{\sin x}{x}$, which has the property $\int_{-\infty}^{\infty} si(x)dx = \pi$, see Section 7.3.7. It seems to be quite typical that those results are discussed without any reference to probability theory.

Finally, note that even a factor such as $2/\pi$ that shows up again and again (e.g., upon going from $C(0,1)$ to $U(0,1)$) is not as harmless as it might first appear. There is Vieta's product,

$$\frac{2}{\pi} = \cos\frac{\pi}{4} \cdot \cos\frac{\pi}{8} \cdot \cos\frac{\pi}{16} \cdots$$

which is a consequence of the 'half-angle formula' $\sin z = 2\sin\frac{z}{2}\cos\frac{z}{2}$ of the sine (Olver et al. (2010), p. 117, formula 4.21.6), or just equation (7.36), p. 399, for the sine cardinalis with $z = \pi/2$.

Sliced Circle

Since the circle is an elementary geometric object, it is possible to generalise it in various directions. Above, we went from the circle to the hyperbola and the lemniscate, which, as already mentioned, may be summarised under the topic of (one-parameter) "sinusoidal spirals".

It may also be noted that the conform mapping $g(z) = 1/z$ demonstrates that the hyperbola and the lemniscate are related in exactly the same way, as circles are. That is, if the circle $\delta B_r(c)$ does not contain the origin, then $g(\delta B_r(c))$ is also a circle. Likewise, if $y = a/x$, where $a > 0$, is a hyperbola in the (x, y) plane, then $(u, v) = g(z) = g(x, y)$ is defined by $u = x/(x^2 + y^2)$ and $v = -y/(x^2 + y^2)$. Putting $y = a/x$ in the latter equations leads to

$$a(u^2 + v^2)^2 = -uv,$$

which defines a lemnicate in the (u, v) plane.

A rather brute method leads to another unimodal density: If one takes the circle $B_{1/2}((0, 1/2))$ in the plane, cuts it at the lowest point (i.e., in the origin) and turns the lower arcs outside, one gets:

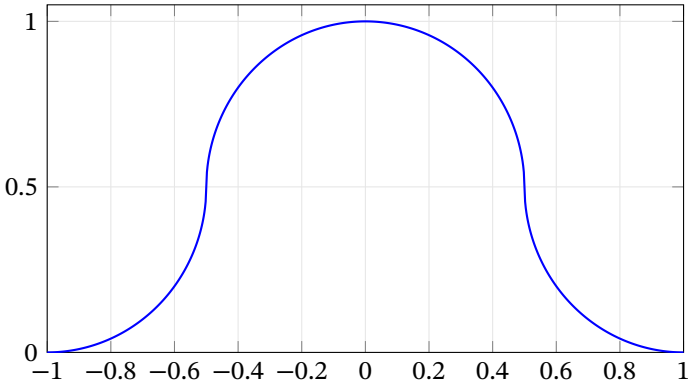

Figure 2.23: The pdf of the 'Sliced Circle' distribution.

Formally, the density is defined by (see Fig. 2.23)

$$f(x) = \begin{cases} 1/2 - \sqrt{1/4 - (x+1)^2} & \text{if} \quad -1 \le x \le -1/2 \\ 1/2 + \sqrt{1/4 - x^2} & \text{if} \quad -1/2 < x \le 1/2 \\ 1/2 - \sqrt{1/4 - (x-1)^2} & \text{if} \quad 1/2 < x \le 1 \end{cases}$$

Note that $\int_0^{1/2} 1/2 + \sqrt{1/4 - x^2}\,\mathrm{d}x = (4+\pi)/16$ and $\int_{1/2}^1 1/2 - \sqrt{1/4 - (x-1)^2}\,\mathrm{d}x = (4-\pi)/16$. Moreover, $\int_0^{1/2} (-\ln x)(1/2 + \sqrt{1/4 - x^2})\,\mathrm{d}x = 1/4 + \pi/32 + (\pi \ln 2)/8 + (\ln 2)/4 \approx 0.794$. Since every realisation adds a positive amount of information, this distribution's expected information is rather high,

$$E_I \quad = \quad 2\left(\int_0^1 (-\ln x) f(x) dx\right) = 2\,(0.794 + 0.027) \approx 1.641$$

Similarly, since densities have no atoms, the transformation $1/X$ cuts along the y axis, and exchanges centre and periphery on every side separately.