Information as Distinctions: New Foundations for Information Theory

David Ellerman
University of California/Riverside

January 23, 2013

Abstract

The logical basis for information theory is the newly developed logic of partitions that is dual to the usual Boolean logic of subsets. The key concept is a "distinction" of a partition, an ordered pair of elements in distinct blocks of the partition. The logical concept of entropy based on partition logic is the normalized counting measure of the set of distinctions of a partition on a finite set–just as the usual logical notion of probability based on the Boolean logic of subsets is the normalized counting measure of the subsets (events). Thus logical entropy is a measure on the set of ordered pairs, and all the compound notions of entropy (join entropy, conditional entropy, and mutual information) arise in the usual way from the measure (e.g., the inclusion-exclusion principle)–just like the corresponding notions of probability. The usual Shannon entropy of a partition is developed by replacing the normalized count of distinctions (dits) by the average number of binary partitions (bits) necessary to make all the distinctions of the partition.

Contents

1 Introduction

Information is about making distinctions or differences. In James Gleick's book, *The Information: A History, A Theory, A Flood*, he noted the focus on differences in the seventeenth century polymath, John Wilkins, who was a founder of the Royal Society. In 1641, the year before Newton was born, Wilkins published one of the earliest books on cryptography, *Mercury or the Secret and Swift Messenger*, which not only pointed out the fundamental role of differences but noted that any (finite) set of different things could be encoded by words in a binary code.

> For in the general we must note, That whatever is capable of a competent Difference, perceptible to any Sense, may be a sufficient Means whereby to express the Cogitations. It is more convenient, indeed, that these Differences should be of as great Variety as the Letters of the Alphabet; but it is sufficient if they be but twofold, because Two alone may, with somewhat more Labour and Time, be well enough contrived to express all the rest. [30, Chap. XVII, p. 69]

Wilkins explains that a five letter binary code would be sufficient to code the letters of the alphabet since $2^5 = 32$.

> Thus any two Letters or Numbers, suppose A.B. being transposed through five Places, will yield Thirty Two Differences, and so consequently will superabundantly serve for the Four and twenty Letters... .[30, Chap. XVII, p. 69]

As Gleick noted:

> Any difference meant a binary choice. Any binary choice began the expressing of cogitations. Here, in this arcane and anonymous treatise of 1641, the essential idea of information theory poked to the surface of human thought, saw its shadow, and disappeared again for [three] hundred years. [12, p. 161]

In this paper, we will start afresh by deriving an information-as-distinctions notion of logical entropy [7] from the new logic of partitions [8] that is mathematically dual to the usual Boolean logic of subsets. Then the usual Shannon entropy [27] will be essentially derived from the concepts behind logical entropy as another way to measure information-as-distinctions. This treatment of the various notions of Shannon entropy (e.g., mutual, conditional, and joint entropy) will also explain why their interrelations can be represented using a Venn diagram picture [5].

2 Logical Entropy

2.1 Partition logic

The logic normally called "propositional logic" is a special case of the logic of subsets originally developed by George Boole [4]. In the Boolean logic of subsets of a fixed non-empty universe set U, the variables in formulas refer to subsets $S \subseteq U$ and the logical operations such as the join $S \vee T$, meet $S \wedge T$, and implication $S \Rightarrow T$ are interpreted as the subset operations of union $S \cup T$, intersection $S \cap T$, and the conditional $S \Rightarrow T = S^c \cup T$. Then "propositional" logic is the special case where $U = 1$ is the one-element set whose subsets $\emptyset$ and 1 are interpreted as the truth values 0 and 1 (or false and true) for propositions.

In subset logic, a *valid formula* or *tautology* is a formula such as $[S \wedge (S \Rightarrow T)] \Rightarrow T$ where for any non-empty U, no matter what subsets of U are substituted for the variables, the whole formula evaluates to U by the subset operations. It is a theorem that if a formula is valid just for the special case of $U = 1$ (i.e., as in a truth table tautology), then it is valid for any U. But in today's textbook treatments of so-called "propositional" logic, the truth-table version of a tautology is usually given as a definition, not as a theorem in subset logic.

What is lost by restricting attention to the special case of propositional logic rather than the general case of subset logic? At least two things are lost, and both are relevant for our development.

- Firstly if it is developed as the logic of subsets, then it is natural, as Boole did, to attach a quantitative measure to each subset S of a finite universe U, namely the normalized counting measure $\frac{|S|}{|U|}$ which can be interpreted as the *logical probability* $\Pr(S)$ (where the elements of U are assumed equiprobable) of randomly drawing an element from S.
- Secondly, the notion of a subset (unlike the notion of a proposition) has a mathematical dual in the notion of a quotient set, as is evidenced by the dual interplay between subobjects (subgroups, subrings,...) and quotient objects throughout abstract algebra.

This duality is the "turn-around-the-arrows" category-theoretic duality, e.g., between monomorphisms and epimorphisms, applied to sets [20]. The notion of a quotient set of U is equivalent to the notion of an equivalence relation on U or a partition $\pi = \{B\}$ of U. When Boole's logic is seen as the logic of subsets (rather than propositions), then the notion arises of a dual logic of partitions which has now been developed [8].

2.2 Logical Entropy

A partition $\pi = \{B\}$ on a finite set U is a set of non-empty disjoint subsets B ("blocks" of the partition) of U whose union is U. The idea of information-as-distinctions is made precise by defining a *distinction* or *dit of a partition* $\pi = \{B\}$ of U as an ordered pair (u, u') of elements $u, u' \in U$ that are in different blocks of the partition. The notion of "a distinction of a partition" plays the analogous role in partition logic as the notion of "an element of a subset" in subset logic. The set of distinctions of a partition π is its *dit set* $\operatorname{dit}(\pi)$. The subsets of U are partially ordered by inclusion with the universe set U as the top of the order and the empty set $\emptyset$ as the bottom of the order. A partition $\pi = \{B\}$ *refines* a partition $\sigma = \{C\}$, written $\sigma \preceq \pi$, if each block $B \in \pi$ is contained in some block $C \in \sigma$. The partitions of U are partially ordered by refinement which is equivalent to the inclusion ordering of dit sets. The discrete partition $\mathbf{1} = \{\{u\}\}_{u \in U}$, where the blocks are all the singletons, is the top of the order, and the indiscrete partition $\mathbf{0} = \{U\}$ (with just one block U) is the bottom. Only the self-pairs $(u, u) \in \Delta \subseteq U \times U$ of the diagonal Δ can never be a distinction. All the possible distinctions $U \times U - \Delta$ are the dits of $\mathbf{1}$ and no dits are distinctions of $\mathbf{0}$ just as all the elements are in U and none in $\emptyset$.

In this manner, we can construct a table of analogies between subset logic and partition logic.

	Subset logic	Partition logic
'Elements'	Elements u of S	Dits (u, u') of π
Order	Inclusion $S \subseteq T$	Refinement: $\operatorname{dit}(\sigma) \subseteq \operatorname{dit}(\pi)$
Top of order	U all elements	$\operatorname{dit}(\mathbf{1}) = U^2 - \Delta$, all dits
Bottom of order	$\emptyset$ no elements	$\operatorname{dit}(\mathbf{0}) = \emptyset$, no dits
Variables in formulas	Subsets S of U	Partitions π on U
Operations	Subset ops.	Partition ops. [8]
Formula $\Phi(x, y, ...)$ holds	u element of $\Phi(S, T, ...)$	(u, u') dit of $\Phi(\pi, \sigma, ...)$
Valid formula	$\Phi(S, T, ...) = U$, $\forall S, T, ...$	$\Phi(\pi, \sigma, ...) = \mathbf{1}$, $\forall \pi, \sigma, ...$

Table of analogies between subset and partition logics

A dit set $\operatorname{dit}(\pi)$ of a partition on U is a subset of $U \times U$ of a particular kind, namely the complement of an equivalence relation. An *equivalence relation* is reflexive, symmetric, and transitive. Hence the complement is a subset $P \subseteq U \times U$ that is:

1. irreflexive (or anti-reflexive), $P \cap \Delta = \emptyset$;
2. symmetric, $(u, u') \in P$ implies $(u', u) \in P$; and
3. anti-transitive (or co-transitive), if $(u, u'') \in P$ then for any $u' \in U$, $(u, u') \in P$ or $(u', u'') \in P$,

and such binary relations will be called *partition relations* (also called *apartness relations*).

Given any subset $S \subseteq U \times U$, the *reflexive-symmetric-transitive (rst) closure* $\overline{S^c}$ of the complement S^c is the smallest equivalence relation containing S^c, so its complement is the largest partition relation contained in S, which is called the *interior* $\operatorname{int}(S)$ of S. This usage is consistent with calling the subsets that equal their rst-closures *closed subsets* of $U \times U$ (so closed subsets = equivalence relations) so the complements are the *open subsets* (= partition relations). However it should be noted that the rst-closure is not a topological closure since the closure of a union is not necessarily the union of the closures, so the "open" subsets do not form a topology on $U \times U$.

The interior operation $\operatorname{int} : \wp(U \times U) \to \wp(U \times U)$ provides a universal way to define operations on partitions from the corresponding subset operations:

> apply the subset operation to the dit sets and then, if necessary, take the interior to obtain the dit set of the partition operation.

Given partitions $\pi = \{B\}$ and $\sigma = \{C\}$ on U, their *join* $\pi \vee \sigma$ is the partition whose dit set $\operatorname{dit}(\pi \vee \sigma)$ is the interior of $\operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)$ (since the union $\cup$ is the subset join operation). But the union of partition relations (open subsets) is a partition relation (open subset) so that:

$$\operatorname{dit}(\pi \vee \sigma) = \operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma).$$

This gives the same join $\pi \vee \sigma$ as the usual definition which is the partition whose blocks are the non-empty intersections $B \cap C$ for $B \in \pi$ and $C \in \sigma$. To define the *meet* $\pi \wedge \sigma$ of the two partitions, we apply the subset meet operation of intersection to the dit sets and then take the interior (which is necessary in this case):

$$\operatorname{dit}(\pi \wedge \sigma) = \operatorname{int}[\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)].$$

This gives the same result as the usual definition of the partition meet in the literature.[1] Perhaps surprisingly, the other logical operations such as the implication do not seem to be defined for partitions in the literature. Since the subset operation of implication is $S \Rightarrow T = S^c \cup T$, we define the *partition implication* $\sigma \Rightarrow \pi$ as the partition whose dit set is:

$$\operatorname{dit}(\pi \Rightarrow \sigma) = \mathit{int}\,[\operatorname{dit}(\sigma)^c \cup \operatorname{dit}(\pi)].^2$$

The refinement partial order $\sigma \preceq \pi$ is just inclusion of dit sets, i.e., $\sigma \preceq \pi$ iff $\operatorname{dit}(\sigma) \subseteq \operatorname{dit}(\pi)$. If we denote the lattice of partitions (using the refinement ordering) as $\Pi(U)$, then the mapping:

$$\operatorname{dit} : \Pi(U) \to \wp(U \times U)$$

Dit set representation of partition lattice

represents the lattice of partitions as the lattice $\mathcal{O}(U \times U)$ of open subsets (under inclusion) of $\wp(U \times U)$.

For any finite set X, a (finite) *measure* μ is a function $\mu : \wp(X) \to \mathbb{R}$ such that:

1. $\mu(\emptyset) = 0$,
2. for any $E \subseteq X$, $\mu(E) \geq 0$, and
3. for any disjoint subsets E_1 and E_2, $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2)$.

[1] But note that many authors think in terms of equivalence relations instead of partition relations and thus reverse the definitions of the join and meet. Hence their "lattice of partitions" is really the lattice of equivalence relations, the opposite of the partition lattice $\Pi(U)$ defined here with refinement as the ordering relation.

[2] The equivalent but more perspicuous definition of $\sigma \Rightarrow \pi$ is the partition that is like π except that whenever a block $B \in \pi$ is contained in a block $C \in \sigma$, then B is 'discretized' in the sense of being replaced by all the singletons $\{u\}$ for $u \in B$. Then it is immediate that the refinement $\sigma \preceq \pi$ holds iff $\sigma \Rightarrow \pi = \mathbf{1}$, as we would expect from the corresponding relation, $S \subseteq T$ iff $S \Rightarrow T = S^c \cup T = U$, in subset logic.

Any finite set X has the *counting measure* $|\ | : \wp(X) \rightarrow \mathbb{R}$ and *normalized counting measure* $\frac{|\ |}{|X|} : \wp(X) \rightarrow \mathbb{R}$ defined on the subsets of X. Hence for finite U, we have the counting measure $|\ |$ and the normalized counting measure $\frac{|\ |}{|U \times U|}$ defined on $\wp(U \times U)$. Boole used the normalized counting measure $\frac{|\ |}{|U|}$ defined on the power-set Boolean algebra $\wp(U)$ to define the logical probability $\Pr(S) = \frac{|S|}{|U|}$ of an event $S \subseteq U$.[4] In view of the analogy between elements in subset logic and dits in partition logic, the construction analogous to the logical probability is the normalized counting measure applied to dit sets. That is the definition of the:

$$h(\pi) = \frac{|\text{dit}(\pi)|}{|U \times U|}$$

Logical entropy of a partition π.

Thus the logical entropy function $h()$ is the dit set representation composed with the normalized counting measure:

$$h : \Pi(U) \rightarrow \mathbb{R} = \Pi(U) \xrightarrow{\text{dit}} \wp(U \times U) \xrightarrow{\frac{|\ |}{|U \times U|}} \mathbb{R}.$$

Logical entropy function

One immediate consequence is the inclusion-exclusion principle:

$$\frac{|\text{dit}(\pi) \cap \text{dit}(\sigma)|}{|U \times U|} = \frac{|\text{dit}(\pi)|}{|U \times U|} + \frac{|\text{dit}(\sigma)|}{|U \times U|} - \frac{|\text{dit}(\pi) \cup \text{dit}(\sigma)|}{|U \times U|} = h(\pi) + h(\sigma) - h(\pi \vee \sigma)$$

which provides the motivation for our definition below of $\frac{|\text{dit}(\pi) \cap \text{dit}(\sigma)|}{|U \times U|}$ as the "logical mutual information" of the partitions π and σ.

In a random (i.e., equiprobable) drawing of an element from U, the event S occurs with the probability $\Pr(S)$. If we take two independent (i.e., with replacement) random drawings from U, i.e., pick a random ordered pair from $U \times U$, then $h(\pi)$ is the probability that the pair is a distinction of π, i.e., that π distinguishes. These analogies are summarized in the following table which uses the language of probability theory (e.g., set of outcomes, events, the occurrence of an event):

	Subset logic	Partition logic
'Outcomes'	Elements u of S	Ordered pairs $(u, u') \in U \times U$
'Events'	Subsets S of U	Partitions π of U
'Event occurs'	$u \in S$	$(u, u') \in \text{dit}(\pi)$
Norm. counting measure	$\Pr(S) = \frac{\lvert S\rvert}{\lvert U\rvert}$	$h(\pi) = \frac{\lvert\text{dit}(\pi)\rvert}{\lvert U \times U\rvert}$
Interpretation	Prob. event S occurs	Prob. partition π distinguishes

Table of quantitative analogies between subset and partition logics.

Thus logical entropy $h(\pi)$ is the simple quantitative measure of the distinctions of a partition π just as the logical probability $\Pr(S)$ is the quantitative measure of the elements in a subset S. In short, information theory is to partition logic as probability theory is to ordinary subset logic.

To generalize logical entropy from partitions to finite probability distributions, note that:

$$\operatorname{dit}(\pi) = \{B \times B' : B, B' \in \pi, B \neq B'\} = U \times U - \{B \times B : B \in \pi\}.$$

Using $p_B = \frac{|B|}{|U|}$, we have:

$$h(\pi) = \frac{|\operatorname{dit}(\pi)|}{|U \times U|} = \frac{|U|^2 - \sum_{B\in\pi}|B|^2}{|U|^2} = 1 - \sum_{B\in\pi}\left(\frac{|B|}{|U|}\right)^2 = 1 - \sum_{B\in\pi} p_B^2.$$

An ordered pair $(u, u') \in B \times B$ for some $B \in \pi$ is an *indistinction* or *indit* of π where $\operatorname{indit}(\pi) = U \times U - \operatorname{dit}(\pi)$. Hence in a random drawing of a pair from $U \times U$, $\sum_{B\in\pi} p_B^2$ is the probability of drawing an indistinction, while $h(\pi) = 1 - \sum_{B\in\pi} p_B^2$ is the probability of drawing a distinction.

Entropies will be defined both for partitions on finite sets and for finite probability distributions (i.e., finite random variables). Given a random variable u with the probability distribution $p = (p_1, ..., p_n)$ over the n distinct values $U = \{u_1, ..., u_n\}$, a distinction of the discrete partition on U is just a pair (u_i, u_j) with $i \neq j$ and with the probability $p_i p_j$. Applying the previous notion to the logical entropy of a partition to this case with $p_B = p_i$ (where $B = \{u_i\}$), we have the:

$$h(p) = 1 - \sum_i p_i^2 = \sum_i p_i (1 - p_i)$$

Logical entropy of a finite probability distribution p.[3]

Since $1 = \left(\sum_{i=1}^n p_i\right)^2 = \sum_i p_i^2 + \sum_{i\neq j} p_i p_j$, we again have the logical entropy $h(p)$ as the probability $\sum_{i\neq j} p_i p_j$ of drawing a distinction in two independent samplings of the probability distribution p. This is also clear from defining the product measure on the subsets $S \subseteq U \times U$:

$$\mu(S) = \sum\{p_i p_j : (u_i, u_j) \in S\}$$

Product measure on $U \times U$

Then the logical entropy $h(p) = \mu(\mathbf{1}_U)$ is just the product measure of the dit set of the discrete partition on U. There is also the obvious generalization to consider any partition π on U and then define for each block $B \in \pi$, $p_B = \sum_{u_i \in B} p_i$. Then the logical entropy $h(\pi) = \mu(\operatorname{dit}(\pi))$ is the product measure of the dit set of π (so it is

[3] This could be taken as the logical entropy $h(u)$ of the random variable u but since the values of u are irrelevant (other than being distinct for $i \neq j$), we can take the logical entropy $h(p)$ as a function solely of the probability distribution p of the random variable.

still interpreted as the probability of drawing a distinction of π) and that is equivalent to $\sum_B p_B(1-p_B)$.

For the uniform distribution $p_i = \frac{1}{n}$, the logical entropy has its maximum value of $1-\frac{1}{n}$. Regardless of the first draw (even for a different probability distribution over the same n outcomes), the probability that the second draw is different is $1-\frac{1}{n}$. The logical entropy has its minimum value of 0 for $p=(1,0,...,0)$ so that:

$$0 \leq h(p) \leq 1-\tfrac{1}{n}.$$

An important special case is a set U of $|U| = N$ equiprobable elements and a partition π on U with n equal-sized blocks of N/n elements each. Then the number of distinctions of elements is $N^2 - n\left(\frac{N}{n}\right)^2 = N^2 - \frac{N^2}{n}$ which normalizes to the logical entropy of $h(\pi) = 1-\frac{1}{n}$ and which is independent of N. Thus it holds when $N=n$ and we take the elements to be the equal blocks themselves. Thus for an equal-blocked partition on a set of equiprobable elements, the normalized number of distinctions of elements is the same as the normalized number of distinctions of blocks, and that quantity is the:

$$h(p_0) = 1-p_0 = 1-\tfrac{1}{n}$$

Logical entropy of an equiprobable set of n elements.

2.3 A statistical treatment of logical entropy

It might be noted that no averaging is involved in the interpretation of $h(\pi)$. It is the number of distinctions $|\operatorname{dit}(\pi)|$ normalized. The definition of the logical entropy $h(p) = \sum_{i=1}^n p_i h(p_i) = \sum_{i=1}^n p_i(1-p_i)$ of a probability distribution $p=(p_1,...,p_n)$ is in the form of the average value of the random variable which has the value $h(p_i) = 1-p_i$ with the probability p_i.

Hence the formula can be arrived at by applying the law of large numbers in the form where the finite random variable X takes the value x_i with probability p_i:

$$\lim_{N\to\infty} \tfrac{1}{N}\sum_{j=1}^N x_j = \sum_{i=1}^n p_i x_i.$$

At each step j in repeated independent sampling $u_1u_2...u_N$ of the probability distribution $p=(p_1,...,p_n)$, the probability that the j^{th} result u_j was *not* u_j is $1-\Pr(u_j)$ so the *average* probability of the result being different than it was at each place in that sequence is:

$$\tfrac{1}{N}\sum_{j=1}^N (1-\Pr(u_j)).$$

In the long run, the *typical* sequences will dominate where the i^{th} outcome is sampled p_iN times so that we have the value $1-p_i$ occurring p_iN times:

$$\lim_{N\to\infty} \tfrac{1}{N}\sum_{j=1}^N (1-\Pr(u_j)) = \tfrac{1}{N}\sum_{i=1}^n p_iN(1-p_i) = h(p).$$

The logical entropy $h(p) = \sum_i p_i (1 - p_i) = \sum_{i \neq j} p_i p_j$ is usually interpreted as the *pair-drawing probability of getting distinct outcomes* from the distribution $p = (p_1, ..., p_n)$. Now we have a different interpretation of logical entropy as *the average probability of being different.*

2.4 A brief history of the logical entropy formula

The logical entropy formula $h(p) = \sum_i p_i (1 - p_i) = 1 - \sum_i p_i^2$ is the probability of getting distinct values $u_i \neq u_j$ in two independent samplings of the random variable u. The complementary measure $1 - h(p) = \sum_i p_i^2$ is the probability that the two drawings yield the same value from U. Thus $1 - \sum_i p_i^2$ is a measure of heterogeneity or diversity in keeping with our theme of information as distinctions, while the complementary measure $\sum_i p_i^2$ is a measure of homogeneity or concentration. Historically, the formula can be found in either form depending on the particular context. The p_i's might be relative shares such as the relative share of organisms of the i^{th} species in some population of organisms, and then the interpretation of p_i as a probability arises by considering the random choice of an organism from the population.

According to I. J. Good, the formula has a certain naturalness:

> If $p_1, ..., p_t$ are the probabilities of t mutually exclusive and exhaustive events, any statistician of this century who wanted a measure of homogeneity would have take about two seconds to suggest $\sum p_i^2$ which I shall call ρ. [14, p. 561]

As noted by Bhargava and Uppuluri [3], the formula $1 - \sum p_i^2$ was used by Gini in 1912 ([10] reprinted in [11, p. 369]) as a measure of “mutability” or diversity. But another development of the formula (in the complementary form) in the early twentieth century was in cryptography. The American cryptologist, William F. Friedman, devoted a 1922 book ([9]) to the "index of coincidence" (i.e., $\sum p_i^2$). Solomon Kullback (of the Kullback-Leibler divergence treated later) worked as an assistant to Friedman and wrote a book on cryptology which used the index. [19]

During World War II, Alan M. Turing worked for a time in the Government Code and Cypher School at the Bletchley Park facility in England. Probably unaware of the earlier work, Turing used $\rho = \sum p_i^2$ in his cryptoanalysis work and called it the *repeat rate* since it is the probability of a repeat in a pair of independent draws from a population with those probabilities (i.e., the identification probability $1 - h(p)$). Polish cryptoanalyists had independently used the repeat rate in their work on the Enigma [25].

After the war, Edward H. Simpson, a British statistician, proposed $\sum_{B \in \pi} p_B^2$ as a measure of species concentration (the opposite of diversity) where π is the partition of animals or plants according to species and where each animal or plant is considered as equiprobable. And Simpson gave the interpretation of this homogeneity measure as "the probability that two individuals chosen at random and independently from the population will be found to belong to the same group."[29, p. 688] Hence $1-\sum_{B\in\pi} p_B^2$ is the probability that a random ordered pair will belong to different species, i.e., will be distinguished by the species partition. In the biodiversity literature [26], the formula is known as "Simpson's index of diversity" or sometimes, the *Gini-Simpson index [23]*. However, Simpson along with I. J. Good worked at Bletchley Park during WWII, and, according to Good, "E. H. Simpson and I both obtained the notion [the repeat rate] from Turing." [13, p. 395] When Simpson published the index in 1948, he (again, according to Good) did not acknowledge Turing "fearing that to acknowledge him would be regarded as a breach of security." [14, p. 562]

In 1945, Albert O. Hirschman ([17, p. 159] and [18]) suggested using $\sqrt{\sum p_i^2}$ as an index of trade concentration (where p_i is the relative share of trade in a certain commodity or with a certain partner). A few years later, Orris Herfindahl [16] independently suggested using $\sum p_i^2$ as an index of industrial concentration (where p_i is the relative share of the i^{th} firm in an industry). In the industrial economics literature, the index $H = \sum p_i^2$ is variously called the Hirschman-Herfindahl index, the HH index, or just the H index of concentration. If all the relative shares were equal (i.e., $p_i = 1/n$), then the identification or repeat probability is just the probability of drawing any element, i.e., $H = 1/n$, so $\frac{1}{H} = n$ is the number of equal elements. This led to the "numbers equivalent" interpretation of the reciprocal of the H index [2]. In general, given an event with probability p_0, the *numbers-equivalent interpretation* of the event is that it is 'as if' an element was drawn out of a set U_{1/p_0} of $\frac{1}{p_0}$ equiprobable elements (it is 'as if' since $1/p_0$ need not be an integer). This interpretation will be used later in the dit-bit connection.

In view of the frequent and independent discovery and rediscovery of the formula $\rho = \sum p_i^2$ or its complement $1-\sum p_i^2$ by Gini, Friedman, Turing, Hirschman, Herfindahl, and no doubt others, I. J. Good wisely advises that "it is unjust to associate ρ with any one person." [14, p. 562]

Two elements from $U = \{u_1, ..., u_n\}$ are either identical or distinct. Gini [10] introduced d_{ij} as the "distance" between the i^{th} and j^{th} elements where $d_{ij} = 1$ for $i \neq j$ and $d_{ii} = 0$. Since $1 = (p_1 + ... + p_n)(p_1 + ... + p_n) = \sum_i p_i^2 + \sum_{i\neq j} p_i p_j$, the logical entropy, i.e., Gini's index of mutability, $h(p) = 1 - \sum_i p_i^2 = \sum_{i\neq j} p_i p_j$, is the average logical distance between a pair of independently drawn elements. But one might generalize by allowing other distances $d_{ij} = d_{ji}$ for $i \neq j$ (but always $d_{ii} = 0$) so that $Q = \sum_{i\neq j} d_{ij} p_i p_j$ would be the average distance between a pair of independently drawn elements from U. In 1982, C. R. (Calyampudi Radhakrishna) Rao introduced precisely this concept as *quadratic entropy* [23]. In many domains, it is quite reasonable to move beyond the bare-bones *logical distance* of $d_{ij} = 1$ for $i \neq j$ (i.e., the complement $1 - \delta_{ij}$ of the Kronecker delta) so that Rao's quadratic entropy is a useful and easily interpreted generalization of logical entropy.[4]

[4] Rao's treatment also includes (and generalizes) the natural extension to continuous (square-integrable) probability density functions $f(x)$: $h(f) = 1 - \int f(x)^2 dx$.

3 Shannon Entropy

3.1 Shannon-Hartley entropy of a set

The Shannon entropy will first be motivated in the usual fashion and then developed from the basic logical notion of entropy. Shannon, like Ralph Hartley [15] before him, starts with the question of how much "information" is required to single out a designated element from a set U of equiprobable elements. This is often formulated in terms of the search [24] for a hidden element like the answer in a Twenty Questions game or the sent message in a communication. But being able to always find the designated element is equivalent to being able to distinguish all elements from one another. That is, if the designated element was in a set of two or more elements that had not been distinguished from one another, then one would not be able to single out the designated element. Thus "singling out" or "identifying" an element in a set is just another way to conceptualize "distinguishing" all the elements of the set.

Intuitively, one might measure "information" as the minimum number of yes-or-no questions in a game of Twenty Questions that it would take in general to *distinguish* all the possible "answers" (or "messages" in the context of communications). This is readily seen in the simple case where $|U| = 2^m$, i.e., the size of the set of equiprobable elements is a power of 2. Then following the lead of Wilkins over three centuries earlier, the 2^m elements could be encoded using words of length m in a binary code such as the digits $\{0,1\}$ of binary arithmetic (or $\{A,B\}$ in the case of Wilkins). Then an efficient or minimum set of yes-or-no questions needed to single out the hidden element is the set of m questions:

"Is the j^{th} digit in the binary code for the hidden element a 1?"

for $j = 1, ..., m$. Each element is distinguished from any other element by their binary codes differing in at least one digit. The information gained in finding the outcome of an equiprobable binary trial, like flipping a fair coin, is what Shannon calls a *bit* (derived from "binary digit"). Hence the information gained in distinguishing all the elements out of 2^m equiprobable elements is:

$$m = \log_2\left(2^m\right) = \log_2\left(|U|\right) = \log_2\left(\frac{1}{p_0}\right) \text{ bits}$$

where $p_0 = \frac{1}{2^m}$ is the probability of any given element (henceforth all logs to base 2).

This is usefully restated in terms of partitions. Given two partitions $\pi = \{B\}$ and $\sigma = \{C\}$ of U, their *join* $\pi \vee \sigma$ is the partition of U whose blocks are the non-empty intersections $B \cap C$ for $B \in \pi$ and $C \in \sigma$. The determination of the j^{th} digit in the binary code for the hidden element defines a binary partition π_j of U. Then to say that the answers to the m questions above distinguish all the elements means that the join, $\bigvee_{j=1}^{m} \pi_j = \mathbf{1}$, is the discrete partition on the set U with cardinality 2^m. Thus we could also take $m = \log\left(\frac{1}{p_0}\right)$ as the minimum number of binary partitions necessary to distinguish the elements (i.e., to single out any given element).

In the more general case where $|U| = n$ is not a power of 2, we extrapolate to the definition of $H(p_0)$ where $p_0 = \frac{1}{n}$ as:

$$H(p_0) = \log\left(\frac{1}{p_0}\right) = \log(n)$$

Shannon-Hartley entropy for an equiprobable set U of n elements.

The definition is further extrapolated to the case where we are only given a probability p_0 so that we say that $H(p_0) = \log\left(\frac{1}{p_0}\right)$ binary partitions are needed to distinguish a set of $\frac{1}{p_0}$ elements when $\frac{1}{p_0}$ is not an integer.

3.2 Shannon entropy of a probability distribution

This interpretation of the special case of 2^m or more generally $1/p_0$ equiprobable elements is extended to an arbitrary finite probability distribution $p = (p_1, ..., p_n)$ by an averaging process. For the i^{th} outcome $(i = 1, ..., n)$, its probability p_i is "as if" it were drawn from a set of $\frac{1}{p_i}$ equiprobable elements (ignoring that $\frac{1}{p_i}$ may not be an integer for this averaging argument) so the Shannon-Hartley information content of distinguishing the equiprobable elements of such a set would be $\log\left(\frac{1}{p_i}\right)$. But that occurs with probability p_i so the probabilistic average gives the usual definition of the:

$$H(p) = \sum_{i=1}^{n} p_i H(p_i) = \sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Shannon entropy of a finite probability distribution p.

For the uniform distribution $p_i = \frac{1}{n}$, the Shannon entropy has it maximum value of $\log(n)$ while the minimum value is 0 for the trivial distribution $p = (1, 0, ..., 0)$ so that:

$$0 \leq H(p) \leq \log(n).$$

3.3 A statistical treatment of Shannon entropy

Shannon makes this averaging argument rigorous by using the law of large numbers. Suppose that we have a three-letter alphabet $\{a, b, c\}$ where each letter was equiprobable, $p_a = p_b = p_c = \frac{1}{3}$, in a multi-letter message. Then a one-letter or two-letter message cannot be exactly coded with a binary $0, 1$ code with equiprobable 0's and 1's. But any probability can be better and better approximated by longer and longer representations in the binary number system. Hence we can consider longer and longer messages of N letters along with better and better approximations with binary codes.

The long run behavior of messages $u_1u_2...u_N$ where $u_i \in \{a,b,c\}$ is modeled by the law of large numbers so that the letter a will tend to occur $p_aN = \frac{1}{3}N$ times and similarly for b and c. Such a message is called *typical.*

The probability of any one of those typical messages is:

$$p_a^{p_aN}p_b^{p_bN}p_c^{p_cN} = [p_a^{p_a}p_b^{p_b}p_c^{p_c}]^N$$

or, in this case,

$$\left[\left(\frac{1}{3}\right)^{1/3}\left(\frac{1}{3}\right)^{1/3}\left(\frac{1}{3}\right)^{1/3}\right]^N = \left(\frac{1}{3}\right)^N.$$

Hence the number of such typical messages is 3^N.

If each message was assigned a unique binary code, then the number of $0,1$'s in the code would have to be X where $2^X = 3^N$ or $X = \log\left(3^N\right) = N\log(3)$. Hence the number of equiprobable binary questions or bits needed per letter of the messages is:

$$N\log(3)/N = \log(3) = 3 \times \frac{1}{3}\log\left(\frac{1}{1/3}\right) = H(p).$$

This example shows the general pattern.

In the general case, let $p = (p_1,...,p_n)$ be the probabilities over a n-letter alphabet $A = \{a_1,...,a_n\}$. In an N-letter message, the probability of a particular message $u_1u_2...u_N$ is $\Pi_{i=1}^N \Pr(u_i)$ where u_i could be any of the symbols in the alphabet so if $u_i = a_j$ then $\Pr(u_i) = p_j$.

In a *typical* message, the i^{th} symbol will occur p_iN times (law of large numbers) so the probability of a typical message is (note change of indices to the letters of the alphabet):

$$\Pi_{k=1}^n p_k^{p_kN} = [\Pi_{k=1}^n p_k^{p_k}]^N.$$

Since the probability of a typical message is P^N for $P = \Pi_{k=1}^n p_k^{p_k}$, the typical messages are equiprobable. Hence the number of typical messages is $\left[\Pi_{k=1}^n p_k^{-p_k}\right]^N$ and assigning a unique binary code to each typical message requires X bits where $2^X = \left[\Pi_{k=1}^n p_k^{-p_k}\right]^N$ where:

$$\begin{aligned} X = \log\left\{\left[\Pi_{k=1}^n p_k^{-p_k}\right]^N\right\} &= N\log\left[\Pi_{k=1}^n p_k^{-p_k}\right] \\ = N\textstyle\sum_{k=1}^n \log\left(p_k^{-p_k}\right) &= N\textstyle\sum_k -p_k\log(p_k) \\ = N\textstyle\sum_k p_k\log\left(\frac{1}{p_k}\right) &= NH(p). \end{aligned}$$

Hence the Shannon entropy $H(p) = \sum_{k=1}^n p_k \log\left(\frac{1}{p_k}\right)$ is interpreted as the limiting *average number of bits necessary per letter in the message.* In terms of distinctions, this is the *average number of binary partitions necessary per letter to distinguish the messages.* It is this averaging result that allows us to consider "the number of binary partitions it takes to distinguish the elements of U" when $|U|$ is not a power of 2 since "number" is interpreted as "average number."

3.4 Shannon entropy of a partition

Shannon entropy can also be defined for a partition $\pi = \{B\}$ on a set U. If the elements of U are equiprobable, then the probability that a randomly drawn element is in a block $B \in \pi$ is $p_B = \frac{|B|}{|U|}$. In a set of $\frac{1}{p_B}$ equiprobable elements, it would take (on average) $H(p_B) = \log\left(\frac{1}{p_B}\right)$ binary partitions to distinguish the elements. Averaging over the blocks, we have the:

$$H(\pi) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right)$$

Shannon entropy of a partition π.

3.5 Shannon entropy and statistical mechanics

The functional form of Shannon's formula is often further "justified" or "motivated" by asserting that it is the same as the notion of entropy in statistical mechanics, and hence the name "entropy." The name "entropy" is here to stay but the justification of the formula by reference to statistical mechanics is not quite correct. The connection between entropy in statistical mechanics and Shannon's entropy is only via a numerical approximation, the Stirling approximation, where if the first two terms in the Stirling approximation are used, then the Shannon formula is obtained.

The first two terms in the Stirling approximation for $\ln(N!)$ are: $\ln(N!) \approx N\ln(N) - N$. The first three terms in the Stirling approximation are: $\ln(N!) \approx N(\ln(N)-1) + \frac{1}{2}\ln(2\pi N)$.

If we consider a partition on a finite U with $|U| = N$, with n blocks of size $N_1, ..., N_n$, then the number of ways of distributing the individuals in these n boxes with those numbers N_i in the i^{th} box is: $W = \frac{N!}{N_1! \times ... \times N_n!}$. The normalized natural log of W, $S = \frac{1}{N}\ln(W)$ is one form of entropy in statistical mechanics. Indeed, the formula "$S = k\log(W)$" is engraved on Boltzmann's tombstone.

The entropy formula can then be developed using the first two terms in the Stirling approximation.

$$\begin{gathered}
S = \frac{1}{N}\ln(W) = \frac{1}{N}\ln\left(\frac{N!}{N_1!\times ...\times N_n!}\right) = \frac{1}{N}\left[\ln(N!) - \sum_i \ln(N_i!)\right] \\
\approx \frac{1}{N}\left[N\left[\ln(N) - 1\right] - \sum_i N_i\left[\ln(N_i) - 1\right]\right] \\
= \frac{1}{N}\left[N\ln(N) - \sum N_i\ln(N_i)\right] = \frac{1}{N}\left[\sum N_i\ln(N) - \sum N_i\ln(N_i)\right] \\
= \sum \frac{N_i}{N}\ln\left(\frac{1}{N_i/N}\right) = \sum p_i\ln\left(\frac{1}{p_i}\right) = H_e(p)
\end{gathered}$$

where $p_i = \frac{N_i}{N}$ (and where the formula with logs to the base e only differs from the usual base 2 formula by a scaling factor). Shannon's entropy $H_e(p)$ is in fact an excellent numerical approximation to $S = \frac{1}{N}\ln(W)$ for large N (e.g., in statistical mechanics).

But the common claim is that Shannon's entropy has the *same functional form* as entropy in statistical mechanics, and that is simply false. If we use a three-term Stirling approximation, then we obtain an even better numerical approximation:[5]

$$S = \frac{1}{N}\ln(W) \approx H_e(p) + \frac{1}{2N}\ln\left(\frac{2\pi N^n}{(2\pi)^n \Pi p_i}\right)$$

but no one would suggest using that "more accurate" entropy formula in information theory. Shannon's formula should be justified and understood by the arguments given previously, and not by over-interpreting the approximate relationship with entropy in statistical mechanics.

3.6 The basic dit-bit connection

The basic datum is "the" set U_n of n elements with the equal probabilities $p_0 = \frac{1}{n}$. In that basic case of an equiprobable set, we can derive the dit-bit connection, and then by using a probabilistic average, we can develop the Shannon entropy, expressed in terms of bits, from the logical entropy, expressed in terms of (normalized) dits, or vice-versa.

Given U_n with n equiprobable elements, the number of dits (of the discrete partition on U_n) is $n^2 - n$ so the normalized dit count is:

$$h(p_0) = h\left(\frac{1}{n}\right) = 1 - p_0 = 1 - \frac{1}{n} \text{ normalized dits.}$$

That is the dit-count or logical measure of the information is a set of n distinct elements.[6]

But we can also measure the information in the set by the number of binary partitions it takes (on average) to distinguish the elements, and that bit-count is:

$$H(p_0) = H\left(\frac{1}{n}\right) = \log\left(\frac{1}{p_0}\right) = \log(n) \text{ bits.}$$

By solving the dit-count and the bit-count for p_0 and equating, we can derive each measure in terms of the other:

$$H(p_0) = \log\left(\frac{1}{1-h(p_0)}\right) \text{ and } h(p_0) = 1 - \frac{1}{2^{H(p_0)}}$$

The dit-bit conversion formulas.

[5] For the case $n = 2$, MacKay [21, p. 2] also uses Stirling's approximation to give a "more accurate approximation" (using the next term in the Stirling approximation) to the entropy of statistical mechanics than the Shannon entropy.

[6] The context will determine whether "dit-count" refers to the "raw" count $|\text{dit}(\pi)|$ or the normalized count $\frac{|\text{dit}(\pi)|}{|U \times U|}$.

The common thing being measured is an equiprobable U_n where $n = \frac{1}{p_0}$. The dit-count for U_n is $h(p_0) = 1 - \frac{1}{n}$ and the bit-count for U_n is $H(p_0) = \log\left(\frac{1}{p_0}\right)$, and the bit-dit connection gives the relationship between the two counts. Using this dit-bit connection between the two different ways to measure the "information" in U_n, each entropy can be developed from the other.

We start with the logical entropy of a probability distribution $p = (p_1, ..., p_n)$: $h(p) = \sum_{i=1}^{n} p_i h(p_i)$. It is expressed as the probabilistic average of the dit-counts or logical entropies of the sets U_{1/p_i} with $\frac{1}{p_i}$ equiprobable elements.[7] But if we switch to the binary-partition bit-counts of the information content of those same sets U_{1/p_i} of $\frac{1}{p_i}$ equiprobable elements, then the bit-counts are $H(p_i) = \log\left(\frac{1}{p_i}\right)$ and the probabilistic average is the Shannon entropy: $H(p) = \sum_{i=1}^{n} p_i H(p_i)$. Both entropies have the mathematical form:

$$\sum_i p_i \left(\text{measure of info. in } U_{1/p_i}\right)$$

and differ by using either the dit-count or bit-count to measure the information in U_{1/p_i}.

Clearly the process is reversible, so one can use the dit-bit connection in reverse to develop the logical entropy $h(p)$ from the Shannon entropy $H(p)$. Thus the two notions of entropy are simply two different ways, using distinctions (dit-counts) or binary partitions (bit-counts), to measure the information in a probability distribution.

Moreover the dit-bit connection carries over to the compound notions of entropy so that the Shannon notions of conditional entropy, mutual information, and joint entropy can be developed from the corresponding notions for logical entropy. Since the logical notions are the values of a probability measure, the compound notions of logical entropy have the usual Venn diagram relations such as the inclusion-exclusion principle. There is a well-known analogy between the "Venn diagram" relationships for the Shannon entropies and the relationships satisfied by any measure on a set ([1], [5]). As L. L. Campbell puts it, the analogy:

> suggests the possibility that $H(\alpha)$ and $H(\beta)$ are measures of sets, that $H(\alpha, \beta)$ is the measure of their union, that $I(\alpha, \beta)$ is the measure of their intersection, and that $H(\alpha|\beta)$ is the measure of their difference. The possibility that $I(\alpha, \beta)$ is the entropy of the "intersection" of two partitions is particularly interesting. This "intersection," if it existed, would presumably contain the information common to the partitions α and β.[5, p. 113]

All of Campbell's desiderata are precisely true when:

[7] Starting with the datum of the probability p_i, there is no necessity that $n = \frac{1}{p_i}$ is an integer so the dit-counts for U_{1/p_i} are extrapolations while the bit-counts or binary partition counts for U_n are already extrapolations even when n is an integer but not a power of 2.

- "sets" = dit sets, and
- "entropies" = normalized counting measure of the (dit) sets, i.e., the logical entropies.

Since the logical entropies are the values of a measure, by developing the corresponding Shannon notions from the logical ones, we have an explanation of why the Shannon notions also exhibit the same Venn diagram relationships.

The expository strategy is to first develop the Shannon and logical compound notions of entropy separately and then to show the relationship using the dit-bit connection.

4 Conditional entropies

4.1 Logical conditional entropy

Given two partitions $\pi = \{B\}$ and $\sigma = \{C\}$ on a finite set U, how might one measure the new information that is provided by π that was not already in σ? Campbell suggests associating sets with partitions so the conditional entropy would be the measure of the difference between the sets. Taking the information as distinctions, we take the difference between the dit sets, i.e., $\operatorname{dit}(\pi) - \operatorname{dit}(\sigma)$, and then take the normalized counting measure of that subset of $\operatorname{dit}(\pi) - \operatorname{dit}(\sigma) \subseteq U \times U$:

$$h(\pi|\sigma) = \frac{|\operatorname{dit}(\pi) - \operatorname{dit}(\sigma)|}{|U|^2}$$

Logical conditional entropy of π given σ.

When the two partitions π and σ are joined together in the join $\pi \vee \sigma$, whose blocks are the non-empty intersections $B \cap C$, their information as distinctions is also joined together as sets, $\operatorname{dit}(\pi \vee \sigma) = \operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)$ (the "union" mentioned by Campbell), which has the normalized counting measure of:

$$h(\pi \vee \sigma) = \frac{|\operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)|}{|U|^2} = \sum_{B \in \pi, C \in \sigma} p_{B \cap C} \left[1 - p_{B \cap C}\right]$$

logical entropy of a partition join $\pi \vee \sigma$.

This logical entropy is interpreted as the probability that a pair of random draws from U will yield a π-distinction *or* a σ-distinction (where "or" includes both).

Then the relationships between the logical entropy concepts can be read off the Venn diagram inclusion-exclusion principle for the dit sets:

$$|\operatorname{dit}(\pi)| + |\operatorname{dit}(\sigma)| = |\operatorname{dit}(\pi \vee \sigma)| + |\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)|$$

so that

$$|\mathrm{dit}(\pi)-\mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| - |\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi\vee\sigma)| - |\mathrm{dit}(\sigma)|.$$

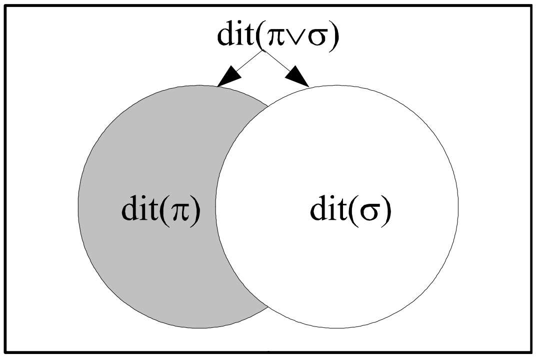

Figure 1: Venn diagram for subsets of $U \times U$

The shaded area in the Venn diagram has the dit-count measure:

$$|\mathrm{dit}(\pi)-\mathrm{dit}(\sigma)| = |(\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma))| - |\mathrm{dit}(\sigma)|$$
$$h(\pi|\sigma) = h(\pi\vee\sigma) - h(\sigma).$$

For the corresponding definitions for random variables and their probability distributions, consider a random variable (x,y) taking values in the product $X \times Y$ of finite sets with the joint probability distribution $p(x,y)$, and thus with the marginal distributions: $p(x) = \sum_{y\in Y} p(x,y)$ and $p(y) = \sum_{x\in X} p(x,y)$. For notational simplicity, the entropies can be considered as functions of the random variables or of their probability distributions, e.g., $h(p(x,y)) = h(x,y)$. For the joint distribution, we have the:

$$h(x,y) = h(p(x,y)) = \sum_{x\in X, y\in Y} p(x,y)[1-p(x,y)]$$

logical entropy of the joint distribution

which is the probability that two samplings of the joint distribution will yield a pair of *distinct* ordered pairs (x,y), $(x',y') \in X \times Y$, i.e., with an X-distinction $x \neq x'$ *or* a Y-distinction $y \neq y'$.

For the definition of the conditional entropy $h(x|y)$, we simply take the product measure of the set of pairs (x,y) and (x',y') that give an X-distinction but not a Y-distinction. Thus given the first draw (x,y), we can again use a Venn diagram to compute the probability that the second draw (x',y') will have $x' \neq x$ but $y' = y$.

To illustrate this using Venn diagram reasoning, consider the probability measure defined by $p(x,y)$ on the subsets of $X \times Y$. Given the first draw (x,y), the probability of getting an (x,y)-distinction on the second draw is $1 - p(x,y)$ and the probability of getting a y-distinction is $1 - p(y)$. A draw that is a y-distinction is, a fortiori, an (x,y)-distinction so the area $1 - p(y)$ is contained in the area $1 - p(x,y)$. Then the probability of getting an (x,y)-distinction that is not a y-distinction on the second draw is the difference: $(1 - p(x,y)) - (1 - p(y)) = p(y) - p(x,y)$.

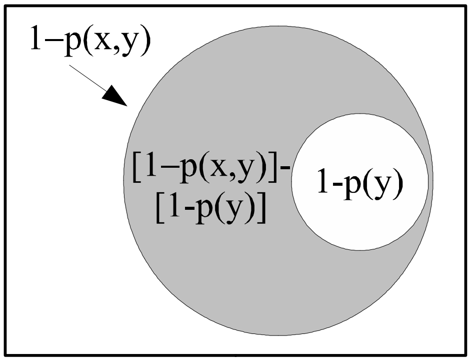

Figure 2: $(1-p(x,y))-(1-p(y))$
$=$ probability of an x-distinction but not a y-distinction on $X \times Y$.

Since the first draw (x,y) was with probability $p(x,y)$, we have the following as the product measure of the subset of $[X \times Y]^2$ of pairs $[(x,y),(x',y')]$ that are X-distinctions but not Y-distinctions:

$$h(x|y) = \sum_{x,y} p(x,y)\left[(1-p(x,y))-(1-p(y))\right]$$

logical conditional entropy of x given y.

Then a little algebra quickly yields:

$$h(x|y) = \sum_{x,y} p(x,y)\left[(1-p(x,y))-(1-p(y))\right]$$
$$= \left[1-\sum_{x,y} p(x,y)^2\right] - \left[1-\sum_{y} p(y)^2\right] = h(x,y)-h(y).$$

The summation over $p(x,y)$ recasts the Venn diagram to the set $(X \times Y)^2$ where the product probability measure (for the two independent draws) gives the logical entropies:

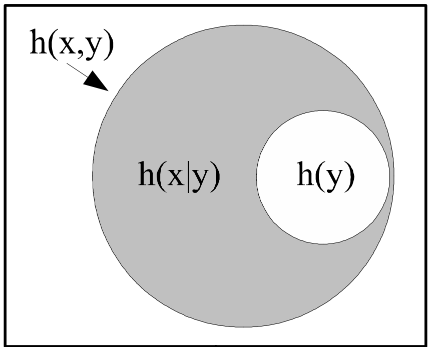

Figure 3: $h(x|y) = h(x,y)-h(y)$.

It might be noted that the logical conditional entropy, like the other logical entropies, is not just an average; the conditional entropy is the product probability measure of the subset:

$$\{[(x,y),(x',y')] : x \neq x', y = y'\} \subseteq (X \times Y) \times (X \times Y).$$

4.2 Shannon conditional entropy

The Shannon conditional entropy for partitions π and σ is based on subset reasoning which is then averaged over a partition. Given a subset $C \in \sigma$, a partition $\pi = \{B\}_{B\in\pi}$ induces a partition of C with the blocks $\{B \cap C\}_{B\in\pi}$. Then $p_{B|C} = \frac{p_{B\cap C}}{p_C}$ is the probability distribution associated with that partition so it has a Shannon entropy which we denote: $H(\pi|C) = \sum_{B\in\pi} p_{B|C} \log\left(\frac{1}{p_{B|C}}\right) = \sum_B \frac{p_{B\cap C}}{p_C} \log\left(\frac{p_C}{p_{B\cap C}}\right)$. The Shannon conditional entropy is then obtained by averaging over the blocks of σ:

$$H(\pi|\sigma) = \sum_{C\in\sigma} p_C H(\pi|C) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_C}{p_{B\cap C}}\right)$$

Shannon conditional entropy of π given σ.

Since the join $\pi \vee \sigma$ is the partition whose blocks are the non-empty intersections $B \cap C$,

$$H(\pi \vee \sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{1}{p_{B\cap C}}\right).$$

Developing the formula gives:

$$H(\pi|\sigma) = \sum_C \left[p_C \log(p_C) - \sum_B p_{B\cap C} \log(p_{B\cap C})\right] = H(\pi \vee \sigma) - H(\sigma).$$

Thus the conditional entropy $H(\pi|\sigma)$ is interpreted as the Shannon-information contained in the join $\pi \vee \sigma$ that is not contained in σ.

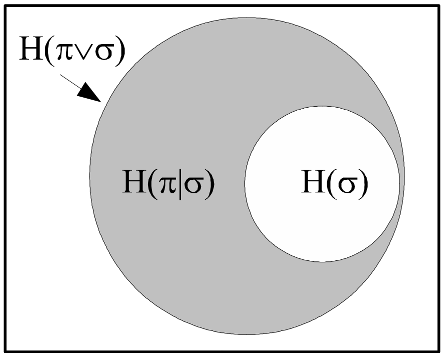

Figure 4: $H(\pi|\sigma) = H(\pi \vee \sigma) - H(\sigma)$
"Venn diagram picture" for Shannon conditional entropy of partitions

Given the joint distribution $p(x,y)$ on $X \times Y$, the conditional probability distribution for a specific $y_0 \in Y$ is $p(x|y_0) = \frac{p(x,y_0)}{p(y_0)}$ which has the Shannon entropy: $H(x|y_0) = \sum_x p(x|y_0) \log\left(\frac{1}{p(x|y_0)}\right)$. Then the conditional entropy is the average of these entropies:

$$H(x|y) = \sum_y p(y) \sum_x \frac{p(x,y)}{p(y)} \log\left(\frac{p(y)}{p(x,y)}\right) = \sum_{x,y} p(x,y) \log\left(\frac{p(y)}{p(x,y)}\right)$$

Shannon conditional entropy of x given y.

Expanding as before gives $H(x|y) = H(x,y) - H(y)$ with a similar Venn diagram picture (see below).

4.3 Shannon conditional entropy from logical conditional entropy

Now we can develop the Shannon conditional entropy from the logical conditional entropy and thereby explain the Venn diagram relationship. The logical conditional entropy is:

$$h\left(x|y\right)=\sum_{x,y}p\left(x,y\right)\left[\left(1-p\left(x,y\right)\right)-\left(1-p\left(y\right)\right)\right]$$

where $1-p\left(x,y\right)$ is the normalized dit count for the discrete partition on a set $U_{1/p(x,y)}$ with $\frac{1}{p(x,y)}$ equiprobable elements. Hence that same equiprobable set requires the bit-count of $\log\left(\frac{1}{p(x,y)}\right)$ binary partitions to distinguish its elements. Similarly $1-p\left(y\right)$ is the normalized dit count for (the discrete partition on) a set $U_{1/p(y)}$ with $\frac{1}{p(y)}$ equiprobable elements, so it requires $\log\left(\frac{1}{p(y)}\right)$ binary partitions to make those distinctions. Those binary partitions are included in the $\log\left(\frac{1}{p(x,y)}\right)$ binary partitions (since a y-distinction is automatically a (x,y)-distinction) and we don't want the y-distinctions so they are subtracted off to get: $\log\left(\frac{1}{p(x,y)}\right)-\log\left(\frac{1}{p(y)}\right)$ bits. Taking the same probabilistic average, the average number of binary partitions needed to make the x-distinctions but not the y-distinctions is:

$$\sum_{x,y}p\left(x,y\right)\left[\log\left(\frac{1}{p(x,y)}\right)-\log\left(\frac{1}{p(y)}\right)\right]=\sum_{x,y}p\left(x,y\right)\log\left(\frac{p(y)}{p(x,y)}\right)=H\left(x|y\right).$$

Replacing the dit-counts by the bit-counts for the equiprobable sets, and taking the probabilistic average gives the same Venn diagram picture for the Shannon entropies.

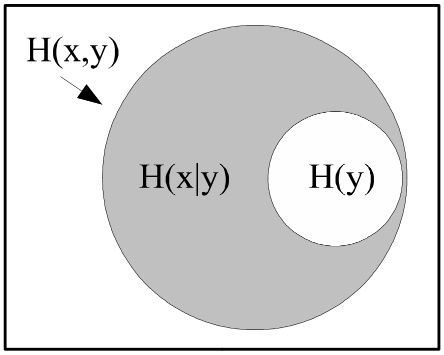

Figure 5: $H\left(x|y\right)=H\left(x,y\right)-H\left(y\right)$.

5 Mutual information for logical entropies

5.1 The case for partitions

If the "atom" of information is the distinction or dit, then the atomic information in a partition π is its dit set, $\operatorname{dit}(\pi)$. Following again Campbell's dictum about the mutual information, the information common to two partitions π and σ would naturally be the intersection of their dit sets:

$$\text{Mut}(\pi, \sigma) = \text{dit}(\pi) \cap \text{dit}(\sigma)$$

Mutual information set.

It is an interesting and not completely trivial fact that as long as neither π nor σ are the indiscrete partition $\mathbf{0}$ (where $\text{dit}(\mathbf{0}) = \emptyset$), then π and σ have a distinction in common.

Proposition 5.1 (Non-empty dit sets intersect) *Given two partitions π and σ on U with non-empty dit sets,* $\text{dit}(\pi) \cap \text{dit}(\sigma) \neq \emptyset$.[8]

Since π is not the indiscrete partition, consider two elements u and u' distinguished by π but identified by σ [otherwise $(u, u') \in \text{dit}(\pi) \cap \text{dit}(\sigma)$]. Since σ is also not the indiscrete partition, there must be a third element u'' not in the same block of σ as u and u'. But since u and u' are in different blocks of π, the third element u'' must be distinguished from one or the other or both in π. Hence (u, u'') or (u', u'') must be distinguished by both partitions and thus must be in their mutual information set $\text{Mut}(\pi, \sigma) = \text{dit}(\pi) \cap \text{dit}(\sigma)$.□

The dit sets $\text{dit}(\pi)$ and their complementary indit sets (= equivalence relations) $\text{indit}(\pi) = U^2 - \text{dit}(\pi)$ are easily characterized as:

$$\begin{aligned} \text{indit}(\pi) &= \bigcup_{B \in \pi} B \times B \\ \text{dit}(\pi) &= \bigcup_{B \neq B'; B, B' \in \pi} B \times B' = U \times U - \text{indit}(\pi) = \text{indit}(\pi)^c . \end{aligned}$$

The mutual information set can also be characterized in this manner.

Proposition 5.2 (Structure of mutual information sets) *Given partitions π and σ with blocks $\{B\}_{B \in \pi}$ and $\{C\}_{C \in \sigma}$, then*

$$\text{Mut}(\pi, \sigma) = \bigcup_{B \in \pi, C \in \sigma} (B - (B \cap C)) \times (C - (B \cap C)) = \bigcup_{B \in \pi, C \in \sigma} (B - C) \times (C - B).$$

The union (which is a disjoint union) will include the pairs (u, u') where for some $B \in \pi$ and $C \in \sigma$, $u \in B - (B \cap C)$ and $u' \in C - (B \cap C)$. Since u' is in C but not in the intersection $B \cap C$, it must be in a different block of π than B so $(u, u') \in \text{dit}(\pi)$. Symmetrically, $(u, u') \in \text{dit}(\sigma)$ so $(u, u') \in \text{Mut}(\pi, \sigma) = \text{dit}(\pi) \cap \text{dit}(\sigma)$. Conversely if $(u, u') \in \text{Mut}(\pi, \sigma)$ then take the B containing u and the C containing

[8] The contrapositive of the "non-empty dit sets intersect" proposition is also interesting. Given two equivalence relations $E_1, E_2 \subseteq U^2$, if every pair of elements $u, u' \in U$ is equated by one or the other of the relations, i.e., $E_1 \cup E_2 = U^2$, then either $E_1 = U^2$ or $E_2 = U^2$.

u'. Since (u,u') is distinguished by both partitions, $u \notin C$ and $u' \notin B$ so that $(u,u') \in (B-(B\cap C)) \times (C-(B\cap C))$.□

The probability that a pair randomly chosen from $U \times U$ would be distinguished by π *and* σ would be given by the normalized counting measure of the mutual information set which is the:

$$m(\pi,\sigma) = \frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2} = \text{probability that } \pi \text{ and } \sigma \text{ distinguishes}$$

Mutual logical information of π *and* σ.

By the inclusion-exclusion principle:

$$|\mathrm{Mut}(\pi,\sigma)| = |\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| + |\mathrm{dit}(\sigma)| - |\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|.$$

Normalizing, the probability that a random pair is distinguished by both partitions is given by the inclusion-exclusion principle:

$$\begin{aligned} m(\pi,\sigma) &= \frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2} \\ &= \frac{|\mathrm{dit}(\pi)|}{|U|^2} + \frac{|\mathrm{dit}(\sigma)|}{|U|^2} - \frac{|\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|}{|U|^2} \\ &= h(\pi) + h(\sigma) - h(\pi\vee\sigma). \end{aligned}$$

Inclusion-exclusion principle for logical entropies of partitions

This can be extended after the fashion of the inclusion-exclusion principle to any number of partitions. It was previously noted that the intersection of two dit sets is not necessarily the dit set of a partition, but the interior of the intersection is the dit set $\mathrm{dit}(\pi\wedge\sigma)$ of the partition meet $\pi\wedge\sigma$. Hence we also have the:

$$h(\pi\wedge\sigma) \leq h(\pi) + h(\sigma) - h(\pi\vee\sigma)$$

Submodular inequality for logical entropies.

5.2 The case for joint distributions

Consider again a joint distribution $p(x,y)$ over $X \times Y$ for finite X and Y. Intuitively, the mutual logical information $m(x,y)$ in the joint distribution $p(x,y)$ would be the probability that a sampled pair (x,y) would be a distinction of $p(x)$ *and* a distinction of $p(y)$. That means for each probability $p(x,y)$, it must be multiplied by the probability of not drawing the same x *and* not drawing the same y (e.g., in a second independent drawing). In the Venn diagram, the area or probability of the drawing that x or that y is $p(x)+p(y)-p(x,y)$ (correcting for adding the overlap twice) so the probability of getting neither that x nor that y is the complement $1-p(x)-p(y)+p(x,y) = [1-p(x)]+[1-p(y)]-[1-p(x,y)]$.

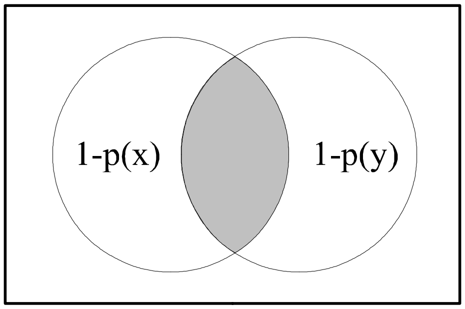

Figure 6: $[1 - p(x)] + [1 - p(y)] - [1 - p(x, y)]$
$=$ shaded area in Venn diagram for $X \times Y$

Hence we have:

$$m(x, y) = \sum_{x,y} p(x, y) [[1 - p(x)] + [1 - p(y)] - [1 - p(x, y)]]$$

Logical mutual information in a joint probability distribution.

The probability of two independent draws differing in *either* the x *or* the y is just the logical entropy of the joint distribution:

$$h(x, y) = h(p(x, y)) = \sum_{x,y} p(x, y) [1 - p(x, y)] = 1 - \sum_{x,y} p(x, y)^2.$$

Using a little algebra to expand the logical mutual information:

$$\begin{aligned} m(x, y) &= \left[1 - \sum_{x,y} p(x, y) p(x)\right] + \left[1 - \sum_{x,y} p(x, y) p(y)\right] - \left[1 - \sum_{x,y} p(x, y)^2\right] \\ &= h(x) + h(y) - h(x, y) \end{aligned}$$

Inclusion-exclusion principle for logical entropies of a joint distribution.

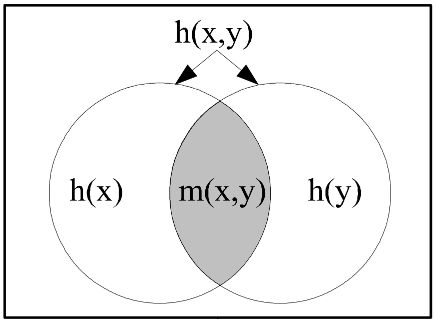

Figure 7: $m(x, y) = h(x) + h(y) - h(x, y)$
$=$ shaded area in Venn diagram for $(X \times Y)^2$.

It might be noted that the logical mutual information, like the other logical entropies, is not just an average; the mutual information is the product probability measure of the subset:

$$\{[(x, y), (x', y')] : x \neq x', y \neq y'\} \subseteq (X \times Y) \times (X \times Y).$$

6 Mutual information for Shannon entropies

6.1 The case for partitions

The usual heuristic motivation for Shannon's mutual information is much like its development from the logical mutual information so we will take that approach at the outset. The logical mutual information for partitions can be expressed in the form:

$$m(\pi,\sigma) = \sum_{B,C} p_{B\cap C}\left[(1-p_B)+(1-p_C)-(1-p_{B\cap C})\right]$$

so if we substitute the bit-counts for the dit-counts as before, we get:

$$I(\pi,\sigma) = \sum_{B,C} p_{B\cap C}\left[\log\left(\frac{1}{p_B}\right)+\log\left(\frac{1}{p_C}\right)-\log\left(\frac{1}{p_{B\cap C}}\right)\right] = \sum_{B,C} p_{B\cap C}\log\left(\frac{p_{B\cap C}}{p_B p_C}\right)$$

Shannon's mutual information for partitions.

Keeping the log's separate gives the Venn diagram picture:

$$\begin{aligned} I(\pi,\sigma) &= \sum_{B,C} p_{B\cap C}\left[\log\left(\frac{1}{p_B}\right)+\log\left(\frac{1}{p_C}\right)-\log\left(\frac{1}{p_{B\cap C}}\right)\right] \\ &= H(\pi)+H(\sigma)-H(\pi\vee\sigma) \end{aligned}$$

Inclusion-exclusion analogy for Shannon entropies of partitions.

6.2 The case for joint distributions

To move from partitions to probability distributions, consider again the joint distribution $p(x,y)$ on $X\times Y$. Then developing the Shannon mutual information from the logical mutual information amounts to replacing the block probabilities $p_{B\cap C}$ in the join $\pi\vee\sigma$ by the joint probabilities $p(x,y)$ and the probabilities in the separate partitions by the marginals (since $p_B=\sum_{C\in\sigma}p_{B\cap C}$ and $p_C=\sum_{B\in\pi}p_{B\cap C}$), to obtain:

$$I(x,y) = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Shannon mutual information in a joint probability distribution.

Then the same proof carries over to give the:

$$I(x,y) = H(x)+H(y)-H(x,y)$$

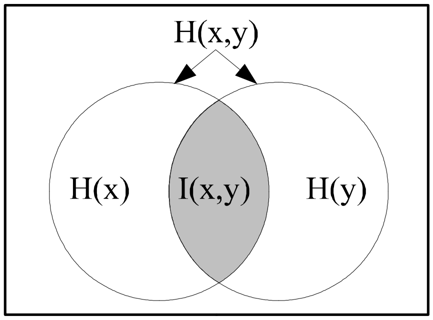

Figure 8: Inclusion-exclusion "picture" for Shannon entropies of probability distributions.

The logical mutual information formula:

$$m(x,y) = \sum_{x,y} p(x,y)\left[[1-p(x)] + [1-p(y)] - [1-p(x,y)]\right]$$

develops via the dit-count to bit-count conversion to:

$$\sum_{x,y} p(x,y)\left[\log\left(\frac{1}{p(x)}\right) + \log\left(\frac{1}{p(y)}\right) - \log\left(\frac{1}{p(x,y)}\right)\right] = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = I(x,y).$$

Thus the genuine Venn diagram relationships for the product probability measure that gives the logical entropies carry over, via the dit-count to bit-count conversion, to give a similar Venn diagram picture for the Shannon entropies.

7 Independence

7.1 Independent Partitions

Two partitions π and σ are said to be (stochastically) *independent* if for all $B \in \pi$ and $C \in \sigma$, $p_{B\cap C} = p_B p_C$. If π and σ are independent, then:

$$I(\pi,\sigma) = \sum_{B\in\pi, C\in\sigma} p_{B\cap C}\log\left(\frac{p_{B\cap C}}{p_B p_C}\right) = 0 = H(\pi) + H(\sigma) - H(\pi \vee \sigma),$$

so that:

$$H(\pi \vee \sigma) = H(\pi) + H(\sigma)$$

Shannon entropy for partitions additive under independence.

In ordinary probability theory, two events $E, E' \subseteq U$ for a sample space U are said to be *independent* if $\Pr(E \cap E') = \Pr(E)\Pr(E')$. We have used the motivation of thinking of a partition-as-dit-set $\operatorname{dit}(\pi)$ as an "event" in a sample space $U \times U$ with the probability of that event being $h(\pi)$, the logical entropy of the partition. The following proposition shows that this motivation extends to the notion of independence.

Proposition 7.1 (Independent partitions have independent dit sets) *If π and σ are (stochastically) independent partitions, then their dit sets $\operatorname{dit}(\pi)$ and $\operatorname{dit}(\sigma)$ are independent as events in the sample space $U \times U$ (with equiprobable points).*

For independent partitions π and σ, we need to show that the probability $m(\pi,\sigma)$ of the event $\operatorname{Mut}(\pi,\sigma) = \operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)$ is equal to the product of the probabilities $h(\pi)$ and $h(\sigma)$ of the events $\operatorname{dit}(\pi)$ and $\operatorname{dit}(\sigma)$ in the sample space $U \times U$. By the assumption of stochastic independence, we have $\frac{|B \cap C|}{|U|} = p_{B\cap C} = p_B p_C = \frac{|B||C|}{|U|^2}$ so that $|B \cap C| = |B|\,|C| \,/\, |U|$. By the previous structure theorem for the mutual information set: $\operatorname{Mut}(\pi,\sigma) = \bigcup_{B\in\pi, C\in\sigma} (B - (B\cap C)) \times (C - (B \cap C))$, where the union is disjoint so that:

$$\begin{aligned}
|\operatorname{Mut}(\pi,\sigma)| &= \textstyle\sum_{B\in\pi,C\in\sigma} (|B| - |B\cap C|)\,(|C| - |B \cap C|) \\
&= \textstyle\sum_{B\in\pi,C\in\sigma} \left(|B| - \frac{|B|\,|C|}{|U|}\right)\left(|C| - \frac{|B|\,|C|}{|U|}\right) \\
&= \frac{1}{|U|^2} \textstyle\sum_{B\in\pi,C\in\sigma} |B|\,(|U| - |C|)\,|C|\,(|U| - |B|) \\
&= \frac{1}{|U|^2} \textstyle\sum_{B\in\pi} |B|\,|U - B| \textstyle\sum_{C\in\sigma} |C|\,|U - C| \\
&= \frac{1}{|U|^2} |\operatorname{dit}(\pi)|\,|\operatorname{dit}(\sigma)|
\end{aligned}$$

so that:

$$m(\pi,\sigma) = \tfrac{|\operatorname{Mut}(\pi,\sigma)|}{|U|^2} = \tfrac{|\operatorname{dit}(\pi)|}{|U|^2}\tfrac{|\operatorname{dit}(\sigma)|}{|U|^2} = h(\pi)\,h(\sigma).\square$$

Hence the logical entropies behave like probabilities under independence; the probability that π *and* σ distinguishes, i.e., $m(\pi,\sigma)$, is equal to the probability $h(\pi)$ that π distinguishes times the probability $h(\sigma)$ that σ distinguishes:

$$m(\pi,\sigma) = h(\pi)\,h(\sigma)$$

Logical entropy multiplicative under independence.

It is sometimes convenient to think in the complementary terms of an equivalence relation "equating" or “identifying” rather than a partition distinguishing. Since $h(\pi)$ can be interpreted as the probability that a random pair of elements from U are distinguished by π, i.e., as a distinction probability, its complement $1 - h(\pi)$ can be interpreted as an *identification probability*, i.e., the probability that a random pair is equated by π (thinking of π as an equivalence relation on U). In general,

$$[1 - h(\pi)]\,[1 - h(\sigma)] = 1 - h(\pi) - h(\sigma) + h(\pi)\,h(\sigma) =$$
$$[1 - h(\pi \vee \sigma)] + [h(\pi)\,h(\sigma) - m(\pi,\sigma]$$

which could also be rewritten as:

$$[1-h(\pi \vee \sigma)]-[1-h(\pi)][1-h(\sigma)]=m(\pi,\sigma)-h(\pi)h(\sigma).$$

Thus if π and σ are independent, then the probability that the join partition $\pi \vee \sigma$ identifies is the probability that π identifies times the probability that σ identifies:

$$[1-h(\pi)][1-h(\sigma)]=[1-h(\pi \vee \sigma)]$$

Multiplicative identification probabilities under independence.

7.2 Independent Joint Distributions

A joint probability distribution $p(x,y)$ on $X \times Y$ is *independent* if each value is the product of the marginals: $p(x,y)=p(x)p(y)$.

For an independent distribution, the Shannon mutual information

$$I(x,y)=\sum_{x\in X,y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

is immediately seen to be zero so we have:

$$H(x,y)=H(x)+H(y)$$

Shannon entropies for independent $p(x,y)$.

For the logical mutual information, independence gives:

$$\begin{aligned} m(x,y) &= \sum_{x,y} p(x,y)[1-p(x)-p(y)+p(x,y)] \\ &= \sum_{x,y} p(x)p(y)[1-p(x)-p(y)+p(x)p(y)] \\ &= \sum_{x} p(x)[1-p(x)]\sum_{y} p(y)[1-p(y)] \\ &= h(x)h(y) \end{aligned}$$

Logical entropies for independent $p(x,y)$.

This independence condition $m(x,y)=h(x)h(y)$ plus the inclusion-exclusion principle $m(x,y)=h(x)+h(y)-h(x,y)$ also implies that:

$$\begin{aligned} [1-h(x)][1-h(y)] &= 1-h(x)-h(y)+h(x)h(y) \\ &= 1-h(x)-h(y)+m(x,y) \\ &= 1-h(x,y). \end{aligned}$$

Hence under independence, the probability of drawing the same pair (x,y) in two independent draws is equal to the probability of drawing the same x times the probability of drawing the same y.

8 Cross-entropies and divergences

Given two probability distributions $p = (p_1, ..., p_n)$ and $q = (q_1, ..., q_n)$ on the same sample space $\{1, ..., n\}$, we can again consider the drawing of a pair of points but where the first drawing is according to p and the second drawing according to q. The probability that the points are distinct would be a natural and more general notion of logical entropy that would be the:

$$h(p\|q) = \sum_i p_i(1 - q_i) = 1 - \sum_i p_i q_i$$

Logical cross entropy of p and q

which is symmetric. The logical cross entropy is the same as the logical entropy when the distributions are the same, i.e., if $p = q$, then $h(p\|q) = h(p)$.

The notion of *cross entropy* in Shannon entropy can be developed by applying dit-bit connection to the logical cross entropy $\sum_i p_i(1 - q_i)$ to obtain:

$$H(p\|q) = \sum_i p_i \log\left(\frac{1}{q_i}\right)$$

which is not symmetrical due to the asymmetric role of the logarithm, although if $p = q$, then $H(p\|q) = H(p)$. Since the logical cross entropy is symmetrical, it could also be expressed as $\sum_i q_i(1 - p_i)$ which develops to the Shannon cross entropy $H(q||p) = \sum_i q_i \log\left(\frac{1}{p_i}\right)$ so it might be more reasonable to use a *symmetrized cross entropy*:

$$H_s(p||q) = \frac{1}{2}\left[H(p||q) + H(q||p)\right].$$

The *Kullback-Leibler divergence* (or *relative entropy*) $D(p\|q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ is defined as a measure of the distance or divergence between the two distributions where $D(p\|q) = H(p\|q) - H(p)$. A basic result is the:

$$D(p\|q) \geq 0 \text{ with equality if and only if } p = q$$

Information inequality [6, p. 26].

Given two partitions π and σ, the inequality $I(\pi, \sigma) \geq 0$ is obtained by applying the information inequality to the two distributions $\{p_{B\cap C}\}$ and $\{p_B p_C\}$ on the sample space $\{(B, C) : B \in \pi, C \in \sigma\} = \pi \times \sigma$:

$$I(\pi, \sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right) = D(\{p_{B\cap C}\} \,\|\, \{p_B p_C\}) \geq 0$$

with equality iff independence.

In the same manner, we have for the joint distribution $p(x, y)$:

$$I\left(x,y\right)=D\left(p\left(x,y\right)||p\left(x\right)p\left(y\right)\right)\geq 0$$

with equality iff independence.

The *symmetrized Kullback-Leibler divergence* is:

$$D_s(p||q)=\frac{1}{2}\left[D\left(p||q\right)+D\left(q||p\right)\right]=H_s\left(p||q\right)-\left[\frac{H(p)+H(q)}{2}\right].$$

But starting afresh, one might ask: "What is the natural measure of the difference or distance between two probability distributions $p=(p_1,...,p_n)$ and $q=(q_1,...,q_n)$ that would always be non-negative, and would be zero if and only if they are equal?" The (Euclidean) distance between the two points in $\mathbb{R}^n$ would seem to be the "logical" answer—so we take that distance (squared with a scale factor) as the definition of the:

$$d\left(p\|q\right)=\frac{1}{2}\sum_i\left(p_i-q_i\right)^2$$

Logical divergence (or *logical relative entropy*)[9]

which is symmetric and we trivially have:

$$d\left(p||q\right)\geq 0 \text{ with equality iff } p=q$$

Logical information inequality.

We have component-wise:

$$0\leq\left(p_i-q_i\right)^2=p_i^2-2p_iq_i+q_i^2=2\left[\frac{1}{n}-p_iq_i\right]-\left[\frac{1}{n}-p_i^2\right]-\left[\frac{1}{n}-q_i^2\right]$$

so that taking the sum for $i=1,...,n$ gives:

$$\begin{aligned}d\left(p\|q\right) &= \frac{1}{2}\sum_i\left(p_i-q_i\right)^2\\ &= \left[1-\sum_i p_iq_i\right]-\frac{1}{2}\left[\left(1-\sum_i p_i^2\right)+\left(1-\sum_i q_i^2\right)\right]\\ &= h\left(p\|q\right)-\frac{h\left(p\right)+h\left(q\right)}{2}.\end{aligned}$$

Logical divergence $=$ *Jensen difference* [23, p. 25] between probability distributions.

Then the information inequality implies that the logical cross-entropy is greater than or equal to the average of the logical entropies:

$$h\left(p||q\right)\geq\frac{h(p)+h(q)}{2} \text{ with equality iff } p=q.$$

[9] In [7], this definition was given without the useful scale factor of 1/2.

The half-and-half probability distribution $\frac{p+q}{2}$ that mixes p and q has the logical entropy of

$$h\left(\frac{p+q}{2}\right)=\frac{h(p||q)}{2}+\frac{h(p)+h(q)}{4}=\frac{1}{2}\left[h\left(p||q\right)+\frac{h(p)+h(q)}{2}\right]$$

so that:

$$h(p||q)\geq h\left(\frac{p+q}{2}\right)\geq\frac{h(p)+h(q)}{2} \text{ with equality iff } p=q.$$

Mixing different p and q increases logical entropy.

The logical divergence can be expressed as:

$$d\left(p\|q\right)=\frac{1}{2}\left[\sum_i p_i\left(1-q_i\right)+\sum_i q_i\left(1-p_i\right)\right]-\frac{1}{2}\left[\left(\sum_i p_i\left(1-p_i\right)\right)+\left(\sum_i q_i\left(1-q_i\right)\right)\right]$$

that develops via the dit-bit connection to:

$$\frac{1}{2}\left[\sum_i p_i\log\left(\frac{1}{q_i}\right)+\sum_i q_i\log\left(\frac{1}{p_i}\right)-\sum_i p_i\log\left(\frac{1}{p_i}\right)-\sum_i q_i\log\left(\frac{1}{q_i}\right)\right]$$
$$=\frac{1}{2}\left[\sum_i p_i\log\left(\frac{p_i}{q_i}\right)+\sum_i q_i\log\left(\frac{q_i}{p_i}\right)\right]=\frac{1}{2}\left[D\left(p||q\right)+D\left(q||p\right)\right]$$
$$=D_s\left(p||q\right).$$

Thus the logical divergence $d\left(p||q\right)$ develops via the dit-bit connection to the symmetrized version of the Kullback-Leibler divergence.

9 Summary and concluding remarks

The following table summarizes the concepts for the Shannon and logical entropies. We use the case of probability distributions rather than partitions, and we use the abbreviations $p_{xy}=p(x,y)$, $p_x=p(x)$, and $p_y=p\left(y\right)$.

	Shannon Entropy	Logical Entropy
Entropy	$H(p)=\sum p_i\log\left(1/p_i\right)$	$h\left(p\right)=\sum p_i\left(1-p_i\right)$
Mutual Info.	$I(x,y)=H\left(x\right)+H\left(y\right)-H\left(x,y\right)$	$m\left(x,y\right)=h\left(x\right)+h\left(y\right)-h\left(x,y\right)$
Independence	$I\left(x,y\right)=0$	$m\left(x,y\right)=h\left(x\right)h\left(y\right)$
Indep. Relations	$H\left(x,y\right)=H\left(x\right)+H\left(y\right)$	$1-h\left(x,y\right)=\left[1-h\left(x\right)\right]\left[1-h\left(y\right)\right]$
Cond. entropy	$H\left(x\vert y\right)=\sum_{x,y}p_{xy}\log\left(\frac{p_y}{p_{xy}}\right)$	$h\left(x\vert y\right)=\sum_{x,y}p_{xy}\left[\left(p_y-p_{xy}\right)\right]$
Relationships	$H\left(x\vert y\right)=H\left(x,y\right)-H\left(y\right)$	$h\left(x\vert y\right)=h\left(x,y\right)-h\left(y\right)$
Cross entropy	$H\left(p\Vert q\right)=\sum p_i\log\left(1/q_i\right)$	$h\left(p\Vert q\right)=\sum p_i\left(1-q_i\right)$
Divergence	$D\left(p\Vert q\right)=\sum_i p_i\log\left(\frac{p_i}{q_i}\right)$	$d\left(p\Vert q\right)=\frac{1}{2}\sum_i\left(p_i-q_i\right)^2$
Relationships	$D\left(p\Vert q\right)=H\left(p\Vert q\right)-H\left(p\right)$	$d\left(p\Vert q\right)=h\left(p\Vert q\right)-\left[h\left(p\right)+h\left(q\right)\right]/2$
Info. Inequality	$D\left(p\Vert q\right)\geq 0$ with $=$ iff $p=q$	$d\left(p\Vert q\right)\geq 0$ with $=$ iff $p=q$

Table of comparisons between Shannon and logical entropies

The above table shows many of the same relationships holding between the various forms of the logical and Shannon entropies due ultimately to the dit-bit connection. The dit-bit connection between the two notions of entropy is based on them being two different measures of the "amount of information-as-distinctions," the dit-count being the normalized count of the distinctions and the bit-count being the number of binary partitions required (on average) to make the distinctions.

Logical entropies arise naturally as the normalized counting measure for partition logic just as probabilities arise as the normalized counting measure for subset logic, where the two logics are dual to one another. All the forms of logical entropy have simple interpretations as the probabilities of distinctions. Shannon entropy is a higher-level and more refined notion adapted to the theory of communications and coding where it can be interpreted as the average number of bits necessary per letter to code the messages, i.e., the average number of binary partitions necessary per letter to distinguish the messages.

References

[1] Abramson, Norman 1963. *Information Theory and Coding.* New York: McGraw-Hill.

[2] Adelman, M. A. 1969. Comment on the H Concentration Measure as a Numbers-Equivalent. *Review of Economics and Statistics.* 51: 99-101.

[3] Bhargava, T. N. and V. R. R. Uppuluri 1975. On an Axiomatic Derivation of Gini Diversity, With Applications. *Metron.* 33: 41-53.

[4] Boole, George 1854. *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities.* Cambridge: Macmillan and Co.

[5] Campbell, L. L. 1965. Entropy as a Measure. *IEEE Trans. on Information Theory.* IT-11 (January): 112-114.

[6] Cover, Thomas and Joy Thomas 1991. *Elements of Information Theory.* New York: John Wiley.

[7] Ellerman, David 2009. Counting Distinctions: On the Conceptual Foundations of Shannon's Information Theory. *Synthese.* 168 (1 May): 119-149. Download at: *www.ellerman.org.*

[8] Ellerman, David 2010. The Logic of Partitions: Introduction to the Dual of the Logic of Subsets. *Review of Symbolic Logic.* 3 (2 June): 287-350. Download at: *www.ellerman.org.*

[9] Friedman, William F. 1922. *The Index of Coincidence and Its Applications in Cryptography.* Geneva IL: Riverbank Laboratories.

[10] Gini, Corrado 1912. *Variabilità e mutabilità.* Bologna: Tipografia di Paolo Cuppini.

[11] Gini, Corrado 1955. Variabilità e mutabilità. In *Memorie di metodologica statistica.* E. Pizetti and T. Salvemini eds., Rome: Libreria Eredi Virgilio Veschi.

[12] Gleick, James 2011. *The Information: A History, A Theory, A Flood.* New York: Pantheon.

[13] Good, I. J. 1979. A.M. Turing's statistical work in World War II. *Biometrika.* 66 (2): 393-6.

[14] Good, I. J. 1982. Comment (on Patil and Taillie: Diversity as a Concept and its Measurement). *Journal of the American Statistical Association.* 77 (379): 561-3.

[15] Hartley, Ralph V. L. 1928. Transmission of information. *Bell System Technical Journal.* 7 (3, July): 535-63.

[16] Herfindahl, Orris C. 1950. *Concentration in the U.S. Steel Industry.* Unpublished doctoral dissertation, Columbia University.

[17] Hirschman, Albert O. 1945. *National power and the structure of foreign trade.* Berkeley: University of California Press.

[18] Hirschman, Albert O. 1964. The Paternity of an Index. *American Economic Review.* 54 (5): 761-2.

[19] Kullback, Solomon 1976. *Statistical Methods in Cryptanalysis.* Walnut Creek CA: Aegean Park Press.

[20] Lawvere, F. William and Robert Rosebrugh 2003. *Sets for Mathematics.* Cambridge: Cambridge University Press.

[21] MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms.* Cambridge UK: Cambridge University Press.

[22] Patil, G. P. and C. Taillie 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association.* 77 (379): 548-61.

[23] Rao, C. R. 1982. Diversity and Dissimilarity Coefficients: A Unified Approach. *Theoretical Population Biology.* 21: 24-43.

[24] Rényi, Alfréd 1970. *Probability Theory.* Laszlo Vekerdi (trans.), Amsterdam: North-Holland.

[25] Rejewski, M. 1981. How Polish Mathematicians Deciphered the Enigma. *Annals of the History of Computing*. 3: 213-34.

[26] Ricotta, Carlo and Laszlo Szeidl 2006. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*. 70: 237-43.

[27] Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*. 27: 379-423; 623-56.

[28] Shannon, Claude E. and Warren Weaver 1964. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

[29] Simpson, Edward Hugh 1949. Measurement of Diversity. *Nature*. 163: 688.

[30] Wilkins, John 1707 (1641). *Mercury or the Secret and Swift Messenger*. London.