An Introduction to Logical Entropy and Its Relation to Shannon Entropy

David Ellerman
University of California at Riverside

January 16, 2014

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 **2**

2 Shannon Entropy **2**
2.1 Shannon-Hartley information content 2
2.2 Shannon entropy of a probability distribution 3
2.3 A statistical treatment of Shannon entropy 3
2.4 Shannon entropy of a partition 4
2.5 Whence "entropy"? 5

3 Logical Entropy **5**
3.1 Partition logic 5
3.2 Logical Entropy 6
3.3 A statistical treatment of logical entropy 8
3.4 A brief history of the logical entropy formula 9

4 Mutual information for Shannon entropies **10**
4.1 The case for partitions 10
4.2 The case for joint distributions 11

5 Mutual information for logical entropies **11**
5.1 The case for partitions 11
5.2 The case for joint distributions 13

6 Independence **15**
6.1 Independent Partitions 15
6.2 Independent Joint Distributions 16

7 Conditional entropies **17**
7.1 Conditional entropies for partitions 17
7.2 Conditional entropies for probability distributions 18

8 Cross-entropies and divergences **19**

9 Summary and concluding remarks **21**

1 Introduction

Information is about making distinctions or differences. In James Gleick's magisterial 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 alphabet.

> 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. [29, 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. [10, p. 161]

We will focus on two notions of information content or entropy, the relatively new logic-based notion of logical entropy [5] and the usual Shannon entropy in Claude Shannon's founding paper, *A Mathematical Theory of Communication* [26]. Both entropy concepts will be explained using the basic idea of distinctions. Shannon's notion of entropy is well adapted to the theory of communications, as indicated by the title of his original article and his later book [27], while the notion of logical entropy arises out of the new logic of partitions [6] that is mathematically dual to the usual Boolean logic of subsets [3].

2 Shannon Entropy

2.1 Shannon-Hartley information content

Shannon, like Ralph Hartley [13] before him, starts with the question of how much "information" is required to distinguish from one another all the elements in a set U of equiprobable elements.[1]

[1] This is often formulated in terms of the search [23] for a designated 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

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 $|U| = 2^m$, the size of the set of equiprobable elements is a power of 2. Then following the lead of Wilkins three centuries earlier, the 2^m elements could be encoded using words of length m in a binary alphabet 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 it takes in general to distinguish the elements are the 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:

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

where $p = \frac{1}{2^m}$ is the probability of any given element. In the more general case where $|U| = n$ is not a power of 2, then the *Shannon-Hartley information content for an equiprobable set U* gained in distinguishing all the elements is taken to be $\log_2 (n) = \log_2 \left(\frac{1}{p}\right)$ bits where $p = \frac{1}{n}$.

2.2 Shannon entropy of a probability distribution

This interpretation of the special case of 2^m or more generally n equiprobable elements is extended to an arbitrary finite probability distribution $p = (p_1, ..., p_n)$ by an averaging process (where $|U| = n$). 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 $\log_2 \left(\frac{1}{p_i}\right)$. But that occurs with probability p_i so the probabilistic average gives the usual definition of the:

$$H(p) = \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right) = -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

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 $\log_2 (n)$ while the minimum value is 0 for the trivial distribution $p = (1, 0, ..., 0)$ so that:

$$0 \leq H(p) \leq \log_2 (n).$$

2.3 A statistical treatment of Shannon entropy

Shannon makes this heuristic 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

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.

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:

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

or, in this case,

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

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 $X = \log_2\left(3^N\right) = N\log_2(3)$. Hence the number of equiprobable binary questions or bits needed per letter of the messages is:

$$N\log_2(3)/N = \log_2(3) = 3 \times \tfrac{1}{3}\log_2\left(\tfrac{1}{1/3}\right) = H(p).$$

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):

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

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:

$$\begin{aligned} X &= \log_2\left\{\left[\Pi_{k=1}^n p_k^{-p_k}\right]^N\right\} = N\log_2\left[\Pi_{k=1}^n p_k^{-p_k}\right] \\ &= N\sum_{k=1}^n \log_2\left(p_k^{-p_k}\right) = N\sum_k -p_k\log_2(p_k) \\ &= N\sum_k p_k\log_2\left(\frac{1}{p_k}\right) = NH(p). \end{aligned}$$

Hence the Shannon entropy $H(p) = \sum_{k=1}^n p_k\log_2\left(\frac{1}{p_k}\right)$ 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.*

2.4 Shannon entropy of a partition

Entropy can also be defined for a partition on a set. A partition $\pi = \{B\}$ on a finite set U is a set of non-empty disjoint subsets of U whose union is U. If the elements of U are equiprobable, then the probability that a randomly drawn element is in a block $B \in \pi$ is $p_B = \frac{|B|}{|U|}$. Then we have the:

$$H(\pi) = \sum_{B\in\pi} p_B\log_2\left(\frac{1}{p_B}\right)$$

Shannon entropy of a partition π.

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 most refined partition is the *discrete partition* $\mathbf{1} = \{\{u\}\}_{u \in U}$ of singleton blocks $\{u\}$ and the least refined partition is the *indiscrete partition* $\mathbf{0} = \{U\}$ whose only block is all of U. The special case of $\pi = \mathbf{1}$ gives the Hartley information content or Shannon entropy $\log_2(n)$ of a set of equiprobable elements. In the more general case where the elements of $U = \{u_1, ..., u_n\}$ are considered as the distinct values of a random variable u with the probabilities $p = (p_1, ..., p_n)$, the induced block probabilities would be $p_B = \sum_{u_i \in B} p_i$ and then the Shannon entropy of the discrete partition $\pi = \mathbf{1}$ is the same as the Shannon entropy of the probability distribution p.

2.5 Whence "entropy"?

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 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.

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

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:[2]

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

but no one would suggest using that "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 Logical Entropy

3.1 Partition logic

The logic normally called "propositional logic" is a special case of the logic of subsets originally developed by George Boole [3]. In the Boolean logic of subsets of a fixed non-empty universe set

[2] MacKay [20, p. 2] uses Stirling's approximation to give another "more accurate approximation" to the entropy of statistical mechanics than the Shannon entropy for the case $n = 2$.

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. It is a theorem that if a formula is valid just for the special case of $U = 1$, then it is valid for any U. But in "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 using the special case of propositional logic rather than Boole's original version 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 its relative cardinality $\frac{|S|}{|U|}$ 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 [19]. 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 only recently been developed [6].

3.2 Logical Entropy

Going back to the original idea of information as making distinctions, a *distinction* or *dit of a partition* $\pi = \{B\}$ of U is 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. The partitions of U are partially ordered by refinement, which is just the inclusion of dit sets, with the discrete partition $\mathbf{1}$ as the top of the order and the indiscrete partition $\mathbf{0}$ as 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	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.
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

But for our purposes here, the key analogy is the quantitative measure $\Pr(S) = \frac{|S|}{|U|}$, the normalized number of elements in a subset S for finite U. Let $\operatorname{dit}(\pi)$ denote the set of distinctions or dits of π, i.e.,

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

In view of the analogy between elements in subset logic and dits in partition logic, the construction analogous to the logical probability $\Pr(S) = \frac{|S|}{|U|}$ as the normalized number of elements of a subset would be the normalized number of distinctions of a partition π on a finite U. That is the definition of the:

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

Logical entropy of a partition π.

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.

	Subset logic	Partition logic
'Outcomes'	Elements u of S	Ordered pairs $(u,u') \in U^2$
'Events'	Subsets S of U	Partitions π of U
'Event occurs'	$u \in S$	$(u,u') \in \operatorname{dit}(\pi)$
Quant. measure	$\Pr(S) = \frac{\lvert S\rvert}{\lvert U\rvert}$	$h(\pi) = \frac{\lvert\operatorname{dit}(\pi)\rvert}{\lvert U \times U\rvert}$
Random drawing	Prob. event S occurs	Prob. partition π distinguishes

Table of quantitative analogies between subset and partition logics

Thus we might say that the logical entropy $h(\pi)$ of a partition π is to partition logic as the logical probability $\Pr(S)$ of a subset S is to 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 $B \in \pi$ 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, which agrees with $h(\pi) = 1 - \sum_{B \in \pi} p_B^2$ being the probability of drawing a distinction.

In the more general case, we assume a random variable u with the probability distribution $p = (p_1, ..., p_n)$ over the n values $U = \{u_1, ..., u_n\}$. Then with the usual $p_B = \sum_{u_i \in B} p_i$, we have the notion $h(\pi) = 1 - \sum_{B \in \pi} p_B^2$ of the logical entropy of a partition π on a set U with the point probabilities $p = (p_1, ..., p_n)$. Note that the probability interpretation of the logical entropy still holds (even though the pairs (u,u') are no longer equiprobable) since:

$$p_B^2 = \left(\sum_{u_i \in B} p_i\right)^2 = \sum_{u_i, u_j \in B} p_i p_j$$

is the probability of drawing an indistinction from $B \times B$. Hence $\sum_{B \in \pi} p_B^2$ is still the probability of drawing an indistinction of π, and the complement $h(\pi)$ the probability of drawing a distinction.

In the case of the discrete partition, we have the:

$$h\left(p\right)=1-\sum_{i}p_{i}^{2}=\sum_{i}p_{i}\left(1-p_{i}\right)$$

Logical entropy of a finite probability distribution p.

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, the probability that the second draw is different is $1-\frac{1}{n}$), and the logical entropy has its minimum value of 0 for $p=(1,0,...,0)$ so that:

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

The two entropies of a probability distribution p or generally of a partition π with given point probabilities p can now be compared:

$$H\left(\pi\right)=\sum_{B\in\pi}p_{B}\log_{2}\left(\tfrac{1}{p_{B}}\right)\text{ and }h\left(\pi\right)=\sum_{B\in\pi}p_{B}\left(1-p_{B}\right).$$

If we define the *Shannon set entropy* as $H\left(B\right)=\log_{2}\left(\frac{1}{p_{B}}\right)$ (the Shannon-Hartley information content for the set B) and the *logical set entropy* as $h\left(B\right)=1-p_{B}$, then each entropy is just the average of the set entropies weighted by the block probabilities:

$$H\left(\pi\right)=\sum_{B\in\pi}p_{B}H\left(B\right)\text{ and }h\left(\pi\right)=\sum_{B\in\pi}p_{B}h\left(B\right)$$

where the set entropies are precisely related: $h\left(B\right)=1-\frac{1}{2^{H(B)}}$ and $H\left(B\right)=\log_{2}\left(\frac{1}{1-h(B)}\right)$.

3.3 A statistical treatment of logical entropy

It might be noted that no averaging is involved in the interpretation of $h\left(\pi\right)$. It is the number of distinctions normalized for the equiprobable elements of U, and, in the more general case, it is the probability that two independent samplings of the random variable u give a distinction of π. But we can nevertheless mimic Shannon's statistical rendering of his entropy formula $H\left(p\right)=\sum_{i}p_{i}\log_{2}\left(\frac{1}{p_{i}}\right)$.

Shannon's use of "typical sequences" is a way of applying the law of large numbers in the form where the finite random variable X takes the value x_i with probability p_i:

$$\lim_{N\to\infty}\tfrac{1}{N}\sum_{j=1}^{N}x_{j}=\sum_{i=1}^{n}p_{i}x_{i}.$$

Since logical entropy $h\left(p\right)=\sum_{i}p_{i}\left(1-p_{i}\right)$ has a similar probabilistic definition, it also can be rendered as a long run statistical average of the random variable $x_i=1-p_i$ which is the probability of being different than the i^{th} outcome.

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 $1-\Pr\left(u_j\right)$ so the *average* probability of the result being different than it was at each place in that sequence is:

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

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:

$$\lim_{N\to\infty}\tfrac{1}{N}\sum_{j=1}^{N}\left(1-\Pr\left(u_{j}\right)\right)=\tfrac{1}{N}\sum_{i=1}^{n}p_{i}N\left(1-p_{i}\right)=h\left(p\right).$$

The logical entropy $h\left(p\right)=\sum_{i}p_{i}\left(1-p_{i}\right)$ 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.*

3.4 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." [12, p. 561] As noted by Bhargava and Uppuluri [2], the formula $1-\sum p_i^2$ was used by Gini in 1912 ([8] reprinted in [9, 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 ([7]) to the "index of coincidence" (i.e., $\sum p_i^2$). Solomon Kullback (see the Kullback-Leibler divergence treated later) worked as an assistant to Friedman and wrote a book on cryptology which used the index. [18]

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 [24].

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."[28, 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 [25], the formula is known as "Simpson's index of diversity" or sometimes, the "Gini-Simpson diversity index." 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." [11, 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." [12, p. 562]

In 1945, Albert O. Hirschman ([15, p. 159] and [16]) 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 [14] 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 [1]. 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 of $\frac{1}{p_0}$ equiprobable elements (it is 'as if' since $1/p_0$ need not be an integer).

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." [12, p. 562]

Two elements from $U = \{u_1, ..., u_n\}$ are either identical or distinct. Gini [8] 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* [22]. 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 Mutual information for Shannon entropies

4.1 The case for partitions

Given two partitions $\pi = \{B\}$ and $\sigma = \{C\}$ on a set U, their *join* $\pi \vee \sigma$ is the partition on U whose blocks are the non-empty intersections $B \cap C$. The join $\pi \vee \sigma$ is the least upper bound of both π and σ in the refinement ordering of partitions on U.

To motivate's Shannon's treatment of mutual information, we might apply some Venn diagram heuristics using a block $B \in \pi$ and a block $C \in \sigma$. We might take the block entropy $H(B) = \log\left(\frac{1}{p_B}\right)$ as representing 'the information contained in B' and similarly for C while $H(B \cap C) = \log\left(\frac{1}{p_{B \cap C}}\right)$ might be taken as the 'union of the information in B and in C' (the more refined blocks in $\pi \vee \sigma$ makes more distinctions). Hence the overlap or "mutual information" in B and C could be motivated, using the inclusion-exclusion principle,[3] as the sum of the two informations minus the union (all logs to base 2):

$$I(B, C) = \log\left(\frac{1}{p_B}\right) + \log\left(\frac{1}{p_C}\right) - \log\left(\frac{1}{p_{B \cap C}}\right) = \log\left(\frac{1}{p_B p_C}\right) + \log(p_{B \cap C}) = \log\left(\frac{p_{B \cap C}}{p_B p_C}\right).$$

Then the *Shannon mutual information* in the two partitions is obtained by averaging over the mutual information for each pair of blocks from the two partitions:

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

The mutual information can be expanded to obtain the inclusion-exclusion principle built into the Venn diagram heuristics:

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

Inclusion-exclusion analogy for Shannon entropies of partitions

[3] The inclusion-exclusion principle for the cardinality of subsets is: $|B \cup C| = |B| + |C| - |B \cap C|$.

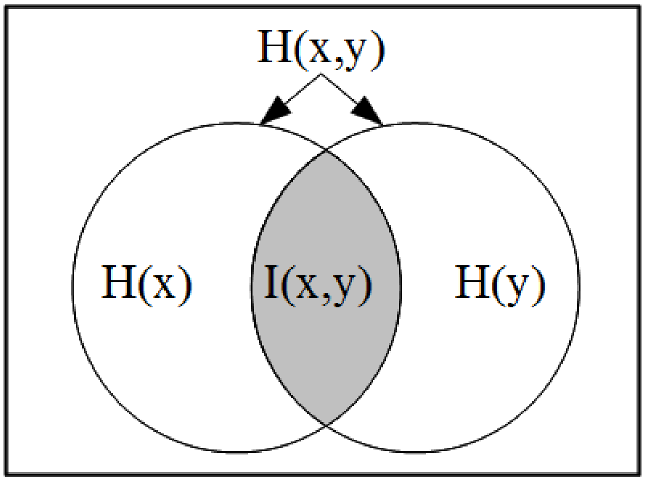
4.2 The case for joint distributions

To move from partitions to probability distributions, consider two finite sets X and Y, and a joint probability distribution $p(x,y)$ where $\sum_{x\in X,y\in Y} p(x,y)=1$ with $p(x,y)\geq 0$, i.e., a random variable with values in $X\times Y$. The marginal distributions are defined as usual: $p(x)=\sum_{y\in Y} p(x,y)$ and $p(y)=\sum_{x\in X} p(x,y)$. Then 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 $p_C=\sum_{B\in\pi} p_{B\cap C}$), we have the definition:

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

Shannon mutual information in a joint probability distribution.

Then the same proof carries over to give [where we write $H(x)$ instead of $H(p(x))$ and similarly for $H(y)$ and $H(x,y)$]:



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

Figure 1: Inclusion-exclusion analogy for Shannon entropies of probability distributions.

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, $\operatorname{dit}(\pi)$. The information common to two partitions π and σ, their *mutual information set*, would naturally be the intersection of their dit sets (which is not necessarily the dit set of a partition):

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

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

Theorem 1 *Given two partitions π and σ on U with $\pi\neq\mathbf{0}\neq\sigma$, then $Mut(\pi,\sigma)\neq\emptyset$.*

Proof: Since π is not the indiscrete partition, consider two elements u and u' distinguished by π but identified by σ [otherwise $(u,u')\in\operatorname{Mut}(\pi,\sigma)$]. 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 $\mathrm{Mut}(\pi,\sigma)$.□[4]

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

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

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

Theorem 2 *Given partitions* π*and* σ *with blocks* $\{B\}_{B\in\pi}$ *and* $\{C\}_{C\in\sigma}$*, then*

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

Proof: 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 $(u,u')\in\mathrm{dit}(\pi)$. Symmetrically, $(u,u')\in\mathrm{dit}(\sigma)$ so $(u,u')\in\mathrm{Mut}(\pi,\sigma)=\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)$. Conversely if $(u,u')\in\mathrm{Mut}(\pi,\sigma)$ then take the B containing u and the C containing 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 relative cardinality of the mutual information set which is the:

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

Mutual logical information of π *and* σ.

Then we may make a non-heuristic application of the inclusion-exclusion principle to obtain:

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

It is easily checked that the dit set $\mathrm{dit}(\pi\vee\sigma)$ of the join of two partitions is the union of their dits sets: $\mathrm{dit}(\pi\vee\sigma) = \mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)$.[5] 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

[4] The contrapositive of this proposition is also interesting. Given two equivalence relations $E_1, E_2 \subseteq U^2$, if $E_1\cup E_2 = U^2$, then $E_1 = U^2$ or $E_2 = U^2$.

[5] But *nota bene*, the dit sets for the other partition operations are not so simple.

This can be extended after the fashion of the inclusion-exclusion principle to any number of partitions.

The mutual information set $\operatorname{Mut}(\pi,\sigma)$ is not necessarily the dit set of a partition. But given any subset $S \subseteq U \times U$ such as $\operatorname{Mut}(\pi,\sigma)$, there is a unique largest dit set contained in S which might be called the *interior* $\operatorname{int}(S)$ *of* S. As in the topological context, the interior of a subset is defined as the "complement of the closure of the complement" but in this case, the "closure" is the reflexive-symmetric-transitive (rst) closure and the "complement" is within $U \times U$. We might apply more topological terminology by calling the binary relations $E \subseteq U \times U$ *closed* if they equal their rst-closures, in which case the closed subsets of $U \times U$ are precisely the indit sets of some partition or in more familiar terms, precisely the equivalence relations on U. Their complements might thus be called the *open* subsets which are precisely the dit sets of some partition, i.e., the complements of equivalence relations which might be called *partition relations*. Indeed, the mapping $\pi \rightarrow \operatorname{dit}(\pi)$ is a representation of the lattice of partitions on U by the open subsets of $U \times U$. While the topological terminology is convenient, the rst-closure operation is not a topological closure operation since the union of two closed sets is not necessarily closed. Thus the intersection of two open subsets is not necessarily open as is the case with $\operatorname{Mut}(\pi,\sigma) = \operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)$. But by taking the interior, we obtain the dit set of the *partition meet*:

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

In general, the partition operations corresponding to the usual binary subset operations of subset logic can be defined by applying the subset operations to the dit sets and then taking the interior of the result so that, for instance, the *partition implication* operation can be defined by:

$$\operatorname{dit}(\sigma \Rightarrow \pi) = \operatorname{int}[\operatorname{dit}(\sigma)^c \cup \operatorname{dit}(\pi)].^6$$

Since $|\operatorname{int}[\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)]| \leq |\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)|$, normalizing yields the:

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

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:

$$1 - p(x) - p(y) + p(x,y) = (1 - p(x)) + (1 - p(y)) - (1 - p(x,y))$$

where $1 - p(x,y)$ is the area of the union of the two circles.

[6] 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.

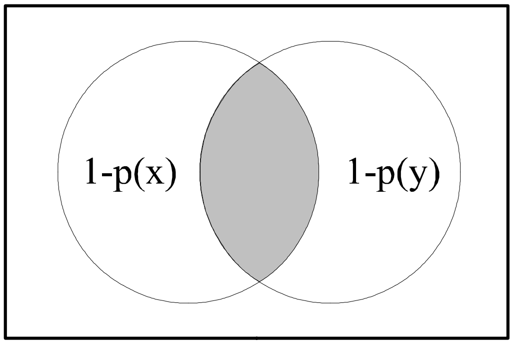


Figure 2: $[1-p(x)]+[1-p(y)]-[1-p(x,y)]$
$=$ shaded area in Venn diagram for $X\times Y$

Hence we have:

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

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:

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

Using a little algebra to expand the logical mutual information:

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

Inclusion-exclusion principle for logical entropies of joint distributions.

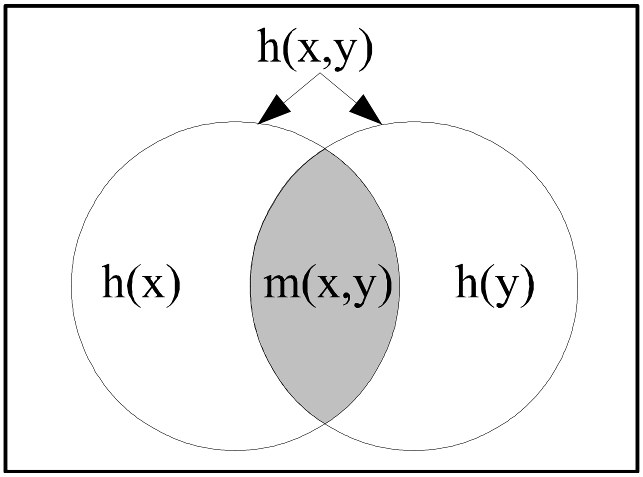


Figure 3: $m(x,y)=h(x)+h(y)-h(x,y)$
$=$ shaded area in Venn diagram for $(X\times Y)^2$.

6 Independence

6.1 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 $\mathrm{dit}(\pi)$ 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.

Theorem 3 *If π and σ are (stochastically) independent partitions, then their dit sets $\mathrm{dit}(\pi)$ and $\mathrm{dit}(\sigma)$ are independent as events in the sample space $U \times U$ (with equiprobable points).*

Proof: For independent partitions π and σ, we need to show that the probability $m(\pi,\sigma)$ of the event $\mathrm{Mut}(\pi,\sigma) = \mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)$ is equal to the product of the probabilities $h(\pi)$ and $h(\sigma)$ of the events $\mathrm{dit}(\pi)$ and $\mathrm{dit}(\sigma)$ 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: $\mathrm{Mut}(\pi,\sigma) = \bigcup_{B\in\pi, C\in\sigma} (B - (B \cap C)) \times (C - (B \cap C))$, where the union is disjoint so that:

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

so that:

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

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 "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 identified 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.

6.2 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 $m(x,y) = h(x) + h(y) - h(x,y)$ 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.

7 Conditional entropies

7.1 Conditional entropies for partitions

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.

Developing the formula gives:

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

so that the inclusion-exclusion formula then yields:

$$H(\pi|\sigma) = H(\pi) - I(\pi;\sigma) = H(\pi \vee \sigma) - H(\sigma).$$

Thus the conditional entropy $H(\pi|\sigma)$ is interpreted as the Shannon-information contained in π that is not mutual to π and σ, or as the combined information in π and σ with the information in σ subtracted out. If one considered the Venn diagram heuristics with two circles $H(\pi)$ and $H(\sigma)$, then $H(\pi \vee \sigma)$ would correspond to the union of the two circles and $H(\pi|\sigma)$ would correspond to the crescent-shaped area with $H(\sigma)$ subtracted out, i.e., $H(\pi \vee \sigma) - H(\sigma)$.

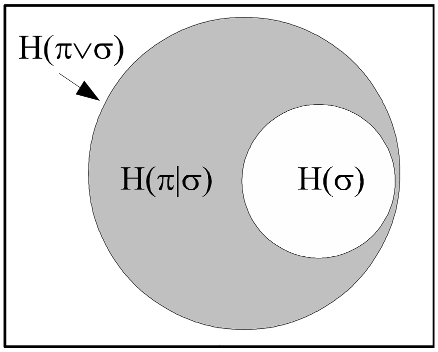


Figure 4: Venn diagram heuristics for Shannon conditional entropy

The logical conditional entropy of a partition π given σ is simply the extra logical-information (i.e., dits) in π not present in σ, so it is given by the difference between their dit sets which normalizes to:

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

Logical conditional entropy of π given σ.

Since these notions are defined as the normalized size of subsets of the set of ordered pairs U^2, the Venn diagrams and inclusion-exclusion principle are not just heuristic. For instance,

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

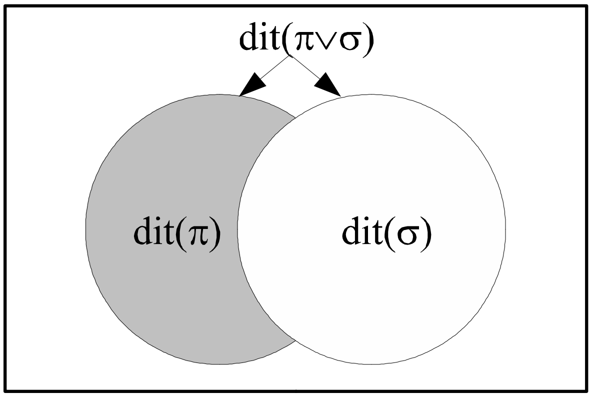


Figure 5: Venn diagram for subsets of $U \times U$

Then normalizing yields:

$$h(\pi|\sigma) = h(\pi) - m(\pi, \sigma) = h(\pi \vee \sigma) - h(\sigma).$$

7.2 Conditional entropies for probability distributions

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

$$H(x|y) = H(x) - I(x, y) = H(x, y) - H(y).$$

The logical conditional entropy $h(x|y)$ is intuitively the probability of drawing a distinction of $p(x)$ which is not a distinction of $p(y)$. Given the first draw (x, y), the probability of getting an (x, y)-distinction 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: $(1 - p(x, y)) - (1 - p(y)) = p(y) - p(x, y)$.

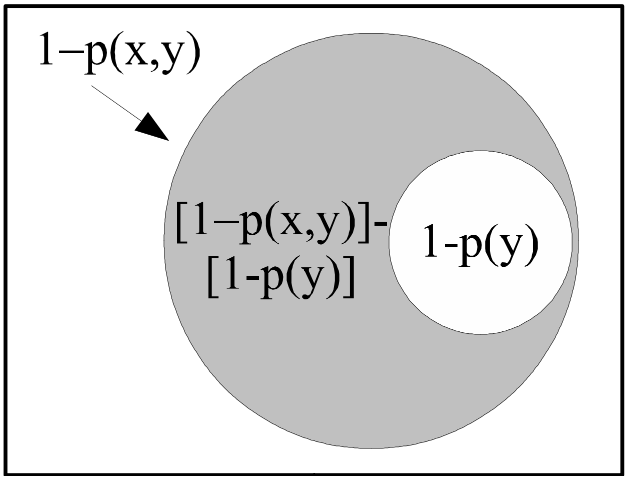


Figure 6: $(1-p(x,y))-(1-p(y))$
$=$ probability of an x-distinction but not a y-distinction on $X\times Y$.

Since the first draw (x,y) was with probability $p(x,y)$, we have the following as the probability 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.

Expanding gives the expected relationships:

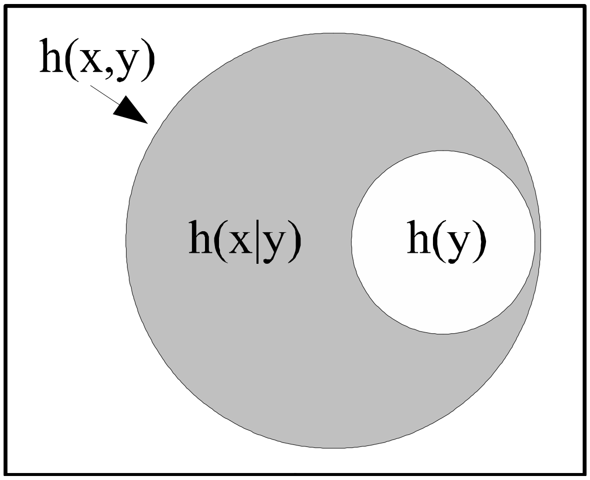


Figure 7: $h(x|y)=h(x)-m(x,y)=h(x,y)-h(y)$.

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 pair of points is distinct would be a natural and more general notion of logical entropy that would be the:

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

$$\text{if } p=q\text{, then } h(p\|q)=h(p).$$

The notion of *cross entropy* in Shannon entropy is: $H(p\|q)=\sum_i p_i\log\left(\frac{1}{q_i}\right)$ which is not symmetrical due to the asymmetric role of the logarithm, although if $p=q$, then $H(p\|q)=H(p)$.

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 [4, 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 $p(x,y)$:

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

with equality iff independence.

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:

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

Logical divergence (or *logical relative entropy*)[7]

which is symmetric and we trivially have:

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

Logical information inequality.

We have component-wise:

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

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

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

Logical divergence = *Jensen difference* [22, 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:

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

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

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

so that:

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

Mixing different p and q increases logical entropy.

[7] In [5], this definition was given without the useful scale factor of 1/2.

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 $p_y = p(y)$.

	Shannon Entropy	Logical Entropy
Entropy	$H(p) = \sum p_i \log(1/p_i)$	$h(p) = \sum p_i (1 - p_i)$
Mutual Info.	$I(x,y) = H(x) + H(y) - H(x,y)$	$m(x,y) = h(x) + h(y) - h(x,y)$
Independence	$I(x,y) = 0$	$m(x,y) = h(x) h(y)$
Indep. Rel.	$H(x,y) = H(x) + H(y)$	$1 - h(x,y) = [1 - h(x)][1 - h(y)]$
Cond. entropy	$H(x \mid y) = \sum_{x,y} p_{xy} \log\left(\frac{p_y}{p_{xy}}\right)$	$h(x \mid y) = \sum_{x,y} p_{xy} [p_y - p_{xy}]$
Relationships	$H(x \mid y) = H(x,y) - H(y)$	$h(x \mid y) = h(x,y) - h(y)$
Cross entropy	$H(p \Vert q) = \sum p_i \log(1/q_i)$	$h(p \Vert q) = \sum p_i (1 - q_i)$
Divergence	$D(p \Vert q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$	$d(p \Vert q) = \frac{1}{2} \sum_i (p_i - q_i)^2$
Relationships	$D(p \Vert q) = H(p \Vert q) - H(p)$	$d(p \Vert q) = h(p \Vert q) - \frac{1}{2}[h(p) + h(q)]$
Info. Ineq.	$D(p \Vert q) \geq 0$ with $=$ iff $p = q$	$d(p \Vert q) \geq 0$ 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. What is the connection? The connection between the two notions of entropy is based on them being two different measures of the "amount of distinctions," i.e., the quantity of information-as-distinctions.

This is easily seen by going back to the original example of a set of 2^n elements where each element has the same probability $p_i = \frac{1}{2^n}$. The Shannon set entropy is the minimum number of binary partitions it takes to distinguish all the elements which is:

$$n = \log_2\left(\frac{1}{1/2^n}\right) = \log_2\left(\frac{1}{p_i}\right) = H(p_i).$$

The Shannon entropy $H(p)$ for $p = \{p_1, ..., p_m\}$ is the probability-weighted average of those binary partition measures:

$$H(p) = \sum_{i=1}^{m} p_i H(p_i) = \sum_i p_i \log_2\left(\frac{1}{p_i}\right).$$

Rather than measuring distinctions by counting the binary partitions needed to distinguish all the elements, let's count the distinctions directly. In the set with 2^n elements, each with probability $p_i = \frac{1}{2^n}$, how many distinctions (pairs of distinct elements) are there? All the ordered pairs except the diagonal are distinctions so the total number of distinctions is $2^n \times 2^n - 2^n$ which normalizes to:

$$\frac{2^n \times 2^n - 2^n}{2^n \times 2^n} = 1 - \frac{1}{2^n} = 1 - p_i = h(p_i).$$

The logical entropy $h(p)$ is the probability-weighted average of these normalized dit counts:

$$h(p) = \sum_{i=1}^{m} p_i h(p_i) = \sum_i p_i (1 - p_i).$$

Thus we see that the two notions of entropy are just two different quantitative measures of:

$$\textit{Information} = \textit{distinctions}.$$

Logical entropy arises naturally out of partition logic as the normalized counting measure of the set of distinctions in a partition. Logical entropy is simpler and more basic in the sense of the logic of partitions which is dual to the usual Boolean logic of subsets. 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 identify a message, i.e., the average number of binary partitions necessary per letter to distinguish the messages.

References

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

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

[3] 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.

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

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

[6] 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.

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

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

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

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

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

[12] 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.

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

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

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

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

[17] Kullback, Solomon 1968. *Information Theory and Statistics*. New York: Dover.

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

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

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

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

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

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

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

[25] 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.

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

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

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

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