

The Relationship of Sample Space to Sample Size in the Calculation of Entropy for Random Binary Symbol Strings

1.0 Introduction

The relationship of sample space to sample size for random binary data can be shown in every case to take the form of a curve in which the percentage difference between the calculated entropy and the sample size decreases as the sample size increases. *What this means is that for any binary symbol string, even quantum random binary data, the data can be organized in a number of ways, each of which results in a calculation of entropy (as measured in bits) that is less than the number of bits in the original string.* This applies to all random binary symbol strings, whether they are of finite or infinite length.

By sample space is meant the total number of possible events or outcomes for a given set or collection of events. For example, the possible outcomes when rolling a die are {1, 2, 3, 4, 5, 6} or 6. The possible outcomes for a binary digit are {0, 1} or 2. The possible outcomes for a binary word of two bits in length are {00, 01, 10, 11} or 4. By sample size is meant the actual number of events that are measured. Three rolls of a die have a sample size of 3. Five flips of a coin have a sample size of 5.

2.0 Example

As an example, the following model describes a string of random binary digits or bits as a series of five (5) 1-bit words. For any string of random binary data used as input to the model, the percent difference between the size in bits of the original data and the resulting entropy is approximately 12.2%, that is, the calculated entropy in bits is roughly 12% smaller than the actual number of bits in the sample space.

Example:

Given a string S of random binary digits 5 bits or greater in length,

where:

the string S is divisible into one or more substrings (i.e., $S_1, S_2, S_3, \dots, S_n$) each with a length of 5 bits and a single remaining substring with a length of 0 through 4 bits and each substring of 5 bits is divisible in turn into two sets of binary symbols of like type (i.e., 1s and 0s),

the entropy calculations for each possible set of 5-bit substrings and their sum is contained in the following table:

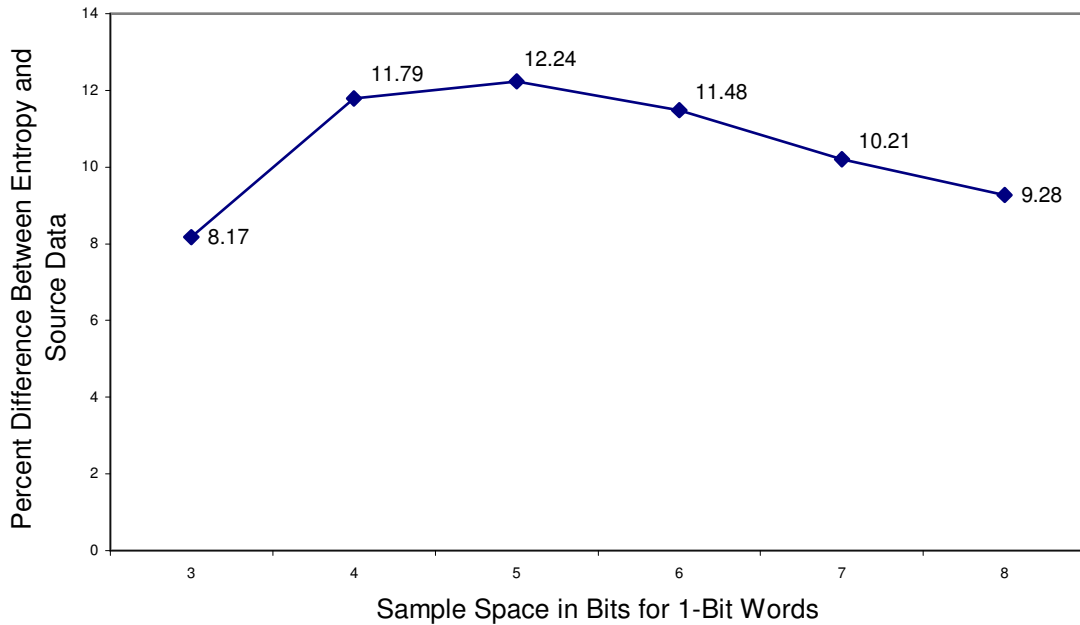
Substring	Entropy	Substring	Entropy	
00000(1)	3.60964	10000	3.60964	
00001	3.60964	10001	4.854753	160Total Bits
00010	3.60964	10010	4.854753	140.41Total Entropy
00011	4.854753	10011	4.854753	12.24%Percent Difference
00100	3.60964	10100	4.854753	
00101	4.854753	10101	4.854753	
00110	4.854753	10110	4.854753	
00111	4.854753	10111	3.60964	
01000	3.60964	11000	4.854753	
01001	4.854753	11001	4.854753	
01010	4.854753	11010	4.854753	
01011	4.854753	11011	3.60964	
01100	4.854753	11100	4.854753	
01101	4.854753	11101	3.60964	
01110	4.854753	11110	3.60964	
01111	3.60964	11111(0)	3.60964	

Those substrings with values of 00000 and 11111 (whose members are of like type and the calculation of whose entropy, because it requires division by 0, is undefined), can be included in the calculations either through a) the change of a single symbol to one of unlike type within the string or b) the addition of a single symbol of unlike type to the string. In the first case, the substring “00000” once converted to “00001” will result in an entropy of 3.60964. Likewise, “11111” once converted to “11110” will result in an entropy of 3.60964. In the second case, the substring “00000” once converted to “000001” will result in an entropy of 3.900135 and “11111” once converted to “111110” will result in an entropy of 3.900135.

In a random collection of symbols, the frequency of each symbol is roughly equivalent to that of every other symbol. This equivalence increases with sample size. Therefore, by assuming a fully random distribution of these substrings within a larger random binary string and by calculating the entropy for each substring independently of the others and summing the results, the total entropy for the larger random binary string can be calculated. In the above example, for any large random binary string, the entropy will always be in the range of approximately 87.76% of the size of the original (source) string for a difference of 12.24%.

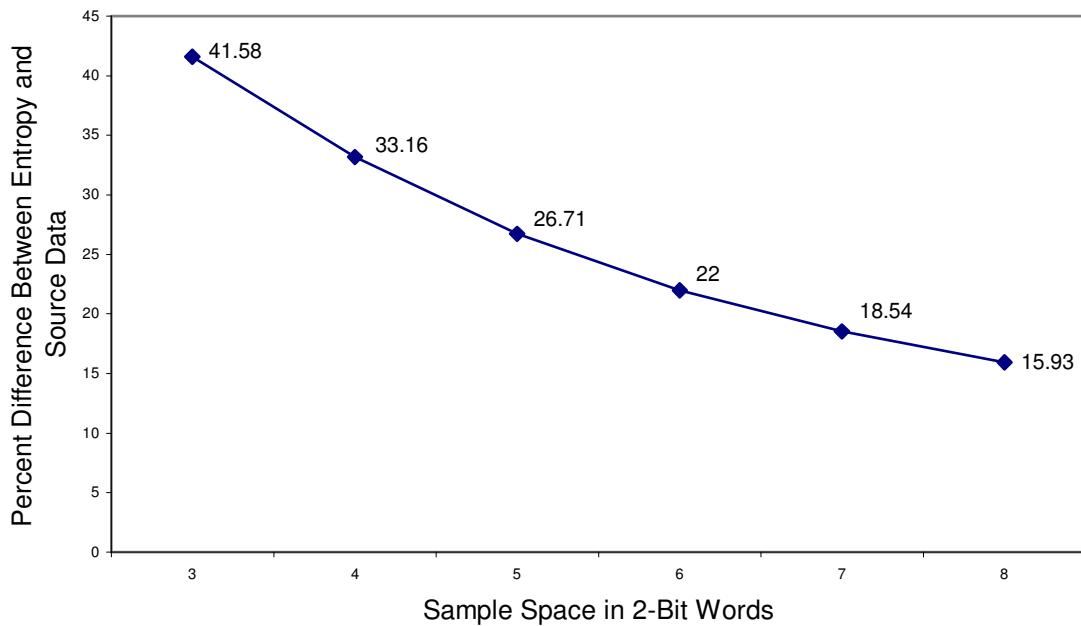
Furthermore, the percent difference between the size of the original string of random binary digits and the calculated entropy varies with sample size. The following graph shows the percent difference between the calculated entropy for a series of 1-bit words with different sample sizes. In this instance the percentage difference increases from approximately 8% for a 3-bit sample space (i.e., a random string of bits divided into groups of 3 1-bit words with each group having 2^3 or 8 possible combinations) to 12.24% for a 5-bit sample space (i.e., a random string of bits divided into groups of 5 1-bit words

with each group having 2^5 or 32 possible combinations) after which it gradually decreases as the sample space increases.



3.0 Varying Sample Space and Sample Size

The calculated entropy for a random binary string varies with sample size as well as with sample space. In all cases, the same number of total bits will have different values of entropy based on sample space and sample size. What this means is that any change in the number of bits per word will change the entropy and any change in the number of words in a block (i.e, the sample size) will change the entropy. For example, the total possible combinations of binary symbol strings 6 bits in length represent a sample space of 2^6 or 64. Multiplying 64 by 6 (bits) results in a total of 384 bits. Each 6-bit binary symbol string can in turn be organized into 6 1-bit words, 3 2-bit words and 2 3-bit words. Even though the total number of (random) bits is the same for all three models, the percent difference between the entropy and the actual number of bits for each model is 11.48%, 41.58% and 66.67% respectively. The following graph shows the percent difference between the calculated entropy for a series of 2-bit words with different sample sizes.



In the above graph, the same data is subdivided into groups of 3, 4, 5, 6, 7 and 8 2-bit words. When calculated as a series of 3 2-bit words, the percent difference between the actual number of bits in the source data and the calculated entropy is 41.58%. When calculated as a series of 4 2-bit words, the same data results in a percent difference of 33.16%. The percent difference between the actual number of bits and the calculated entropy decreases as the number of words used in the calculation increases.

By changing the sample size (i.e, the number of binary words used in the calculation of entropy) and the sample space (i.e, the total possible combinations of bits based on the number of bits in a word), it is possible to vary the calculation of entropy and maximize the percent difference between the number of bits in the data and the calculated entropy for that data. The following table shows the total bits, calculated entropy and percent difference for various combinations of 1, 2, 3 and 4-bit words.

Word Length In Bits	Block Length in Bits	Words per Block	Total Bits	Entropy	Percent Difference
1	3	3	24	22.03	8.17
1	4	4	64	56.45	11.79
1	5	5	160	140.41	12.24
1	6	6	384 (1)	339.89	11.48
1	7	7	896	804.49	10.21
1	8	8	2048 (2)	1857.75	9.28
2	6	3	384 (1)	224.31	41.58
2	8	4	2048 (2)	1368.74	33.16
2	10	5	10240	7504.51	26.71
2	12	6	49152 (3)	38335.14	22
2	14	7	229376	186843.74	18.54

2	16	8	1048576	881495	15.93
3	6	2	384 (1)	128	66.67
3	9	3	4096	2082.5	54.8
3	12	4	49152 (3)	26960.86	45.14
3	15	5	491520	304601.54	38.02
4	8	2	2048 (2)	512	75
4	12	3	49152 (3)	18004.01	63.37
4	16	4	1048576	47476447.22	54.56

In this graph, numbers in parenthesis in the Total Bits column show how even random data can have different values of entropy when divided into different word and block lengths. With few exceptions, for a given word length in bits, the larger the resulting value of Word Length in Bits divided by Words per Block, the greater the value of Percent Difference.

4.0 Summary

Since any collection of data can be represented in binary form, it follows that any collection of data can be shown, using the above described methods, to have an information content or entropy that is smaller in size than the original collection of data. This is true even if the original collection of data is random in nature. This means that all information, even that information that describes a physical model such as the universe, can be shown to have multiple expressions, many of which occupy a smaller space than the original expression. The same methods can be applied in turn to these resulting expressions. The ability to reiteratively and recursively reduce the entropy for any collection of data supports the idea that all information is in some sense holographic, that is, it contains within itself the information necessary to describe much "larger" and much "smaller" systems of information as well as the information necessary to establish models of equivalence among those systems. It also supports the corollary idea that any collection of information can act as a source for the generation of an infinite number of classes, attributes and relations.