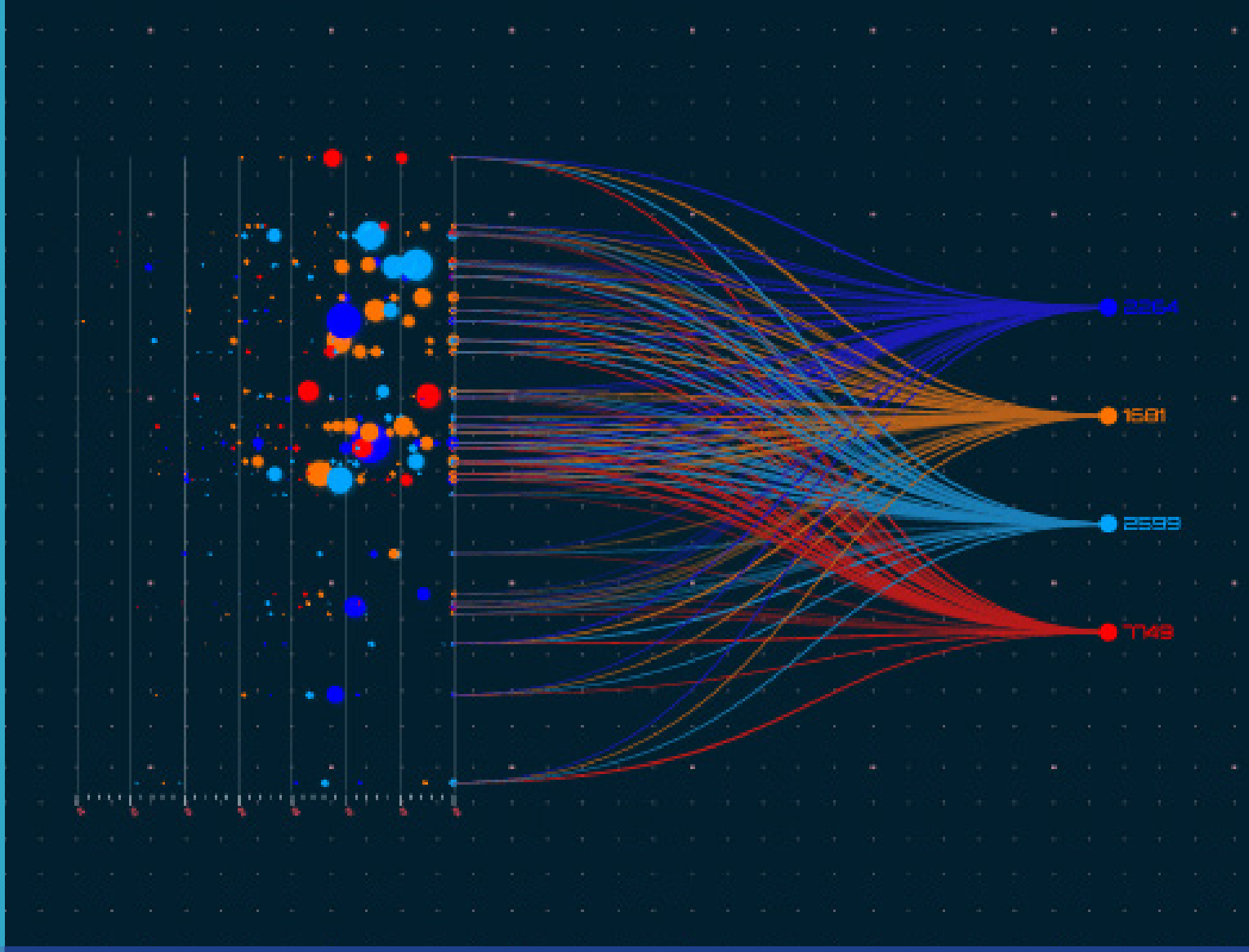


NOVEMBER 2025

Measuring Statistical Bias in Data Using Entropy

American Academy of Actuaries
Data Science and Analytics Committee

American Academy of Actuaries

actuary.org

Drafted by members of the
Data Science and Analytics Committee:
Dorothy Andrews MAAA, ASA, PhD
Dave Sandberg MAAA, FSA
Paul Meixler MAAA,EA, FCA
Robert Gomez MAAA, FSA
Leonard Reback MAAA, FSA
Alexander Esche MAAA, ASA

The American Academy of Actuaries is a 20,000-member professional association whose mission is to serve the public and the U.S. actuarial profession. For 60 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.

American Academy of Actuaries

AMERICAN ACADEMY OF ACTUARIES
1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036
202-223-8196 | WWW.ACTUARY.ORG

November 2025

Any references to current laws, regulations, or practice guidelines are correct as of the date of publication.

Measuring Statistical Bias in Data Using Entropy

Introduction

Statistical, cognitive, and social systemic biases are the latest threats to the stability of actuarial models, garnering unwanted attention from regulators. Measuring statistical bias poses fewer challenges than measuring social systemic or cognitive biases, the latter of which is nearly impossible to measure. Some of the bias testing approaches can detect bias toward protected classes, a primary interest of regulators, but such testing requires identifying protected class attributes and attaching them to outcomes for bias testing. Insurers do not collect race and many other protected class attributes, and existing inference methods have proven to be unacceptably inaccurate by most standards. This paper will not solve all these issues, but, instead, will demonstrate how entropy, a measure of information content, can be used to quantify the level of homogeneity in a data field. The more homogeneous the values in a data field, the more biased the elements in the field toward one value, and the lower the entropy metric. Similarly, the less homogeneous the values in a field, the more diverse the values, and the higher the entropy metric. High entropy may be appropriate depending on the expected values for a field. Likewise, low entropy may be appropriate and desired for a given field. This paper simply demonstrates how entropy metric can be used to quantify the level of bias or diversity in data, against acceptable tolerances.

Foundations of Entropy

Entropy has its roots in the information theory that underpins early digital communication. History recognizes four pioneers in the founding of information theory: Harry Nyquist, Ralph Hartley, Nobert Weiner, and Claude Shannon. Unsurprisingly, all four had connections to Bell Labs, which is considered the birthplace of information theory.[1] Harry Nyquist authored “Certain Factors Affecting Telegraph Speed,” in 1924. The paper theorized how to transmit the maximum amount of information over a circuit[2] and contains a theoretical section quantifying "intelligence" and the "line speed" at which it can be transmitted by a communication system. Ralph Hartley authored “Transmission of information” in 1928.[3] In the paper, Hartley derives the following formulaic measure of information:

$$H = n \log S = \log S^n.$$

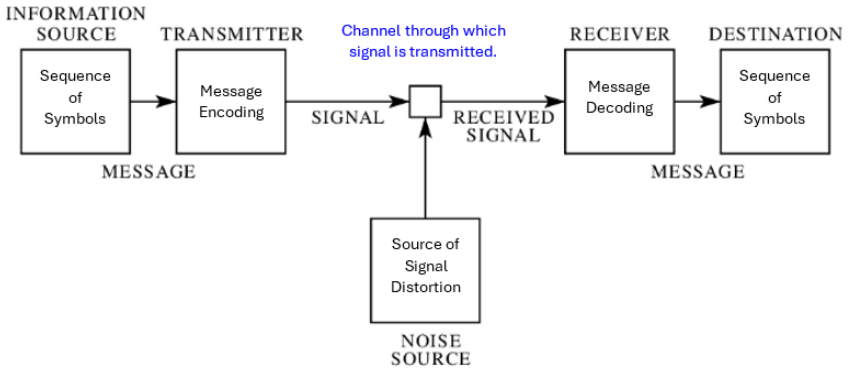
In this formula, S represents the number of possible symbols that can form a transmission, n represents the actual number of symbols in a transmission, and H is the measure of information in a transmission. Nobert Weiner authored, “Cybernetics: Or Control and Communication in the Animal and the Machine,” in 1948, which describes the probability density function for

[1] Georgescu, I. (2022). Bringing back the golden days of Bell Labs. *Nature Reviews Physics*, *4*(2), 76-78.
[2] Nyquist, H. (1924). Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, *43*, 412-422.
[3] Hartley, R. V. (1928). Transmission of information 1. *The Bell System Technical Journal*, 7(3), 535-563.

continuous information.[4] Weiner believed that information was measurable and could be studied using statistics.[5]

Despite the contributions of Nyquist, Hartley, and Weiner, Shannon is widely acknowledged as the "Father of Information Theory," mainly because he devised a complete framework for describing digital communications.[6] Shannon published "A Mathematical Theory of Communication"[7] in 1948 clearly distinguishing the delivery of a message from its meaning as depicted below.

The process starts with a sequence of symbols that the sender wishes to transmit to the receiver at a specific destination. The symbols must be digitized and passed through an encoding algorithm.

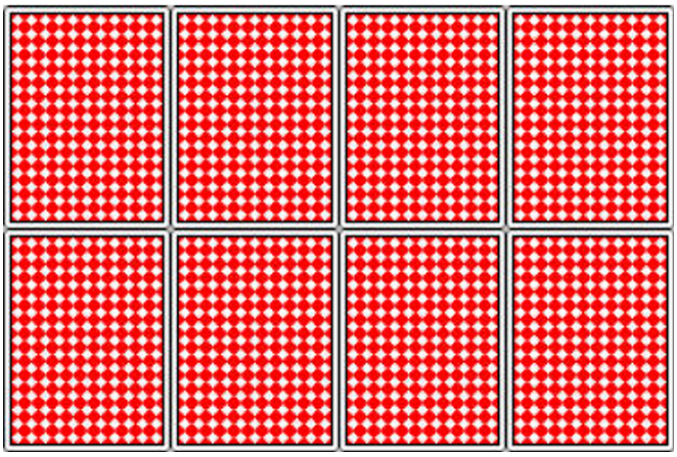


The encoded message is then passed through a medium such as wires, cable, phone lines, etc., to a decoding algorithm that converts the signal back into the original sequence of symbols. There is the potential for noise to interfere with the transmittal of the encoded message which would result in a distorted message once decoded.

The encoding of a message involves determining the least amount of information needed to correctly transmit a message and have it correctly interpreted after transmission. Shannon theorizes that any message be correctly determined using a series of questions where the only two responses are "yes" or "no," which can be encoded as either a "0" or a "1," i.e., binary digits or bits. The first use of the term "bit" can be traced back to Shannon's treatise on the theory of communications.[8] The minimum number of bits required to reduce the uncertainty of a message by one-half at each stage of questioning until uncertainty is zero, is the measure of information contained in a message. For example, let's say we have eight cards from a deck all face down—only one card is a King, and we seek to determine which one.

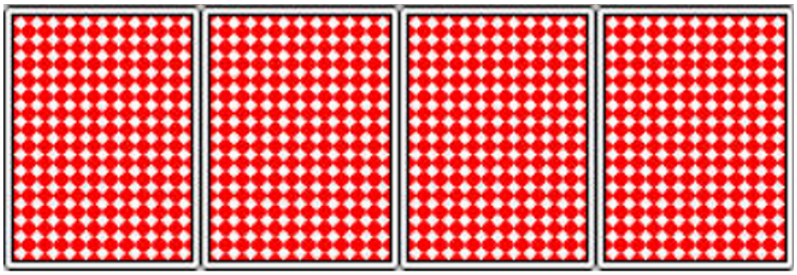
---

[4] "Norbert Wiener Issues *Cybernetics*, the First Widely Distributed Book on Electronic Computing"; History of Information; 2012.
[5] "Cybernetics"; *Bulletin of the American Academy of Arts and Sciences*; April 1950.
[6] "Claude Shannon"; *Encyclopedia Britannica*; April 26, 2025.
[7] "A Mathematical Theory of Communication (Shannon, C. E.)"; *The Bell System Technical Journal*; July and October 1948.
[8] "A Mathematical Theory of Communication (Shannon, C. E.)"; *The Bell System Technical Journal*; July and October 1948.

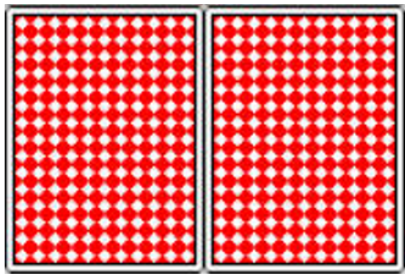


We could turn over one card at a time until the King is found, but it could take one, if you are lucky, or seven tries before the King is found. Shannon proposed a more systematic way to determine the maximum number of tries needed to determine the King with certainty. The approach might ask the following questions:

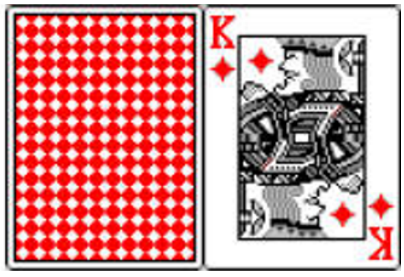
1. Is the King in the top row? If the answer is “No,” then it must be in the bottom row.



2. Is the King one of the two right most cards? If the answer is “No,” then it must be one of the two left most cards.



3. Finally, is the King the left card? If the answer is “No,” then the King must be the right card.



To summarize, we started out with eight cards and asked three "yes" or "no" questions before determining where the King was with certainty. It is no coincidence that we have

$$8 = 2^3.$$

We can rewrite this expression as:

$$\log_2(8) = \log_2(2^3)$$

$$\log_2(8) = 3$$

The calculation says that it takes three bits ("yes" or "no" responses coded as "1" or "0") of information to determine a message with certainty given eight symbols. It is noteworthy that the maximum number of guesses needed is far less than seven.

Shannon's formulation of the same problem defines information in terms of probabilities. We have

$$I(x) = log_s \frac{1}{p(x)} = -log_s p(x)$$

Since the probability of selecting the King is 1/8, the information associated with the selection is

$$I(King) = log_s \frac{1}{p(King)} = -log_s \left[\frac{1}{8}\right] = 3$$

The information associated with not picking the King is

$$I(NotKing) = log_s \frac{1}{p(NotKing)} = -log_s \left[\frac{7}{8}\right] = 0.19265$$

The information is higher when the probability is lower.

What Is Entropy?

The most basic definition of entropy is a measure of disorder or randomness in a system. It is a measure that has been used in thermodynamics, information theory, statistical mechanics, dynamical system theory, fractal geometry, biology, machine learning, economics and finance, and other fields.[9] As previously discussed, the higher the entropy, the more uncertainty (disorder or randomness) in the system. The importance of entropy in the field of information theory has already been discussed. The following discussion will provide insights into the utility of entropy in thermodynamics, statistics, finance and economics, and machine learning. Later in this paper, entropy is described as a measure for diversity (and bias as an opposite concept). The entropy of mortality tables will be discussed in this context.

[9] "Entropy – A Guide for the Perplexed"; *PhilSci Archive* journal; Frigg and Werndl (2010).

Entropy in Thermodynamics

Thermodynamics studies the relationships between work, heat, temperature, and energy. Four laws govern the interplay between these four elements. They are:[10]

1. The zeroth law of thermodynamics—Two systems in thermal equilibrium with a third are in thermal equilibrium with each other.
2. The first law of thermodynamics—Energy cannot be created or destroyed. It is conserved.
3. The second law of thermodynamics—Systems tend toward disorder and will never reverse toward an ordered state. Disorder states have higher entropy than ordered states.
4. The third law of thermodynamics—The entropy of a perfect crystal at absolute zero temperature is zero.

There are three types of systems that are important in thermodynamics: Open, closed, and isolated. Each system type plays a significant role in determining how energy and matter is exchanged within a system and the surrounding environment. An open system allows energy and matter to move from inside the system to outside the system. Boiling water in a kettle is a good example.[11] Water from inside the kettle escapes through the spout of the kettle in the form of steam.

The following graph illustrates the piecewise change in entropy (S(cal/K) for one mole of water that changes from a solid to a liquid and then to a vapor as the temperature increases (Celsius).[12]

A closed system allows energy, but not matter, to be transferred within and with the outside surrounding environment. A closed bottle of water does not allow the water to escape, but the water can change states. It can transition from hot to cold and cold to hot depending on the outside environment. Finally, an isolated system does not allow the exchange of energy or matter with the surrounding environment. The universe is an example of an isolated system, because neither energy nor matter is able to enter or leave. [13]

To gain an appreciation of these four laws, researchers have examined how to apply them to the state of a runner's body,[14] which represents an open system, to understand how the runner's body consumes and transforms energy. A runner's body needs and converts several types of energy from the start to the completion of a run—chemical energy, potential energy, and kinetic energy. The body derives chemical energy from proteins, carbohydrates, and fats.[15] When the body engages in work, it burns chemical energy in the form of calories, giving off heat whose unit is joules. One calorie is roughly 4.186 joules. Running, like any activity, requires chemical energy

---

[10] "Thermodynamics"; *Encyclopedia Britannica*; August 2025.
[11] "30 Examples of Open, Closed and Isolated Systems"; Examples Lab; 2018.
[12] Masterton WL, Solwinski, EJ, (1973), Chemical Principles, WB Saunders Company, p 333.
[13] "The Equation of the Universe (According to the Theory of Relation)"; *Journal of Modern Physics*; 2019.
[14] "Entropy Measures Can Add Novel Information to Reveal How Runners' Heart Rate and Speed Are Regulated by Different Environments"; *Frontiers in Psychology*; June 4, 2019.
[15] "How the Body Uses Energy"; Rockets Sports Medicine Institute; June 3, 2019.

to create the necessary biological work to move the body from point A to point B. During the running process, the human body converts potential energy into kinetic energy back into potential energy and the pattern repeats throughout the run. Potential energy is associated with the position of an object, while kinetic energy is associated with the motion of an object. An object at rest stores potential energy, while an object in motion stores kinetic energy. The two forms of energy are inversely related. When one is high, the other is low.

The human body is in a resting position at the start of a run, harboring a store of potential energy. As the runner commits work to start the run cycle, potential energy is transformed into kinetic energy as the runner climbs in flight to reach the double float position, when both feet are off the ground. The shift back to potential energy occurs during the double float phase of the run, and, almost as fast as it occurred, potential energy is converted back to kinetic energy as gravity brings one foot back in contact with the ground. The process repeats until the run is over and the body returns to a resting, and possibly exhausted, position. Since entropy is a measure of disorder, kinetic energy has higher entropy than potential energy, because an object in motion is more disordered than an object at rest. Researchers found that runners exhibit higher entropy when running on a 400-meter track than running along routes that are more familiar or unusual.[16] The researchers hypothesized that running monotonous routes caused runners to vary speeds more in an attempt to increase their arousal levels to stay engaged. This suggests routes that have more to offer visually, or more unusual terrains, are best for the training regimen of runners, because high entropy was also found to produce more erratic heart rates.

Entropy in Statistics

Rudolf Carnap's critique of classical thermodynamics included the argument that entropy in thermodynamics has the same character as other thermodynamic concepts such as heat, pressure, temperature, etc., which serve "for the quantitative characterization of some objective property of a state of a physical system." From this starting point, his aim was to construct a statistical concept of entropy. He classified concepts along the spectrum of entropy from physical property to a nonphysical concept.[17]

Carnap's system of conceptual classification assigns the elements to be classified into k cells C_j (j=1, … ,k) each with an individual description $D^{concept}$. A quantitative classification description D^{quant} corresponds to each individual description $D^{concept}$. D^{quant} assigns the number n_j of elements classified into each cell C_j. Of particular interest is when there is a uniform distribution of n_j of elements across each cell C_j. Carnap labels this the degree of order of $D^{concept}$ that has the uniform distribution.

[16] Exel, J., Mateus, N., Gonçalves, B., Abrantes, C., Calleja-González, J., & Sampaio, J. (2019). Entropy measures can add novel information to reveal how runners' heart rate and speed are regulated by different environments. *Frontiers in Psychology*, *10*, 1278.

[17] Carnap, R (1952 and 1954) in *Two Essays on Entropy* edited by Shimony, A, (1977), University of California Press.

In Carnap's unpublished paper, "The Concept of Degree of Order," the term randomness (as defined in Statistics) refers to a characteristic of the *procedure of selection* of a sample from a given population. A procedure is defined as random if all possible samples of the same size have uniform distribution, hence randomness is not a characteristic of the mathematical structure of the sample. On the other hand, he concludes that the degree of disorder is a characteristic of mathematical structure. In other words, entropy is a mathematical function.

Entropy, the uncertainty (disorder, diversity, or randomness) function, is a real number associated with any probability distribution on a finite set. A *probability distribution* is defined for $n \geq 1$ on a finite set [1, …, n] is a **p** = (p_1, …, p_n) of real numbers $p_i \geq 0$ such that $\sum p_I = 1$. For $n \geq 1$, write

$$\Delta n = [probability\ distribution\ on\ [1, \ldots, n]],$$

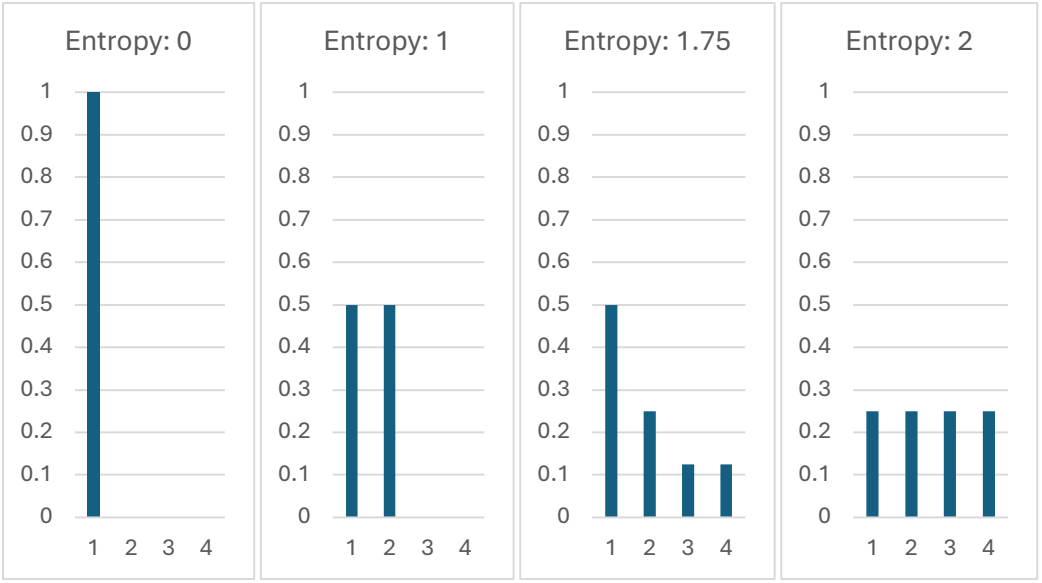
where n is the number of outcomes of the distribution. Appendix A has a coin toss example (n=2), dice example (n=6), and an alphabet example (n=4).

The (Shannon) Entropy of **p** is

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

In case of comparing entropies with different n's, the normalized entropy is $H(\boldsymbol{p})$ / log n, which produces results between zero and one that may be used for comparisons. The normalized entropy is used below in the comparison of the deferred distribution of death n | q(x) at different ages of life table.

Below are four sample distributions with entropy in the range from zero to two.



Decreasing Entropy Increasing Entropy

Certainty	Uncertainty
Order	Disorder
High Bias	Low Bias
Non-Randomness	Randomness
Higher Information	Lower Information

The higher the entropy, the more uncertainty (disorder, diversity, or randomness) in the system. The highest entropy equals two (log base 2 of 4 is 2) for the uniform distribution, which has the most disorder, diversity, or randomness. The lowest entropy above equals zero for a distribution completely in a single classification, which has the most order, bias, and non-randomness. Entropy is a measure for both bias and diversity, but as illustrated above, bias is the opposite of diversity.

Appendix A provides further examples of the calculation of entropy for a coin toss example, dice example, alphabet example, and continuous probability distributions.

Entropy in the Stock Market

There are many examples of the use of entropy in finance and economics. The randomness of the stock market has been extensively studied. All sorts of factors, systematic and unsystematic, influence the price of a stock at any moment in time. Volatility—the standard deviation of a stock's price (or return), denoted by σ (sigma)—is a familiar metric that has been used to estimate the randomness of a stock's price. A companion metric is Beta (β), which measures a stock's systemic risk relative to the stock market as a whole. Since entropy is a measure of disorder, it is not unreasonable to examine its utility to measure the randomness of a stock's price, and many researchers have. Approximate Entropy is "a statistical measure of the level of randomness of a data series which is based on counting patterns and their repetitions."[18] The presence of patterns suggests non-randomness and order. Therefore, low approximate entropy implies more predictability in the movement of a stock's price, whereas high approximate entropy implies less predictability. One issue to be aware of is that approximate entropy is very sensitive to sample size. A small sample size can adversely impact the statistical significance of the Chi-square measure associated with approximate entropy. A sample size of N > 200 is recommended.

[18] "Quantifying the randomness of the stock markets"; Scientific Reports; Sept. 4, 2019.

Approximate Entropy (ApEn) is a measure designed to find patterns of length *m* in a sequence of numbers or time series. When APEn is low, no pattern was discernible. When ApEn is high, a pattern is discernible, predictable, and non-random. Formulaically, it is defined as [19]

$$ApEn(m, r, n) = \phi^m(r) - \phi^{m+1}(r)$$

where,

$$\phi^m(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} log\left(C_i^m(r)\right)$$

$$C_i^m(r) = \frac{number of j such\ that\ d[x(i), x(j)] \leq r}{n}$$

$$d[x(i) = [u(i), u(i+1), \ldots u(i+m-1)]$$

$$where = 1 \leq i \leq n$$

- *x(i) is an m-dimensional vector that contains the run of data starting with u(i).*
- *u(i) represents a data time series equally spaced in time.*

It is necessary to understand the components and parameters of the approximate entropy formula to fully understand how it measures the entropy of a stock's return. Sample size is denoted by n. The r parameter is a similarity threshold for pattern acceptance,[20] where a value greater than r indicates values are related and a value less than r, indicates they are not. In the example below, this parameter will be set equal to the median, since a stock can go up or down. The approximate entropy of a stock is an indication of the predictable up and down pattern in its price. The parameter m is defined as the length of the data segments being compared for similarities The data segment in the time series is defined as

$$x(i) = [u(i), u(i+1), \ldots u(i+m-1)].$$

If $m = 2$, then the data segment is

$$x(i) = [u(i), u(i+1)]$$

Here $x(i)$ represents two consecutive elements in the time series. The function $C_i^m(r)$compares all pairwise differences d. It is called a correlation integral, and it is a measure of closeness between data segments. The numerator of $C_i^m(r)$ counts the number of data segments of consecutive values of length m less than the threshold r.[21]

---

[19] Leinster, T (2021). *Entropy and Diversity, The Axiomatic Approach. p41*. Cambridge University Press.
[20] "Approximate entropy"; *Wikipedia*; 2025.
[21] "A comprehensive comparison and overview of R packages for calculating sample entropy"; *Biology Methods and Protocols*; 2019; Chang Chen, Shixue Sun, Zhixin Cao, Yan Shi, Baoqing Sun, Xiaohua Douglas Zhang.

Example

Let's say a time series of 10 returns is given by:

U = {0.97%, 0.37%, -0.12%, 0.07%, 0.27%, 0.26%, -0.20%, 0.16%, 0.07%, 0.08%}

The sample size is purposefully small to facilitate the application of the formulas. Assume m = 2 and r = 0.2%

We can form the sequence of x(i) vectors as:

$$\mathrm{x}(1) = [\mathrm{u}(1), \mathrm{u}(2)] = [0.97\%, 0.37\%]$$
$$\mathrm{x}(2) = [\mathrm{u}(2), \mathrm{u}(3)] = [0.37\%, -0.12\%]$$
$$\mathrm{x}(3) = [\mathrm{u}(3), \mathrm{u}(4)] = [-0.12\%, 0.07\%]$$
$$\mathrm{x}(4) = [\mathrm{u}(4), \mathrm{u}(5)] = [0.07\%, 0.26\%]$$
. . .

The next step is to calculate all the differences d for all values of dimension equal to 1, m, and $m+1$. The values of d for dimension equal to one are needed to define the r, while the ApEn formula is calculated using $\phi(m)$ and $\phi(m+1)$. Therefore, we need to define x(i) for each ϕ function as follows:

$\phi(m)$: $x(i) = [u(i), u(i+1), \ldots u(i+m-1)]$

$\phi(m+1)$: $x(i) = [u(i), u(i+1), \ldots u(i+(m+1)-1)]$

For m =2, we have,

$\phi(2)$: $x(i) = [u(i), u(i+1)]$

$\phi(3)$: $x(i) = [u(i), u(i+1), u(i+2)]$

The calculations to fill in the lower triangle of a correlation matrix is depicted below.

$d[\mathrm{x}(1), \mathrm{x}(1)]$					
$d[\mathrm{x}(2), \mathrm{x}(1)]$	$d[\mathrm{x}(2), \mathrm{x}(2)]$				
$d[\mathrm{x}(3), \mathrm{x}(1)]$	$d[\mathrm{x}(3), \mathrm{x}(2)]$	$d[\mathrm{x}(3), \mathrm{x}(3)]$			
$d[\mathrm{x}(4), \mathrm{x}(1)]$	$d[\mathrm{x}(4), \mathrm{x}(2)]$	$d[\mathrm{x}(4), \mathrm{x}(3)]$			
…	…	…		$d[\mathrm{x}(9), \mathrm{x}(9)]$	
$d[\mathrm{x}(10), \mathrm{x}(1)]$	$d[\mathrm{x}(10), \mathrm{x}(2)]$	$d[\mathrm{x}(10), \mathrm{x}(3)]$	…	$d[\mathrm{x}(10), \mathrm{x}(9)]$	$d[\mathrm{x}(10), \mathrm{x}(10)]$

The formula for d will verify that all the diagonal elements are equal to zero. The rest of the matrix is filled by reflecting the lower triangular values in the upper triangle. The matrix arithmetic is the same for dimensions m, and m+1. The only difference is the length of the data segments involved in the arithmetic, as follows:

$$\text{For m} = 2\text{:}\, x(i) = [u(i), u(i+1)]$$
$$\text{For m} = 3\text{:}\, x(i) = [u(i), u(i+1), u(i+2)]$$

For m =1, d involves subtracting vectors of length one. For m =2, d involves subtracting vectors of length two, and for m =3, d involves subtracting vectors of length three. The idea is to look for patterns of length one, two, and three in the time series. Take the simple sequence:[22]

U_{51} = {85, 80, 89, 85, 80, 89, 85, 80, 89, 85, 80, 89, …}

For $\text{m} = 2$, we have

$$d[\text{x}(1), \text{x}(1)] = \max|\{85, 80\} - \{85, 80\}| = \max|85 - 85, 80 - 80| = 0$$
$$d[\text{x}(1), \text{x}(2)] = \max|\{85, 80\} - \{80, 89\}| = \max|85 - 80, 80 - 89| = 9$$
$$d[\text{x}(1), \text{x}(3)] = \max|\{85, 80\} - \{89, 85\}| = \max|85 - 89, 80 - 85| = 9$$
$$d[\text{x}(1), \text{x}(4)] = \max|\{85, 80\} - \{85, 80\}| = \max|85 - 85, 80 - 80| = 0$$
. . .

For $\text{m} = 3$, we have

$$d[\text{x}(1), \text{x}(1)] = \max|\{85, 80, 89\} - \{85, 80, 89\}| = \max|85 - 85, 80 - 80, 89 - 89| = 0$$
$$d[\text{x}(1), \text{x}(2)] = \max|\{85, 80, 89\} - \{80, 89, 85\}| = \max|85 - 80, 80 - 89, 89 - 85| = 9$$
$$d[\text{x}(1), \text{x}(3)] = \max|\{85, 80, 89\} - \{89, 85, 80\}| = \max|85 - 89, 80 - 85, 89 - 80| = 9$$
$$d[\text{x}(1), \text{x}(4)] = \max|\{85, 80, 89\} - \{85, 80, 89\}| = \max|85 - 85, 80 - 80, 89 - 89| = 0$$
. . .

[22] "Approximate Entropy and Sample Entropy: A Comprehensive Tutorial"; National Library of Medicine; May 2019.

Since the simple sequence shows a clear pattern, the difference matrices will also show a pattern, and that pattern will be quantified by the correlation integrals, $C_i^m(r)$. The correlation integrals count the number of differences that are within the similarity tolerance, r, and divide the result by the total number of differences. They are the probabilities of differences being less than the similarity tolerance and they are needed for the ϕ function. Each column of the difference matrix produces a $C_i^m(r)$ for each i in $1 \leq i \leq n - m + 1$. The entropy formula is

$$\phi^m(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} log\left(C_i^m(r)\right)$$

And finally, ApEn is given by

$$ApEn(m, r, n) = \phi^m(r) - \phi^{m+1}(r)$$

In our sample of returns, we have

$$ApEn(2,0.2\%, 10) = \phi^2(0.2\%) - \phi^3(0.2\%)$$
$$ApEn(2,0.2\%, 10) = (-1.27) - (-1.56) = 0.29$$

While beyond the scope of this paper to derive, it is possible to calculate the Chi-Squared statistic for this result and its p-value. The Chi-Squared result is

$\chi^2 = 8.01$ on 4 degrees of freedom, p-value = 0.0911

This result is not statistically significant, which means there is no discernible pattern in the movements of the stock's price. The movement in the small sample appears to be random.

For the simple sequence,

U_{51} = {85, 80, 89, 85, 80, 89, 85, 80, 89, 85, 80, 89, …},

the Chi-Squared result is

$\chi^2 = 56.84$ on 4 degrees of freedom, p-value = 1.337e-11

This result is statistically significant and expected since the repeating pattern in the sequence is quite apparent. This result means the pattern is regular and predictable. Additional discussion of this sequence can be found at the link in the footnote.[23] The additional tables supporting the calculation of the Approximate Entropy of the stock returns can be found in Appendix B.

[23] "Approximate entropy"; *Wikipedia*; 2025.

Machine Learning

The breakthrough of the use of entropy in machine learning is credited to Australian computer scientist, John Ross Quinlan. Quinlan is the inventor of the powerful ID3, C4.5 and C5.0 algorithms that are the engines for many decision tree algorithms. The C5.0 algorithm is an improvement over the C4.5 algorithm, and C4.5 is an improvement of the ID3 algorithm. The major difference between C5.0 and C4.5 is that C5.0 builds a rule set that C4.5 does not. Otherwise, the two algorithms leverage the same mathematics to determine how to make optimal splits in data to create homogeneous terminal nodes in a decision tree. A related concept to entropy that is important to discuss before examining how entropy is applied to decision trees is *information gain*.

Information gain is a measure that aids in determining the best feature to split the data at each node of a decision tree. [24] It is calculated as the difference between the entropy for a node before it is split and the probability weighted entropies of children nodes after a proposed split from their parent node. The feature split resulting in the maximum information gain is the best feature for decision tree node splits.

Let's consider the following data that reflects three features (Age, Mileage, Road Tested) and one outcome (Buy Decision) to build a decision tree for the purchase decision of a used car.[25]

Age	Mileage	Road Tested	Buy Decision
Recent	Low	Yes	Buy
Recent	High	Yes	Buy
Old	Low	No	Don't Buy
Recent	High	No	Don't Buy

The decision tree is trying to determine the best splitting rule that will lead to the most homogenous or pure children nodes. The best splitting rule will result in all the "Buy" decisions in one terminal node and all the "Don't Buy" decisions in the other node after the split. The parent node reflects all the decisions before any rules are applied.

---

[24] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

[25] "Decision tree: Part 2/2. Entropy and Information Gain"; *TDS Archive*; Sept. 6, 2019.

We can calculate the entropy of

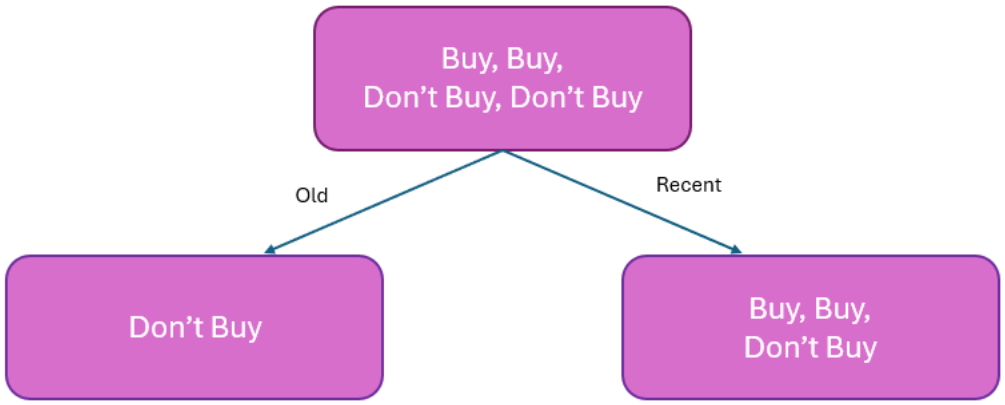
Buy – 2 Instances Don't Buy – 2 Instances Prob(Buy) = 1/2 Prob(Don't Buy) = 1/2

the root node as:

$$Entropy = -Prob(Buy) * log_2(Buy) - Prob(Don'tBuy) * log_2(Don'tBuy)$$
$$= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right)$$
$$= 1.0$$

This value is important for calculating information gain, which is used to determine the best split of the parent node into two child nodes. There are three possibilities for splitting the tree at the parent node. We can split on the Age, Mileage, or Road Tested variables and for each of these variables, there are two choices. It is possible to use continuous variables for splitting decisions as well. For simplicity, this example will focus on the binary choices for each of the three variables. The one that leads to the largest information gain is the optimal variable for splitting the tree at the parent node.

For the Age variable, the visualization of the split is depicted below. Information gain is a splitting criterion based on entropy.



The weighted entropy for the child nodes is given by

$$ChildEntropy = \frac{1}{4} * (0) + \frac{3}{4} * (0.918) = 0.688$$

The information gain is given by

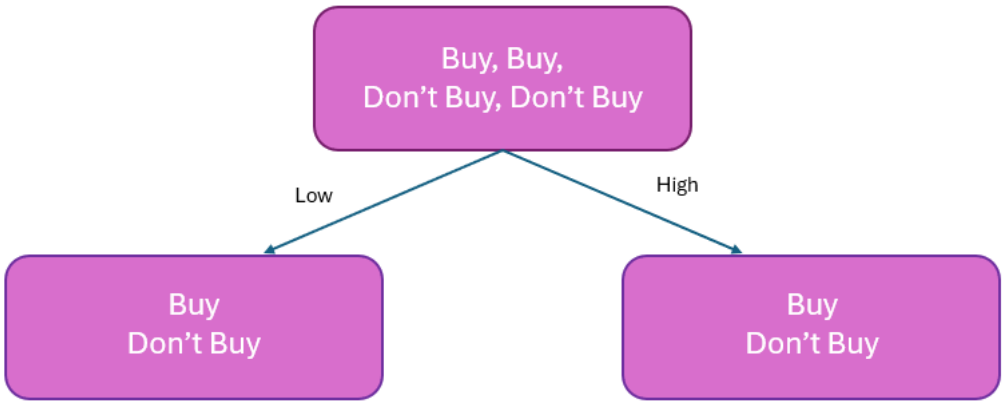
$$InformationGain = 1 - 0.688 = 0.3112$$

We repeat this exercise for the Mileage and Road Tested variables.

The Mileage binary outcomes are

1. Low Mileage
2. High Mileage

If the parent node is classified by the Mileage variable, the composition of children nodes is depicted below.



The weighted entropy for the child nodes is given by

$$ChildEntropy = \frac{1}{2} * (1) + \frac{1}{2} * (1) = 1.00$$

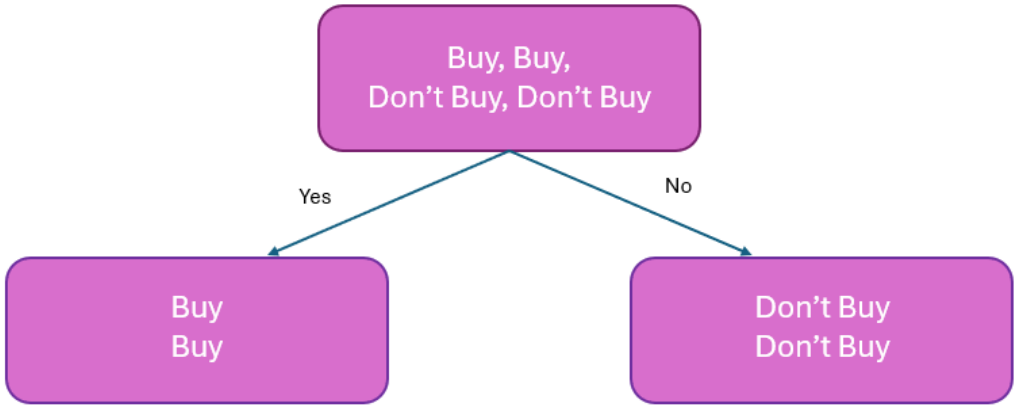
The Information Gain is given by

$$InformationGain = 1 - 1 = 0.0$$

Finally, the Road Tested variable binary outcomes are

1. Yes
2. No

If the parent node is classified by the Road Tested variable, the composition of children nodes is depicted below.



The weighted entropy for the child nodes is given by

$$Child\ Entropy = \frac{1}{2} * (0) + \frac{1}{2} * (0) = 0.00$$

The information gain is given by

$$Information\ Gain = 1 - 0 = 1.0$$

In summary, the information gain from the three variables is summarized below.

Variable	Information Gain
Age	0.3112
Mileage	0.000
Road Tested	1.000

The maximum information gain results from splitting the parent node on the Road Tested variable. This means the most homogeneous groupings on the children node occurs when we split on the Road Tested variable. Whether a car is road tested or not is the best decision rule for predicting buying decision. There is no need to consider the age of the vehicle or the mileage in establishing a decision rule that optimally segments the data. Segmentation like in this example is the goal of a decision tree analysis. The end product is a set of rules that segment the data into homogeneous groups, and entropy is the measure from developing the decision rules.

When there are two outcomes on a variable, the maximum entropy is 1.0. When there are four outcomes, the maximum entropy is two. For eight outcomes, the maximum is 3.0. The pattern follows the following mathematical relationships:

$$\text{Log}_2(2) = 1$$

$$\text{Log}_2(4) = 2$$

$$\text{Log}_2(8) = 3$$

$$\text{Log}_2(16) = 4$$

.
.
.

$$\text{Log}_2(X) = Y$$

The general formula works for all values of $X \subset \{1, 2, 3, 4, 5, 6, \ldots\}$ not just those values that are powers of two.

Maximum Entropy Classifier Model

The maximum entropy classifier model is a generalization of the naïve Bayes classifier model.[26] Instead of using probabilities for the parameters, it uses iterative optimization to find the parameters. It lowers entropy (increases information) with each step in the iteration.

The intuition related to maximum entropy classification is that the model would capture frequencies of the joint variables without making unwarranted assumptions. For example, the model would initially select Distribution 1 even though any of these distributions are correct. Distributions 2 and 3 would reflect assumptions currently not known.

Distribution	A	B	C	D	E
1	20%	20%	20%	20%	20%
2	5%	25%	35%	25%	10%
3	0%	100%	0%	0%	0%

[26] Bird, Klein, Loper. Natural Language Processing with Python. O'Reilly. 2009. Pages 252-253.

The maximum entropy principle states that the distribution chosen, among the distributions consistent with the known information, is the distribution that has the highest entropy. That is, the remaining variables are initially set to the uniform distribution. Again, the model would select Distribution 4 when variable A is known to have a 60% frequency. Distributions 5 and 6 would reflect assumptions currently not known.

Distribution	A	B	C	D	E
4	60%	10%	10%	10%	10%
5	60%	5%	15%	15%	5%
6	60%	3%	5%	7%	25%

The iterative optimization continues to find all the parameters. The calculation of maximum entropy for the distribution uses iterative optimization to find the parameters.

Hospitalization Decision Tree

The goal of using entropy to build classification trees is to split the data such that each subset has lower entropy (more information) than the original set. This has particular importance in health care.

At a health insurance company, nurses must efficiently triage thousands of pre-certification requests for hospital admissions. Traditional guidelines relying on fixed criteria may not capture the complex interplay between clinical factors. Entropy-based classification trees can enhance this process by quantifying the predictive value of specific clinical indicators. Instead of applying uniform and often narrative-based guidelines, we can develop more sophisticated review triggers based on the information gained from various clinical factors. Then we can systematically test these more objectively derived guidelines for biases against certain populations.

From a risk classification perspective, these methods can substantially complement traditional actuarial approaches. While actuaries have historically relied on age-sex factors and broad diagnostic categories for risk adjustment, entropy-based splitting can identify more nuanced clinical patterns. For instance, we might discover that for diabetic patients, the combination of HbA1c levels and medication adherence creates more homogeneous risk pools than the standard complications/no-complications dichotomy used in many risk adjustment models. The implications for pricing follow a similar reasoning.

The following hospitalization decision tree example demonstrates how entropy can systematically identify the most predictive factors for inpatient admission decisions. While this example uses simplified disease categories and severities, in practice, such trees could incorporate more granular clinical indicators like vital signs, lab values, and functional status measures to create evidence-based protocols for medical necessity determination.

Suppose we have the following dataset:

Disease	Severity	Hospitalization?
Disease B	Low	No
Disease B	Low	No
Disease A	Low	Yes
Disease C	Medium	Yes
Disease C	High	Yes
Disease C	High	No
Disease A	High	Yes
Disease B	Medium	No
Disease B	High	Yes
Disease C	Medium	Yes

1. We calculate entropy for the entire dataset.

H(Hospitalization?) = -(P(Yes)$\log_2$P(Yes) + P(No)$\log_2$P(No)) = 0.97

2. We calculate the information gain for both features to determine which one provides a better split.
    a. Feature 1 (Disease: A, B, C)
        i. H(A) = 0
        ii. H(B) = 0.811
        iii. H(C) = 0.811
    b. Feature 2 (Severity: Low, Medium, High)
        i. H(Low) = 0.918[27]
        ii. H(Medium) = 0.918
        iii. H(High) = 0.811

Information Gain = H(Hospitalization) – $\Sigma \frac{Numberofsamples \in value}{Totalsamples} xH(Factor)$

IG(Disease) = 0.970 – (2/10 * 0 + 4/10 * 0.811 + 4/10 * 0.811) = 0.322

IG(Severity) = 0.096

[27] For the entropy values, we have -1*$\log_2$(1) = 0 and –((2/3)*$\log_2$(2/3) + (1/3)*$\log_2$(1/3)) = 0.918.

3. Since “Disease” has the higher information gain, we choose it as the root node for the decision tree and split it along its values.
    a. Branch 1: Disease A – since the entropy is 0 (pure subset with all “Yes”), this leaf node predicts “Yes” for all diseases of this type.
    b. Branch 2: Disease B – since the entropy is not 0, we check if we gain additional information by splitting up by Severity.
        i. H(Low | Disease B) = 0 – this leaf node predicts “No” for Disease B if the severity is low.
        ii. H(Medium | Disease B) = 0 – this leaf node predicts “No” for Disease B if the severity is medium.
        iii. H(High | Disease B) = 0 – this leaf node predicts “Yes” for Disease B if the severity is high.

Calculating the information gain again, we see there is no further improvement to be made, since splitting on “Severity” perfectly classifies the data in this branch.

    c. Branch 3: Disease C – since the entropy is not 0, we check if we gain additional information by splitting up by Severity.
        i. H(Low | Disease C) = 0
        ii. H(Medium | Disease C) = 0
        iii. H(High | Disease C) = 1

The IG from splitting Disease C by Severity is 0.311. Since the IG is positive, we should split Disease C by Severity. Since we cannot split the Disease C & High Severity category further without additional factors, we might choose the majority class (or in case there is an even split like in this case), arbitrarily choose “Yes”.

Final Decision Tree Structure

- Root Node: Disease
    - Disease A: Predict "Yes"
    - Disease B:
        - Severity
            - Low: Predict "No"
            - Medium: Predict "No"
            - High: Predict "Yes"
    - Disease C:
        - Severity
            - Medium: Predict "Yes"
            - High: Predict "Yes" (with note on uncertainty due to mixed class)

This example illustrates how entropy helps in selecting the most informative branch to split on when creating classification trees in a care management setting. We can then inspect these trees for bias in an open and transparent way that we can’t when all the clinical decision making is given to an individual health care professional, with their own hidden biases.

Entropy Balancing Study

In *How do Private Equity Buyouts Affect Employee Pension Plans?* Wensong Zhong uses entropy balancing to analyze the impact of private equity (PE) buyouts on the defined benefit plans of target firms from an Econometrics viewpoint.[28]

A pension buyout is a financial arrangement where an employer with a defined benefit or pension plan makes a one-time payment to an insurance company and transfers some or all the responsibility of paying out future pension benefits to the employees. It releases the employer from their pension liabilities for those employees.

The following is his abstract and paper referring to the third person instead of the first person.

> Using data from the Form 5500 filings, he finds that following a buyout, DB (Defined Benefit) plans are more likely to be frozen or terminated. Regarding the actuarial assumption the pension characteristics, he finds an increase in the pension liability discount rate and decreases in the projected benefit obligations, pension assets, and contributions, but he did not find significant effects on funding ratio. Additionally, he finds that investment strategies for these plans become riskier, with a higher allocation to equities and lower allocations to cash, government securities, insurance accounts, and mutual funds. However, there is no significant effect on realized returns. These results suggest that private equity buyouts may negatively affect the retirement incomes of plan participants of target firms.
>
> For robustness check, he used entropy balancing to make PE-backed firms similar to the control firms in terms of the covariates. More specifically, he used the data of all covariates in year t − 1 which is one year before the buyout and generated the entropy balancing weight by each cohort. Then, he repeated the test with the entropy balancing weight. The results are generally consistent to the main results. This outcome suggests that his results are not driven by the selection bias.

The entropy balancing method is based on J. Hainmueller' paper "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies."[29]

Entropy balancing is used on data where there is a binary division in the data. For example, in the pension plan buyout paper, the data contain pension plans that had a buyout and pension plans that did not have a buyout. Entropy balancing adjusts the data so that certain features of the two groups are the same. This adjustment is done to try and identify the relationship of the binary characteristics to other characteristics that are in the data independent of already known

---

[28] "How Do Private Equity Buyouts Affect Employee Pension Plans?"; Zhong, Wensong; Dec. 19, 2024.

[29] Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*. 2012;20(1):25-46. doi:10.1093/pan/mpr025

relationships. In the pension plan buyout paper, entropy balancing was used to examine the relationship between PE buyouts and the likelihood of DB plan termination or freeze.

The following hypothetical example illustrates the concept of entropy balancing. Assume that there is an already known difference in the likelihood of a small pension plan or a large pension plan having a DB plan termination or freeze for mature companies.[30] Also suppose that in the pension plan data, small pension plans are significantly more likely than large plans to have a plan termination and there are many more small plans then there are large plans. When looking at the data for pension plans that terminated, the data would mostly be for small pension plans and vice versa. Without entropy balancing, the difference in the rates of DB termination in the subsets of the data that had a buyout and those that did not have a buyout could mostly be due to there being more small pension plans that terminated and vice versa. Entropy balancing adjusts the data. The entropy balancing produces a set of data that has about the same proportion of small pension plans in both the subset that had a buyout and the subset that did not. This may be done by applying different weights to the smaller plans than the larger plans.

After the entropy balancing is done, the rates of DB termination in the subset of the entropy-balanced data that had a buyout and the subset that did not, can then be compared to identify the impact of the buyout independent of the size of the pension plan. For example, Zhong says that PE buyouts increase the likelihood of DB plan termination or freeze by 14.2 percentage points. Our interpretation is that if there were two comparable plans, and one had a PE Buyout and the other did not, the data shows that the plan with the buyout will have a 14.2 percentage point higher chance of DB plan termination or freeze.

Life Table Entropy

Leonard Hayflick in *Longevity Determination and Aging*[31] states a thermodynamic type of viewpoint that "the aging of living things is not unlike the aging of everything in the universe including the universe itself. The molecular disorder that defines biological aging might occur passively by increasing decrements in the energy necessary to maintain molecular fidelity or actively through, for example, the action of reactive oxygen species (free radicals). Although biological aging occurs in an open system, the Second Law of Thermodynamics applies in that entropy increases despite the constant availability of energy in the form of food. Entropy increases in biological systems because natural selection has not favored systems that can maintain molecular fidelity indefinitely. Energy is better spent on strategies that ensure reproductive success in order to perpetuate the species rather than spending it on post reproductive longevity that has little species survival value."

To test Hayflick's statement, we calculated the normalized entropy for the Life Table for the Projected Total Population in the United States: 2020 Census. The entropy for the deferred distribution of death n|qx for decadal ages is provided below. The results show the pattern of increasing entropy as Hayflick described. The exception begins at age 80 for the males and 90 for females. The decreases at those ages appear to be due to truncating the table age 100 and

[30] Most private equity firms and funds invest in mature companies rather than startups to increase their worth or to extract value before exiting the investment years later.

[31] "Longevity Determination and Aging"; Society of Actuaries; Sept. 25, 2001.

accumulating all the deaths in age 100 for 100 and over, as illustrated in the following table below where the where the accumulated deaths are highlighted in yellow**.**

As a comparison, we also calculated the normalized entropy for the mortality table in IRS Notice 2019-26 417(e)(3) 2020.[32] It appears that the IRS mortality table has more certainty (order, bias, non-randomness), than the Life Table, which may be a result of using more refined mortality data from pension plan sponsors rather than general U.S. population data. However, the premature peaking of the normalized entropy around age 60 seems to be contrary to the pattern of increasing entropy as Hayflick described. The calculation of entropy for the deferred distribution of death provides new insights.

In a peer review of this paper, the suggestion[33] was made to adjust the IRS table for the significant differences between the Life Table for ages greater than 100. The IRS table has probabilities of death to age 120, where the Life Table puts all deaths after age 100 into age 100. We pooled all the deaths after age 100 into age 100 for the IRS table. The entropy numbers at each age in the revised IRS table decreased when compared to the table below. However, the maximum entropy remained at age 60 as in the IRS table below. We concluded that it does not appear that decisions made in the IRS table on (a) graduation and (b) smoothing the rates for over age 95 to achieve a targeted 50% max at age 115 explains the decline in entropy at the older ages as observed within the IRS table.

	Life Table for the Projected Total Population in the United States: 2020		**IRS Notice 2019-26 417(e)(3) 2020**
	Normalized Entropy		**Normalized Entropy**
Age	**Male**	**Female**	**Unisex**
0	0.866022	0.8307002	
10	0.881290	0.8454024	
20	0.900687	0.8655545	0.81325011
30	0.918786	0.8872633	0.82940574
40	0.940104	0.9115660	0.84743485
50	0.961831	0.9369014	0.86670623
60	0.977194	0.9607660	0.88099269
70	0.985273	0.9817410	0.87610506
80	0.980782	0.9912299	0.83852813
90	0.971539	0.9827134	0.74455053

[32] "Updated Mortality Improvement Rates and Static Mortality Tables for Defined Benefit Pension Plans for 2020"; Internal Revenue Bulletin 2019–26; June 24, 2019.

[33] Suggestion from Timothy Geddes.

Age	Male				Female			
	Death probability (q_x)	Number of lives (l_x)	Number of deaths (d_x)	Life expectancy (ex)	Death probability (qx)	Number of lives (lx)	Number of deaths (d_x)	Life expectancy (ex)
90	0.14125	22,021	3,110	4.73	0.11537	33,929	3,914	5.34
91	0.15417	18,911	2,915	4.43	0.12781	30,015	3,836	4.97
92	0.16789	15,996	2,686	4.14	0.14125	26,179	3,698	4.62
93	0.18240	13,310	2,428	3.88	0.15569	22,481	3,500	4.30
94	0.19767	10,882	2,151	3.63	0.17113	18,981	3,248	4.00
95	0.21366	8,731	1,865	3.40	0.18752	15,733	2,950	3.73
96	0.23030	6,866	1,582	3.19	0.20484	12,783	2,619	3.47
97	0.24754	5,284	1,308	3.00	0.22301	10,164	2,267	3.24
98	0.26527	3,976	1,054	2.82	0.24194	7,897	1,910	3.02
99	0.28343	2,922	829	2.66	0.26152	5,987	1,566	2.83
100+	1.00000	2,093	2,093	2.51	1.00000	4,421	4,421	2.65

Conclusion

Entropy is a concept that dates back to the pioneering days of communication where it was used to determine the minimal amount of information (or entropy) necessary to code and decode a digital transmission. Shannon devised the methodology to improve the efficiency of digital communication by minimizing transmission traffic without compromising accuracy encoding human language and decoding the signals after transmission. The entropy of a message is the expected value of its information content. Higher entropy means a greater level of uncertainty and lower interpretability of a given message. Lower entropy means more predictability in the message content and more accuracy in decoding it into understandable human language. It is this characteristic of entropy, as a measure of order, disorder, uncertainty, and chaos that has led to its utility in the fields of thermodynamics, dynamical system theory, fractal geometry, biology, machine learning, economics and finance, among others.

Entropy as a measure of uncertainty has been discussed in this paper using applications in thermodynamics, statistics, portfolio analysis, machine learning, data rebalancing, and life expectancy analysis. The common observation in all these applications is that entropy is a measure of disorder, which is related to bias and diversity. In fact, it has been demonstrated in this paper that bias is the opposite of diversity, and this is the first paper to define this relationship in terms of low and high entropy. When entropy is low, bias is high, and diversity is low. When entropy is high, bias is low, and diversity is high. Interpreting entropy in terms of bias and diversity provides a computational method for examining data elements in large data sets. For example, it can be determined whether the distribution of levels on categorical variables are diverse enough to prevent bias in outcomes. The theoretical maximum entropy, which is the natural logarithm of the number of levels, can be determined for a categorical variable against which the actual entropy on the variable can be compared for acceptability. If it is unacceptable, it could indicate that additional sampling is necessary. On the other hand, a low entropy number may be desired if more homogeneous data is needed for an analysis. If the level of entropy is too high, then some observations may need to be excluded from the modeling data to improve its homogeneity. These relationships were best observed in the decision tree example discussed above.

As the mitigation of bias in data from all sources becomes increasingly important to insurers and regulators, entropy is likely to become a reliable metric to gain a better understanding of bias and diversity in data. While bias has already been defined in statistical terms, the term diversity has not. The analyses presented in this paper have given statistical meaning to the term diversity as the antonym to bias, and a tool to measure the balance of modeling data elements using the classical notion of entropy dating back to the early days of digital communication. The old has become new again.

Appendix A—Coin Toss Example (n=2), Dice Example (n=6), Alphabet Example (n=4), and Continuous Probability Distributions.

This appendix provides further examples of the calculation of entropy for a coin toss example, dice example, alphabet example, and continuous probability distributions.

Coin toss example

A simple example of entropy is tossing a fair coin. The coin has two possible outcomes, heads or tails, and that the probability of each is 0.5. To calculate the entropy, we take:

$$\text{Entropy} = -\Pr(\text{Heads}) * \log_2\Pr(\text{Heads}) - \Pr(\text{Tails}) * \log_2\Pr(\text{Tails})$$

$$= -0.5 * \log_2(0.5) - 0.5 * \log_2(0.5)$$

$$= -0.5 * (-1) - 0.5 * (-1)$$

$$= 0.5 + 0.5$$

$$= 1$$

So, a fair coin toss has one bit of entropy. The coin toss generates one bit of information.

If we used a two-sided coin, say a two-headed coin, the entropy would be as follows:

$$\text{Entropy} = -\Pr(\text{Heads}) * \log_2\Pr(\text{Heads}) - \Pr(\text{Tails}) * \log_2\Pr(\text{Tails})$$

$$= -1 * \log_2(1)$$

$$= -1 * 0$$

$$= 0$$

The entropy is zero because there is no information to be gained by tossing the coin.

We can calculate the entropy for intermediate degrees of “unfairness”. For example, if the coin had a 70% probability of coming up heads and a 30% probability of coming up tails, we would get:

$$\text{Entropy} = -\Pr(\text{Heads}) * \log_2\Pr(\text{Heads}) - \Pr(\text{Tails}) * \log_2\Pr(\text{Tails})$$

$$= -0.7 * \log_2(0.7) - 0.3 * \log_2(0.3)$$

$$= -0.7 * (-0.51) - 0.3 * (-1.74)$$

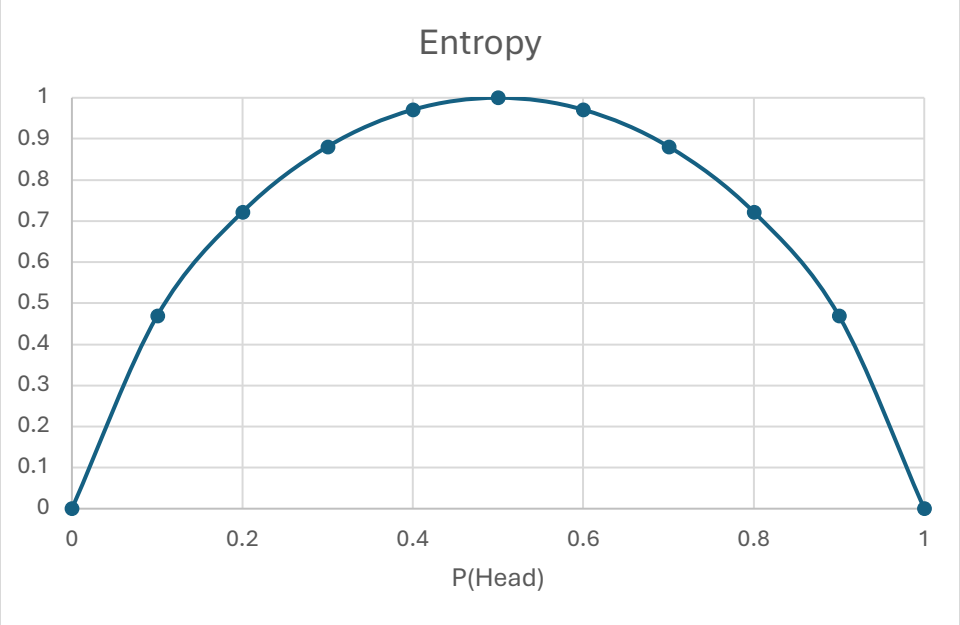
$$= 0.36 + 0.52$$

$$= 0.88$$

So, the entropy is a little less than one. We gain a little less information from a coin toss when the coin is biased than from a tossing a fair coin. This is consistent with the notion that maximum entropy is achieved when the distribution is uniform.

If we continue the calculations for the biased coins, we get the following table:

Pr(Heads)	Entropy
0%	0.00
10%	0.47
20%	0.72
30%	0.88
40%	0.97
50%	1.00
60%	0.97
70%	0.88
80%	0.72
90%	0.47
100%	0.00



Dice example

A similar example of entropy would be tossing a fair die. Each number from one to six has a 1/6 probability of being achieved. Thus, the entropy of one toss can be calculated as:

$$\text{Entropy} = -Pr(1) * \log_2 Pr(1) - Pr(2) * \log_2 Pr(2) - Pr(3) * \log_2 Pr(3) - Pr(4) * \log_2 Pr(4) - Pr(5) * \log_2 Pr(5) - Pr(6) * \log_2 Pr(6)$$

$$= 6 * -(1/6) * \log_2(1/6))$$

$$= 6 * -(1/6) * (-2.58)$$

$$= 2.58$$

So, the entropy from tossing a fair die is 2.58, which is greater than the entropy from tossing a fair coin. This is because the die has six possible results rather than just two, so the toss of a die produces more than one bit of information.

Note that if the dice had four sides, the entropy would be two (see the alphabet example below). If the dice had eight sides, the entropy would be three. In this case, with six sides, the entropy is between that of a four-sided die and that of an eight-sided die.

If the die was weighted, then the entropy would be less than 2.58. Let's say the die was weighted so that there was a 95% probability of rolling a six, and a 1% probability of rolling each of the other numbers. Then the entropy would be calculated as:

$$\text{Entropy} = 5 * -(.01) * \log_2(.01) - (.95) * \log_2(.95)$$

$$= 5 * -(.01) * (-6.64) - (.95) * (-0.07)$$

$$= 0.33 + .07$$

$$= 0.40$$

So, this rather extreme weighting generates much less entropy than a fair die, but since numbers other than six are still possible the entropy is greater than zero.

As another example of a weighted die, let's say the die was weighted so that there was a 50% probability of rolling a one and a 50% probability of rolling a six, but no chance of any other numbers coming up. In that case the entropy would be identical to that of a fair coin, since there are two possible results, each with an equal probability:

$$\text{Entropy} = -Pr(1) * \log_2 Pr(1) - Pr(6) * \log_2 Pr(6)$$

$$= -0.5 * \log_2(0.5) - 0.5 * \log_2(0.5)$$

$$= -0.5 * (-1) - 0.5 * (-1)$$

$$= 0.5 + 0.5$$

$$= 1$$

Alphabet example

Let's assume we have a four-letter alphabet. The only four letters are A, B, C and D. Assume that each letter has an equal probability of being used in any given word. The entropy of this alphabet is:

$$\text{Entropy} = -Pr(A) * \log_2 Pr(A) - Pr(B) * \log_2 Pr(B) - Pr(C) * \log_2 Pr(C) - Pr(D) * \log_2 Pr(D)$$

$$= -0.25 * \log_2(0.25) - 0.25 * \log_2(0.25) - 0.25 * \log_2(0.25) - 0.25 * \log_2(0.25)$$

$$= -0.25 * (-2) - 0.25 * (-2) - 0.25 * (-2) - 0.25 * (-2)$$

$$= 0.5 + 0.5 + 0.5 + 0.5$$

$$= 2$$

The entropy of two can be interpreted as saying that we require two bits of information to reflect any letter of the alphabet. We cannot reduce the number of bits used beyond something like:

A = 00, B =01, C = 10, D = 11

But let's say that the probabilities of each letter being used were not equal. Let's say the probability of an A is 60%, the probability of a B is 25%, the probability of a C is 10% and the probability of a D is 5%.

Now the entropy is:

$$\text{Entropy} = -Pr(A) * \log_2 Pr(A) - Pr(B) * \log_2 Pr(B) - Pr(C) * \log_2 Pr(C) - Pr(D) * \log_2 Pr(D)$$

$$= -0.6 * \log_2(0.6) - 0.25 * \log_2(0.25) - 0.1 * \log_2(0.1) - 0.05 * \log_2(0.05)$$

$$= -0.6 * (-0.74) - 0.25 * (-2) - 0.1 * (-3.32) - 0.05 * (-4.32)$$

$$= 0.44 + 0.5 + 0.33 + 0.22$$

$$= 1.49$$

In this case the entropy is less than two, and so we may be able to represent the alphabet in bits more efficiently. For example, let's assign:

A = 0, B = 10, C = 110, D = 111

Because the letter A is used more often and is only assigned one bit, this representation is more efficient than simply assigning two bits to each letter. Under this mapping, the average number of bits used for each letter is equal to:

60% * 1 + 25% * 2 + 10% * 3 + 5% * 3 = 1.55 bits.

Continuous probability distributions

Uniform distribution

Entropy can also be calculated for continuous probability distributions. This is often referred to as differential entropy, because it measures the entropy in comparison to that of a continuous uniform distribution on [0,1], which has entropy of zero. A distribution that has lower entropy than a uniform distribution on [0,1] will have negative differential entropy.

For example, let's take a uniform distribution over the interval [0,b]. To simplify the integration, we will first calculate the entropy using the natural logarithm, which will produce the entropy in units of nats, and then convert to bits.

The entropy of this uniform distribution (in nats) is equal to:

$$Entropy = -\int_0^b p(x) ln\big(p(x)\big) dx$$

$$Entropy = -\int_0^b \frac{1}{b} ln\left(\frac{1}{b}\right) dx$$

$$Entropy = -ln\left(\frac{1}{b}\right)$$

$$Entropy = ln\,(b)$$

To convert the units of entropy to bits, we just have to divide by ln(2), so we get the entropy in bits for a uniform distribution as:

Entropy = ln (b) / ln(2)

We can see that entropy increases as b increases.

Note that if b=2, the entropy is equal to one bit, consistent with the fair coin toss example. If b=4, the entropy is two bits, consistent with the alphabet example where four letters were equally likely.

Normal Distribution

We can also calculate the entropy for a normal distribution. Again, we will simplify the calculation by calculating the entropy in units of nats by using the natural logarithm and then convert the units to bits.

$$Entropy = -\int_0^b p(x) ln\big(p(x)\big) dx$$

So, the entropy is equal to the negative expected value of the log of a normally distributed variable $N(\mu,\sigma^2)$:

Entropy = $-E[\ln((2\pi\ \sigma^2)^{-0,5} * \exp(-(x-\mu)^2 / 2\sigma^2))]$

= $-\ln(2\pi\ \sigma^2)^{-0,5} * E[(-(x-\mu)^2 / 2\sigma^2)]$

= $0.5 * \ln(2\pi\ \sigma^2) + \sigma^2 / 2\sigma^2$

= $\ln(2\pi\ \sigma^2) / 2 + ½$

Since ½ = ln $(e^{(1/2)})$, we get the entropy in nats as:

Entropy = $\ln(2\ e\ \pi\ \sigma^2) / 2$

Converting to bits, we get the entropy as:

Entropy = $\ln(2\ e\ \pi\ \sigma^2) / 2 \ln 2 = \log_2(2\ e\ \pi\ \sigma^2) / 2$

We can see that entropy increases as σ^2 increases.

Now we can compare the entropy of a uniform distribution with variance equal to 1 with the entropy of a normal distribution with variance equal to one.

The entropy of a normal distribution with variance equal to 1 is:

$\ln(2 e \pi \sigma^2) / 2 \ln 2 = \ln(2 e \pi) / 2 \ln 2 = \ln (17.08) / 2\ln2 = 2.84/1.39 = 2.05$ bits

A uniform distribution with variance equal to one would have $b = 12^{0.5} = 3.464$

The entropy of a uniform distribution with b=3.464 is:

$\ln(3.464) / \ln 2 = 1.24/0.69 = 1.79$ bits

Although the variances are equal, the normal distribution has greater entropy than the uniform distribution. This is because the uniform distribution is restricted to the range [0, 3.464], while the normal distribution can take on any real number, giving the normal distribution more opportunity for a “surprise”. But among all distributions restricted to the same range, the uniform distribution would have the greatest entropy.

Appendix B: Stock Return Entropy Calculations

Difference Tables for m = 1

m = 1	1	2	3	4	5	6	7	8	9	10
1	0.00%	0.40%	0.04%	1.70%	1.49%	0.15%	0.75%	1.08%	1.20%	0.10%
2	0.40%	0.00%	0.44%	1.30%	1.89%	0.25%	1.15%	1.48%	1.60%	0.30%
3	0.04%	0.44%	0.00%	1.74%	1.45%	0.19%	0.71%	1.04%	1.16%	0.14%
4	1.70%	1.30%	1.74%	0.00%	3.19%	1.55%	2.45%	2.78%	2.90%	1.60%
5	1.49%	1.89%	1.45%	3.19%	0.00%	1.64%	0.74%	0.41%	0.29%	1.59%
6	0.15%	0.25%	0.19%	1.55%	1.64%	0.00%	0.90%	1.23%	1.35%	0.05%
7	0.75%	1.15%	0.71%	2.45%	0.74%	0.90%	0.00%	0.33%	0.45%	0.85%
8	1.08%	1.48%	1.04%	2.78%	0.41%	1.23%	0.33%	0.00%	0.12%	1.18%
9	1.20%	1.60%	1.16%	2.90%	0.29%	1.35%	0.45%	0.12%	0.00%	1.30%
10	0.10%	0.30%	0.14%	1.60%	1.59%	0.05%	0.85%	1.18%	1.30%	0.00%

Difference Tables & Natural Log of Correlation Integrals for m = 2

m = 2	1	2	3	4	5	6	7	8	9
1	0.00%	0.44%	1.30%	1.89%	1.49%	1.15%	1.48%	1.60%	1.20%
2	0.44%	0.00%	1.74%	1.45%	1.89%	0.71%	1.15%	1.48%	1.60%
3	1.30%	1.74%	0.00%	3.19%	1.55%	2.45%	2.78%	2.90%	1.60%
4	1.89%	1.45%	3.19%	0.00%	3.19%	1.55%	2.45%	2.78%	2.90%
5	1.49%	1.89%	1.55%	3.19%	0.00%	1.64%	1.23%	1.35%	0.29%
6	1.15%	0.71%	2.45%	1.55%	1.64%	0.00%	0.90%	1.23%	1.35%
7	1.48%	1.15%	2.78%	2.45%	1.23%	0.90%	0.00%	0.33%	1.18%
8	1.60%	1.48%	2.90%	2.78%	1.35%	1.23%	0.33%	0.00%	1.30%
9	1.20%	1.60%	1.60%	2.90%	0.29%	1.35%	1.18%	1.30%	0.00%
log	-1.50	-1.10	-2.20	-2.20	-1.50	-1.10	-1.10	-1.50	-1.50

$\phi^2(0.97\%) = -1.52$

Difference Tables & Natural Log of Correlation Integrals for m = 3

m = 3	1	2	3	4	5	6	7	8
1	0.00%	1.74%	1.45%	1.89%	1.49%	1.15%	1.48%	1.60%
2	1.74%	0.00%	3.19%	1.55%	2.45%	2.78%	2.90%	1.60%
3	1.45%	3.19%	0.00%	3.19%	1.55%	2.45%	2.78%	2.90%
4	1.89%	1.55%	3.19%	0.00%	3.19%	1.55%	2.45%	2.78%
5	1.49%	2.45%	1.55%	3.19%	0.00%	1.64%	1.23%	1.35%
6	1.15%	2.78%	2.45%	1.55%	1.64%	0.00%	0.90%	1.23%
7	1.48%	2.90%	2.78%	2.45%	1.23%	0.90%	0.00%	1.30%
8	1.60%	1.60%	2.90%	2.78%	1.35%	1.23%	1.30%	0.00%
log	-2.08	-2.08	-2.08	-2.08	-2.08	-1.39	-1.39	-2.08

$\phi^3(0.97\%) = -1.91$

1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036

202-223-8196 | **ACTUARY.ORG**