The Portent of Uncertainty

And the Mathematics of the Unknown

An Information Theory Primer

Revision 10

Grant Holland
Santa Fe, NM
March 2013

Contents

Prologue

This primer on information theory has been developed for readers who not only want to understand the basic mathematical concepts that constitute this rich and elegant discipline, but at the same time are interested in developing an intuitive understanding of its principles. Beyond learning the ideas, it would also be highly desirable if the reader can gain insight into why information theory can be so pertinent to how the world around us operates. These are the goals of this primer.

To pursue this end, this primer presents an intuitive, learn-by-discovery approach that leads subsequently to formal definitions and descriptions. As such, it is intended for an undergraduate course in information theory. More advanced texts in information theory that initially present the concepts in a more formal and abstract manner may be more appropriate for advanced undergraduate or graduate mathematics majors. Such sources are referenced throughout this text and in the references section at the end.

The available literature on information theory tends to be either introductory, or light reading, aimed at a popular audience or, at the other extreme, formal graduate level or advanced texts and academic articles. This primer aims at a middle ground and intends to provide a more mathematical presentation than a popular expose', but more explanation and insight than a graduate or advance undergraduate mathematics text.

The reason that some people find information theory interesting - even exciting - is because it is so natural in addressing the unknown. Indeed, information theory is the study of uncertainty [Kleeman 2010, lecture 1, p. 1]. Information theory exists, firstly, to ascertain whether and when the uncertain has portent; and, secondly, what is portended by the uncertain; and, thirdly, to what degree something is portended. Thus the potential offered by information theory is "conditioned predictability".

That the subject of information theory is uncertainty may be at odds with the expectations of many readers. Admittedly, the term "information" often denotes factuality – not uncertainty. And, the phase "information" may sound dry, unexciting, factual - even data-ish. And, stylistically, the very term "information" may sound "so second millennium". But nothing could be further from the truth. Information theory is the mathematics of the uncertain and the unknown.

From what we are currently seeing in the world of contemporary scientific and popular literature, information theory is quickly gaining a new recognition; and it just may herald an altered understanding of intellectual pursuit into the twenty-first century. Instead of treating uncertainty as an enemy that needs to be ignored, feared or obliterated, a new understanding of information theory may be able to show us how to pursue uncertainty opportunistically.

Prerequisites

As we shall make clear below, it is reasonable to say that information theory is an extension to probability theory. In fact, information theory is essentially probability theory with a new measuring function added – a measure named *entropy* [Khinchin 1057, p.1]. This new function measures the degree of uncertainty inherent in any "situation" that is describable by probabilities.

Thus, this primer requires that the reader have some basic understanding of elementary probability theory. But fear not! The primer will present most of the basics of probability theory needed to understand the ideas involved. Otherwise, very little knowledge of probability theory is assumed.

In fact, the primer assumes only that the reader has an intuitive or basic understanding of what "the probability of an event" means[1], that the value of the probability of an event is a real number that ranges between zero (0) and one (1) inclusive, and that the sum of all of the possible outcomes, or events, of "a situation" is established (by arbitrary agreement) to be one (1). (We shall not bother at this point to define these ideas with precision.)

Otherwise, no other conceptual prerequisites are assumed.

Historical Backdrop

In the 1940s, Claude Shannon at Bell Telephone Laboratories was developing the mathematics to help the phone company optimize their telephone networks. He named his work "The Mathematical Theory of Communications". [Shannon 1948].

He was aware that the basis of his work needed to be probability theory, since the phenomena involved were not strictly deterministic. However, he noticed that probability theory does not directly define a measure of uncertainty – which he needed for his mathematical theory of communications.

But, he remembered that such a measure had already been developed for another application: namely statistical mechanics. He took a look at what physicist J. Willard Gibbs had done in that regard as early as 1902. He found that Gibbs had already defined a measuring function for the amount of uncertainty inherent in a probability distribution – even though Gibbs, of course, had developed it specifically for statistical mechanics and was applying it only to certain probability distributions of interest to physics.

Thus, Shannon had discovered a second application (his theory of communications) that needed the same concept that was also needed by Gibbs' statistical mechanics. The concept we are talking about is a measure for the degree of uncertainty inherent in a probability distribution. As a mathematician, Shannon realized that if there are two applications that require a measure of uncertainty, then there are probably many more that do as well. Shannon realized that this discovery argued for a more general theory of uncertainty.

Therefore, Shannon decided to invent this new general theory. Once he had developed the new general theory of uncertainty, he could then apply it as the foundation of his new mathematical theory of communications.

So, he started working to develop his new theory of uncertainty. Right away he realized the importance Gibbs' work in statistical mechanics toward this development. He immediately found that Gibbs had already produced almost exactly what he needed for a general theory of uncertainty.

Gibbs had developed a function to measure the uncertainty inherent in the probability distribution that he was using to model the behavior of particle systems, such as the molecules in an ideal gas. But Shannon found that this function was almost exactly what was needed for the more general case of measuring the degree of uncertainty inherent in any probability distribution – not just the ones that Gibbs was working with in statistical mechanics.

---

[1] A reasonable working definition of "the probability of an event" for this primer is: "the fraction of times that the event occurs of the total number of occurrences of all events - after a large number of repeated trials". The probability of an event is often described as the "relative frequency of the event".

Shannon noticed that Gibbs' definition was articulated as a formula that was multiplied by a particular constant named "K", which had a particular meaning to statistical physics. Shannon realized that this constant K worked as a choice of scale in Gibbs' formula (just as "meters" or "yards" are optional scales for measuring length). In other words, one could substitute any other constant in place of "K" in Gibbs' formula, and the resulting formula would still measure of uncertainty in the same way – only using a different scaling factor.

Thus, Shannon chose to substitute the value 1 for the factor K in Gibbs formula, with the result of the "K" disappearing from the formula. Thus a simpler formula was born for measuring the amount of uncertainty in any probability distribution, regardless of whether the distribution related to physics or to any other domain of interest.

The next problem for Shannon was to decide what to name his new formula. He decided to retain the same name for this function as Gibbs had used – since the definition of the function by Shannon required only a minor and simplifying alteration to Gibbs measure. Gibbs had called it entropy. Thus, so did Shannon.

When entropy is added to probability theory, the result is Shannon's new theory of uncertainty. His new theory, by all rights, could have been called "uncertainty theory", and perhaps it should have. But for reasons we shall introduce next, it has come to be called "information theory".

[Some authors conflate Shannon's two theories – *information theory* and *communications theory*. But this author finds utility in distinguishing them. The reader is referred to *Appendix 2: Information Theory vs Communications Theory* for further considerations.]

Uncertainty and Information

Shannon further showed that information is what you get when you remove uncertainty from a situation! Therefore, rather than naming his new theory "uncertainty theory", it came to be called "information theory". Thus, information theory was born. However, it is important to understand that the "information" in information theory is based upon uncertainty.

In fact, as we shall see, in information theory, one never works with information directly – only with uncertainty! The whole theory is about how to measure uncertainty in a variety of circumstances. Then that same value is applied to the information that results when the uncertainty is removed from the situation!

The reason for this indirect approach to measuring information is that we know how to work with uncertainty. It is through probability theory. (This was Gibbs' and Shannon's great discovery.) But we do not know how to measure information directly. And, we now have this correspondence between uncertainty and information that we just described.

Therefore, we use probability theory to measure uncertainty, and then apply that same measurement to the information that results by the removal of the uncertainty that we measured using probability theory [Khinchin 1957, p. 7]. This is why information theory is simply an extension to probability theory.

Shannon and others further developed information theory chiefly by embellishing it with other ideas from probability theory such as stochastic processes and their various behaviors. The result is the information theory that we have today. This theory has been applied to many areas of science and engineering, and has enjoyed a recent wave of applications with fields such as quantum information theory.

Uncertainty and Predictability

We live in a world of uncertainty. And uncertainty often changes to certainty when "the die is cast". Just so, certainty often then changes back to uncertainty with the circumstances. So, certainty/uncertainty generally has a dynamical aspect – constantly changing between the two.

Of course, this brings up the need to predict where uncertainty will land during times of certainty (after "the die is cast"). Thus, anytime we find ourselves trying to make predictions, then we know that we have a situation that has an aspect of uncertainty about it. It is in these times that information theory is called for, because it pertains to uncertainty and issues of predictability.

So any situation that is concerned with uncertainty and predictability may benefit from the application of information theory. Of course, these issues often arise in the sciences. Thus information theory has found many applications in the sciences in recent years.

It is significant that uncertainty comes in degrees. Sometimes we are completely certain. At other times we are completely uncertain. Usually, though, we are somewhere in-between: a little bit certain, a little bit uncertain or some intermediate degree of certainty/uncertainty. Thus, we often ask the question "Just how certain is it, anyway, that <some event> will happen?"

So, the possible degrees of uncertainty range from complete certainty to complete uncertainty. Complete uncertainty is often called *random* and complete certainty is often called determined or *deterministic*.

When we are "in the world of uncertainty", it would be useful to have a measuring function that gives us a way to evaluate the uncertainty in terms of, at any moment, its degree.

Information theory provides such measure. The purpose of information theory is to measure the degree of uncertainty of any "situation", and to explain the dynamics of uncertainty as its degree changes over time.

The Minimum Requirements of Information Theory

Of course, being a branch of mathematics, information theory does not operate in a total vacuum. It must assume that it knows at least something about the situation it is addressing. It must have something to work with. In fact, information theory must initially possess a minimum of two types of information pertaining to a situation that it wants to address.

First, it must be focused on some happening, some event, of unknown outcome. However, what those *possible* outcomes are must be known. That is, the set of all possible outcomes of the event in question must be known and well defined in order for information theory to be able to work with the event to assess its degree of uncertainty.

In other words, information theory must, I the beginning, have a list of possible outcomes to work with. It need not initially know which of these possible outcomes will be manifest by the event. But it needs to know specifically, in an identifiable manner, what that set of possible outcomes is. This set of possible outcomes of an event is called the *sample space of the event*.

Identifying the sample space is the first type of information that information theory needs in order to begin. If you cannot identify a sample space for the event in question, then you cannot use information theory. Maybe you can try mysticism, hope,

or wishful thinking – but without a sample space you cannot use information theory to probe the unknown.

But having a sample space is not enough. To use information theory you have to know another type of information. You have to have a sense of relative likelihood of occurrence of each one of the possibilities, or sample points, of the sample space. These are the second type of information that information theory must have in order to get started. These are called the *probabilities of the outcomes*, or sample points, of the sample space.

Together, these two types of information can be described as a set of paired values, where the first value in each pair is a logically possible outcome (or sample point), and the second value in each pair is the probability of that sample point. This set of pairs is called a *probability distribution* of the situation[2].

Of course, requiring that information theory must have available to it a probability distribution for any situation that it intends to work with is, admittedly, a very strong requirement. But such a requirement is what makes information theory a mathematical field, rather than, say, magic!

It is worth pointing out that, even though you have a probability distribution for a situation, you are still a long way from knowing what the actual outcome is. But, information theory is a critical, or formal, approach to 1) ascertaining whether the situation has any predictability to it (that is, its degree of uncertainty), and 2) discerning what can be said, if anything, regarding prediction of an outcome for the situation.

It should be clear from what was just said that information theory is heavily based on probability theory. In fact, from what has been said it appears that information theory *is* probability theory. Actually, we shall show that information theory adds a significant measuring function to probability theory – a function that measures the degree of uncertainty of a probability distribution. We have mentioned that the name of this measure is *entropy*.

Thus, it is reasonable to say that information theory is an extension to probability theory in which *entropy*, and some other measures of uncertainty that are elaborations of entropy, are considered. Such is the point of view adopted by this primer.

[2] A convention that has been established in probability theory is that these probabilities are all non-negative real numbers, and that they collectively sum to 1. This convention allows us to use the same probability scale from 0 to 1 for all distributions.

Preview

Information theory is the mathematical investigation of uncertainty. As such, it defines the notion of uncertainty in terms of other, simpler, mathematical ideas; and then it develops these ideas into a full theory of uncertainty and predictability.

This primer is organized into three parts. Part I presents the mathematical foundations that information theory chooses to use to develop its notions of uncertainty. Part I introduces the notions of chance variation and probability theory as its starting point, and defines the central construct of information theory, *entropy*, as a function of probability measure. Subsequently, all essential ideas of information theory are defined in terms of entropy. Information theory can be characterized as *variations on the theme of entropy*.

Part II introduces the idea that two chance variables can be involved in an uncertain situation; and that, when they are, there are conditions under which one of them may portend the outcomes of the other. A special form of entropy measures this degree of portent, or dependency.

Part III extends the ideas of Part II to enable prediction over time. Again, a special form of entropy characterizes the predictability of such processes over time.

Part I: The Valuation of Uncertainty and Information

Part I establishes that information theory is an extension to probability theory. Probability theory represents the idea of chance variation, wherein the same procedure, called a trial, may produce a different result each time that it is processed. The set of possible results, or outcomes, must be established in advance of processing a trial. This set is referred to as the sample space of the trial. Thus, repeated trials may produce different results, or sample points, from the same sample space. In addition, each sample point is assigned a measure of likelihood “p”, called the probability of the sample point. A set of all pairings of sample points and their probabilities is called a probability distribution[3] of the trial. A given sample space may have many possible probability distributions, exactly one of which applies for a given experiment of that trial. A sequence of repeated trials of a sample space that all use the same probability distribution is called an experiment of the trial.

As indicated, probability theory brings with it a measuring function named probability. Probability is a function that measures the degree of likelihood of a sample point. In so doing, it maps, or associates, a single number (between 0 and 1 inclusive) to each sample point of the sample space. Each of these numbers is called the probability of the sample point. By convention, in order to normalize these assignments, these assignments ensure that the probabilities of all of the sample points sum to 1.

[3] In probability theory, the concept of a *probability distribution* is represented by a number of different mathematical entities (PDF, CDF, PMF etc.) Technically, what was described here as “a set of all pairings of sample points and their probabilities” for discrete sample spaces is known as a *categorical distribution*. Since this primer deals exclusively with discrete sample spaces (whose sample points are not assumed to have been assigned numeric values), then the categorical distribution will adequately represent our probability distributions throughout. Therefore, we shall use the phrases “probability distribution” and “probability distribution function” to mean “categorical distribution” in this primer.

Measuring the Degree of Uncertainty of a Sample Point

Information theory extends all of this by adding a new measuring function named uncertainty. Like probability, uncertainty also assigns a real number to each sample point. However, instead of measuring the degree of likelihood of a sample point, uncertainty measures the degree of uncertainty of a sample point.

In information theory, the notion of uncertainty is different from but directly related to likelihood. Since it is directly related to likelihood, then the uncertainty of a sample point can be defined in terms of the probability of that same sample point. In other words, uncertainty is a function of probability.

But the question then becomes, "What is this relationship between the idea of probability and the idea of uncertainty?" Since ancient times, it has often been expressed that "the lower the probability, the higher the uncertainty, and vice versa". Therefore, the measure of the uncertainty of a sample point should be inversely related to the measure of probability of that same sample point.

However, the mathematical expression that provides this inverse relationship has the property that it gets too large too fast as probability gets smaller. Therefore, we need to "do something" to "slow it down" as the probability gets very small. The result is a revised definition of the uncertainty measure that takes the logarithm of the inverse. This "correction" still insures that uncertainty varies inversely as probability, but it also insures that uncertainty does not get "too large too fast" as probability gets smaller.

Measuring the Degree of Uncertainty of a Probability Distribution

We have said that, in probability theory, a sample space can have more than one probability distribution – each of which describes how relative likelihood is distributed across the sample space at any point in time.

We have already seen that information theory is focused on the idea of *uncertainty*. Information theory notices that these probability distributions - in distinction from their individual sample points - can also be characterized by their *amounts of uncertainty*. That is, each of these probability distributions has an *inherent degree of uncertainty* – and that amount of uncertainty can be different from one probability distribution to another, even if all of these distributions are defined on the same sample space! This is the essential discovery of information theory – one not usually pointed out in probability theory.

Thus, information theory defines the idea of the *degree of uncertainty inherent in a probability distribution*. The measure of this degree of uncertainty inherent in a probability distribution is called *entropy*. Entropy is the central notion of information theory. Once entropy has been defined, information theory becomes a study in variations on the theme of entropy. All other constructs in information theory are various forms of entropy, which we shall call *entropic measures*.

Ultimately, information theory needs to be able to show how any mathematical function can be "wrapped" with this "veil of uncertainty" – entropy - in a way that can show how the normal behavior of that function can be modified to account for varying amounts of uncertainty that it may take on as a result of chance variation. This idea is the key to applying information theory to any scientific, or other, domain that needs to account for the uncertainty of chance variation.

But before any of this is possible, information theory must show how to define entropy to be a *measure of uncertainty of an entire probability distribution.* So far, we have defined an uncertainty measure for single sample points of a sample space. But we

must now somehow extend that definition to the entire sample space and its probability distribution.

But this is easy! We simply take the average! In other words, the entropy of a probability distribution is the mean uncertainty of its sample points.

Relative Entropy

Relative entropy is a simple mechanism provided by information theory to apply the idea of *degree of uncertainty* to a slightly more complex situation than what we have done with the calculation of the entropy of a chance variable, including the joint entropy of a joint sample space.

Relative entropy is interested in comparing two different probability distributions, call one of them p(x) and the other q(x) – both of which have the same sample space. The approach of *relative entropy* is to compare the *degrees of uncertainty* of these two distributions.

Relative entropy works by looking at each sample point in the space one at a time and then calculating two different uncertainty values for it. One of these uncertainty values if $u_p(x)$ and the other is $u_p(x)$. Relative entropy takes the difference of these two uncertainties for each sample point, and then calculates the mean of the difference across all sample points. This mean is the *relative entropy* of these two probability distributions, p and q, over the same sample space.

Thus, *relative entropy* is the mean difference in uncertainty as measured by two probability distributions across all sample points of a distribution. Thus, relative entropy is a way of measuring the difference in the amounts of uncertainty inherent in two probability distributions on the same sample space.

Entropic Measures

We have often conveyed above that information theory is based on a concern for the degree of uncertainty inherent in “situations” – as represented by their probability distributions. This degree of uncertainty is measured in information theory by the *entropy* measure.

A surprisingly large number of real-world phenomena can be seen as a function of uncertainty. In fact, the dynamic behavior of pretty much all phenomena can be “wrapped” in a concern for their degrees of uncertainty. In this way, information theory is able to represent the impact of *chance variation* in any domain of interest.

So far, we have used entropy to directly assess the amount of uncertainty of chance variables. However, information theory needs to be able to apply this idea to more complex situations that make changes to chance variables. We would like to be able to predict the affects on the uncertainty of these chance variables after they have been changed, or transformed.

In other words, we would like to be able to perform transformations on the direct uncertainties that we have been calculating using entropy, and then see what the resulting uncertainties are after the transformation have taken place. We call this “functions of uncertainties of chance variables”. And they produce a new level of entropy for the transformed result.

We call these transformations *entropic measures*.

We just saw a simple example of an entropic measure in the form of *relative entropy*. Relative entropy starts with a single chance variable – namely “X”, with its probability distribution p(x). It then makes the situation more complex by “transforming it“ – by

adding a second chance variable to the situation – namely the same sample space X, but with a different probability distribution q. What relative entropy does is to compare the degrees of uncertainty of these two chance variables.

So, *relative entropy* is a simple transformation of the "entropy of one chance variable" to the "entropy of the difference of uncertainties of two chance variables".

However, in information theory, one is allowed to get arbitrarily complex in these transformations – as long as they constitute a *generalization of the concept of entropy*.

The idea of *entropic measures* is that we can generalize the concept of entropy by transforming it in a consistent way. Recall that we have described entropy as the "mean uncertainty of the sample points of a probability distribution".

Entropic measures generalizes this idea by substituting the phrase "function of uncertainty" in place of the word "uncertainty" in this description of entropy. As a result, we obtain a reasonable high-level description of an *entropic measure* as:

> Entropic measure: The mean of a function of uncertainty of the sample points of a probability distribution.

In Part I of this primer, we start with the mathematical definition of *entropy* and make the appropriate substitution consistent with what we just discussed so as to produce a mathematical definition of *entropic measure*.

With this notion of *entropic measure*, then, we can generalize the notion of entropy to be able to "wrap" almost any mathematically described phenomenon with the necessary information theoretic machinery to able to apply information theory to account for the uncertainty of chance variation in that domain of interest.

Part II: The Portent of Uncertainty

Chance happenings contain a lot of uncertainty. But even so, sometimes the occurrence of one particular outcome of a chance variable has at least something to say about the likelihood of various possible outcomes of another chance variable.

For example, take the weather. We all know that the weather is not very predictable. A lot of uncertainty is involved. The success of the weatherman, or lack thereof, attests to that.

But the weather is not *totally* uncertain. For example, if it rains today, that says at least *something* about whether it will rain tomorrow. For example, the weather comes in "weather systems" that sometimes last for several days in a row. This means that rain often follows rain for some number of days in a row. So too with sunshine.

No matter what the reason, though, we know from experience that knowing what the weather is for today can give us a hint as to what the weather might be tomorrow. The way we state this in probability theory is to say "the probability distribution that describes our best guess as to what the weather will be tomorrow changes, depending upon what the weather is today."

In other words, using this "extra information" about what the weather is today can *reduce the degree of uncertainty* about what the weather will be tomorrow.

Colloquially, we can say that, even in the midst of uncertainty, some knowledge can *portend* something about the outcome of tomorrow. This *portent* can reduce uncertainty.

However, this reduction of uncertainty, this *portent*, does not always hold true. In fact, it only holds true under certain conditions. These conditions require that there exists a

dependency between two chance variables that can be exploited by probability theory in order to result in a reduction in uncertainty.

For example, in the above "weather" example, the two chance variables are "the weather today" (call it X) and "the weather tomorrow" (call it Y). In that particular example, there is a probability relationship between these two chance variables. This probability relationship is essentially the fact that the probability distribution that describes the outcome of Y (tomorrows weather) *depends upon* the outcome of X (today's weather). In fact, *depending upon* the outcome of today's weather, the probability distribution that describes the outcome of tomorrow's weather will be a different probability distribution.

This is the nature of probabilistic dependency – or, as we say "statistical dependence" or "stochastic dependence". Stochastic dependence (our preferred term) is a relationship between two chance variables! And it exists in degrees! Some pairs of chance variables are *more stochastically dependent* on each other than are other pairs. In fact, some pairs of chance variables are not stochastically dependent at all. These we say are *stochastically independent*. Chance variables that are stochastically dependent are, to some degree of other, predictors of each other. Chance variables that are stochastically independent are *not* predictors of each other at all.

Information theory is very interested in the certainty with which we can say that one chance variable is, or is not, stochastically dependent on another. Of course, this question as we have described it always takes place in conjunction with two (or more) chance variables.

Consequently, in order to pursue an investigation of stochastic dependence, statistical portent and predictability, we must apply what we investigated in Part I to a type of probability space that involves two (or more) chance variables – in the same probability space!

We call this type of probability space a *joint probability space*. From joint probability spaces, we can then develop the machinery to consider *stochastic dependence*. Finally, we shall then begin to as the question of how can we measure the degree of stochastic dependence between two (or more) chance variables.

This question will result in our construction of a function to measure such a degree of stochastic dependence – a measure named *mutual information*. Mutual information will then be the Launchpad from which we consider Part III – the issue of predictability and prediction in time. Of course, in Part III, we shall use all of the information theory machinery that we have developed in Part II in order to develop a capability to do prediction mathematically.

Joint Sample Spaces

Of course, all of this machinery begins with the idea of a probability space that involves two or more chance variables – a joint space. But first, we must discuss the beginnings of such a space – the joint sample space.

It possible to combine two chance variables into a single event, or happening, where an "outcome", of sample point, is embodied in a single pair of the outcomes of the two individual (or "component") chance variables. The new "combined" happening is, itself, a new chance variable. Its sample space consists of *pairs* of the sample points of the two component sample spaces. For example one chance variable, call it "X", might be "rolling a die" and a second chance variable Y might be "flipping a coin".

While the sample space of X consists of the six sides of the die {1, 2, 3, 4, 5, 6}, and the sample space of Y consists of {H, T), the new sample space will consist of all possible pairings of the six die faces with the set {H, T}. We shall name this new "joint sample space" (X, Y). An example sample point of this joint sample space is (5, H), indicating that the die came up "5" and the coin came up heads. Obviously, this new joint sample space (X, Y) has 6*2 = 12 pairs as its sample points.

Since we are naming the joint sample space with a paired nomenclature, (X, Y) where "X" is the name if the first chance variable and the "Y" is the name of the second chance variable that we combined to make the new sample space, then we shall conventionally use the lower case pairing, (x, y) to represent an abstract sample point (X, Y).

Joint Probability Distributions

Since a joint sample space is, in fact, a sample space, then it can have a probability distribution. That is, each sample point (x, y) of a joint sample space (X, Y) will be assigned a probability.

And, of course, when you take all of these assignments together, then you have a *joint probability distribution* on the sample space (X, Y) that associates a probability with each (x, y) that is a member of (X, Y).

Notice that we are discussing three different probability distributions here: 1) a probability distribution on X, call it p(x), 2) a probability distribution on Y, call it p(y), and a third probability distribution on (X, Y), call it p(x,y).

Since we have defined these probability distributions on X, on Y and on (X, Y), then we have promoted all three sample spaces to the lofty position of *probability spaces*. Specifically, (X, Y) is a *joint probability space*, while X and Y are its *component* probability spaces.

Note also that the probability distribution of probability space X could possible change over time, depending on circumstances. Therefore, sample space X could have many different probability distributions. Also, Y can, too, have many different probability distributions – but only one at a time. The same is also true of joint probability space (X, Y).

Moreover, suppose we are given probability spaces X and Y with specific probability distributions p(x) and p(y). Then, the probability distribution p(x, y) for joint sample space (X, Y) is not specifically determined by the probability distributions p(x) and p(y). It is true that both p(x) and p(y) do impose some limitations on what p(x, y) can be. But, it is possible that the joint probability distribution p(x, y) could be a large number of different distributions and still comply with p(x) and p(y).

The conclusion of this is that, for any given component probability spaces X and Y, the probability space (X, Y) might have many possible probability distribution. p(x, y) is *not uniquely determined* by p(x) and p(y).

In fact, Part II of this primer can be understood as an investigation of what happens when the joint distribution changes for a given pair of component distributions. Some of these possible joint distributions for a given pair of component distributions will yield a very predictable situation, while other possible joint distributions for the same pair of component distributions will yield very unpredictable situations.

This continuum of predictability amount across all possible joint distributions for a given pair of component distributions engenders information theory with powerful

applicability across a large number of application domains of interest in science, technology, the arts and the humanities.

Joint Entropy

Any joint distribution is, in fact, a probability distribution, complete with sample points, each of which has its own probability. Consequently, just like any probability distribution, it has an entropy value – which measures its *degree of uncertainty*.

In fact, the entropy of a joint distribution is calculated in the same manner as any other distribution. Recall that we do this by calculating the *uncertainty* of each of its sample points, and then calculating the mean uncertainty of all of those sample points.

The entropy of a joint distribution is called *joint* entropy. So, like the entropy of any other probability distribution, *joint* entropy is the mean uncertainty of the joint sample points of a joint distribution.

Conditional Probability Distributions

Consider the following scenario. For convenience, lets put the possible states of the weather into four categories: rain, snow, clear, cloudy. Consider the following two chance variables: X is the weather on any given day. Y is the weather on the following day. For example purpose, lets assume that each day has exactly one weather state.)

Consider the joint probability space (X, Y) whose sample points consist of all pairs whose first entry is the weather on one day and whose second entry is the weather on the following day.

Now, consider the sample point that represents today and tomorrow's weather, (x, y). We know what the x value of (x, y) is, because we know today's weather. Suppose that today's weather is "cloudy". Now it turns out that knowing that today's weather is cloudy gives us a hint as to what tomorrow's weather will be. That is, in this example, knowing the value of x helps us to guess the value of y.

The reason for this is that the various possible values of tomorrow's weather have a different probability *depending upon* the value of today's weather!

In other words, our notion of the probability distribution for "y" changes – depending upon the outcome of x! If x = "cloudy", our probability distribution for Y is one set of probabilities. If x = "rain", our probability distribution for Y is another set of probabilities. In other words, each possible outcome of x has its own individual probability distribution for Y!

This situation is called *conditional probability*. A *conditional probability distribution* is the collection of all of these probability distributions for Y – one for each possible outcome of X – put together into a single mathematical construct.

Obviously, then, a "conditional probability distribution" is not really a single probability distribution, but rather a collection of individual probability distributions with one for each possible outcome of X. Nevertheless, we treat it as a single construct – a matrix – because as such it forms interesting and consistent numerical relationships with other probability distributions on the same sample space.

However, we do treat this matrix as a single distribution. It is called the *conditional probability distribution of Y given X*, where X and Y are two chance variables in a joint sample space. We symbolize this distribution by (Y|X) and its probability distribution as p(X|Y).

Of course, for any given joint sample space (X, Y) we can also consider the probability of X given Y, as well. This we symbolize as (X|Y) and its distribution as p(X|Y).

It turns out that this concept of conditional probability is the backbone of all predictability in information theory. Almost all constructs in information theory discussed afterwards in this primer build upon the idea of conditional probability.

Conditional Entropy

The two conditional probability distributions discussed in the previous section, p(Y|X) and p(X|Y) can also be defined to have an entropy value in the usual way: as the mean uncertainty of (y|x) and the mean uncertainty of (x|y) respectively. These entropies are described as H(Y|X) and H(X|Y).

It turns out that the values of H(X), H(Y), H(X, Y), H(Y|X) and H(X|Y) enjoy a number of interesting, surprising and significant numerical interrelationships which furthers the significance and applicability of information theory to a number of application domains.

Mutual Information

In Part II, we have shown how conditional probability can portray different degrees of dependency between chance variables that are related by joint distributions. But, ultimately, we need a way to measure this degree of dependency on a scale that runs from zero for no dependence to larger values for more dependence.

The measure of stochastic dependence developed by information theory is called *mutual information*. We shall see that it also leverages the idea of entropy.

To understand the strategy used by mutual information the measure the degree of stochastic dependence of a joint distribution, one must first understand that for the given two component chance variables X and Y of the there are many possible joint probability distributions besides the one of interest. Of all of these, there is only one that is stochastically *independent*. All the rest exhibit varying degrees of stochastic dependence. Of course, another one of these possible joint distributions is the one whose degree of dependence we are trying to measure.

The approach of mutual information is to measure “how far away” the joint distribution we are trying to measure is from the on and only stochastically independent joint distribution on these two chance variables.

The strategy works by first looking at each sample point of the joint sample space. The uncertainty value for the sample point is calculated twice – once using the joint distribution we are measuring (the target distribution) and a second time using the stochastically independent distribution. Then the uncertainty value based on the target distribution is subtracted from the uncertainty value based on the stochastically independent distribution.

This calculation is then repeated for every sample point in the joint sample space – yielding a set a numbers, each of which is the difference of uncertainty values of its sample point based in those two distributions. Finally, the mean of all of these differences of uncertainty are calculated. This result is the mutual information of the joint distribution of interest.

The fact that we are utilizing the differences between the stochastically independent distribution and the target distribution accounts for why this measure is concerned with “how far away” the target distribution is from stochastic independence – and therefore constitutes a measure of stochastic dependence.

And, the fact that we are calculating uncertainty values before we take the difference is the par that is “entropy-like”. This part accounts for the fact that these differences may not be consistent. Some of them may be large differences while others might be small. Therefore, the act of calculating their uncertainties and then taking their difference, as well as calculating the mean of all of these differences in uncertainty, factors in the uncertainty that may be involved in these calculations.

The astute reader will also have realized that the description for the calculation of mutual information just described is a special case of *relative entropy*. Recall that relative entropy, described earlier, is an entropic measure of “how far away” one probability distribution is from another, when both distributions have the same sample space.

But, mutual information also works by comparing two probability distributions on the same sample space. However, this time, both distributions are joint distributions, And, while one of them is the distribution whose degree of dependence is being measured, the other is the joint distribution on the same joint sample space that is stochastically independent.

Therefore, mutual information applies relative entropy to measure “how far away” a joint distribution of interest is from stochastic independence – and therefore, how stochastically dependent the target joint distribution is.

Multi-dimensional Joint Probability Spaces

In Part II, we initially examine the idea creating a new kind of a probability space by jointing the individual sample spaces of two chance variables into a single sample space whose sample points are pairs that combine the sample points of the two individual component spaces.

But this idea can be generalized to include the combining of, not only two component chance variables, but of 3, 4, 10, 120 or any number of component sample spaces into a single joint sample space of any number of component chance variables. Each sample point of such a joint probability space is an n-tuple with as many elements as there are chance variables being joined.

Moreover, all of the machinery that we developed for the 2-D joint probability spaces above can be generalized to apply to joint probability spaces of more than 2 chance variables. These include:

> Joint entropy, conditional probability distributions, conditional entropy, and mutual information.

These concepts and construct provide the foundational machinery to develop an information theory of predictability and prediction, which is the subject of Part III.

Part III: Prediction: The Meaningfulness of Uncertainty in Time

<>

Part I: The Valuation of Uncertainty and Information

Information is often confused with data. But data is static – it just sits there doing nothing. Information is dynamic – it does something. It operates…as we shall see.

Information Theory's Environment

Information theory operates in the world of uncertainty. Usually uncertainty is problematic for us. Very often we would like to change uncertainty so that it becomes certainty. If we cannot remove uncertainty immediately, then at least we would like to decrease its degree so that we are moving in the direction of certainty.

In any event, we must first have some way to measure the degree of uncertainty in order to be able to consider its decrease. This is the general problem that information theory addresses – measuring the degree of uncertainty inherent in a "situation".

The General Mileu

In order to address this issue, we must first ascertain what kind of environment information lives in. Who are its players and what are their interactions?

1. There is an observer (you or I) that is observing a *happening*, occurrence or *event*. We shall use the name *trial* to refer to an instance of such an event.
2. A trial takes place in time, and can be analyzed into three phases. In the beginning of a trial, the observer is aware of a *sample space* of possible outcomes (*sample points*), each of which has the possibility of becoming manifest during the trial. The observer also knows that exactly one of these sample points will become manifest during the trial, but she does not know which one. This sample space is sometimes referred to as a *state space*, and the sample points are also sometimes referred to as *states* of the trial. Logical *combinations of the sample points* are also of interest. They are called *events*. (These logical combinations include events in which exactly one of the sample points has occurred – "singleton events".)
3. The trial will result in the selection or manifestation of exactly one of the sample points. This manifestation is called the *realization of the trial*, and the manifested sample point is called the *realized* sample point.
4. After realization, the observer becomes aware of the outcome – which is exactly one of the sample points.

For example, consider the trial wherein a single six-sided die – as is used in gambling games – is rolled. Each of the six sides has a well-defined identity – which is a number of dots between 1 and 6 inclusive. We shall declare that the sample space of this trial is these six numbers. When the die stops rolling after it has been tossed, exactly one of its six sides will be facing up. This manifestation of exactly one of the sample points is called *realization*.

Each of these six sample points has a *probability* value that is a measure of its likelihood of being the sample point that is realized during the trial. It is conventional to assign a probability value to a sample point that is between zero (0) and one (1), inclusive. In addition, this convention specifies that the sum of the probability values of all of the sample points of the same sample space sum to 1. Obviously, the higher the probability of a sample point, the more likely is that sample point to be realized within the trial.

Regarding our example sample space consisting of the gaming die, normally, we consider that all six faces should have the same probability as each other if the die is considered to be “fair”. This means that each die face has a probability of 1/6.

We can also define “compound events” of the existing sample points by the use of logical combinations of the sample points. Then we could calculate probabilities for the compound events as well as for the sample points. For example, from our dice example, we could define the event named “die face is even” to mean the “logical or” of the three events 2, 4 and 6 dots. If the probability of each of these three sample points is 1/6, then the probability of the compound event “die face is even” would be 1/6 + 1/6 +1/6 = ½. This approach enables us to expand our notion of probabilities to more happenings.

We can even alter our view of a trial by rearranging the sample space to obtain a different sample space – and thus a different probability distribution. In our die example, our sample space was { {1}, {2}, {3}, {4}, {5}, {6} }. However, we could create a new sample space of the die be reorganizing this sample space – as long as the resulting sample space is a partitioning of the outcomes 1, 2, 3, 4, 5, 6. An instance of this is:

{ {1, 2, 3}, {4}, {5, 6} }.

This new sample space has three sample points. The first is {1, 2, 3}. The second is {4}. The third is {5, 6}. And, the new sample space has a new set of probabilities. The sample point {1, 2, 3} has the probability of 3/6, or ½. The sample point {4} has probability 1/6. And the sample point {5, 6} has the probability 2/6 = 1/3. Therefore, any trial has many possible sample spaces and concomitant probability distributions. We get to use which one of these we wish to represent real life applications of interest. This fact will come into play during Part II of this primer.

Right now we have been concentrating on a single trial. However, we could be interested in repeating the same trial several times – that is, repeated trials. We shall use the term experiment to mean a set of repeated trials. Often the trials in an experiment use the same sample space and probability distribution. For example, rolling a die several different times, or making different purchases of stock over several different times constitute an experiment.

Sometimes in this primer we shall be interested in a single trial. At other times, especially in Part II, we shall become interested in repeated trials, or experiments.

Degrees of Uncertainty of a Trial

As far back as Aristotle it was recognized that an event is considered to be “uncertain” because it is unlikely. That is, “the more unlikely, the more uncertain”. In other words, uncertainty is directly related to unlikelihood [Vedral 2010]. This means that whenever unlikelihood goes up, so does uncertainty; and, whenever unlikelihood goes down, so does uncertainty.

We could reword this observation by saying, uncertainty is *indirectly related* to likelihood. We make this slight change in wording in order to more easily translate this idea into probability theory – which is the language of information theory. Probability theory uses the idea of the probability of an event, which is a measure of its likelihood.

So, the lower the probability of an event, the higher we want its degree of uncertainty to be – and conversely. Mathematically, this means uncertainty and probability need to be inversely related. As one goes up, the other goes down, and conversely.

Thus, our first task in information theory is to *find a function* that measures the uncertainty of an event. And we now know something about what that measuring function must be: it must be a mathematical expression in which uncertainty varies indirectly as probability.

Whatever our definition of the uncertainty of an event turns out to be, it must give larger values for the measure of uncertainty whenever the probability measure of the event is smaller, and vice versa.

This task of defining a function that measures the degree of uncertainty of a sample point (or of an event) can be viewed as inventing yet another way to measure – or assign a number to – a sample point. We already have one way to measure a sample point: the probability of that sample point. Probability measures the degree of likelihood of the sample point. But now we want a second way to measure a sample point: the degree of uncertainty of the sample point. For want of a better name, we shall name this new measure the *uncertainty* of the sample point.

In addition, from what we said at the beginning of this section, we want to define this uncertainty measure of a sample point so that it ends up being inversely related to the probability measure of the same sample point. This means that, when we get around to defining mathematically this uncertainty measure of a sample point, that we can define it in terms of the probability measure of the same sample point – in such a way that the two are inversely related.

The Three Phases of a Trial

Before we proceed to developing the function that measures the degree of uncertainty of a sample point, lets look more closely at the anatomy of a trial. This anatomy will help us see where uncertainty and where information enter into the discussion of a trial.

From the above description of the players in information theory and their interactions, we can surmise that a trial has three phases:

1: Uncertainty phase: The observer is uncertain about which sample point is going to be *realized* during this trail. However, there are two things that the observer *does* know about. 1) The observer knows the specific sample points that constitute the sample space of the trial. 2) The observer knows the probabilities of each of these sample points. We can summarize both of these points by saying that the observer knows the *probability distribution* of the trial - which is a construct that collectively describes both the sample points and their probabilities.

2. Realization phase: Exactly one sample point of the sample space is “made manifest”, or *realized*. This is the point at which, as we say, “the die has been cast.”

3. Information phase: The observer becomes aware of which sample point has been *realized*. At this phase, uncertainty changes to certainty.

It is when an observer is in phase 1 – the uncertainty phase of a trial – she “has a problem”. The “problem” is that the observer is in a state of uncertainty. This is the issue that information theory addresses.

Therefore, it is the uncertainty phase that gets most of the attention of information theory. The “problem” lies in the uncertainty phase.

Another reason that uncertainty gets most of the attention is that it is richer and more interesting than the other two phases. The second two phases are not so interesting: they both have only one thing in them each – the single realized outcome. But the uncertainty phase contains a very rich apparatus – the probability distribution,

including all of the sample points in its sample space as well as the probabilities of each sample point. This is a lot of richness to work with mathematically, and information theory puts all of it to good use – as we shall soon see.

Because of the paucity of information available in the information phase of a trial, there is very little to work with for the purpose of defining a measure of the amount of information that resides there. However, because of the richness of information available in the uncertainty phase, there is plenty to work with for the purpose of defining how much uncertainty is evident in the uncertainty phase of the trial.

Furthermore, it is reasonable to claim that the value inherent in the information phase is immediately related to the amount of uncertainty in the uncertainty phase. The argument is that the value of the realized information in the information phase is due to the relief that the observer fells by virtue of the *removal of the uncertainty* (of the first phase) that results in the information in the third phase.

In other words, the "value" attributed to the resulting information is "how much pain it relieves" by "removing the uncertainty of the first phase.

An analogy would be the question "How much is an aspirin worth?" The answer depends on "How badly does your headache hurt? The value attributed to *information* is measured by "how much pain the uncertainty was causing that was removed to result in the information produced by the trial".

Thus, the value received during the information phase, when the observer becomes aware of which outcome has been realized, is the result of the relief given by the removal of the uncertainty that existed during the uncertainty phase.

Therefore, it is reasonable to define the value of the information resulting in the information phase to equal exactly the amount of uncertainty in the uncertainty phase.

Consequently, the strategy of information theory is to assign the same value to the measure of the information produced by the information phase of the trial as is assigned to the degree of uncertainty inherent in the uncertainty phase of the trial.

So, the measure of the degree of information of a trial is equal to the measure of the degree of uncertainty of the trial.

In short, the amount of information in a trial is equal to the amount of uncertainty of the trial.

Deterministic Variation

Science and mathematics are typically fond of processes and procedures that are totally predictable. Total predictability can be articulated by saying "If a particular procedure is repeated, it produces the same results every time." This concept is called *determinism*, because the procedure *determines the results, or* outcome.

And if the procedure is permitted to behave differently with different "inputs", or *initial conditions*, then we still call it *determinism* by augmenting the above statement to read: "If a particular procedure is repeated using the same initial conditions each time, then it produces the same results every time." This is a slightly more flexible form of determinism.

This fondness of intellectual pursuit for determinism may explain the central role that is played by the concept of *function* in mathematics. Functions are often used in mathematics to model deterministic processes. This works because *functions* represent procedure that take initial conditions, called input parameters, and produce the same result for the same input parameter each time the function is processed.

In the language used by functions, these input parameters are called “variables” and are typically represented by alphabetic letters, such as “x”. Thus, for a given value of “x”, the result is always the same every time a function has the same initial value of this “x”. We shall call such a variable a *deterministic variable*.

In order to account for variation within a deterministic framework, one allows the (deterministic) variable to take on differing input values each time the function is processed. Of course, for any given input value of the variable, the function produces the same results every time. But variation can still be achieved by letting the same variable, say “x”, take on different values each time the function is processed, or “run”.

This is deterministic variation.

Chance Variation

Chance variation breaks with this pattern. Chance variation represents procedures that can result in *different results* each time the procedure is processed – even though the input parameter is the same value!

We would also have *chance variation* if the procedure has no input parameters, and yet produces different results each time the procedure is processed.

Let revise the previous statement somewhat. What actually occurs in chance variation is the procedure *may* produce different results each time the procedure is processed. It is allowable for the procedure to produce the same results sometimes – just as long as it *may* produce different results sometimes also.

The accurate way to describe *chance variation* is that each time the procedure is processed with the same input parameters (initial conditions), it produces exactly one member of a specific set – called the sample space, and never produces any result that is not in the sample space.

In fact, within this definition – one could say that *deterministic variation* is a special case of *chance variation* where the size of the sample space is 1. Thus, *chance variation* is the more general case where actual variation is allowed, and *deterministic variation* is the special case of chance variation where the result set (sample space) is limited to just one possibility – the single member of the sample space.

Chance Variables

Regarding the input parameters, or variables, of deterministic variation, from basic algebra we know that several possibilities are allowed: 1) no input parameters, 2) exactly one input parameter, or 3) multiple input parameters. Not surprisingly, chance variation also allows for these same cases. Chance variation, too, represents these case with no variables, one variable or multiple variables – also just like deterministic variation.

Whereas in deterministic variation, we called the variables “deterministic variables”, in chance variation, we call them *chance variables*. Thus, a *chance variable* always results in exactly one *outcome* each time that its procedure is processed. However, unlike a deterministic variable, the outcome may be different each time the procedure is processed.

Just as we call such a variable a *chance variable*, we call such a procedure a *chance procedure*, or a *stochastic procedure*. Thus, a *chance procedure* uses *chance variables* and represents *chance variation*.

As an example, consider any situation that involves randomness – such as rolling a six-sided die used in games of chance. From the perspective of an observer (people

playing the game), rolling a die is the repeated performance of the same chance procedure – that of actually throwing the die. This is a procedure that has no input parameters. The player, from the point of view of the player, simple performs the same procedure every time – rolling the die. However, there are six possible outcomes – exactly one of which turns up each time the die is thrown. These six possibilities constitute the sample space.

Like a deterministic procedure, we can still think of a *chance procedure* as a mathematical function that produces a *sample space*. Thus, a deterministic procedure always produces the same result for each input parameter, or value of its deterministic variables, or for each input parameter. But a *chance procedure* produces the same *sample space* for each input parameter, or value of its chance variables.

Chance Variables and Probability Distributions

We said in the previous section that the difference between a *deterministic procedure* and a *chance procedure* is that, whereas a *deterministic procedure* associates exactly one *result* with each value of its deterministic input parameters, a *chance procedure* associates exactly one *set of possible results* – called a *sample space* – with each value of its chance input parameters.

However, there is more difference between *deterministic* and *chance* procedures than the result of a deterministic procedure being a single entity and the result of a random procedure being an entire set of multiple possibilities (a *sample space*).

In addition, each of the sample points of the sample space must be assigned a *probability* – which, as we have already discussed is a numerical value that represents the *degree of likelihood* of its sample point being realized. Additionally, in probability theory, there is a convention (Kolmogorov's postulates) that "normalizes" the collection of probability values of all of the sample points of a sample space. This convention states that all of these probabilities are non-negative and sum to 1.

So, we have just enhanced our sample space so that all of its sample points are assigned a probability value. The result is an even more complex construct that includes all of the sample points plus their probabilities. This can be articulated as the set of all pairing of the sample points and their probabilities. This set of pairs is called the *probability distribution* of the chance procedure[4].

We have already seen that we can abbreviate a chance procedure with a chance variable, such as "X" – just as we abbreviate deterministic procedures and variables, such as "x". Thus, we can also associate a probability distribution with a chance variable – which we shall also do as an abbreviation.

---

[4] Technically, "the set of all pairings of the sample points and their probabilities" is defined as a *categorical distribution*, which is one way to represent a probability distribution. In this primer, we shall be dealing almost exclusively with discrete probability spaces. For those spaces, the categorical distribution is typically used to characterize distributions. Thus, for simplicity we shall use the phrase *probability distribution* to mean the categorical distribution.
In most probability theory applications, the sample points are all associated with some numerical value in addition to a probability value. Subsequently, these numerical values are assigned the same probabilities as their sample points and are used in place of their sample points. Thus, a "new sample space" of real numbers replaces the initial sample space – thus ensuring that arithmetic can be performed on the new sample space, since it is numeric. The resulting probability space is called a *random variable*.
However, the mathematics of information theory – always involving some notion of *entropy* – does not make use of this extra numerical assignment values, and only uses the probability assignments in its calculations. Thus, the use of *random variables* over the simpler *chance variables* is essentially ignored by information theory.

We can now express the difference between a *deterministic procedure* and a *chance procedure* by saying the following: We expect a *deterministic procedure* and its initial conditions (input parameters) to produce a single *result*, our outcome. And this outcome will be the same each time the procedure is performed with the same initial conditions. On the other hand, we expect a *chance procedure* and its initial conditions (input parameters) to produce a single *result*, our outcome. But this outcome can be different each time the procedure is performed with the same initial conditions. However, we require that the variable outcomes of a chance procedure are all some particular well-specified set – called the *sample space*.

Thus, a deterministic procedure produces the same outcome each time it is performed. But a chance procedure can produce different outcomes each time it is performed. However, it is known in advance what the possible outcomes for the chance procedure are – they are members of a well defined set named the sample space of the procedure. If the sample space is not predefined, then the procedure is too vaguely defined and does not qualify as a chance procedure.

Also, it is worth noting that a deterministic procedure can be treated as a special case of a chance procedure where in the resulting probability distribution has exactly one sample point whose probability is 1 – and the probabilities of the remaining sample points are 0.)

Random Variables

We have been using the phrase “chance variation” to mean a situation in which repeating the same procedure may produce multiple distinct (inconsistent) results. (Or at least it appears so from the point of view of some observer.) And, we have been using the phrase “chance variable” to identify such a procedure type.

And we have used the term *chance variable* as a variable symbol (as opposed to a constant symbol) that takes on exactly one of these variable values at some point in time. The outcome of a chance variable is some “happening” or event in which some sample point, or logical combination of sample points, is realized. The “value” of a chance variable at some point in time is always exactly one of the sample point of the sample space of interest.

However, in probability theory, one often encounters the phrase *random variable*. A *random variable* is another name for a specific kind of *chance variable* – a chance variable that has been assigned a real number – or that already is a real number. When an outcome of a random variable is realized, the real number that has been associated with the realized sample point is used, rather than the sample point itself.

When random variables are used, then it is assured that arithmetic can be performed from the outcomes of the random variables. This fact makes random variables very popular in applications that involve chance variation. For example, it is possible to calculate functions such as mean and variance whenever *random variables* are being used. However, with chance variables that are not also random variables, then the calculation of means and variances, etc., are not defined because they are not possible. In fact, these functions make no sense to applications whose chance variables are not also random variables.

Many branches of probability theory – such as statistics - cannot do very much with a chance variable that is not also a random variable. From the very start, they need to be able to perform arithmetic upon the outcomes of chance variation, and they cannot work with a sample space unless its sample points have real numbers meaningfully assigned to them. They *require* random variables. But this requirement imposes a

severe restriction upon these branches of probability theory, and puts many applications that involve randomness out of their reach.

But, this is not a restriction for information theory! This is true because information theory does not require its chance variables to also be random variables. All of the important and characterizing functionals and measures that are defined by information theory make use of the probabilities of the sample space. But none of them refer to any value function on the sample space! That is, the sample points of sample spaces that information theory works with do not have to be real numbers, and do not have to have their sample points meaningfully assigned real number values (other than their probabilities). This avails information theory to many applications (especially complex ones) that disciplines such as statistics have limited usefulness with. We shall have more to say about this momentarily. (We should also point out that the chance variables used by information theory *can* also be random variables, but they do not *have* to be.)

Of course, random variables usually assign some real number values to its sample points that have some meaning to its application domain. For example, if the sample points are, say, students, the assigned real numbers may be their *test scores*, or grades. If the sample points are athletes, their assigned numbers might be their *average point scores per game.* Or, if the sample points are commercial corporations, their assigned numbers could be their current stock market prices. These numerical assignments to each of the sample points are something extra that *chance variables* do not need, but that *random variables* must have. So, a *random variable* is a chance variable whose sample points are all assigned some real number value.

We know that all *chance* variables have a *probability function* that assigns a probability to each of its sample points. But in addition to the probability function, a *random variable* also has a second function – a *value function* that assigns some real number to each of its sample points. So we can say that a *random variable* is a *chance variable* that also has a *value function* as well as a *probability function*. On the other hand, chance variables are only required to have probability functions.

This distinction can become important, because if the sample points all have numbers assigned to them, then arithmetic can be performed on the sample points as well as on their probabilities. Lets look at a couple of examples.

In our first example, we want to perform an experiment to ascertain whether 3-year old boys have a preference for any particular primary colors. To test this, we use three balls, each being the same except for their colors: red, green and blue. We perform an experiment by selecting 100 boys at random from some defined population, and present each of them with the three balls to determine which color the boy selects. After the experiment is done, we can assign probabilities to each of the three colors via the relative frequencies that each was selected.

However, suppose someone asks “What is the average, or mean, of the balls chosen?” Clearly, the question makes no sense – is undefined. The reason for this is that no mean can be calculated because no numeric value has been associated with each of the three colors. Recall that the formula used in probability and statistics for calculating the mean of a probability distribution is the multiply the *value* of each sample point (colored ball) by its probability, and then to sum those products. But these balls do not have an associated value! Therefore, it is not applicable to multiply a ball’s probability by its value! In fact, there is no “average ball” in this case. The concept is undefined because the balls do not have any natural “value” in this experiment. Of course, we could have arbitrarily assigned some number to each ball –

such as 0 to red, 1 to blue and 2 to green. (But why not 11 to red, 17 to blue and 153 to green?) But, if we had done so just to able to calculate a mean, then the mean would be a meaningless number (pun intended).

The above example used a very simple sample space – almost too simple to have a meaningful mapping into the real numbers. Information theory can do a great deal with this simple space because it has *entropy*; whereas, the discipline of statistics has little to say about it, since no mean, variance or moments can be calculated for it.

Very complex sample spaces also often resist any simplistic associations with number values to their complex sample points. This is especially true whenever the sample points are entities that form very complex patterns. An example of this is biochemical molecules. Suppose that we are interested in the probability distribution that is formed by pairs of biochemical molecules randomly encountering each other within a liquid environment (e.g. a cell) and then incurring an oxidation-reduction reaction. In this case, each sample point is a pair of molecules. And, our sample space is the set of all pairs of biomolecules. The probability that we would associate with each pair is its observed likelihood of reacting within a chemical space. Obviously, this is a very complex space whose sample points are very complex entities. In such an application, it does not really apply to ask "What is the average pair of molecules". The reason is because there is no meaningful assignment, or mapping, of real number to each of the pairs of molecules in this complex sample space. In fact, attempting to force a real number assignment to each member of this sample space would have a "flattening" effect. That is, a very complex multidimensional space would have been reduced to the one-dimensionality of the real numbers – and would actually result in a loss if information.

Therefore, this biomolecular space is represented by a *chance variable* that is not also a *random variable*. And, information theory is able to provide a rich model of this space, whereas statistics would have trouble getting off the ground with it, since no mean, variance or other moments exist for it.

Often, however, in science, business and engineering, we are careful to articulate our probability models in a manner that ensures that each sample point has meaningful numerical value. To wit, market valuations, speed, monetary value, grade point average, pounds per square inch, etc.

If our sample space is initially a chance variable that is not meaningfully "associatable" to real numbers, then we often in many applications figure out some way to change it so that it is. For example, the probability space that models the flipping of a coin is a chance variable that is not also a random variable. This is easily seen because its sample points are "heads" and "tails" – and these are not numbers and do not have any natural numerical value. However, each of these sample points is easily and naturally assigned a probability. Since it is not a random variable, then we cannot calculate it mean or perform other arithmetic operations in the sample points.

However, it is quite easy to transform this probability space to a related probability space that is! The technique is to convert this probability space into a "counting problem". Here is how it is done. Define a new single "trial" to consist of flipping a coin, say, 10 times. The outcome, then, would be a list (or "tuple") of ten outcomes – i.e. (H, T, T, T, H, T, H, H, T, H, T, T). This 10-tuple is an example sample point from our new sample space. We would assign a probability to this sample point either empirically or theoretically.

But we still do not have a numeric "value" assignment to this sample point. But here is how we can meaningfully define one: Simply count the number of heads in the tuple.

For the particular 10-tuple presented above, there are four heads. Therefore its *value* is 4. Thus, every point in this new probability space has both a *probability assignment* and a real number *value assignment*. Therefore, our initial chance variable probability space has been converted into a random variable probability space.

Of course, there is no semantically meaningful reason to perform such a conversion to our biomolecular example probability space that we described earlier. It is already as complex as it needs to be as a chance variable; and nothing is served by imposing a numerical assignment on it just to be able to say that it is a random variable.

It turns out that Statistics, as a mathematics discipline, almost always uses random variables because it needs to calculate statistics such as *mean*, *median*, *variance*, *standard deviation*, and the so-called *moments* and *central moments* – all of which make use of these *number values*, as well as the *probabilities* that are also associated with the sample points.

So, the discipline of statistics has little use for chance variables that are not also random variables, because statistics almost exclusively deals with calculations, such as mean, median, variance and standard deviation that involve both these numerical values as well as their probabilities. In other words, if you have a probability application whose chance variables are not – and do not need to be – random variables, then the discipline of statistics may not have any usefulness to your application.

Of course, such applications include ones that are very simple – like the colored balls example above – that their sample points do not need numerical value assignment in addition to probability assignments. However, it also includes applications – like the biomolecules one above – whose sample points are already so rich and complex that they do not admit to being "collapsed" into the real number line, together with the concomitant loss of information with such a mapping. It is particularly for these complex sample spaces (state spaces) that information theory may prove to be a better tool than that of statistics.

However, information theory is characterized almost exclusively by calculations and mathematical constructs that *do not use a numerical value function on sample spaces* at all, and *only use their probability assignments*! For example, the central theme of information theory is that of the uncertainty inherent in a probability distribution as measured by the *entropy functional*.

Upon inspection it becomes clear that the entropy definition uses only the probability of each sample point as its only input – and does not refer to a value function at all – as do the mean and central moment functions of statistical theory. Therefore, information theory does not need the extra equipment (the numerical value assignment) of a random variable, and can make do just as well with *chance variables* that are not also *random variables*.

The reader should also know that the distinction being made here between *chance variables* and *random variables* is an issue raised, perhaps, only in this primer. However, one of the first authors to use the phase "chance variable" with these semantics was Claude Shannon [Shannon 1948, p 11]. On the other hand, many important and significant tombs on probability theory, statistics or information theory do not make this distinction[5] and refer only to random variables but not chance variables.

But it is important to understand that the ability of information theory to work with probability spaces that are not required to possess a value function on its sample

[5] On the other hand, one of the first authors to use the phase "chance variable" with these semantics was Claude Shannon [Shannon 1948, p 11]>.

points opens up information theory to work with many very complex applications that whose accessibility to the discipline of statistics is extremely limited. This limitation of statistics is due to the fact that its essential arsenal of tools - which include mean, variance, the moments and central moments and its moment generating functions – all completely depend upon the existence of a value function that maps its sample points into the real numbers. But information theory has no dependency and only requires probabilities.

Unfortunately, too many texts in probability theory fail to make this important distinction, and instead sometime injudiciously use the phrase “random variable” to include both cases, and at other times use the term more carefully but fail to appreciate the importance of probability spaces that have no value functions. In any event, it is often that references are encountered to the more restrictive “random variables” where the more inclusive “chance variables” would have been more appropriate.

For example, most definitions of discrete stochastic processes describe such constructs as “a sequence of random variables”. It would have been more inclusive – and more correct – to have said “a sequence of chance variables”. Indeed, the prediction theory aspects of information theory (see Part III) utilize stochastic processes very heavily. Yet they are not required to restrict their attention to random variables – since the more general chance variables are of interest there.

For example, [Kleeman 2012, Lecture I, p. 1], explaining the purpose of information theory, says

> “The central idea of information theory is to measure the uncertainty associated with random variables.”

While I completely agree with Kleeman that *information theory information theory* is the is the mathematics of uncertainty, I would have preferred that the more general and inclusive phrase “chance variables” had been used in place of the phrase “random variables” in his above utterance – a substitution that I expect would be entirely consistent with his meaning.

Of course, information theory can just as well work with chance variables that are also random variables. However, the mathematic constructs of information theory – entropy and its related entropic functionals – generally make no use of a value function, thus allowing information theory to serve many new classes of applications.

More likely, the common usage of the phrase “random variable” is actually ambiguous[6] – sometimes taking the more restrictive meaning that I suggested above, and sometimes taking the more general meaning that I am attributing to the phrase “chance variable”.

Nevertheless, my fear is that many, if not most, of the probability spaces considered by information theory could be left out of the conversation if the more restrictive meaning is used. For this reason, I am making an issue of this distinction more than most probabilists bother to do, and resurrecting the phrase “chance variable” (from

[6] To complicate matters, the phrase “random variable” enjoys inconsistent usage in the scientific literature. It is true that probability theorists and other mathematicians are pretty consistent – using the phrase strictly to require real numbers as sample points. However, other scientists are often less disciplined, and sometimes use the phrase to include all chance variables. It must also be understood that the phrase “chance variable” is not widely used. I have adopted it in this primer in order to be able to make the distinction between the two cases described.

Shannon) for the purpose of clearing it up. And, I shall continue with this distinction throughout the remainder of this primer.

Strategy to Measure the Values of Uncertainty and Information

We have said quite a bit about uncertainty, information and the relationship between the two so far. Since this relationship is the essence of information theory, lets pause and summarize what we have said.

1. Our initial goal is to develop a measuring function to calculate the amount, or degree, of information that is inherent in a trial.
2. A trial consists of a set of possible outcomes, called a *sample space* or *population*, and produces exactly one of them during the trial.
3. Accordingly, a trial has three phases in time: 1) the *uncertainty phase*, 2) the *realization phase*, and 3) the *information phase*.
4. In the uncertainty phase, there are some things that are known, but the outcome is not known. What is known is the set of possible outcomes. This set is also known by the terms *state space*, *sample space*, *population* and *alternatives*. We shall primarily use the term sample space. Each possibility in the sample space is known as a logical possibility, a sample point, a state and an event. We shall primarily use the term sample point.
5. There are even more things beyond this that are known in the uncertainty phase. Specifically, for each sample point, its *probability* of being the sample point that is *made manifest* during the realization phase is also known. The pairing of all of these sample points with their respective probabilities is called a *probability distribution* – which is also known during the uncertain phase of a trial. In fact, the *probability distribution* captures in one construct all that is known during the uncertainty phase.
6. However, what we really want to know – the outcome of the trial – is not known during the uncertainty phase of a trial.
7. In the realization phase, the outcome is made manifest, or *realized*. That is, exactly one of the sample points of the sample space is "selected". The "die is cast", so to speak. So the outcome is made manifest during the realization phase.
8. In the information phase, the observer (remember, there is an observer) becomes aware of which sample point was realized. Thus, the outcome is known during the information phase.
9. The uncertainty phase is very rich with mathematical constructs as compared to the other two phases. It has a lot of "equipment": a sample space with several sample points, and a probability distribution. In fact, this "equipment" is plenty to base a calculation of the *value of uncertainty* upon.
10. But it is the *realized outcome in the information phase* that we want to establish a value for, not the uncertain phase!
11. But, we have established an equivalence relationship between the uncertainty in the uncertain phase and the realized outcome in the information phase. This equivalence relationship says that the value of the realized outcome (or information) that is present in the information phase is equal to the relief received by the observer by the removal of any uncertainty that was present in the uncertainty phase.

12. Therefore, in order to calculate the "amount of information inherent in a trial", we can instead calculate the "amount of uncertainty that is inherent in the uncertainty phase of the trial". We can then assign that degree of uncertainty to the value of the information produced by the trial. We take this approach because there is enough mathematical material available in the uncertainty phase of the trial to make such a calculation. Whereas, there is not enough information in either the realization phase or the information phase to make such a calculation.

In the next few sections, we shall begin to implement this strategy for defining the measure of the degree if information of a trial that is used by information theory.

Specifically, we shall develop a measuring function (or measure) for calculating the degree of uncertainty found in the uncertain phase of a trial. Having proclaimed that this value should also apply as a measure of the amount of information in the information phase, then we shall specify that this measure is a measure of both the uncertainty and the information of a trial of an experiment.

Probability Spaces and Information Spaces

So far, we have been discussing a lot of "things": trials, sample spaces, sample points, probabilities, probability distributions, uncertainty, information, etc. Putting all of these together into a big "space, we arrive at the idea of an *information space* – which happens to be the same as the idea of a *probability space* with something new added. We shall delve into all of this in the present section.

Trials, Chance Variables and Probability Distributions

From what has been said, we can see that probability theory is the study of trials and experiments. But a trial is characterized by a probability distribution, which contains everything needed to calculate the degree of uncertainty of each possible outcome, or sample point, of the trial.

What we want to do next is to take this measure of uncertainty of single sample points and apply it to all of the sample points collectively of a sample space. That is, we want to spread out the notion of *uncertainty of a single* trial so as to apply it to all of the points of a sample space collectively.

This would be much like taking the idea of "the height of a child in a classroom" and generalizing it to measure the "height of all the children in the classroom". Of course, we normally do this by *taking the average*, or *mean*, of all the children – and thus arrive at a meaningful number that represents the heights of the classroom as a whole.

In the case of an *information space*, we want to take the idea of "the degree of uncertainty of a single sample point of the sample space of a probability distribution, and then generalize it so that it characterizes the overall *uncertainty* of the entire sample space of the probability distribution, rather than of just one sample point at a time. And, as with the children in the classroom, we shall take the *average*, or *mean*, of all of the individual sample points in order to arrive at an *overall* measure of the degree of uncertainty inherent in the probability distribution.

This overall measure of uncertainty inherent in the whole probability distribution shall be named the entropy of the probability distribution.

Thus, henceforth in the remainder of Part I, we shall develop this idea of the *entropy* of a probability distribution. And, we shall come to understand entropy as a function, which measures both the degree of uncertainty and concomitantly the degree of information inherent in the probability distribution of a trial.

Moreover, we shall eventually show how this idea of *entropy* gets re-applied and generalized in many possible ways by information theory to measure a very large number of probabilistic phenomena. In fact, it is fair to characterized information theory as the study, development, theory and application of the concept of *entropy*!

In keeping with the parlance of probability theory, we shall occasionally refer to a trial or its probability distribution as a *chance variable*. For example we may often refer to the *entropy of a chance variable* or the *entropy of a probability distribution*. This terminology is used interchangeably.

Probability Spaces

In the previous section, we established that everything we need to know about a trial is contained in the probability distribution of that trial. However, the formal term for a *trial* in probability theory is *probability space*, and the formal definition of a probability space takes apart the concept of probability distribution and identifies it as three distinct parts. These are captured in the following definition.

Probability space (preliminary definition):

An ordered triplet (S, E, p), where:
S is a set of elements called sample points. S is called a sample space, or population.
E is a set of subsets of S. These subsets are called *events*.
p is a probability measure that assigns probabilities to all of the elements of E – which of course are also subsets of S.

Thus, if A is a member of E, then we shall write p(A) to mean "the probability of A".

To make this definition useful for our needs, we shall add a couple of stipulations.

First, note that we have just defined *probabilities* on subsets of S – not on the sample points of S themselves. This contradicts our earlier usage – wherein we have been saying that probabilities are defined directly on sample points.

We shall now amend this earlier usage to comply with our definition immediately above. We shall still allow the idea of "the probability of a single sample point". However, we shall amend that idea to mean "the probability of the set whose only member is that sample point".

In other words, we have just switched our focus from "sample points" to "sets of sample points" – which we are calling "*events*". Of course, we are still including single sample points in our consideration, because an *event* can be a set whose only element is a single sample point. However, probability theory finds it more useful to switch to an emphasis on *events* – and to include "individual sample points" in the conversation by treating them as "singleton events", which we shall also do henceforth.

That is, if x is a sample point in S, then we define "the probability of x" to mean p({x}) – or, the probability of the set whose only member is x. This definition is now consistent with our definition of *probability space* in the present section – because {x} actually *is* a *subset* of S, as it needs to be; whereas "x" is an *element* of S and *not* a subset of S.

Second, we cannot allow E to be just any set of subsets of S. Rather, we need to be able to combine any of the subsets of E in order to produce other subsets of E. This trait will ensure that, after the combing has occurred, that the resulting subset will also have a probability defined for it. The kind of "combining" that we are talking about is the Boolean operators "and", "or" and "not".

In order to assure that this kind of "combining" of events always produces another set that has a probability measure defined on it (that is, it is also an event), probability

theory imposes some rules on the combining operations. These "rules" are called a *sigma-algebra* – or sometimes a *sigma-field*.

However, to avoid going into the theory of sigma-algebras in this primer, we shall instead describe some stipulations that, in a non-rigorous manner, amount to the same result as the theory of sigma-algebras. These rules are slightly more restrictive (especially rule #1) than those of an actual sigma-algebra. But in the interest of simplicity, we shall use them for this exposition nevertheless.

So, to achieve this result in a simpler description, we shall require the following:

1. For every sample point x in S, we shall require that the set {x} is an element of E. (Lets call the set "{x}" the "singleton set" of x.)
2. In addition, we want to be able to logically combine any existing elements of E and have confidence that the result has a probability – is also an element of E. By "logically combine" we mean the set-theoretic operations of union, intersection and complementation.
3. It can be demonstrated that, taken together, the above two specifications imply that S itself is a member of E, and also that the empty set Φ is a member of E.

As we said above, this specification is stated in probability theory by saying that together, (S, E) constitutes a sigma-algebra.

Having explained all of this, we can now restate our definition of probability space so that it includes these considerations and stipulations.

Probability space (final definition):

An ordered triplet (S, E, p), where:
S is a set of elements called sample points. (S is called a sample space, or *population*).
E is a set of subsets of S. That is, E is an *event* of S. Further, it is required that (S, E) constitute a sigma-algebra.
p is a probability measure that assigns probabilities to all of the elements of E – which of course are also subsets of S.

It is important to understand that, in probability theory, a probability distribution is an "abbreviated version" of all of the information contained in a probability space. Another way to say this is that a probability space is a more detailed breakdown of a probability distribution. In probability theory, we most often refer to the probability distribution, and only occasionally – when we need to engage in more critical thinking – do we need to refer to the probability space version.

Information Spaces

Obviously, probability spaces are an essential foundation of information theory. Notice that one of the three essential elements of a probability space is "p", the probability measure. Note that any probability measure is a function whose domain space is E and whose result (codomain) is a real number (the probability). Specifically, for any probability space, this result lies between 0 and 1, inclusive, and these probabilities must sum to 1.

But, as we have said, information theory starts with a probability space and then adds a significant new measure – that of *uncertainty*. This fact will be reflected in our formal definition of *information space*. In fact, we shall define *information space* in this section by adding a single measure to our definition from the previous section of *probability space*.

Like a probability measure, *uncertainty* is also a function that maps an event from E to a real number. But, whereas probability is a *measure of the likelihood of an event*, *uncertainty is a measure of the uncertainty of an event*.

An *information* space is, then, a probability space with the addition of a new measuring function - uncertainty. So, in information theory, each event in E has both a *probability measure* and an *uncertainty measure*.

This means that, for the purposes of information theory, we need to enlarge our definition of probability space to also include an uncertainty measure as well as the existing probability measure. We shall call this new, enhanced, probability space an *information space*.

Thus, we shall define information space as follows.

Information space:

An ordered quadruplet (S, E, p, u), where:
S is a set of elements called sample points. (S is called a sample space, or *population*).
E is a set of subsets of S. That is, E is an *event* of S. Further, it is required that (S, E) constitute a sigma-algebra.
p is a probability measure that assigns probabilities to all of the elements of E – which of course are also subsets of S.
u is an uncertainty measure that assigns a degree of uncertainty to all elements of E.

Obviously, the definition of information space is the same as the definition of probability space with the addition of the measure of uncertainty “u”. What we must do now is to define the uncertainty measure “u” as a well-defined mathematical function that works for any information space.

Steps to Realizing Our Goal

Our goal at this point is to develop a function that measures the degree of information that is inherent in a trial. We have determined that the material that we have available to us in order to make such a calculation is the information space – which is a mathematical depiction of the three phases of a trial. Unfortunately, those three phases are rich in “intellectual equipment” about the *uncertainty* aspects of the trial and paltry concerning the *information* of the trial.

Fortunately, though, we have discovered a relationship between the *uncertainty* aspects and the *information* aspects that enables us to equate their values. We have shown that the *amount of relief given by the removal of the uncertainty* actually *constitutes the value of the information* of the trial [Khinchin 1957, p. 7]. Therefore, if we can calculate the amount of uncertainty of the trial, we can reasonably apply that value as the value of the information of the trial.

Thus, this approach converts our problem of calculating the amount information in a trial to one of calculating the amount of uncertainty inherent in a trial.

In the previous section, we discussed the possibility of defining the degree of uncertainty of the individual sample points of the uncertainty phase of a trial. We have not yet defined such a measuring function, but we did decide that it is possible to do so.

However, even if we were to develop a measuring function for the amount of uncertainty inherent in each sample point, this would not be enough to address our problem – because we need to measure the amount of uncertainty inherent in the trial as a whole – not just its individual sample points.

Therefore, the work that we have cut out for ourselves can be addressed in two steps:

1. To develop a function that measures the degree of uncertainty inherent in each *event* of a trial. (An *event* can consist of a single sample point.) *We* have previously given the name *uncertainty* to this measure, and symbolized it by "u" when we defined information space.
2. To use this measuring function "u" to develop a second function that measures the amount of information inherent in the whole probability distribution (or probability space) of the trial – rather than just one of its events at a time. Historically, in information theory, this function has been given the name *entropy*, and symbolized by "H".[7]

The remainder of Part I of this primer is the definition of these two measuring functions, "u" and "H", and the leveraging of them to develop a scheme to apply a definition of value to the concept of the information of a trial.

Measuring the Uncertainty of an Event

In this section, we are going to present a function that measures the amount of uncertainty inherent in any single *event* of a probability distribution. We shall name this measure "*uncertainty*", which is an abbreviation for the "degree of uncertainty inherent in an event of a probability space". And we shall symbolize this measuring function by "u(x)" – or "u of x" – where "u" is the symbol of the uncertainty function and "x" is an event of S. "u(x)" is also called "the uncertainty of event x".

The development of this measure u(x) of the uncertainty of a single event of a probability space lies at the very heart of information theory. Pretty much every consideration that information theory makes after this will be defined in terms of this measuring function.

Even entropy itself – the most essential element of information theory - which measures the degree of uncertainty of the entire probability distribution - will be defined in the next section as the average of this measure over all of the singleton sample points (singleton events) of a probability distribution[8]. And after that, pretty much every significant idea in information theory is defined in terms of entropy.

---

[7] Historically, the idea of *entropy* was invented in physics by Clausius first (in the mid 19th century), re-defined by Boltzmann for statistical mechanics in the late 19th century and then generalized to apply to any situation with a probability distribution in information theory in the mid-20th century. It was Boltzmann who first introduced the letter "H" to symbolize his statistical ideas. However, the "H" of information theory that we shall be using has the negative value of the "H" of Boltzmann. (As Boltzmann's "H" goes down, information theory's "H" goes up.) In fact, Boltzmann's H is equal to –k time the information theory H, where k is a constant scaling factor, or conversion factor, that can be ignored for information theory purposes [Tolman 1937; Shannon 1948].

[8] Technically, in probability theory, *probability* is a measuring function that is defined on (measures) *events*, not on sample points. We have said that an *event* is a set of sample points of a probability space. However, in normal parlance, we often talk of "the probability of a sample point". In order to be able to use this language with our definition of probability being defined for events, we shall stipulate that the phrase "the probability of a sample point" actually means "the probability of the singleton event whose only member is the sample point mentioned". By doing this, we remain consistent with our specification that *probabilities* are defined only on *events*, but also support the idea of "the probability of a sample point". Thus whenever s is a sample point, then "p(s)" really means "p({s})".

In this section, we shall present both a verbal and a mathematical definition of "the uncertainty of event x", or u(x).

However, this section will not present a motivating discussion that explains why the definition that is presented makes intuitive sense. Such a presentation is left for Appendix 1 – which is dedicated to such a discussion. It is strongly recommended that any reader who wants to develop a good understanding of information theory take the opportunity at this time to read and understand Appendix 1 – if not now, then at some point later.

The Measure of the Uncertainty of a Sample Point

In this section we shall define the function that is used in information theory to measure the degree of uncertainty of a single sample point of a probability space. It is:

Measure of the uncertainty of a sample point:

$$u(x) \begin{cases} = \log(1/p(x)) & \text{for x where } p(x) \neq 0, \text{ and} \\ = 0 & \text{for x where } p(x) = 0. \end{cases}$$

Notice that we had to make a special case for x where p(x) is zero, because the expression "log(1/p(x))" entails division by zero for that case. In other words, u(x) is define by the expression "log(1/p(x))" as long as p(x) is not zero. But if p(x) is zero, then the expression would involve division by zero, which is not defined. Therefore we cannon use the above expression when p(x) is 0. For that case, we define u(x) = 0.

This function is foundational in information theory, because it, together with probability theory, is the root of all that follows.

You may be wondering why it is necessary to use logarithms to define the idea of uncertainty. Why can't we use a simpler expression? The answer to that question is explored in Appendix 1. Because of the importance of this measure in information theory, it is strongly suggested that the reader consult this appendix!

Surprisingly, this measuring function has not been assigned a widely adopted name! Although one researcher tried giving it the name "surprisal", that name has not enjoyed universal adoption. For this text, as mentioned above, we shall refer to it as the "uncertainty of a sample point of a probability space".

Actually, most texts on information theory begin by defining the notion of *entropy* from the start. However, this author has found these concepts easier to understand if the uncertainty function is defined first, and then *entropy* is defined as the *average uncertainty*.

In addition, the more primitive notion of *the uncertainty of an event* is the unifying idea that weaves together an entire constellation of other measuring functions – functions that are actually generalizations of entropy, and that ultimately constitute the more advanced topics of information theory. We shall call these the *entropic measures* and dedicate most of the discussion of this primer to them.

Specifying u(x) for a Specific Distribution

Notice that u(x) depends upon the probability distribution p. Therefore, if there are two probability distributions on the same sample space, say p and q, then there will be two distinct measures of uncertainty for the sample points "x" of the space. We shall name them $u_p(x)$ and $u_q(x)$.

In this case,

$u_p(x) = \log(1/p(x))$ for x where $p(x) \neq 0$
$u_q(x) = \log(1/q(x))$ for x where $q(x) \neq 0$

In fact, there may be any number of probability distributions defined on the same sample space. And, each will have its own uncertainty function[9].

The necessity to distinguish between two or more probability distributions on the same sample space could occur, for example, if the probabilities of the sample points change from time to time. This distinction will become necessary in the section below on *relative entropy*. In that case, we shall compare the *actual* probability distribution that we observe on a sample space to a probability distribution that we anticipated or expected.

If we are not comparing two such distributions, then we can simply use the symbol "u(x)", because it will be implicitly clear which probability distribution is being used. But if two or more probability distributions on the same sample space are being compared, then we can use the subscripted notation "$u_p(x)$" or "$u_q(x)$" as discussed.

Simplifying u(x)

The function u(x) is pretty simple as it stands. But we shall show here that there is a simpler equivalent expression still:

$u(x) = \log(1/p(x))$
$= \log(1) - \log(p(x))$
$= 0 - \log(p(x))$
$= -\log(p(x))$

The form of the second line is due to the fact that the log of a quotient is the difference between the log of the numerator and the log of the denominator. The form of the third line is due to the fact that log(1) is 0.

Thus

$u(x) = -\log(p(x))$

One most often encounters this latter form in information theory references because of its simplicity and reduction in the number of calculations. However, it is one step removed from the initial intuitive definition and meaning of Aristotle's idea of the inverse relationship between probability and uncertainty. Thus, this primer will use both forms, but will more often use the first form in order to remind the reader of the inverse relationship between probability and uncertainty.

Choice of log base for u(x)

The astute reader will note that something is missing from our definition of u(x) – namely that we have not specified which base to use for the logarithm function.

[9] Any sample space may have more than one probability distribution – generally an infinite number. Of course, each of these probability distribution can be used to define its own uncertainty function u(x). As well, each probability distribution defines its own probability space. And each probability distribution together with its unique uncertainty measure defines its own information space.

It is true that some log base must be specified before a calculation can be completed all the way to producing a numerical result.

However, we did not specify a log base as of yet because everything we have said about u(x) so far is true regardless of which log base one selects. Thus, as it stands, our definition of u(x) permits the user to select any log base that is desired. Of course, the choice of distinct log bases results in distinct numerical answers. So allowing the users to select their choice of log base works as long as the same log base is used when making comparisons between different calculation of u(x).

Popular log bases are 2, 10 and e. However, computer scientists are partial to the use of the log base 2, since it results in uncertainty values that are modeled by "bits" in computer architectures.

In this primer, we shall primarily use a log base of 2 with examples that involve calculations all the way to a numerical answer. Thus, for these cases, one can specify the definition of u(x) as:

$u(x) = \log_2(1/p(x))$

But in more abstract discussion, we shall omit the specification of a log base in order to emphasize that what is being said is true regardless of the selection of log base.

Example

Suppose the probability of catching exactly 3 fish when one goes deep sea fishing with a particular charter service has been observed to be 1/16 – or one out of every 16 fishing outings. Then what is the uncertainty of catching exactly 3 fish?

(We shall use the log base of 2 in this and all following examples in this primer – unless otherwise noted.)

This is calculated by substituting 1/16 into our definition of u(x). The symbol "3fish" shall represent the event that we catch exactly three fish.

$u(3fish) = \log_2(1/p(3fish))$

$= \log_2(1/(1/16))$

$= \log_2(16)$

$= 4.$

We could also have started with the equivalent expression for u(x):

$u(x) = - \log_2(p(x))$

which would have given us:

$u(3fish) = -\log_2(p(3fish))$

$= -\log_2(p(3fish))$

$= -\log_2(1/16)$

$= -(-4) = 4.$

Either way, our measure of the degree of uncertainty inherent in the sample point "exactly 3 fish are caught" is 4.

Entropy: Measuring the Uncertainty of a Probability Distribution

In this section, we shall execute the second step of our task of defining "uncertainty": developing a function that measures the amount of uncertainty inherent in the whole probability distribution (or probability space)[10].

As we have said, historically this function has been given the name *entropy*, and symbolized in information theory by "H".

Recall that we have already decided above how we shall approach this. We shall use the measure u(x) of a single event or sample point that we just developed. Specifically, we shall define entropy as the mean, or average, u(x) measure across all of the sample points in the distribution.

In other words, we shall apply our expression

$u(x) = \log(1/p(x))$

to each sample point in the distribution, and then take their average.

One can take an average of a set of numbers by adding up the numbers and then dividing that sum by the number of numbers. But in probability theory, there is an equivalent but more efficient way to calculate an average. This way is called the *mean*. To calculate the mean, you multiply each distinct number in the set by its probability, yielding a product. Then you add all of these products together and obtain the mean. We shall use that approach to calculating the mean of all of our "u(x)" values.

So, for all sample points, x_1, x_2, x_3,…, x_n, we want to multiply the probability of that sample point "x" times "u(x)" for that sample point. This gives the expression:

$\text{entropy}(p) = p(x_1)*u(x_1) + p(x_2)*u(x_2) + p(x_3)*u(x_3) + \ldots + p(x_1)*u(x_n)$

Notice that we used "entropy(p)" – the entropy of probability distribution p – to emphasize that entropy is a measure of an entire probability distribution. Specifically, entropy is the mean uncertainty of the sample points of a probability distribution.

Moreover, the argument p of the entropy function specifies precisely which probability distribution of the sample space of interest is being used to calculate the entropy. It is possible that more that one probability distribution is defined on the same sample space – each being applied at distinct times.

We can re-write the above expression by substituting the definition of u(x), which is "log(1/p(x)", in place of every "u(x)" in the above expression for entropy. At the same time, we shall use the symbol H for entropy as described above, giving:

$H(p) = p(x_1)*\log(1/p(x_1)) + p(x_2)* \log(1/p(x_2)) + \ldots + p(x_n)* \log(1/p(x_n))$

We can write this definition in a more manageable form using the summation symbol $\sum$. This refined definition of entropy is:

[10] Admittedly, Shannon introduces and defines the concept of entropy first [Shannon 1948, p.11], and then subsequently points out that entropy is simply the average of a more fundamental quantity, which he calls the "entropy H_i of each state" [Shannon 1948, p.13]. Shannon's "entropy H_i of each state i" is what this primer has named the "uncertainty of an event", and symbolized by the functional u(x). Clearly, the "entropy H_i of each state i" is a more fundamental quantity than the entropy of a probability distribution H - since H is the average of all the "H_i"s, as Shannon points out. It is the view of this primer that much of the mystery surrounding *entropy* is dispelled by introducing the more fundamental uncertainty function ("u(x)", or "H_i") first, and then defining H, or "entropy", subsequently as the average of the more fundamental quantity. Some contemporary texts [Vedral 2010] take the same approach in this regard as this primer, while others do not [Cover and Thomas 2006].

$$H(p) = \sum_{i \in S} p(x_i) * \log(1/p(x_i))$$

H(p) is called the entropy of probability distribution p, or entropy of chance variable P, or even the entropy of the probability space P.

We can also use the simplified version of our definition of u(x) in the previous section to articulate the corresponding simplified version of H(p). It is:

$$H(p) = -\sum_{i \in S} p(x_i) * \log(p(x_i))$$

As before, we shall use both definitions in this primer. The first is held to be more intuitive and as reflecting the meaning of uncertainty (per Aristotle), while the second is a simpler expression.

Interpretations of Entropy

To recapitulate, our definition of the entropy of probability distribution p is

$$H(p) = \sum_{i \in S} \log(1/p(x_i))$$

Where

S is a discrete sample space
p is a probability distribution on S
x_i is the i-th sample point in S
$p(x_i)$ is the probability of x_i according to p.

Also, since we are interchangeably referring to probability distributions as chance variables, then we could replace our references to “p” with references to “X”, the chance variable terminology for p. This is the terminology that is most often used in the literature. Thus, alternatively we can say that the entropy of chance variable X is given by

$$H(X) = \sum_{i \in S} \log(1/p(x_i))$$

Where

S is a sample space
X is a chance variable on S
x_i is the i-th sample point in S
$p(x_i)$ is the probability of x_i according to X.

Entropy is a measure of the average uncertainty that is inherent in a probability distribution[11]. This is made clear by looking at its mathematical definition above, and

[11] The careful reader may have noticed the use of the words “mean” and “average” when characterizing the notion of the entropy of a probability distribution. And, in so noticing, the question may have arisen; “If we are calculating a mean, does this not imply that a *random variable* is involved in information theory – rather than merely requiring a *chance variable*”? More specifically, the implication is that since the entropy is the mean of the uncertainties u(x) of each of the sample then this u(x) must be a real valued function (which it is) and therefore a *random variable* – not merely a *chance variable*. Does this fact - together with the insistence that information theory is characterized as the study of entropy - not undermine the claims of this primer thus far that information theory deals with the more general idea of *chance variables* rather than the more restrictive idea of *random variables*? We shall now argue that, even so, it is misleading to characterize

understanding that entropy is the mean uncertainty u(x) of all of the sample points of the sample space of the distribution.

In fact, entropy is information theory's measure-of-choice of the uncertainty inherent in a probability distribution. Further, it is reasonable to say that information theory is an extension to probability theory wherein entropy has been added [Khinchin 1957, p. 1]. That is, information theory = probability theory + entropy + resulting consequences.

An Alternative Interpretation of Entropy: Randomness

Another reasonable interpretation of the definition of the entropy of a probability distribution is:

> Entropy: The degree of randomness, or non-determinism across the sample points of the distribution.

"Randomness" is another word for *chance variation*. And it is reasonable to define *determinism* as the absence of chance variation. Further, it is also reasonable to understand determinism as "zero amount of chance variation.

From this we can conclude that both *randomness* and *determinism* are points on the same continuous measuring scale – the scale of chance variation. Of course, *determinism* occupies exactly one point on this scale – the point that has zero amount of chance variation. On the other hand, all other points on this scale represent varying degrees, or positive amounts, of chance variation.

And, of course, *entropy* provides precisely this measure. So this explanation provides another reasonable interpretation of entropy a measuring the degree of randomness inherent in a probability distribution.

Another Alternative Interpretation of Entropy: Spread

Another reasonable interpretation of the definition of the entropy of a probability distribution is:

> Entropy: The degree of the spread of probability across the sample points of the distribution.

---

the uncertainty of an event u(x) as a random variable – and we shall refrain from doing so here. Tbis argument follows:
While it is true that u(x) actually fits the definition of a *random variable*, its usage in information theory is so different from a typical random variable that it would stretch the credulity of the term beyond broad acceptability. This is true on two accounts.
The first is that a typical random variable on a probability space is *not* defined in terms of the probabilities of the space - but instead stands alone and independent of them. On the other hand, u(x) has as its only input parameter the probability of the sample point that it is defined on.
Secondly, the utility of recognizing a value function as a random variable is so that the whole arsenal of functionals defined on random variables can be brought to bear on the space. This arsenal includes functionals such as the mean, variance and all of the central moments. However, in the case of an information space, the only one of these functionals that is used is the *mean* (used to define the entropy), which is already defined outside of the notion of random variables anyway.
Therefore, while the readers would technically be within their reasonable bounds to argue that the functional u(x) is in fact a random variable, such u(x) exhibits two properties that are so unlike typical random variables that this primer finds it more useful to not count u(x) within that category.
Therefore, the author stands by his earlier rant claiming that information theory does not require the more restrictive *random variables* and is able – unlike the discipline of statistics – to handle the more inclusive category of chance variables, thereby availing itself to many interesting and complex applications that are not so available to statistics.

Looking at the measurements for various probability distributions over the same sample space shows this.

For example, for a given sample space, the distribution with the maximum entropy is the uniform distribution – where the probabilities are evenly spread across all the sample points. That is, all of the sample points are equiprobable. This is the "most democratic" spreading around of the probabilities across the sample points. It is also the distribution with the highest degree of uncertainty.

So, in a sense one could say, *entropy* is a measure of how "democratically" the probabilities are "spread". For example, in applications that involve chance variation in marketing or other propagandistic enterprises, entropy could reasonably provide a measure of the adoption rate of those ideas.

On the other hand, for a given sample space, the distribution with the minimum entropy (zero) is the constant distribution – where one sample point has the probability of 1, and all the other sample points have probability zero. This not only has the most certainty, but also the least amount of "spread" of probabilities.

This characteristic makes statistical entropy a very good analog of thermodynamic entropy, which some say measures "energy spread" [Leff 2012].

Notice that I did not say that these two versions of the word "entropy" are exactly the same. They are not. However, thermodynamic entropy and statistical entropy can be shown to be mathematical analogs of each other. In the statistical mechanics analog of thermodynamics, the spread of energy across the spectrum of possible states is probabilistic; and statistical entropy turns out to be a good analog of thermodynamic entropy. R.C Tolman authored an important book about this subject entitled "The Principles of Statistical Mechanics" [Tolman 1938].

However, information theory is about statistical entropy (as is statistical mechanics), whereas Thermodynamics is a branch of physics that defines the notion of thermodynamic entropy. Therefore, except for occasional references, we shall deal almost exclusively with statistical entropy on the present information theory primer.

So, we have established that a second reasonable interpretation of statistical entropy is that of the degree of the spread of probability across the sample points of a distribution.

However, we can directly derive a third interpretation, mathematically, from this second interpretation. It is: "a measure of the *degree of uncertainty* across the sample points of a probability distribution.

Thus we now have three reasonable and mathematically equivalent interpretations of our mathematical definition of statistical entropy, as follows:

Three interpretations of statistical entropy:

1. A measure of the degree of uncertainty of a probability distribution.
2. A measure of the degree of the uniformity of spread of the probabilities across the sample points of the distribution.
3. A measure of the degree of the uniformity of spread of the uncertainties u(x) across the the sample points of the distribution.

We shall see that each of these interpretations has its applications in information theory.

For conciseness, in what follows we shall refer to the first of these interpretations and not the other two. But the reader should remember that the other two meanings are also reasonable interpretations. For any particular application of these entropy measures, any of these interpretations may be used as appropriate.

Some Probability Distributions are More Uncertain than Others

The fact that the uncertainty of a probability distribution is worth measuring implies that some probability distributions have more uncertainty than others. In fact, some probability distributions have a maximal degree of uncertainty, some have a minimal degree of uncertainty and all the rest have intermediate degrees of uncertainty.

In this primer we are confining our attention to probability distributions with finite sample spaces. (For theory and practice of information spaces with infinite sample spaces, consult [Cover and Thomas 1991, p. 224].) Probability distributions on a finite sample space have a range of possible entropy values that ranges from a value of zero for deterministic distributions to a value of log(N), where N is the size of the sample space).

So, entropy measures this degree of uncertainty of a probability distribution between a maximum and a minimum value. We shall see that the number of sample points in the probability space determines the maximum value. The minimum entropy value for any probability space is zero (0) for deterministic distributions and the maximum entropy value of log(N), where N is the number of sample points, for completely random distributions.

In fact, for a given finite sample space, the probability distribution with the maximum entropy (degree of uncertainty) is the uniform distribution, where all sample points are the same (equally likely). Since all of the probabilities involved are equally likely, then they do not give us any hint at all as to which one will be realized, and therefore this situation represents the most uncertainty.

The probability with the minimum entropy (value 0) is the so-called *constant distribution*. In this distribution, exactly one of the sample points has a probability of 1 and all of the rest have a probability of 0. In this situation, it is absolutely certain that the sample point whose probability is 1 will be realized during the trial, and that none of the others will. Therefore, this situation represents certainty: the certainty that the sample point whose probability is 1 will be realized and also the certainty that none of the others will be realized.

These two distributions have the maximum and the minimum entropies for their sample space size, and any other distributions for that sample space will have an entropy that lies somewhere between the minimum (0) and the maximum.

Entropy as a Statistic

A “statistic” is any measure of a probability distribution[12]. That is, a statistic is a function that associates a single number with a probability distribution and that measures some particular aspect of the distribution. For example, you are probably

[12] Technically, a “statistic” measures some aspect of a *sample* of a probability distribution. The term “parameter” is used in the study of statistics whenever the entity being measured is the entire population, rather than merely a sample. So, we are using the term “statistic” in a more general sense. Formally, we should probably be using the term “population” here rather than “statistic”. Admittedly, information theory – being a “sister science” to Statistics does not make the distinction between “populations” and “samples”, and instead only deals with populations.

familiar with statistics such as mean, median, mode, variance and standard deviation. These are all statistics because they all assign a single number to describe some aspect of a probability distribution.

Different statistics measure different aspects of a distribution. One such aspect is "central tendency" – which is the tendency of some distributions to "crowd around" some center of the distribution. Such as statistics as mean, median and mode are "measures of central tendency". Other aspects are "measures of variation", which are measured by variance and standard deviation.

Entropy is also a "statistic" because it too assigns a single number to a probability distribution. And entropy also describes some particular aspect of the distribution. For entropy this aspect is *uncertainty* (as opposed to "central tendency" or "variation"). Entropy is a measure of the uncertainty inherent in a probability distribution.

However, entropy differs from the other statistics named above because it applies to any probability distribution. On the other hand, the other statistics named only apply to probability distributions whose sample points are numbers. (Algebraically, they form a field.) This is because these other statistics require arithmetic to be performed using the sample points themselves – not just their probabilities. For example, the calculation of the mean requires that you multiply each sample point by its probability. Of course its probability is already a number, but the sample point itself may not be.

Thus, if the sample points of a probability distribution are not numbers, then the mean is not definable for it. Nor are standard deviation, variance or the other statistics named above. Of course, it is possible to associate a number with each sample point, and then use that number in place of the sample point. But if those number assignments are arbitrary and don't make sense to the sample space involved, then the statistics will also be meaningless.

An example of this is the roll of a six-sided die that is used in games. These six die faces are often identified by some unique number of dots on each face – usually 1 to 6 dots. Thus, one could consider the sample space to be the numbers 1 through 6. But taking a mean of these numbers is not meaningful. The reason for this is that those particular numbers do not represent "amounts" or even "order". Rather, for dice, the numbers on their faces merely represent identities. Symbols other than numbers could just as well have been used. The number used on the faces of dice for identification purposes only do not have the full richness of numbers: amount, difference or even ordering (left to right). Thus, the use of numbers to represent the six die faces is for identification only – and not of any of the other aspects of numbers that are the notions of mean, median, variance, etc., utilize.

Thus, if the sample points of a probability distribution are inherently non-numeric, such as dice or even coins, then the application of statistics such as mean, median, variance and standard deviation are meaningless. On the other hand, the idea of the "amount of uncertainty" inherent in the rolling of a die or the flipping of a coin are, indeed, meaningful. The point here is that, even probability spaces for which the typical statistics (mean, median, variance, etc.) are meaningless, the entropy is nevertheless meaningful. In fact, entropy is a meaningful statistic for any probability distribution!

For example, consider an experiment in which a coin is flipped, with heads and tails constituting the sample space. Suppose, for example that the probability distribution is the uniform distribution. That is the probability of both heads and tails is ½. Then in order to calculate the mean, we have to multiply Heads by ½ and Tails by ½ and then add these two products. But you cannot multiply either Heads or Tails by ½ because

neither Heads nor Tails is a number! And, you cannot calculate the mean of this distribution! Or the variance, standard deviation or median!

However, the calculation of the entropy of tossing a coin is defined! This is because the calculation of entropy requires only that each sample point have a probability. It does not require that the sample point itself is a number.

Another example is any sample space whose sample points are each compound combinations of simpler things. Even if the simpler things are numbers, the compound combinations are not. This situation is often encountered in science. For example, in classical and statistical mechanics, the sample points are descriptions of the location and momentum of particles. Even though the components of these sample points are numbers, the sample points themselves are not. They are far richer than numbers. And none of the statistics mentioned are defined for them. However, both u(x) and H are defined for them.

You run into the same problem in one way or another with all of the other statistics mentioned except entropy. Even the mode statistic requires the implied ordering of numbers, even though no arithmetic operations are involved.

But entropy is a special statistic – it does not require that the sample points themselves be numbers.

To see this, just look at the entropy function definition again:

$$H(p) = \sum_{i \in S} p(x_i)*\log(1/p(x_i))$$

The only input values in this definition are the probabilities $p(x_i)$ of the sample points. The sample points themselves do not appear in this function – as they do in the other statistics mentioned such as the mean or the variance. Therefore, the sample points need not be numbers, but can be anything: balloons, balls, monkeys, compound combinations of monkeys or even numbers!

Random Variables

Therefore, there is a need to distinguish between probability distributions (or chance variables) whose sample points are numbers (or fields) and those that are not. This distinction is made with the use of the terms *chance variable* and *random variable*.

Formally, a random variable is a chance variable whose sample points are real numbers. On the other hand, this primer refers to all probability distributions as chance variables. Thus, the concept of random variables is a special case of the concept chance variable.

The significance is that entropy can be calculated for all chance variables, but the other statistics mentioned (mean, median, mode, and the central moments) can be calculated for random variables only, and cannot in general be calculated for (are not defined for) all chance variables.

Unfortunately, the use of the term *chance variable* is not widespread. In fact, it has been defined specifically in this primer in order to make this important distinction. Regrettably, the use of the term random variable in the probability theory literature is ambiguous. Some uses of the term (often by mathematicians) restrict its use to probability distributions whose sample points are real numbers – as I have here.

On the other hand, many references to the term *random variable* apply it to all probability distributions. Even more regrettably, many references assume that all

probability distributions have real numbers a sample points, and that moment-like statistics can be calculated for them all.

In any event, the reader should be aware that the usage of the term *random variable* in the literature is ambiguous. It is often used to mean “any probability distribution”, whereas, formally, a random variable requires real numbers as sample points. This ambiguity is unfortunate, because some sample points are too complex to be real numbers. This is true for certain compound state spaces such as found in classical and statistical mechanics.

However, this primer assumes that the distinction between chance variables and random variables is significant – as is the fact that entropy is distinctive in that it is defined for all chance variables, whereas the other statistics mentioned are only defined for some chance variables.

The Significance of Entropy

Thus, entropy is a statistic that can be applied to any probability distribution no matter what its sample points are. However, all of the other statistics mentioned require a special case: all of the sample points must be numbers (or mapped to numbers).

This makes entropy very special. But there is another reason for the special character of entropy:

> Entropy captures a fundamental attribute of all probability distributions: the notion of uncertainty.

If probability theory is about anything at all, it is about uncertainty. From this statement alone, I would argue that entropy is the most significant statistic of a probability distribution, because it captures the essence of probability theory.

All probability distributions have entropy defined for them, regardless of what their sample points are made of. The degree of uncertainty of a probability distribution is the first thing that you should want to know about it. It is its foundational characteristic.

This is why the addition of entropy to probability theory by Claude Shannon in 1948 was so significant that the enhanced theory was given a new name: information theory. Obviously, information theory is primarily about uncertainty. This is a point made by [Kleeman 2009, p. 1] where he states as the opening sentence to his seminar on information theory and predictability

> “The central idea of information theory is to measure the uncertainty associated with random variables.”

However, as we have already discussed, information results from the removal of uncertainty [Khinchin 1957].

And the value of that information (that resulted through realization) is the same as the value of the uncertainty that was removed to produce it. (The value of a solution to a problem is equal to the value of the amount of “trouble” cause by the problem.) Thus, uncertainty and information are imminently related. Thus, what should perhaps have been called “uncertainty theory” has been named “information theory”.

The Applications of Entropy

As we have said, entropy is a measure of uncertainty. How this idea applies to different situations is always subject to interpretation, as is any other statistic. The maximum uncertainty for any domain occurs whenever the situation is arrived at where

the probability of the system being in any of its possibly defined states is equally likely. For example, uncertainty in the stock market is often interpreted as “volatility”. Uncertainty in dynamical systems engineering is often reasonably interpreted as “instability”.

It is left to the application to determine the interpretation of entropy for its own domain. It is also left to each application to determine how much entropy is desirable for the application. Very often in dynamical systems theories, the desirable degree of uncertainty is intermediate between minimum (entropy = 0) and whatever the maximum value of entropy is for the sample space size (some positive real number). In fact, often the maximum amount of entropy is considered to be “chaos”, while the minimum amount is considered to be “dead”.

In any event, the entropy of a probability distribution may be its most fundamental statistic, as argued above. In addition, entropy has the advantage that it is defined directly for any probability distribution, regardless of whether its sample points are numbers. This fact is particularly useful for very rich sample spaces whose individual sample points are complex multidimensional compound constellations of underlying entities in the first place, and do not benefit from being “flattened” by being transformed to a sample space of real numbers. This is the case for the sample spaces used in Organodynamics, for example. It is also true for statistical mechanics. In these cases, entropy is the only statistic available for characterizing the probability space.

The Value of Information

We have said that a trial has three phases:

1. **Uncertainty phase**. (“Before the die is cast.”) In the uncertainty phase of a trial, the outcome is uncertain. A probability distribution is known for the trial in this phase, containing a specification of the sample space as well as the probabilities of each sample point. But the outcome – which specific sample point will be realized during the trial – is not yet known. The good news is that this probability distribution is a very rich mathematical construct. The bad news is that the outcome is not yet known.
2. **Realization phase**. (“After the die is cast.”) The outcome is manifest. Exactly one of the sample points has been realized. Uncertainty has been removed leaving certainty as the result.
3. **Information phase**. The observer becomes aware of which outcome has been realized. The identity of the outcome is called information. Information is also characterized by certainty.

Information theory takes the position that during the course of a trial, uncertainty has been removed though realization, and that this removal has resulted in information.

Moreover, information theory takes the position that the information (in the information phase) represents a “relief from suffering” that was caused by the uncertainty (in the uncertainty phase). Therefore, the value that should be attributed to the resulting information is equivalent to the value of the relief given by removing the uncertainty whose removal resulted in that information [Khinchin 1957].

Fortunately, the probability distribution in the uncertainty phase provides plenty of “material” with which to measure the value of the uncertainty. In fact, the *entropy function* described in the previous section does precisely that.

Therefore, we shall calculate the value of the entropy of the trial by using the probability distribution belonging to the uncertainty phase. The, we shall apply that entropy value to the information phase as well as to the uncertainty phase. In other words, the entropy function will ultimately measure both the degree of information in a trial as well as the degree of uncertainty in that trial.

In short, then, entropy is a measure of both uncertainty and information. And, the amount of uncertainty in a trial is equal to the amount of information produced by that trial.

Relative Entropy

Suppose we have two distinct probability distributions on the same sample space. Perhaps one of them describes the probabilities of the sample points at one time and the other describes the probabilities of the same sample points at a different time. Or, perhaps one of the distributions is our "best guess" at what the probabilities are for a sample space, while the other distribution is what we find out the probabilities actually are for that sample space – which we determine by making some observations.

For whatever reason, suppose we have two distinct probability distributions on the same sample space. We would like to compare and contrast these two distributions.

We have seen that information theory is primarily concerned with a specific aspect of probability distributions – namely *their degrees of uncertainty* – that we measure with *entropy*.

We also have a couple of other different, but equivalent, interpretations of entropy:

1. The degree of the uniformity of spread of the probabilities across the sample points of the distribution; or
2. The degree of the uniformity of spread of the uncertainties u(x) of the sample points of the distribution.

Above, we began discussing a situation in which we want to see "how different" one probability distribution is from another when both are defined on the same sample space.

We must ask, "How can entropy be used to measure this difference?" We shall find that entropy interpretation #2 above is apropos to our task of measuring "how different" are two probability distributions on the same sample space.

However, instead of being merely focusing on the difference of the uncertainties of the distributions as a whole, we shall focus on the difference of the uncertainties of the sample points. Relative entropy is an entropic statistic that is designed to measure this difference, and is the subject of this section.

Relative entropy is an important statistic in information theory, because it is used in special cases in Parts II and III to define other essential statistics that provide the ultimate power of information theory – namely, the degree to which one chance variable can be a predictor of another.

The Intended Applications of Relative Entropy

The situation, or application, that we are interested in addressing in this section is when you make a "best guess" as to what the probability distribution of a some chance variable is before you find out what the *actual distribution* is by observing it.

Once you have made your best guess, then you subsequently observe the chance variable and collect actual data on the sample space. At that point you will have determined the (an) actual probability distribution of the chance variable.

The two distributions that you are interested in comparing here are the "best guess" distribution and the "actual observed" distribution. What we want to do ultimately is to measure "how far off" the best guess distribution is from the "actual" distribution. In a sense, then, relative entropy is a measure of "how surprised you are", or "how much you learned" from the experiment.

One reason that you might bother making the initial best guess distribution first is that you may be able to have a richer understanding of actual distribution in light of how it compares with the best guess distribution. This is especially true if you are already have experience with the "best guess" distribution and you want to understand the actual distribution in terms of the "best guess" one. In other words, estimating a "best guess" distribution first give you a basis for comparison when you finally ascertain the actual distribution.

Another reason that you might bother to make a best guess distribution is that you might be performing statistical "hypothesis testing" within the scientific method or within some other process that uses statistical inference[13]. In such a case, one typically makes an assertion that one might hope to reject by the use of statistical inference. Such an assertion is the hypothesis that an observed phenomenon could not have happened by chance alone. This is called a null hypothesis. In statistical inference, one already has in mind a particular probability distribution that describes what "chance behavior" looks like. If one can show that the probability of an observed phenomenon, using that particular "chance" distribution is very small, then it is reasonable to conclude that the observed phenomenon is not from that distribution – and that the observed phenomenon "is unlikely to be happening by chance alone". In other words, some agent or influence other than chance must be at work.

This kind of thinking is the basis for the scientific method and traditional applications of statistics in science and engineering. For example, for a major class of experiments, scientists have a good reason to expect that the normal distribution will apply to the experiment they are working on. Therefore, they use the normal distribution as their initial best guess.

However, instead of the phrase "best guess distribution", practitioners use the Latin terms *a priori* distribution, or prior distribution. And, instead of the phrase "actual distribution", they use *a posteriori* distribution.

Researchers are usually not terribly shocked if the *a posteriori* distribution is considerably different from the *a priori* one. The important thing is that they have a better feeling for the actual distribution because of the fact that they have the *a priori* distribution as a context to compare it too.

Anyway, researchers generally "make an initial guess" at to what the probability is going to be before they collect data (the a priori distribution), and then – after they have observed the actual distribution (the a posteriori distribution), they compare the two to see how "far off" their initial best guess was defined to accommodate.

Even though scientists often use the normal distribution as the *a priori*, they sometimes have good reasons to use other distributions. For discrete sample spaces, a frequent example is the uniform distribution. Such an application would be the testing of a die (used in gaming) to see whether it is "fair" – that is, "has a uniform probability

[13] See Appendix 3: Three Approaches to Critical Thinking.

distribution" (equally likely die faces). If a "cheater" has loaded the die with a hidden weight, then the probability distribution may no longer be uniform. In hypothesis testing, other distributions that are frequently used for the a priori distribution are the "Student's t distribution", which is also used to compare two distributions, and the "Chi-squared distribution".

In the case of a "loaded die", the researcher – and the dice player – would be interested in "what is the difference in uncertainty" between a game played with a loaded die versus a game played with a fair die. A "fair die" is the most uncertain die possible, because each face has the same probability as any other face. In fact, the purpose of the game stipulating that a fair die be used is so that maximum possible uncertainty is the case for the game.

However, a loaded die decreases the uncertainty of the game by making it more certain that one particular die face – the one favored by the off-center hidden weight – is more probable than the others.

So, a comparison between a fair die and a loaded die comes down to a difference in the degree of uncertainty between their two probability distributions. The fair die has a higher degree of uncertainty than does the loaded die. This is an important point, and key to understanding the approach taken by information theory to compare probability spaces – actually, to compare information spaces.

Relative entropy is the statistic that information theory has developed for the purpose of measuring that difference – the difference between the degrees of uncertainty of two probability distributions.

Approach to Defining Relative Entropy

So our situation is this: We have two probability distributions on the same sample space whose degrees of uncertainty we want to compare. We shall call these distributions the *a priori* ("best guess") distribution and the *a posteriori* ("actual") distribution in the manner we just discussed. Lets give the name "q" to the *a priori* distribution and the name "p" to the *a posteriori* distribution.

We want to see how the differences in the uncertainties of p and q at each sample point are "spread across" the entire sample space. We are interested in two aspects at the same time. One of theses aspects is "how uniform" is the difference of uncertainty between these two distributions across all of the sample points. The second aspect is "how much difference" in uncertainty is there. We want our measure to take both of these aspects into consideration.

If there are many occasions where this difference is large, then this should contribute to elevating the overall value that we are measuring. And, if there is a lot of uniformity – "unanimity", or "agreement" – across all of the sample points regarding their values, then this should also contribute to elevating the value that we are measuring.

Our strategy, then, for doing this comparison will be this: For each sample point, we shall:

1. Calculate the difference between the uncertainty of that sample point based on distribution q and the uncertainty of that same sample point based upon distribution p.
2. Calculate the mean of these differences of uncertainty across all of the sample points in the sample space.

Defining Relative Entropy

In this section, we shall develop the algorithm for calculating relative entropy according to the two-point strategy we laid out in the previous section.

Recall that we are given two probability distributions, p and q, on the same sample space S. We want to develop an algorithm that places a measure on how these two distributions differ in their relative degrees of uncertainty.

The subsections below develop this algorithm one step at a time.

Calculating the Differences of Uncertainties

This step reads, according to the above section,

> Calculate the difference between the uncertainty of that sample point based on distribution q and the uncertainty of that same sample point based upon distribution p.

First we must articulate expressions for the "uncertainty of a sample point" for each of the two probability distributions p and q. Obviously, these are "$u_p(x)$" and "$u_q(x)$".

Next, we must subtract them. The question immediately arises "In which order?" Do we subtract "$u_q(x) – u_p(x)$" or "$u_p(x) – u_q(x)$"?

The answer to this question depends upon which of these two distributions we consider to be "the baseline distribution" – the one against which we compare all others. We would want to subtract the "other" distribution from the baseline, because that is what we do with baselines. We always want to see "how far away" other things are from a "baseline thing".

The expressions *a priori* and *a posteriori* directly pertain to the issue of "which distribution is the baseline distribution". In our case, we know which one of these two distributions is the "real one" and which one, ultimately, was the "incorrect one". The "real" one is the *a posteriori* one because that is the one that was arrived at by actual observation. On the other hand, the *a priori* distribution is merely our initial "best guess". Even when we established the *a priori* distribution, we did not expect it to necessarily be the ultimate correct distribution. Thus, for these reasons, our "baseline" distribution is the *a posteriori* on, which we have named "p".

This answer determines our direction of subtraction. We subtract our calculation of uncertainty based upon p from our calculation of uncertainty based on q. In other words, our expression for step 1 is:

$$u_q(x) – u_p(x)$$

Finally, for step 1, we calculate this difference of uncertainty values for all sample points of the sample space S. We are left with the set of these difference values, one for each sample point in S.

Calculating the Mean Differences of Uncertainty

This step reads, according to the above section,

> Calculate the mean of these differences of uncertainty across all of the sample points in the sample space.

In the previous step, we calculated a set of differences of the form $u_q(x) – u_p(x)$, one difference value for each sample point in sample space S.

The question is, "What do we do with all of these differences?"

The answer is, "We calculate their average", or mean.

But this raises another question: “What distribution do we use to calculate this mean?”

To answer this, recall that the calculation of a mean of a set of values involves multiplying each of those values by its probability (and then adding all of those products). The question we just asks means “where do we get those probabilities from?”

Two obvious candidates in our case are the *a priori* distribution “q” and the *a posteriori* distribution “p”. But which one do we use? The answer is “We use distribution p” because it is the “actual”, “real” or “baseline” distribution.

Therefore, we shall define the *relative entropy* of these two distributions p and q by calculating the mean difference of their uncertainties for each sample point in S. Moreover, we use the probabilities from distribution p for the purpose of calculating this mean.

We are now in a position to state the results of step 2, and of the desired definition of relative entropy symbolically.

Definition 1: The Relative Entropy of probability distributions p and q with respect to p:

$$D(p||q) = E_p(u_q(x_i) - u_p(x_i)),\ i \in S$$

This says that the relative entropy of probability distributions p and q, with p being the *a posteriori* distribution, is the expected value, or mean, of the all of the differences $u_q(x) - u_p(x)$ across all sample points x of sample space S.

Notice that we have introduce the new symbol “D(p||q)” to mean “the relative entropy of distributions p and q with respect to p”. The “with respect to p” language means that “p” is the *a posteriori distribution*. This means two things. First, it means that $u_q(x) - u_p(x)$ will calculate the difference between the two uncertainty values. Secondly, it means that distribution p will be used for the purpose of calculating the mean of all of these differences.

An equivalent, but more self-contained, definitions is

Definition 2: The Relative Entropy of probability distributions p and q with respect to p:

$$D(p||q) = \sum_{i \in S} p(x_i)*[\ u_q(x) - u_p(x)\ ]$$

Analysis of Relative Entropy

In our definition of D(p || q), we “started from the inside and worked outwards”. This means that we first worked with each individual sample point of our sample space S.

For each of these sample points we then calculated the uncertainty values for each of out two parameters p and q. Since p and q are both probability distributions, then it makes sense to calculate their uncertainty values using our function u(x) that we have been developing throughout Part I.

Next we calculated the difference of these two uncertainty values for this sample point. And then we proceeded to perform this calculation for all of the sample points of S. Finally, then, we took the average of all of these differences. The result was one number, the relative entropy, which we then applied to the pair of distribution p and q to attribute a value measure for “how different” they are regarding uncertainty.

So, our approach started with each individual sample point, then developed a numerical value for each of them, and then finally averaged all of these values to present a single number for the pair of distributions p and q. This is what we mean by the “inside out”.

It is fair to say that this "averaging" approach could have the following interpretation:

The degree of the uniformity of spread of **the difference uncertainties $u_q(x) - u_p(x)$** of the sample points of S.

This interpretation of *relative entropy* takes its meaning from a similarly-worded interpretation of *entropy* that was discussed above:

Entropy Interpretation #3: The degree of the uniformity of spread of the probabilities across the **uncertainties u(x)** of the sample points of the distribution.

In both cases, we are interpreting the averaging of uncertainty values – one uncertainty value for each sample point of sample space S. The difference is that *relative entropy* is a little bit "fancier". It involves two distribution instead of one; and it involves taking the difference between their uncertainty values for each sample point rather than merely taking one of them

In any event, what these two interpretations have in common is that 1) they each begin by looking at one sample point at a time and then calculating a value for that sample point by using the uncertainty measure u(x). 2) Next, the both repeat this calculation for every sample point of X. A third thing that they both have in common is that they then calculate the mean of these calculations. Finally, the interpretations of both speak of measuring "the spread" of their own notion of uncertainty across the entire sample space.

It would be interesting to contrast this "inside out" approach to measuring the "relative uncertainties" of two probability distributions with a simpler, perhaps more obvious, approach to such a calculation.

An alternative approach could have been to calculate the *entropies* of each distribution separately; and then to simply subtract them. Such a formula would be:

Difference in entropies of p and q = H(q) – H(p).

One might call this the "outside in" approach to calculating the difference of uncertainties of two probability distributions.

It turns out that these two approaches generally produce different answers! Thus, there is a need for D(p || q) to be defined the way we did it above. It should be obvious that D(p || q) is a more detailed answer. That is, there are many different pairs of distributions (p, q) that yield the same H(q) – H(p) value but that also yield different D(p || q) values.

Something to Wonder About

However, there are special cases of p and q wherein D(p || q) = H(q) – H(p). Since the calculation of H(q) – H(p) requires less steps than the calculation of D(p || q), it could be useful to know the conditions under which the two formulas yield the same value. It is left to the reader to wonder what the conditions are for two distribution on the same sample space p and q such that D(p || q) = H(q) – H(p).

Comparing Entropy and Relative Entropy

Recall that one of our definitions of entropy of a probability distribution is "the expected value of the uncertainties u(x) of the sample points of the distribution. This is defined as:

$$H(p) = \sum_{i \in S} p(x_i)*u(x_i)$$

It is interesting that we can obtain our definition of relative entropy by substituting "**$u_q(x) - u_p(x)$**" for "**u(x)**" in the above definition of entropy:

$$D(p||q) = \sum_{i \in S} p(x_i)*[u_q(x_i) - u_p(x_i)]$$

And, we have the following interpretation #2 of D(p||q):

The relative entropy of distributions p and q relative to p is the expected value with respect to p of the differences between the uncertainty of p and uncertainty of q for each sample point of the shared sample space of the two distributions.

The conclusion that we can draw from this observation is that *entropy* and *relative entropy* have a lot in common. Whereas, entropy is the expected value of the uncertainties of all of its sample points, relative entropy is the expected value of an expression that involves the uncertainties of its sample points.

Interpretations of Relative Entropy

IN the section above in which we presented and defined the notion of *entropy*, we also provided three different but related interpretations.

Three interpretations of statistical entropy:

1. A measure of the degree of uncertainty of a probability distribution.
2. A measure of the degree of the uniformity of spread of the probabilities across the sample points of the distribution.
3. A measure of the degree of the uniformity of spread of the uncertainties u(x) across the sample points of the distribution.

In the present section, we shall generalize these interpretations and apply them to the notion of *relative entropy*.

Relative Entropy Interpretation #1:

A measure of the average difference in uncertainty of two different probability distributions of the sample points of a sample space.

Relative Entropy Interpretation #2:

A measure of the uniformity of spread of difference in probability of two different probability distributions of the sample points of a sample space.

Relative Entropy Interpretation #3:

A measure of the uniformity of spread of difference in uncertainty of two different probability distributions of the sample points of a sample space.

From these interpretations we can see that two distinct forces can work together to increase the relative entropy of two distributions larger: 1) the magnitude of the differences of the uncertainties, and 2) the extent to which these differences are uniformly spread across the sample space.

In other words, we can think of relative entropy as “the extent to which large differences in the uncertainties of distributions p and q are widespread.”

A Relative Entropy Example

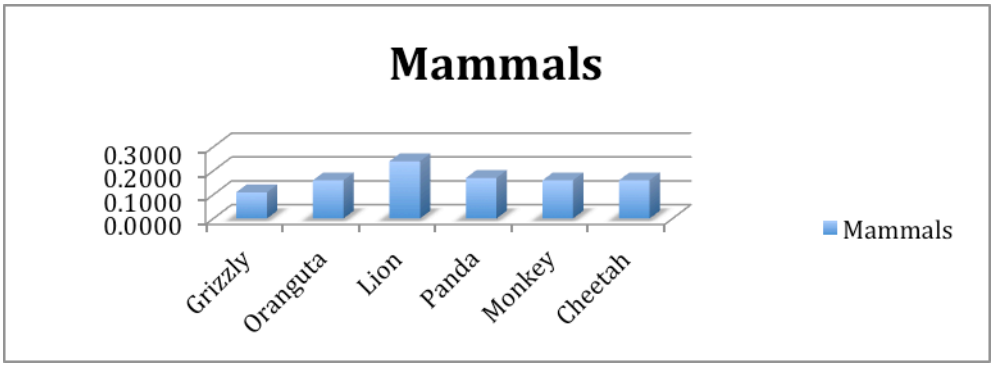
Lets look at the example we mentioned above – the “loaded die”. Someone who wants to cheat at a particular game involving a single six-sided die has decided to secretly substitute a die that has been tampered with in a manner that changes the probabilities of each face. The tampering was accomplished by embedding an off-center hidden weight in the die.

The die used is associated with a particular board game that shows the faces of six particular mammals – rather than distinctive numbers of dots used by standard die. We do this in order to dispel the usual assumption that the outcomes (sample points) of probability distributions must have numerical values. The embedded weight has been placed off-center in a location internal to this die that results in each face having the following probabilities:

A Posteriori Distribution “p”: Loaded “Mammal” Die Face Probabilities

Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total Probabilities
0.1100	0.1600	0.2400	0.1700	0.1600	0.1600	1.0000

Here is a graph of distribution “p”.



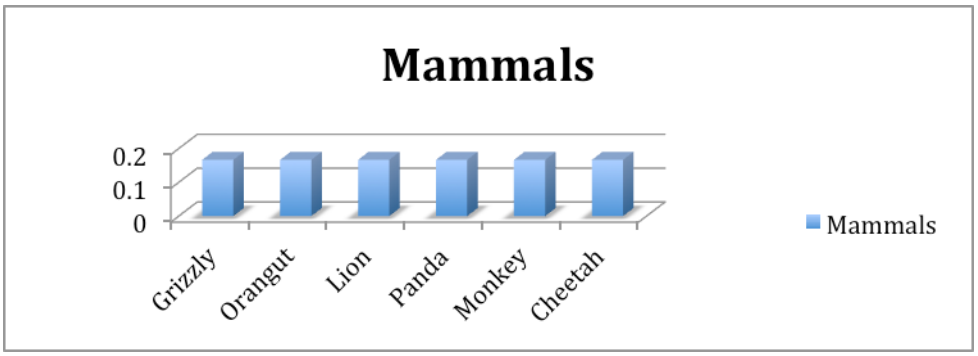
We want to calculate the relative entropy of this distribution of the die faces as compared with the “usual distribution” of die faces, which is the uniform distribution. In other words, for our comparison, the above distribution is the a posteriori distribution, because it is the actual distribution. Since this is the a posteriori distribution in this comparison, we shall name it “p”.

The a priori distribution in this case is the uniform distribution, because what is normally expected of a die is that all if its faces have equally likely probabilities, and is therefore the reasonable “best guess”. Since the uniform distribution in this comparison is the a priori distribution, we shall name it “q”. Its distribution is shown here:

A Priori Uniform Distribution “q”: The Usual “Mammal” Die Face Probabilities

Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total Probabilities
0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	1.0000

Here is a graph of distribution “q”.



Our first step in calculating the relative entropy is - for each face - to compare the uncertainty of that face based on p to the uncertainty of that face based on q. We do this by subtracting $u_p(x)$ from $u_q(x)$. (We use a log base of 2.) We do that here:

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah
$u_q(x_i)$	2.5847	2.5847	2.5847	2.5847	2.5847	2.5847
$u_p(x_i)$	3.1844	2.6439	2.0589	2.5564	2.6439	2.6439
$u_q(x_i)$-$u_p(x_i)$	-0.5998	-0.0592	0.5258	0.0283	-0.0592	-0.0592

The next step in calculating the relative entropy is to calculate the mean of the bottom row of the above table. To do this, we multiply each number in the bottom row by its probability from the a posteriori distribution p – from the first table above. We do this here:

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah
$p(x_i)$	0.1100	0.1600	0.2400	0.1700	0.1600	0.1600
$u_q(x_i)$-$u_p(x_i)$	-0.5998	-0.0592	0.5258	0.0283	-0.0592	-0.0592
$p(x_i)*[u_q(x_i)-u_p(x_i)]$	-0.066	-0.0095	0.1262	0.0048	-0.0095	-0.0095

Recall that our formulation of relative entropy – as we have developed it thus far is:

$\sum_{i \in S} p(x_i)*[u_q(x) - u_p(x)]$

Thus, we must add the entries of the bottom row of the table above to achieve the relative entropy of our loaded die with respect to its above probability distribution p and as compared to the a priori uniform distribution q:

$\sum_{i \in S} p(x_i)*[u_q(x) - u_p(x)] =$ **0.0366**

Relative Entropy Practical Definition

The formulation of relative entropy that we have used so far has a couple of advantages going for it. For one, it expresses the meaning of relative entropy very directly. This formulation makes it apparent that relative entropy is the mean difference between an a posteriori distribution and an a priori distribution, using the probabilities of the a posteriori distribution.

A second advantage is that this formulation suggests how relative entropy is a variation on entropy wherein the "$u_p(x)$" is replaced by "$[u_q(x) - u_p(x)]$"[14].

However, for calculation purposes, this formulation can be improved upon.

In the present section, we shall provide more practical definition than the one given above by starting with this formulation and deriving a more efficient one, which we shall then use in what follows. As one might expect, this more efficient definition is the one usually encountered in mathematical texts on information theory.

Lets first introduce the standard symbol for relative entropy in information theory:

$D(p||q)$

[14] Notice that "$[u_q(x) - u_p(x)]$" is a function of $u_p(x)$.

Which means the relative entropy of a posteriori distribution p as compared with a priori distribution q with respect to distribution p on sample space S.

Lets begin with a restatement of our current formulation, and then derive from it a more efficient formulation:

$D(p||q) =$

$\sum_{i\in S} p(x_i)*[u_q(x) - u_p(x)]$

$= \sum_{i\in S} p(x_i)*[\log(1/q(x)) - \log(1/p(x))]$

$= \sum_{i\in S} p(x_i)*[(\log(1)-\log(q(x)) - (\log(1)-\log(p(x))]$

$= \sum_{i\in S} p(x_i)*[(0-\log(q(x)) - (0-\log(p(x))]$

$= \sum_{i\in S} p(x_i)*[(-\log(q(x)) - (-\log(p(x))]$

$= \sum_{i\in S} p(x_i)*[-\log(q(x) + \log(p(x))]$

$= \sum_{i\in S} p(x_i)*[\log(p(x) - \log(q(x))]$

$= \sum_{i\in S} p(x_i)*[\log(p(x)/q(x))]$

Clearly, the last of these expressions requires the minimal operations. Thus, the practical definition of the relative entropy of distributions p and q with respect to p is:

$D(p||q) = \sum_{i\in S} p(x_i)*[\log(p(x)/q(x))]$

As mentioned above, this more efficient definition is the one usually encountered in mathematical texts on information theory.

Corroboration of the Practical Definition of Relative Entropy

If we were to subject the “mammal die” example above to this formal definition of relative entropy, then we should obtain the same value as before, which was 0.0366.

Let go through the calculation with the new formulation and see if we obtain the same answer:

$p(x_i)/q(x_i)$	0.6599	0.9598	1.4397	1.0198	0.9598	0.9598
$\log_2(p(x_i)/q(x_i))$	-0.5998	-0.0592	0.5258	0.0283	-0.0592	-0.0592
$p(xi)*\log_2(p(xi)/q(xi))$	-0.066	-0.0095	0.1262	0.0048	-0.0095	-0.0095
$\sum_{i\in S} p(x_i)*[\log(p(x)/q(x))]$	**0.0366**					

Yes, we arrive at the same value for relative entropy (**0.0366**) of the two distributions p and q with respect to p regardless of which of the two formulations that we have presented.

Things to Wonder About

Is D(p||q) = D(q||p) for all distributions p and q on the same sample space?

If not, then are there particular distributions p or q for which D(p||q) = D(q||p)?

If so, then what is (are) it (they)?

General Entropic Measures

After we introduced the concept of *relative entropy* above, we compared its definition to that of *entropy*. Here is the side-by-side comparison. We have underlined the parts of the two definitions that are the same.

Entropy:

$$H(p) = \underline{\sum_{i \in S} p(x_i)} * u_p(x_i)$$

Relative Entropy:

$$D(\ p||q\ ) = \underline{\sum_{i \in S} p(x_i)} * [\ u_q(x_i) - u_p(x_i)\ ]$$

We want to point out again that these two definitions are the same except for the expression following the "*" operator.

Also, if we look closely at both definitions, we can see that they have a couple of things in common. The first is that they are both taking the average of the expression following the "*" operator. We know this because the symbols prior to that, "$\sum_{i \in S} p(x_i)$" indicate just that.

The second thing both definitions have in common is that the expression that each is taking the average (mean) of contains "$u_p(x_i)$". In other words, the expression that is being averaged is "a function of $u_p(x_i)$".

We can summarize all of this by saying that both the *entropy* and the *relative entropy* involve taking the average values of some function of u(x) across an entire sample space – using a specified probability distribution p over the sample space.

For *entropy*, the "function of u(x)" in question is, in fact, "u(x)". For *relative entropy*, the "function of u(x)" in question is "[$u_q(x_i) - u_p(x_i)$]".

All of this opens up the possibility that we could formulate other functions of u(x) and take their average as well. And this is exactly what information theory does. It extends this practice inventing a variety of new ways of expressing various "functions of u(x)" and then taking their average. Of course, each time this is done, the resulting definition is given a new name.

This practice produces a kind of "generalization" if the concept of *entropy* – which we shall refer to as *entropic measures*. So *entropy* and *relative entropy* are both examples of *entropic measures*.

So we are introducing here a class of measures called *entropic measures*, all of which have the form:

$$M(p) = \sum_{i \in S} p(x_i) * F(u(x_i)))$$

Where F is some function of u(x).

What these measures have in common is that each of them is the expected value of some expression involving "u(x)". More formally, we say that all of these measures are the expected value of some function of u(x) – or E(F(u(x)), where F(u(x)) is some function of u(x).

In other words, this primer defines an *entropic measure* as the mean value, across a sample space of some function of u(x), relative to a specific probability distribution on that sample space.

For example, consider the definition of entropy:

$$H(p) = \sum_{i \in S} p(x_i)*u_p(x_i).$$

We ask the question: "How does this definition compare with our general pattern of an *entropic measure*:

$$M(p) = \sum_{i \in S} p(x_i)*F(u(x_i)))$$

The answer is that, for *entropy*, "M(p)" is "H(p)", and "$F(u(x_i)))$" is simply "$u_p(x_i)$". In other words, our definition of entropy fits the *entropic functions* pattern. In this case, the "$F(u(x_i)))$" is simply "$u_p(x_i)$".

Lets look at a second example and see how its definition fits the entropic functions pattern. This time we shall look at *relative* entropy, whose symbol is D(p || q). Recall that the formula for D(p || q) is:

$$D(p||q) = \sum_{i \in S} p(x_i)*[u_q(x_i) - u_p(x_i)]$$

But this also fits the *entropic functions* pattern. This time, the "$F(u(x_i)))$" is

$$[u_q(x_i) - u_p(x_i)]$$

which is certainly a function of "$u_p(x_i)$", because "$u_p(x_i)$" appears within the expression "$[u_q(x_i) - u_p(x_i)]$".

Of course, D(p||q) is a little more interesting example of an *entropic measure*, because it "$F(u(x_i)))$" is a little more complicated than the "$F(u(x_i)))$" for entropy.

In the remainder of this primer, we shall see a number of other entropic measures. Two examples are *mutual information*, which we shall see in Part II, and *entropy rate*, which we shall see in Part III. Another example is *conditional entropy*, which we shall also see in Part II.

For *mutual information*, as we shall see in Part II, "F(u(x))" is different from the above expressions, but is an expression involving u(x) nevertheless. The same shall be true for the measure named *entropy rate* that we shall introduce in Part III.

We have said that information theory can be understood as *probability theory* plus *entropy*. But now we can extend that statement further and proclaim that *information theory is probability theory plus entropic measures*.

Another name for *entropic measure* is *entropic functional*. This term is used by [Kleeman 2012, Lecture 2]. The concept that the essential constructs of information theory are functionals is also pointed out by [Cover and Thomas 2006, p.4]. The use of the word *functional* is interesting because it shows what kind of mathematic "creature" these measures are. A *functional* is a somewhat "fancy" function. We typically think of

a function as mapping “some numbers to some other numbers”. For example, the function $f(x) = x^2$ maps 2 to 4, 3 to 9, 5 to 25 and so on.

But a *functional* has bigger ambitions. It maps “complicated things” to numbers. For example, it could map an entire probability distribution to a single number. Certainly this is the case for *entropy*. Entropy maps an entire probability distribution to a single number. And, it does that for *any* discrete probability distribution. All of these other *entropic measures* that we mentioned above – entropy, conditional entropy, relative entropy and mutual information – map complicated expressions (that relate to probabilities and uncertainties) to numbers. This is why Kleeman uses the word *functional*.[15]

Thus, in this primer, we shall use the phrases *entropic measure* and *entropic functional* interchangeably.

In short, we can conclude this section on *entropic measures* by asserting that all of the principle measures of information theory are the mean value of some function of the uncertainty u(x) of a probability distribution.

[15] As a matter of fact, all *statistics*, such as mean, median, mode, variance, standard deviation the *moments* are also functionals on probability distributions. And they are also measures. In fact, measure is a functional on the things that it measures.

Part II: The Portent of Uncertainty

In Part II, we are going to look at how uncertainty, and thus information, can have *meaning* in information theory, and how that kind of meaning can be measured.

We shall see that information theory does not imbue absolute meanings to specific terms, as does lexicography, nor the discipline of semantics. We shall leave those treatments to the philosophers.

Rather, information theory is concerned with a relative form of *meaningfulness*: the portent of one chance variable to another. Information theory does not try to attribute meaning to any chance variable. However, it is profoundly interested in whether, and the extent to which, the outcome of one chance variable portends the outcome of another.

The Meaning of Information

Shannon insisted that the notion of meaning is "irrelevant to the engineering problem" in communications theory [Shannon 1948, p. 1]. And in one way he was correct, not only as relates to communications theory, but also to information theory.

It is true that information theory is not concerned with the assignment of "meaning" to specific pieces of "information". That is the subject of linguistics and its sub-discipline semantics.

But, in another way, the concept of meaningfulness is alive and well in information theory. This is the case when the "happening" of one chance variable portends the "happening" of another chance variable. In probability and information theories, though, we try to be a little more formal than using words like "happening". So we say it this way: The concept of meaningfulness in information theory applies when the outcome of one chance variable portends the outcome of a second chance variable.

So meaningfulness is at work in information theory when there are two chance variables where the outcome of the second one depends upon the outcome of the first one.

This kind of meaningfulness is relative. It describes a dependency relationship between two chance variables. It tells you what one chance variable portends about another – even though it does not tell you what either chance variable means on its own. In this way, it is unlike the concept of meaning in semantics, which is absolute, and defines the meaning of a word or phrase "on its own".

In addition, the meaningfulness we are discussing in information theory pertains to the reduction of uncertainty! Specifically, it pertains to the reduction of uncertainty of one chance variable via the knowledge of the outcome of another chance variable.

In other words, suppose we have a situation between two chance variables in which knowing the outcome of the first happens to reduce our uncertainty about the upcoming outcome of another chance variable. In such a case, we would say that the outcome of the first "means something to us". It is meaningful to us because it "helps us to predict" the outcome of the second.

When this relationship exists between two chance variables, we often say, "The two variables are correlated." When it does not happen between two chance variables, we say, "The two variables are uncorrelated." In information theory we use two other terms more often. If the chance variables are meaningful to each other in this way, we say that they are *stochastically dependent*. If they are not meaningful in this way, we say that they are *stochastically independent*.

Let's take an example from the stock market. It would be useful to us if we could find two stocks whose price fluctuations are stochastically dependent; because then changes in one would be meaningful to changes in the other. In such a case, changes in one could be a predictor of changes in the other.

We may not know anything about either stock in isolation of the other. (That would be semantic knowledge.) But if we know that the two stock price changes are stochastically dependent, then we could have some level of confidence that changes in the price of one can predict changes in price of the other. And this is true even though we may know nothing about either stock – as long as we know something about their dependency relationship!

Of course, there is also the matter of degree – of just *how dependent* the two stock price changes are on each other. The stronger the dependency relationship is, the stronger will be our degree of certainty of using one stock price change as a predictor of the changes in the other stock's price.

The point here is that the concept of degree of dependency and degree of uncertainty come into play. In fact, we are obviously going to want to have a measure this degree of dependency. Moreover, such a measure will need to have something to do with some degree of uncertainty. Of course, we spent the entirety of Part I of this primer building up to a measure of the degree of uncertainty – entropy. Consequently, you might expect that any measure of the degree of dependency between two chance variables that we develop here in Part II will probably have something do to with – or be defined in terms of – our notion of entropy that we developed Part I.

(Of course, for the time being we are ignoring the role that time plays in this stock market example. That is, if one of the stocks is to be used as a predictor of the other, then its correlated change in price must temporally precede that of the other. If not, then the first stock cannot be practically used as a predictor of the second. We shall take up this treatment of time in conjunction with stochastic dependency in Part III of this primer, whose subject is prediction. In the present Part II, we shall concentrate on stochastic dependency, its detection and its measurement.)

In any event, this subject of stochastic dependence obviously arises whenever we have two chance variables. In such a case, it then becomes possible to wonder whether the outcomes of one of them portend the outcomes of the other – or not.

But if this portent between the two chance variables exists, then this means that certain outcomes of one occur whenever certain outcomes of the other occur. In other words, if this portent exists, then we can talk about the occurrence of pairs of outcomes, where the first outcome in each pair is from one of the chance variables, and the second outcome of that pair is from the second chance variable.

Moreover, if this portent exists, then these pairs of the outcomes that portend each other will occur with a higher incidence – higher probability – than pairs that do not portend each other. In fact, this is the essence of portent: that certain values of the two chance variables tend to "pair up" and happen together more often than would be "expected by chance alone".

Of course, we are now talking about what happens when an outcome of one chance variable pairs up with an outcome of a second chance variable. So, we are now talking about *pairs* of outcomes, one from one chance variable and one from another. But this is a new sample space – one whose sample points are pairs of outcomes from the two initial sample spaces, or chance variables. Moreover, these each of these new sample points – the pairs – also has its own probabilities. In other words, we are now dealing with a new probability space with its own probability distribution.

We shall call this new probability distribution – whose sample points are these pairs – the *joint probability distribution* (*joint distribution* for short). The new probability space will be called the *joint probability space* (*joint space*, for short).

And note that the joint space has a lot more sample points (being all possible pairs of the other two sample spaces) than either of the initial sample spaces. Also note that the probabilities of the joint space are *not* mathematically determined by the probabilities of the two initial (or component) spaces. Rather, one has to actually observe the joint space in action to see what its actual joint probabilities are. And – here is the subject of the entirety of Part II – the probabilities of the joint distribution determine the degree of stochastic dependence (or portent) of the two component chance variables.

That is, if the joint space has one probability distribution, then the two component chance variables could be highly dependent. If it has another distribution, then the two component chance variables could be highly independent. If it has yet a third distribution, the two chance variables have a degree of dependency could lay somewhere between the maximum and the minimum. In other words, there is a spectrum of possible degrees of stochastic dependency between two chance variables – depending on what the joint distribution is.

So, the degree of dependency between two chance variables is completely determined by their joint probability distribution. This fact is the subject of Part II. And our task in Part II will be to develop a measure for this degree of dependency between two chance variables. We shall use this as a measure of the meaningfulness or portent of two chance variables with respect to each other. As we shall see, this measure will be named the *mutual information* between the two chance variables.

Even if there is no dependence (or portent, or meaningfulness) between two chance variables (i.e. they are stochastically independent), then we can say that the degree of dependence between them is zero – and therefore, we can still measure its degree of dependence!

In Part I, we talked about the value of uncertainty/information in terms of the degree of uncertainty of outcomes of a single chance variable. We measured this value with the entropy of the chance variable. Now, in Part II, we are talking about the value of uncertainty/information in terms of how meaningful one chance variable can be to another – within a situation where both chance variables occur together in the same joint event.

So, it is reasonable to expect that our measure of meaningfulness between two chance variables – mutual information – will somehow be defined using the idea of entropy. In Part II, we shall do this.

The Economics of Information: Value Based on Supply and Demand

In Part I, we developed a measure of the uncertainty of a single event or sample point of a sample space. This measure of uncertainty assigned a numerical value to an event that attributed to the event worthiness, or value, in information theory.

We named this measure “the uncertainty of a sample point or an event”, and symbolized it with the letter “u”. Specifically, we defined it as a function of sample point x. Thus, “u(x)” is the symbol we use for the measure u of sample point x. We define u(x) as:

$$u(x) = \log(1/p(x)).$$

This measure emerged early in the history of information theory. It turns out to give the properties to this measure of information that one would desire. We shall review some of these now.

This function u(x) gives a higher measure to sample point x if x is more uncertain (has a lower probability) and gives x a lower measure to a sample point if it is more certain (has a higher probability).

Essentially, the measure u(x) gives higher values to rare events and lower values to common events that happen more often. So then it is reasonable to say that u(x) is a measure of the rarity of a single event or sample point.

Economic theory says that the value of a thing is determined by the interplay of two aspects of that thing: 1) its supply, and 2) the demand for it by a population of consumers (observers in statistical terminology).

The approach we have taken to valuing an event so far (using the uncertainty of the event) has been to attribute value to the event based on its rarity, which is a supply aspect of value. The more rare is a sample point, the higher is its uncertainty value. Moreover, our calculation of entropy in Part I is essentially the "average rarity" of the sample points of a probability distribution. Thus, entropy also is a supply-oriented measure of value.

But in order to have a complete theory of the valuation of a sample point or to an entire probability space, we need to include a demand aspect to our consideration of valuation.

In economic theory, demand for a particular outcome pertains to some intrinsic value of that outcome to a consumer (observer). Our approach to the demand aspect of an event will be to consider some aspect of the event that can give it some informational intrinsic value to an observer. (We shall discuss shortly exactly what this intrinsic value of information is.)

Therefore, to calculate an overall value of a sample point, we must develop some way of combining rarity and intrinsic value (supply and demand) to produce an overall value of the information inherent in that sample point. The way that we shall make this combination is to calculate the rarity of the intrinsic value of each sample point of the distribution.

We shall explain in greater detail as we develop this measure exactly how we shall mathematically develop this rarity of intrinsic value of the relationship of two chance variables in order to produce a measure of their meaningfulness to each other – the measure that we shall name mutual information.

Portent Is Intrinsic Informational Value

It is reasonable to assert that the meaningfulness of portent is an intrinsic value of information.

The portent of one chance variable for another, as we have been discussing, has many informational applications, including prediction, organization, stability and others. Of course, portent is a type of meaning that attributes value to relationships between two chance variables, as we have been so far discussing in Part II.

Thus meaningfulness, dependence or portent is a relationship between two chance variables that potentially exhibits informational intrinsic value. In fact, information theory has selected portent of one chance variable for another as its intrinsic-value-of-interest.

As mentioned, portent often goes by other monikers in information theory - especially “stochastic dependence”, or “statistical dependence”. We shall use both phrases interchangeably in this primer.

Therefore, stochastic dependence can, and will, be used as the demand aspect of the value of information.

So, in Part II, we shall be developing a way of valuing information (named *mutual information*) that takes into account both its meaningfulness to other information as well as the rarity of that meaningfulness. And, we shall be gradually developing the mathematical equipment that information theory offers in order measure that value.

Basic Example Joint Experiments

Before we proceed to develop the mathematical constructs put forth by information theory to measure the portent of one chance variable to another, however, it will useful to visit some basic examples of joint chance variables so that we can get a feeling for the issues involved, the patterns that arise and some things to think about.

Once we have done that, then we shall then be in a position to better understand the approach and the mathematics offered by information theory to address those issues, and to develop the mathematical constructs of information theory.

So, this section will present five basic experiments, all dealing with the same two sample spaces - which involve two dice – the kind used in games of chance. The six faces of one of the dice will be one of these two sample spaces, while the six faces of the other die will be the other sample space. Therefore, each experiment can be viewed as a joint probability space. As such, each experiment has one joint distribution and two component distributions.

However, we shall make five different variations on this pair of dice. Each variation will be one of our experiments. All of the variations have to do with changes to the probability distributions involved. The point of these various experiments is to see how changing the joint distribution of a pair of chance variables affects their degree of stochastic dependence. We shall also make some changes to the component distributions to see how that affects the degree of stochastic dependence.

For two of these experiments (Experiments 1 and 3), both component distributions are the uniform distribution. This means that for these two experiments, the faces of the first die are equally likely; and so are the faces of the second die.

But for the other three experiments (Experiments 2, 4 and 5), both component distributions of the two dice are not equally likely. In fact these two distributions are different from each other.

Having discussed the two component distributions for each of the five experiments, lets turn our attention to the joint distributions. We have discussed that it is the joint distribution that contributes mostly to whether a probability space is stochastically independent or stochastically dependent.

Moreover, as we shall see, a certain relationship between the joint distribution and its component distributions determines the degree of stochastic dependency between the two chance variables associated with the component distributions. Therefore, it is a focus on the joint distribution that will reveal a way to measure the degree of stochastic dependence between two chance variables.

Once we have introduced these five examples in this section, and the reader has become familiar with some of the issues involved, we shall then proceed in the next major section to present the mathematical constructs of information theory. We shall, at the same time, utilize the five examples to compare and contrast the values calculated for each of these constructs for each of these examples.

It is hoped that this approach of appealing deeply to intuition first and presenting the formal treatment second will result in a better feeling and understanding for information theory on the part of the reader.

Five Simple Representative Experiments

Lets now set up and consider our five example experiments, all involving joint events. These five examples cover a spectrum of all of the main cases of stochastic

dependence that we wish to discuss in this primer. So we shall be referring to each of these throughout the remainder of Part II.

In fact, as a set of experiments, they will create a uniform context to discuss all of the topics that make up Part II. So for convenience and efficiency, we shall describe each of these five experiments in this section, and then introduce the formal ideas of Part Ii in the following sections. We shall then use these five examples to illustrate these formal ideas in these following sections.

We expect to introduce most of the ideas in information theory in an intuitive manner within our discussion of these five experiments. Then hopefully the reader will be able to take this intuition to our formal treatment of the mathematics of information theory in the next major section of Part II – and will able to comprehend those formalities and constructs with a better understanding.

Consequently, we will delve deeply into each of these examples, since hopefully most of the learning on the part of the reader will occur intuitively when these examples are initially presented. The formal mathematics that appears later will, then, simply be a formal articulation of what the reader has already discovered during the consideration of the examples. Hopefully, then, the formal constructs will “go down with relative ease” as compared with being “thrown at” the reader “without preparation”, as is often done in many mathematical textbooks.

All five examples involve two dice of the kind used in games of chance. And in all five experiments, a trial of the joint experiment will consist of rolling the two dice together. That is, a trial of one of our example experiments will be a joint event consisting of a pair of outcomes of two other component chance variables. One of the component chance variables is the rolling of one die, and the other component chance variable is the rolling of the other die. In our joint event, these two component events will occur together.

This means that each of the component chance variables has its own component probability distribution that describes how the component chance variable behaves in isolation – when a single dice is rolled. Each of those component chance variables has six possible outcomes (sample points) – one for each face of its die. Thus each component chance variable has six outcomes, or sample points – one for each face of its die.

In addition, we have a third chance variable – the joint chance variable. Each outcome of this joint chance variable is a pair whose first entry is a face from the first die and whose second entry is a face from the second die. Therefore, the sample space of the joint chance variable has 6x6 = 36 outcomes (sample points).

It is this joint chance variable that we are focusing on in Part II. Specifically, we shall evaluate at the outcomes of this joint chance variable in a manner that will determine 1) whether one of its component chance variables depends upon (portends something about) the other, and 2) if so, then the degree to which one of the component chance variables depends upon (portends something about) the other.

Primarily, we shall be interested the fact that – for two specific component chance variables - there are many possible ways those joint probability distributions can be defined. In other words, given two component probability distributions, there are many possible joint distributions – each of which depicts its own degree of dependency between the two component distributions. Some of those possible joint distributions indicate a higher level of dependence between the two chance variables; while others indicate a lower level of dependence.

Moreover, no matter what the two individual component probability distributions are, there are still many possible joint probability distributions that can be defined for their joint space. And, it is which of these joint probability distributions that actually applies to the joint space that determines the degree of stochastic dependency of these two chance variables.

The essential implications of all this are that 1) it is possible to have different sets of joint probabilities for the same two component distributions, and 2) the joint probabilities determine the degree of dependence between the two chance variables.

In order to dispel the misconception that sample points must always be numbers[16], we are not going to use ordinary dice. Also, in order to be able to refer to each of the two dice separately, we are going to make them different from each other. Therefore, we are going to use two custom dice. It is easy to purchase custom dice. Simply search for "custom dice" on the Internet, and you will find vendors who will take your order and deliver to you the custom dice – for a price. (These products are often used for promotional and advertising giveaways.)

Each of our two custom dice will be cubical in shape. (You can order shapes other than cubes from the custom dice vendors.) But neither of them will have "numbers of dots" on their faces, as do typical gaming dice. Rather, in our examples, one of our custom dice will have pictures of mammals on its six faces and the other will have pictures of trees.

The mammal die will have two bears, two cats and two apes. Animals of each of these three types will be place on opposite faces of the die. A Grizzly, an Orangutan and a Lion will occupy three faces that share a common vertex. A Panda, a Monkey and a Cheetah will occupy the three faces that share the opposite vertex.

The tree die will have two pines, two oaks and two cedars. Trees of each of these three types will be placed on opposite faces of the die. A Ponderosa, a Live Oak and a Lebanon Cedar will occupy three faces that share a common vertex. A Pinyon, a Pin Oak and a Juniper will occupy the three faces that share the opposite vertex.

This completes our description of the two individual dice – which define our two component chance variables – used for all five of our experiments. Lets now describe the joint chance variable that is developed between these two component chance variables.

The joint sample space for all five experiments will be the set of all possible pairs of die faces where the first face in each pair is from the "mammal die" and the second face of the pair is from the "tree die". Therefore, there are 6x6 = 36 sample points in this joint sample space. An example of one of these sample points is the pair (Orangutan, Ponderosa).

In other words, we have established that the "probability spaces" of all five of our dice experiments have the same sample spaces – component and joint.

---

[16] Another example of probability spaces whose sample points are not numbers is *coin tossing*, where the sample points are "heads" and "tails". Of course, if the sample space does not consist of numbers, then arithmetic cannot be performed on the sample points, and most statistics (e.g. mean, median, variance, standard deviation, etc.) cannot be calculated for the distribution. Most interesting for this article, though, is that *entropy* will still exist for a distribution even if its sample points are not numbers! The reason for this is that the calculation of entropy only uses the probabilities. It does not use the values of the sample points. In fact, entropy exists for any finite probability distribution.

So, if all five experiments have the same component and joint sample spaces, then what makes their probability spaces different from each other? The answer is this: After we have initially introduced all five of these experiments, we are going to change the probability distribution of the joint sample space a number of times – while keeping the individual component distributions the same. (The component sample spaces will also be changed to some extent, as we shall describe below.)

Each time we change the joint probability distribution; we shall also see that the degree of dependency between the two component chance variables also changes.

However, in this section we do not yet have a measuring function that is able to tell us the amount of dependency involved in any one of these joint probability distributions. We shall have to wait until the formal section below before we develop such a measuring function.

For the time being, though, we have a way to use our intuition to make this assessment. This way uses the graph of the joint distribution. In this section, we shall demonstrate how to inspect this (and a related) graph to get an informal assessment of the degree of stochastic dependency between the two chance variables (the two dice) associated with each of the joint distributions.

Later, when we get to the formal section and present the measuring function that information theory uses for this purpose (named *mutual information*), we can then check it against our intuition and ascertain if the measure meets with our expectations.

We are going to present these five experiments in a particular order. In the first two experiments, 1 and 2, we are going to make sure that the two chance variables are stochastically independent. This means that their “amount of dependence” should be “none at all”. (We would hope that their *mutual information* is measured as 0 (zero).)

However, we shall give Experiments 1 and 2 different component distributions. Both component distributions in Experiment 1 will be the uniform distribution. However, we shall give Experiment 2 distributions that are non-uniform. And yet, its joint distribution will also produce stochastic independence. We do this to show that stochastic independence does not require that both component distributions be uniform.

In our next three experiments, 3, 4 and 5, we shall look at 3 differing degrees of stochastic dependence – each experiment being a little more stochastically dependent than the previous. This increased degree of dependency is, of course, the result of changing their joint distribution probabilities in a way that is increasingly “more drastic” – in some manner.

Of course, in order to make all of these probability distributions realistic, we have to physically “tamper with” these two dice in a manner that 1) changes their component probability distributions, and 2) changes their joint probability distribution. Actually, we are going to request of the custom dice vendor that it makes these dice to our specification so that they result in these probability distribution changes. We shall next explain what these physical dice specifications are.

In the two experiments that we want to be stochastically independent, Experiment 1 and Experiment 2, we are going to make sure that the outcome of the mammal die has no effect on the outcome of the tree die. This is the normal way that dice work. Normal dice are statistically independent. We are going to call these two the independent experiments. In these independent experiments, the outcomes of the second die are not influenced by and do not depend upon the outcomes of the first die. There is no meaningfulness or portent or stochastic dependence going on between the two die in Experiment 1 and Experiment 2.

In the other three experiments, we are going to tamper with the two dice in such a way that the outcome of the mammal die is going to influence the probabilities of the tree die. In other words, the way that we are going to tamper with the two dice alters the probabilities of each of the possible pairs of dice when they are thrown together. For example, the probability of, say, the pair (Lion, Pinyon) will be different for the custom dice (after the tampering) than it would be for that same pair (Lion, Pinyon) with ordinary dice that had not be so tampered with.

In fact, this tampering will have the following effect when both dice are tossed together in Experiments 3, 4 and 5:

> The probabilities of the faces of the tree die will be different based on which face of the mammal die lands up. In other words, which face of the tree die lands up is dependent on which face of the mammal die lands up.

In other words, since the two die will be stochastically dependent in these three experiments, then certain mammal die faces and certain tree die faces will show an *affinity* for each other.

More specifically, knowing which face lands up on the mammal die changes the probability distribution for the faces of the tree die. For this reason, we are going to call these three the dependent experiments.

So, in these three dependent experiments, there is some meaningfulness between the two die. If we know which face lands up on the mammal die, it will change our estimate of which face has landed up on the tree die. Therefore, knowing which mammal die face landed up is meaningful to trying to guess which tree die face also landed up.

Lets now look more closely at these two categories of experiment. We shall further break each category down into two other specific experiments and explain how each of them is unique.

Two Example Independent Joint Experiments

In these two experiments, we are going to make sure that the outcome of the tree die has no dependencies over the outcome of the mammal die.

But what do we mean by that? We will answer that question more thoroughly later in this primer. But for the time being, lets loosely say that whenever both dice are tossed together, knowing the outcome of the mammal die does not help one guess the likelihood of any possible outcome of the tree die.

The reason that knowing which mammal shows up in a particular toss is of no help guessing which tree die is has also landed up in that same toss is because no matter which of the mammals occurs, the probability of each tree occurring remains the same.

We shall look at what this means in terms of probabilities later in this primer. But, for the time being, lets just say “being independent chance variables simply means that knowing the outcome of one component event in a joint event does not assist in guessing the outcome of the other component event in the same joint event”.

We are going to look at two different experiments in this category – both of which are independent experiments. The difference between them lies in the individual probabilities of each die taken separately. However, it will be the joint probability distribution that determines that the two chance variables are stochastically independent.

We shall show how we can look at the graph of a joint distribution and discern that its two chance variables are stochastically independent.

Lets now look at each of these two different independent experiments.

Experiment 1: An Independent Joint Experiment with Equal Probabilities

In this experiment, we are going to make sure that the two die are stochastically independent. We do this by giving the joint distribution between the two dice a particular set of probabilities.

We are also going to make sure that both dice are "fair dice". This means that we are going to make sure that each face on each die has the same probability of landing "up" as all of the other faces on the same die.

The Probabilities for each Die Individually

To say that the mammal die is "fair" means that the probabilities of each of the six mammals faces landing up is 1/6, or approximately 0.1667 rounded to 4 decimal places; and the probability of each of the trees landing up is 1/6, or approximately 0.1667.

This is the case for any two normal dice that we would purchase – or approximately the case. Therefore, we do not have to tamper with either of the dice to ensure equiprobability. A finite probability distribution in which all sample points are equally likely is called the *uniform distribution*. All of the faces on a fair die have probabilities of 1/6. Therefore, a fair die exhibits the uniform distribution.

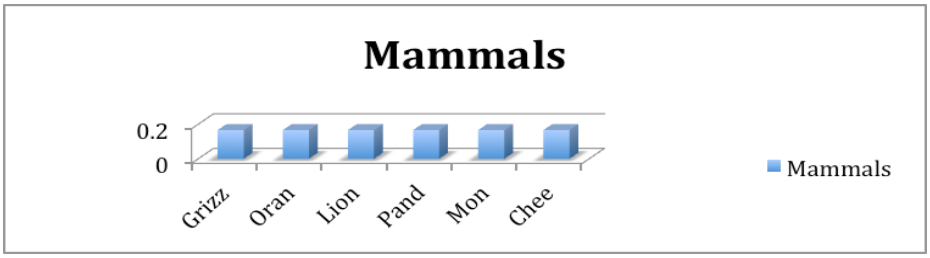
Lets look at the probabilities of the faces of each of these two dice individually – as though we were tossing one die in isolation, rather then tossing them both as a pair. These are called the *component distributions* of the joint distribution. Below are those component distributions in table format – as we shall also show with the other four experiments.

For the mammal die the probabilities of the six faces, since they are equally likely for this experiment, must be as follows (accurate to 4 decimal points):

Mammal Die Component Probability Distribution

Grizzly	Orangutan	Lion	Brown Bear	Monkey	Cheetah	Total
0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	1.0000

And the graph of this uniform (all probabilities are equally likely) distribution is:



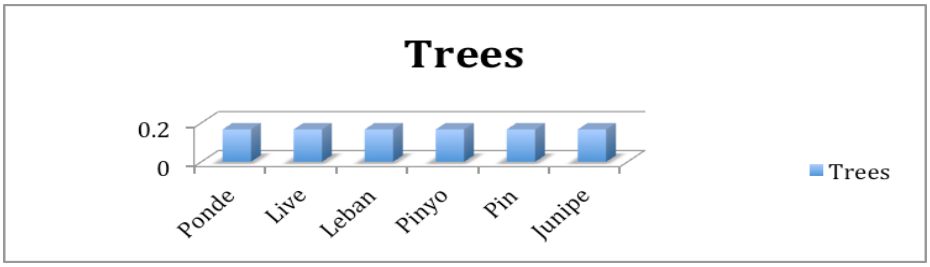
For the tree die the probabilities of the six faces, since they are equally likely for this experiment, must be as follows:

Tree Die Component Probability Distribution

Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total

0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	1.0000

And the graph of this uniform (all probabilities are equally likely) distribution is:



The Probabilities of the Joint Events

All five of the example experiments involve tossing both dice at the same time. These are joint experiments whose sample points are joint events. Those sample points are pairs such as (Orangutan, Pinyon), where the first member is a face from the mammal die and the second member is a face from the tree die.

A good way to make a table for the joint distribution is to lay it out in two dimensions, where the six faces of the mammal die are listed down the left side and the six faces of the tree die are listed across the top.

Then, the probability of a joint sample point, say (Lion, Pinyon), would be found in the cell at the intersection of Lion and Pinyon. We depict such a table here without the probabilities, which we shall fill in next.

So, such a table shows all three distributions that are involved in specifying a joint distribution: the two *component distributions* as well as the *joint distribution*. In the template table below, the probabilities of the *mammal component distribution* are listed in the “Total Mammal” column; the probabilities of the *tree component distribution* are listed in the “Total Tree” row; and the probabilities of the *joint distribution* are listed in the cells of the body of the two dimensional table.

Mammal and Tree dice joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly							
Orangutan							
Lion				<Probability of (Lion,Pinyon) goes here>			
Panda							
Monkey							
Cheetah							
Total Tree							

So, our task is to fill out the empty cells of this table – in such a way that it accords with the description of our first experiment. This description is that the joint events must have probabilities that result in their two component chance variables (the mammal die and the tree die) be "independent of each other".

We'll get into what "independence" requires pretty soon. But for the time being lets just say that that "being independent of each other" is going to end up placing some constraints on what the probabilities of the joint events can be. In other words, if the joint probabilities are defined in a certain way, then the Mammal and the Tree dice will be stochastically independent. But if they are defined in other ways, then the Mammal and the Tree dice will be stochastically dependent.

For experiments 1 and 2, we must define the joint probabilities in such a way that the two dice are stochastically independent. So, our task is to determine what the joint probabilities must be in order for the Mammal and the Tree dice to be stochastically independent.

Now, it is going to turn out that the probabilities of the two component distributions governs what the joint probabilities must be in order for the two dice to be stochastically independent. Therefore, in our attempts presently to figure out what the joint probabilities must be to achieve stochastic independence, we will have to take into account what the two component distributions are.

Therefore, it will be convenient to have the two component probability tables of the mammal and the tree dice, from above, readily available. In that case, let us simply copy the component probability distribution from the above two tables into this table. Then we can attempt to fill out the remainder of this table with joint probabilities that will make the two dice stochastically independent.

We'll put the component probability distribution of the mammal die in the "Total Mammals" column. And we'll put the component probabilities of the tree die in "Total Tree" row. This is illustrated in the table below.

At this point, in preparation for add in the probabilities of our joint events, looks like this.

Mammal and Tree dice joint distribution (X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly							**0.1667**
Orangutan							**0.1667**
Lion							**0.1667**
Panda							**0.1667**
Monkey							**0.1667**
Cheetah							**0.1667**
Total Tree	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**1**

Now, what is left to do is to fill out the remaining cells of the above table with the joint probabilities – in such a way that the two dice will be stochastically independent. This of course, is the main task at hand for developing Experiment 1.

Before we get started with that task, lets make some observation about the table so far. The "1" in the lower right cell indicates that the cells of the "Total Tree" row, and the cells of the "Total Mammal" both sum to one. What is also true is that all of the empty cells will also sum to one - once we have populated them with probabilities. These facts are true because we have three probability distributions represented in this table. The cells in the Total Tree row are the tree die component distribution "X"; the cells in the Total Mammal column are the mammal die component distribution "Y"; and the 36 empty cells are the joint distribution "(X, Y)".

Next, we want to complete the table by adding in the joint probabilities where the empty cells are. Since we are creating this distribution, we are free to put any probabilities we want – but under the constraints we have set out for ourselves.

For one thing, since all of the empty cells constitute a probability distribution, then they must sum to 1. (Even though they are a joint probability distribution).

But, there is another constraint. We have specified that we want the probabilities of this joint distribution to result in the two chance variables involved (the two dice) to be stochastically independent. Therefore, the probabilities we select to place in the table must be "just so". Otherwise the result will not produce stochastic independence.

Unfortunately, we haven't yet stated the criteria for picking the joint probabilities so that they will result in stochastic independence. We shall soon present an intuitive way to look at the graph of a joint distribution and tell whether the distribution is stochastically independent. But we shall have to hold off for now describing the formal criterion for creating the table in the first place so that we have stochastic independence.

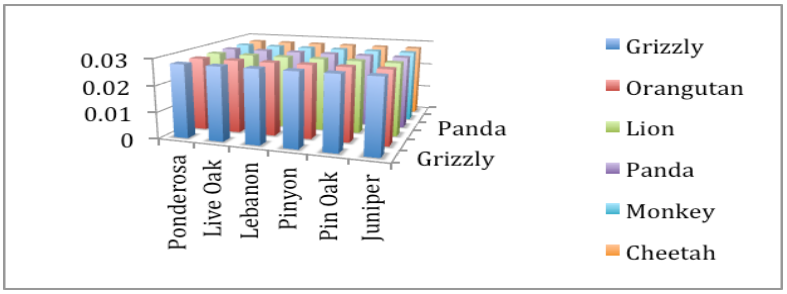
However, as the author of this primer, I already know the answer. Therefore, I'm going to go ahead and populate the above table with the correct probabilities to result in stochastic independence. Then, after we have presented the intuitive way of looking at its graph to test for independence, we will have an example to work with.

So, below, I have filled in the correct probabilities to ensure stochastic independence. We shall explain the intuitive way to ascertain that they are that very soon below.

Mammal and Tree dice joint distribution (X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Orangutan	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Lion	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Panda	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Monkey	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Cheetah	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Total Tree	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**1**

A graph of this perfectly uniform joint distribution looks like this:



The fact that the joint events must have probabilities that result in the mammal die chance variable and the tree die chance variable being “independent of each other” raises a question. The question is, “How does this ‘independence of the mammal and tree dice’ affect what the joint probabilities can be?”

We have actually already explained this above. And we repeat the answer here. What it means is this:

> The tree die being independent of the mammal die simply means that, for a given joint event, knowing the outcome of the mammal die does not assist in guessing the outcome of the tree die.

For example, suppose you roll the two dice, but do not look at the result. Instead you ask a friend to tell you which face landed up on the mammal die. The question is, if he tells you the answer, will that help you guess which face turned up on the tree die? If this information *does help* you guess the tree die result, then the two chance variables are not independent in the joint event (situation). If the information *does not help* you guess the outcome of the second die, then the two chance variables are independent of each other.

We are now going to present the method that we promised above by which one can test out a joint distribution to see if its joint probabilities are such that the two chance variables involved are stochastically independent.

It turns out that this method involves "doing something" to the joint distribution table that changes it – that transforms it into a new table. This new table is called the *conditional probability distribution*, or abbreviated to the *conditional distribution*. It is actually the conditional distribution that exposes whether the two chance variables are stochastically independent.

So, first, we shall show how to transform – or derive, actually – the joint distribution table to the conditional distribution table.

But first lets discuss why the joint distribution by itself is not exactly what we need to tell by inspection whether we have stochastic independence.

Recall our question from above: "How does this 'independence of the mammal and tree dice' affect what the joint probabilities can be?"

We said that what it means is this:

> The tree die being independent of the mammal die simply means that, for a given joint event, knowing the outcome of the mammal die does not assist in guessing the outcome of the tree die.

So we must ask, "Using the above table, how can we ascertain whether knowing which mammal die face landed up helps us to guess which tree die face also landed up?

Consider this. "Knowing what the outcome of the mammal die is" is equivalent to selecting one of the rows of the above table!

For example, if we find out that the mammal die landed with "Panda up", then we know that the Panda row above is the one that matters, and the other rows no longer matter for this trial.

In fact, we could say that all of the other rows have been eliminated – just by knowing that Panda landed up on the mammal die. This observation is very helpful, because would allow us to re-write the above probability picture *simply by eliminating all of the rows except for the Panda row*.

This results in the following revised table:

Revised joint distribution, given that the mammal die landed with Panda up

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Panda	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Total Tree	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**

Notice that the total tree probabilities are now the same as the Panda row, since the finding that the mammal die actually landed with Panda up has eliminated the other rows.

But notice what is wrong with this table. Its total probabilities no longer sum to one! Rather, it sums to **0.1667**. Of course, this is a problem because all probability distributions must sum to one. Consequently, we must take corrective action. What we must do is to "normalize" this table so that it sums to 1.

However, if we "normalize" this table, the relative proportion of the probabilities in the Panda row should remain the same as they are now. To maintain this proportionality we can divide all of the probabilities in the Panda row by their sum, which is **0.1667**.

If we do this – normalize this table by dividing all if the entries in the Panda row by 0.1667, then we get the following normalized table – for the case when we know that the mammal face landed with Panda up:

Normalized Revised joint distribution, given that the mammal die landed with Panda up

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Panda	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Total Tree	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**1**

Dividing each probability in the Panda row, 0.0278, by their row sum of 0.1667 yields 0.1667, as shown in then above normalized table. And, of course, this results in a new row sum for the Panda row of "1". And, of course, the new Total Tree row changes to whatever the Panda row is, since it is the only "mammal faces" row.

Here is what this little table tells us: If we knew that the mammal die landed with Panda up, then the effective joint probability table would be changed, or "collapsed", to the above normalize, revised distribution.

It should be realized that this "normalized revised" distribution is a probability distribution for tree die faces landing "up".

In fact, we shall give this distribution a name (other than "normalized revised"). We shall call it the *conditional probability distribution for the tree die given that the mammal die was Panda*. Now that's a mouthful, so we shall symbolized it as

(Tree | Mammal=Panda)

Again, this is spoken as "the tree die probability distribution given that the mammal die landed with Panda up". Or it is abbreviated "Tree given Mammal is Panda". (The "|" symbol is pronounced, "given".)

But what we really want to know – in order to establish stochastic independence – is whether or not the conditional probability distributions for any other mammal die faces would have been any different than for Panda.

In other words, what we need to do is to calculate the conditional distributions for all six mammal die faces, and see if they are all the same as each other. If they are all the same, then the two dice chance variables are stochastically independent. But if any are different, then the two dice chance variables are stochastically dependent.

So, what we shall now do in order to determine stochastic independence is to calculate the other five of these conditional distributions. However, instead of calculating them as five separate tables, it will be tidier to put them all in a single table. And we can omit the "Tree Total" row, since each row will itself sum to one for this new table.

This new table needs a name. We shall call it the conditional distribution for Y given X, and symbolize it as "(Y | X)".

The conditional distribution (Y | X) is derived from the joint distribution (X, Y) by dividing each row of (X, Y) by its row sum. In order to calculate the conditional

distribution for Experiment 1, it will be convenient to repeat the joint distribution (X, Y) for Experiment 1 from above, which we do here:

Experiment 1 joint distribution (X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Orangutan	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Lion	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Panda	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Monkey	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Cheetah	0.0278	0.0278	0.0278	0.0278	0.0278	0.0278	**0.1667**
Total Tree	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**1**

Thus, to calculate the conditional distribution (Y | X) for Experiment 1, we divide each probability in the above table by its row sum. We also drop the "Total Tree" row. Here is the result:

Experiment 1 conditional distribution (Y | X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Orangutan	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Lion	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Panda	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Monkey	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**
Cheetah	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	**1**

Now, we are finally ready to ascertain whether the chance variable Y is stochastically independent of the chance variable X for this joint distribution.

We do this by comparing the rows of the above conditional distribution. If all of the rows are the same as each other, then we have stochastic independence. And we do have stochastic independence!, because all of the rows are indeed the same as each other.

We know this because no matter which of the mammal die faces landed up, the conditional probability distribution for the tree die – shown by the associated row – would have been the same. Thus, knowing which mammal face landed up gives no useful information to help guess which tree die had landed up.

(Notice that all the columns are the same as each other also. But that fact has no bearing on concluding stochastic independence.)

Summary of Experiment 4

In absence of any formal methods, we have just developed an informal method of looking at a joint probability distribution and ascertaining whether its component chance variables are stochastically independent or stochastically dependent.

First, derive the conditional distribution from the joint distribution (X, Y) by dividing each of its joint probabilities by their row sum. This results in a new table called the conditional distribution (Y | X) of the two chance variables.

Then, inspect the conditional distribution to see if all of its rows are equal to each other. If they are, then the two chance variables are stochastically independent. If the rows are not all the same, then the two chance variables are stochastically dependent.

Although, after we have presented all five of the sample experiments in this section, we shall introduce an intuitive graphical inspection to obtain a sense of relative degrees of dependence of two or more joint distributions. And we shall apply this informal method to our five experiments there. Then later in the formal section, we shall verify this intuition with formal measures of stochastic dependence.

(What we have not yet done is to provide a method for determining the degree of stochastic dependence of two chance variables. Such a method will have to wait until the formal section on information theory mathematical constructs. However, intuitively, you can imagine that if all of the conditional probability distributions - all of the rows of (Y | X) – are conspicuously different from each other, then the degree of dependence should be large. However, if the all of the rows of (Y | X) are nearly the same as each other, then the degree of dependence should be small. The problem, then, is to determine a way to measure the extent to which these rows are different from each other. This is the ultimate task of Part II of this primer.)

Things to Wonder About

Notice that the probability of each joint sample point in the joint probability distribution for Exercise 1 is 0.0278. Notice also that both component probabilities for that joint sample point are 0.1667. Question: How can one use the two component probabilities (0.1667 and 0.1667) in a mathematical expression to obtain 0.0278 as the answer?

The answer is “Multiply them!”

In other words, each of the joint probabilities in the table above is the product of its respective component probabilities. (That is, 0.0278 = 0.1667 x 0.1667, at least accurate to 4 decimal points.) This fact raises two other questions:

1. If all of the probabilities in a joint distribution are the product of their respective component probabilities, then does this fact guarantee that the two component chance variables involved are necessarily stochastically independent?
2. If two chance variables are stochastically independent, then does this fact guarantee that each of its joint probabilities must be the product of the probabilities of its respective component probabilities?

Another question regards the possible symmetry of the independence between two chance variables. We made a good argument above that the table of joint probabilities we ended up with represents the case that the tree die’s outcome is independent of the mammal die’s outcome. In other words we were able to set up the joint probabilities so that the tree die chance variable is independent of the mammal chance variable. But, must it necessarily follow that the mammal die must then also be independent of the tree die?

This question is the same as asking this: Suppose one knows that “Knowing the outcome of the mammal die does not help an observer to guess which face would turn up on the tree die.” Then can one necessarily conclude from this fact that “Knowing the outcome of the tree die does not help an observer to guess which face would turn up on the mammal die?”

In other words, the question is, if one chance variable is independent of a second, then must the second be independent of the first? That is, “Is stochastic independence symmetrical?” Can we dispense with saying “Y is independent of X” and simply say, “X and Y are independent?”

We shall leave the investigation of both of these questions until later – after we have finished describing all five of the example experiments.

Experiment 2: A Joint Independent Experiment with Non-equal Probabilities

In the first experiment, we took two chance variables both with equally likely probabilities, and we were able to form a joint distribution from them in such a way that the second one (the tree dice) was statistically independent of the first.

So, it is fair to ask the following question: “Does the fact that the probabilities of both of the underlying chance variables are equally likely have something to do with their independence?” In other words, if we made both the two chance variables not equally likely (not uniform), then is it still possible to assign probabilities to their joint distribution in such a way that they are independent?

This second example that we provide in this section will prove that we can. That is, we are going to provide two dice whose probabilities are not equally likely. And yet we will still be able to give them a joint distribution that makes them stochastically independent of each other. Let us now proceed to see how we do that.

The Probabilities for each Die Individually

Suppose we asked the custom device vendor to “load both dice”. This would make them “unfair dice”. Secretly loading dice is a way to cheat in dice games. It means that a small weight has been embedded within each die and that the placement of this weight is off-center. This causes the probabilities of the faces of the die to no longer be equally likely. Typically, the cheating player will secretly replace both “fair dice” with a loaded pair. The cheater will know that the probabilities are no longer equally likely – but the other players will not.

We will still use the mammal die and the tree die as before, but the probabilities of their face will no longer be uniform due their loading with an off-center weight.

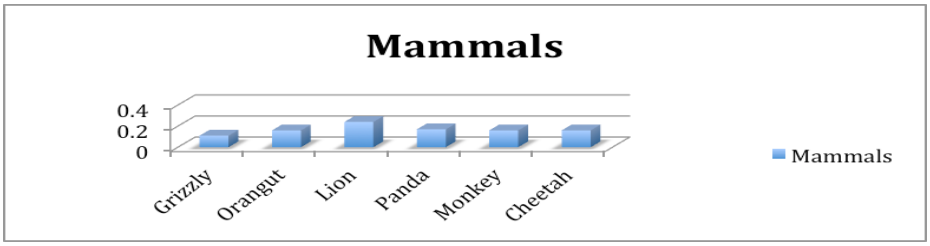
As for the mammal die, we shall assume that the off-center loading produces the following table of probabilities when the die is rolled by itself. (This might result, for example, if the embedded weight more moved from the center of gravity of the die toward the common vertex of the Grizzly, Orangutan and Lion faces.)

We shall also begin to refer to the mammal chance variable as “X”, and the tree chance variable as “Y” for the remainder of this primer.

Mammal Die Faces Component Probability Distribution X

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total
Mammals	0.1100	0.1600	0.2400	0.1700	0.1600	0.1600	1

The probability graph for X is:

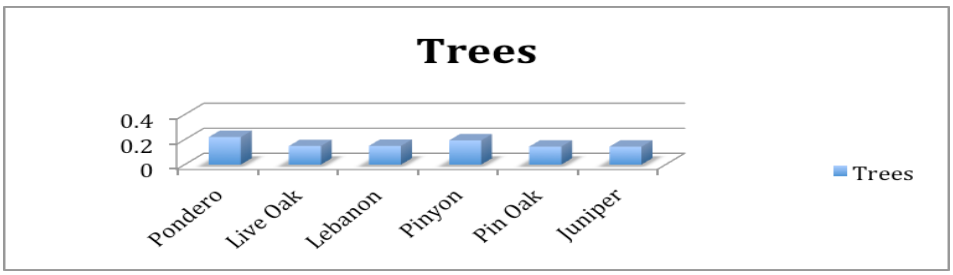


For the tree die, chance variable Y, lets assume the following probabilities (when the die is rolled by itself) are exhibited.

Tree Die Faces Component Probability Distribution Y

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total
Trees	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1

The probability graph for Y is:



The Probabilities of the Joint Events

Just like we did in the first example experiment, we must now see if we can select some probabilities to assign to the 36 joint events in such a way that the two chance variables, the mammals die and the tree die, are *stochastically independent*.

We had pretty good luck with a particular algorithm in the first example. So lets see if the same algorithm works also for the second example – even though the second example does not involve individual distributions that are equally likely.

Recall that the algorithm that we used in the first example experiment was this: For each joint sample point, we multiplied its two component probabilities in order to obtain its joint probability.

Lets translate the above statement into "probability-speak" so that we can start to become accustomed to how mathematicians talk. The advantage this brings is precision of thought. As these examples get more complex, this ability to speak precisely will become invaluable. Here's how we shall say this more precisely:

> For each joint sample point (x, y), where x is a mammal face and y is a tree face; its probability p(x, y) will be p(x)*p(y), where p(x) is the probability of mammal "x", and p(y) is the probability of tree "y" according to the two above tables.

In words, the probability of a joint event will be the product of the probabilities of its two underlying component events. For example, suppose we want to assign a probability to the joint event (Lion, Pinyon) according to this algorithm. Then we look and see that the probability of Lion from the mammal component probability distribution is 0.2400; and we see that the probability of Pinyon from the tree component probability table is

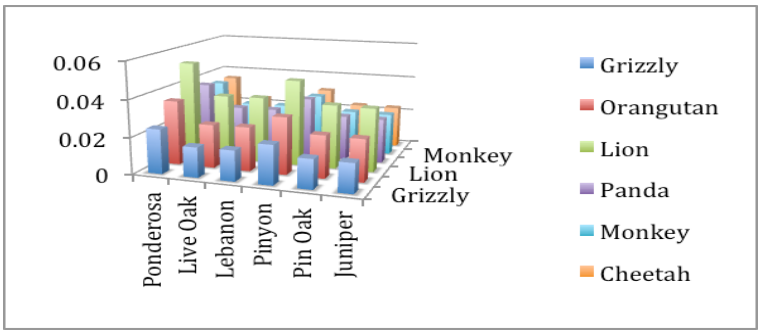
0.1940. Therefore, our algorithm says to assign the product of these two probabilities to the joint event (Lion, Pinyon). Thus the probability of (Lion, Pinyon) will be assigned .0466 (rounded up to 4 decimal places).

If we follow this algorithm for all 36 of the joint events, we get the following table.

Experiment 2's joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0242	0.0163	0.0165	0.0213	0.0159	0.0158	**0.1100**
Orangutan	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Lion	0.0528	0.0356	0.0360	0.0466	0.0347	0.0344	**0.2400**
Panda	0.0374	0.0252	0.0255	0.0330	0.0245	0.0243	**0.1700**
Monkey	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Cheetah	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

The graph of the these joint probabilities is:



The Independence of the Mammal and the Tree Chance Variables

Lets now determine whether the joint distribution for Experiment 2 that we developed above results in the experiment's two chance variables being *stochastically independent*.

Naturally we shall subject Experiment 2's joint probability distribution to the same test that we subjects Experiment 1's joint probability distribution. The reader will recall that this test consists of:

1. Deriving the conditional probability distribution for the experiment from the joint distribution. One can accomplished this by dividing every joint probability in the joint distribution table by its row sum. Also, the conditional distribution does not need the Total Tree row.
2. Once the conditional distribution has been derived from the joint distribution in this manner, then inspect the conditional distribution to see if all of the rows are equal to each other. If so, then the two chance variables are *stochastically independent*. If any of them differ from the others, then the two chance variables are *stochastically dependent*.

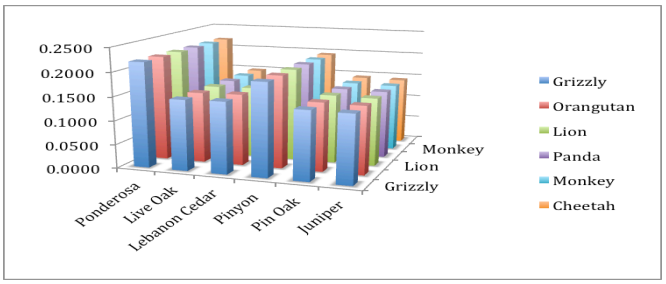
Here is the conditional distribution derived as described from the above joint distribution:

Experiment 2's conditional distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Orangutan	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Lion	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Panda	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Monkey	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Cheetah	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1

Clearly, all of the rows have the same probability distributions as all of the other rows. Therefore, Experiment 2's chance variables are stochastically independent.

The graph of the these conditional probabilities is:



Notice how the conditional distribution for Experiment 2 reveals that, when the "weights" of each of the individual "animal" distributions is normalized (so that the all have the same "weight" as each other), then these animal distributions are the same. In other words, it does not matter which of the animal faces shows up on the mammal die, the probability distributions for the Y= Tree faces is the same distribution.

Things to Wonder About

Thus far we have twice obtained joint probability distributions between two chance variables such that variables turned out to be statistically independent. The algorithm that we used in both cases to obtain the joint probabilities was to multiply their respective component probabilities.

This scheme, or algorithm, worked in the Experiment 1 where the underlying individual chance variable probabilities were equally likely (the uniform distribution). But the same multiplication algorithm also worked in Experiment 2 where the two underlying distribution were not uniform.

So, perhaps this idea of multiplying the two underlying probabilities together to get the joint probabilities is somehow fundamental to the idea of statistically independent chance variables.

This raises the following general question: "It appears that multiplying the component probabilities yields joint probabilities such that the two chance variables are statistically independent. Firstly, is this always true? And, secondly, is this the only algorithm that results in statistically independent chance variables? That is, is the algorithm unique?"

And here is a second question: “If one chance variable is statistically independent of a second, then must the second be statistically independent of the first? In other words, is stochastic independence symmetric? Can we stop saying “Y is statistically independent of X” and simply say, “X and Y are statistically independent?”

We shall leave the investigation of both of these questions until later – after we have finished describing all four of the example experiments.

Be aware that we have not actually proven yet that this multiplication scheme produces stochastic independence. All we have done so far is to show two examples in which we calculated joint probabilities by multiplying their component probabilities. And we subsequently did show that those two example joint distributions happened to result in stochastically independent chance variables. But right now we are just presenting five examples. After we are done with that, then we are going to have a more formal development of all of these ideas – including our “multiplication scheme” for achieving stochastic independence.

Three Example Dependent Joint Experiments

In the three experiments discussed in this section, we are going to set up the joint probabilities so that the outcome of the tree die *does* depend on the outcome of the mammal die.

In other words, suppose that both dice are rolled together; and subsequently someone tells us which face of the mammal die landed up. In these three experiments, our new knowledge of which mammal die face landed up *will help us* to more accurately guess which tree die face also landed up.

This is what it means to say that the outcome of the tree die is stochastically dependent on the outcome of the mammal die; or that the outcome of the mammal die *portends something* about, or *is meaningful to*, the outcome of the tree die.

But what does stochastic dependence mean in terms of probabilities? It means that the probability distribution for which face of the tree die lands up will depend upon which face of the mammal die landed up. In short, it means that the conditional probability distributions for each of the mammal faces are not all the same as each other.

For example, if Lion lands up on the mammal die, then the probabilities of the faces of the tree die will one distribution, but if Monkey lands up on the mammal die, then the probabilities of the faces of the tree die may be a distinct distribution. It means that at least one of the conditional distributions will be different from another.

Recall that the opposite of this is true whenever the joint distribution represents stochastically independent chance variables. In that case, the conditional distributions given that any of the mammals faces lands up are all the same as each other.

In other words, to say that the tree die is stochastically dependent of the mammal die means that the conditional probability distributions for the faces of the tree die change depending on which face of the mammal die landed up. In fact, there might be as many a six different conditional probability distributions for the tree die, one for each face of the mammal die. One thing for certain is that the same probability distribution for the tree die does not work for all six faces of the mammal die.

We are going to provide three different example experiments that are stochastically dependent. The essential difference among these three experiments lies in their degrees of stochastic dependence.

Remember, it is stochastic dependence that gives information theory its sense of meaningfulness, its powers of predictability, and the self-awareness to be able to evaluate its ability to make predictions in any given situation. Thus, obtaining both an intuitive and a formal grasp of stochastic dependency is paramount in information theory.

What we intend to show in this section - once we have described all three of these experiments is:

> Each of these three stochastically dependent experiments exhibits a different degree of stochastic dependency from the other two experiments. In other words, we shall show by example that the idea of "degree of stochastic dependence" is a meaningful concept; and therefore it makes sense to want to measure this degree of stochastic dependence. At this time we shall approach this notion of measuring the degree of dependence of two chance variables in a graphical and intuitive manner.

What we intend to show in the formal section, which follows the introduction of the five example experiments, is:

> There exists mathematically a measure of the degree of stochastic dependence of the two chance variables of a joint distribution. Its name is *mutual information*. We shall show that the mutual information of each of these three experiments – as measured by the formal mathematical definition - meets with the intuition of the reader concerning the three experiments arrived at in this present section.

Lets now describe each of these three experiments; and at the same time, attempt to develop an intuitive sense of

- what stochastic dependency means, and
- what makes one joint distribution more or less stochastically dependent than others

by inspecting graphs related to each of these three experiments.

Experiment 3: A Dependent Joint Experiment with Equal Probabilities

In this experiment, we are going to tamper with each die in manner that results in the dice having an influence on each other. This will mean that whichever face lands up on the mammal die will have an influence on, or an affinity for, which face lands up on the tree die.

In terms of probabilities, this means that whichever face turns up on the mammal die will specify all of the probabilities of the faces of the tree die turning up. And, for each face that turns up on the mammal die, a different probability distribution may be specified for the six tree die faces.

This is what makes Experiment 3 different from Experiments 1 and 2. In those experiments, it did not help us to know which face turned up on the mammal die, because the resulting probability distribution for all six faces of the tree die was the same, no matter which face turned up on the mammal die.

The way we are going to induce such a situation is to have the custom dice vendor, when it manufactures our dice, to embed a small magnet inside of each die. Of course, each magnet will have a "north" and a "south" magnetic pole. And how the vendor chooses to orient the magnets inside of their respective die will determine which faces on the mammal die are magnetically attractive –or repellent - to which faces in the tree die.

This orientation of these hidden magnets will then concomitantly determine the probabilities of how the faces landing up on the mammal die influence the faces landing up on the tree die. The way we say this in probability language is that the joint probabilities are determined by the orientations of the hidden magnets within the respective dice.

Now, in this particular experiment, we are going to specify to the vendor that we want both magnets perfectly centered within their respective die – so that the magnets will not act as a load to make the dice unfair.

In fact, each dice individually will remain fair! This is because the centering of the magnet will not affect the probabilities of each face when a die is tossed by itself.

It is only when the two die are rolled together that we see unexpected behavior – and unexpected probabilities! This is because only when rolled together and in near proximity within the influence of the magnetic fields of the two dice will the magnets in the two die have a chance to have a influence upon each other.

In fact, to make this experiment consistent every time the pair of dice is rolled, we will have to specify that they be tossed into a confined area that keeps them in close proximity – such as a small (say 1meter) box[17].

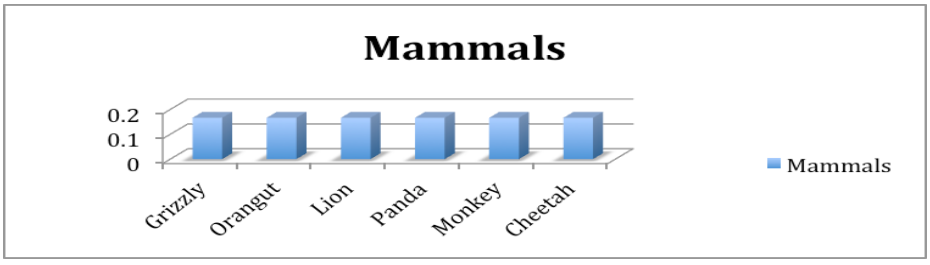
The Component Probability Distributions

As we have already established for this experiment, the magnets are perfectly centered within both dice with the result that the probabilities of each tie – when tossed in isolation of the other die. This means that the faces of both dice are equally likely, and therefore have probability of 1/6 = 0.1667 – accurate to 4 decimal places. The probabilities for the faces of each die in this experiment are listed in the two tables below.

Mammal Die Faces Probabilities X

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah
Mammals	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

The probability graph for X is:



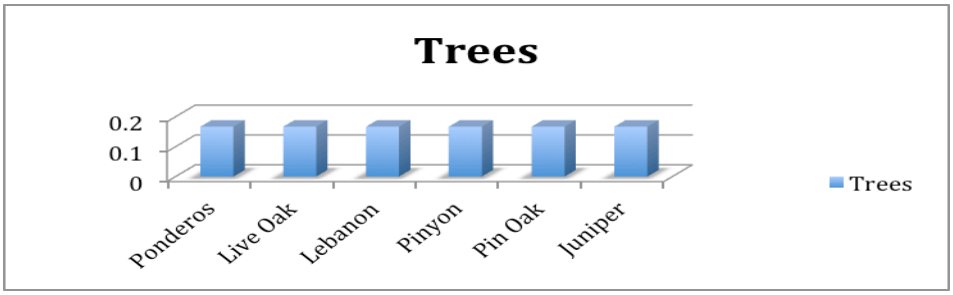
[17] If we did not confine the two dice to a small area, such as a 1 meter box, then dice could sometimes roll away from each other, and the interplay of the two magnetic fields could have different strengths each time the dice were rolled. Of course, this would cause the joint probability distribution to change with each roll of the dice. As we shall see in the next chapter, such changing of the probability distribution with each trial is called a time-inhomogeneous stochastic process – which is more complicated than we want to work with in this chapter.

For the tree die the probabilities of the six faces, since they are equally likely for this experiment, must be as follows (accurate to 4 decimal places):

Tree Die Faces Probabilities Y

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper
Trees	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

The probability graph for X is:



The Joint Probability Distribution

We can give this Experiment 3 any joint probabilities that we want. But we are limited by a couple of constraints that we imposed upon ourselves above.

The first constraint is that this joint distribution must be stochastically dependent. This fact means that it *cannot be stochastically independent*, such as Experiment 1 – which, in fact, shares the same two component distributions as the present experiment. So we want the joint distribution for this Experiment 3 to be different from the joint distribution of Experiment 1.

Now since Experiment 1 is stochastically independent, then it should exhibit no (degree of zero) stochastic dependency. But, of course, we want Experiment 3 to exhibit some positive amount of stochastic dependency.

Therefore, any amount of difference between the joint distribution of Experiment 3 and the joint distribution of Experiment 1 should represent the positive amount of stochastic dependency that is exhibited by Experiment 3.

Consequently, a reasonable strategy for inventing the joint distribution for Experiment 3 would be to make some alterations to the joint distribution of Experiment 1. And, intuitively, the “greater” are the alterations we make to Experiment 1 in order to produce Experiment 3, the greater will be the degree of dependency of Experiment 3!

This fact will form the basis for developing our intuition for how to compare two or more joint distributions in order to ascertain their relative degrees of stochastic dependence. That is, in order to achieve a joint distribution that is stochastically dependent, we shall begin with one that is stochastically independent and then apply some changes to its joint probabilities. And, the greater the amount of changes that we make, the greater we would expect the degree of stochastic dependence to be. The difficulty, of course, is to figure out how to measure this “amount of change”. Such is the subject of the formal development of mutual information later in Part II.

But of course, we can’t make just any changes to it. We have to make sure that the these three things are preserved across these changes:

1. The total sum of all of the joint probabilities must remain 1, because a joint distribution is, after all, a probability distribution.
2. All of the rows of the new joint distribution must sum to the same value as same row for Experiment 1. The reason for this is because these row sums happen to be from the Mammal component probability distribution. And the mammal component probability distributions are the same for both experiments.
3. All of the columns of the new joint distribution must sum to the same value as same column for Experiment 1. The reason for this is because these column sums happen to be from the Tree component probability distribution. And the tree component probability distributions are the same for both experiments.

Obviously, creating the new joint distribution for Experiment 3 by making some changes to the joint distribution of Experiment 1 will not be easy. But I have already done it for us and shall present it momentarily.

The second constraint that we have laid out for ourselves is that we want Experiments 3, 4 and 5 to provide a gradual increase in their degrees of stochastic dependency.

This means, intuitively, that that:

- Experiment 3 should be just a little different from Experiment 1.
- Experiment 4 should be more different from Experiment 1 than Experiment 3 was.
- And Experiment 5 should be more different from Experiment 1 than Experiment 4 was.

As indicated above, the author has already worked out a joint distribution for Experiment 3 that satisfies these constraints and shall present it here.

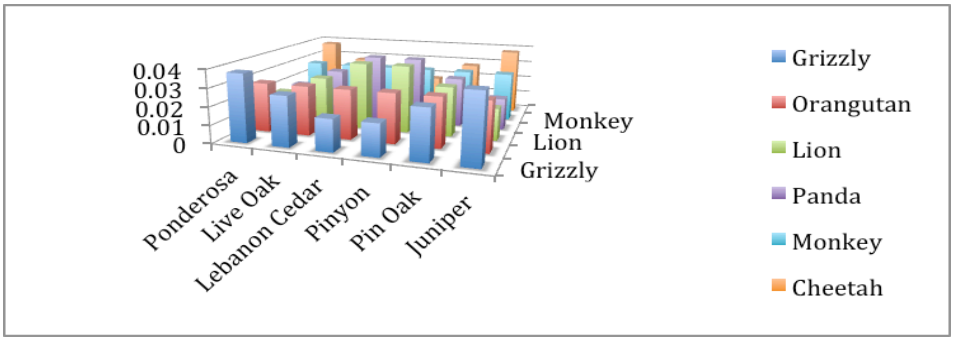
Recall that this particular joint distribution for Experiment 3 will have been the result of the way that the custom dice vendor oriented the north-south poles of the magnets that it embeds inside of the two dice.

Notice that many of the joint probabilities have the value 0.0277, as did all of the probabilities of the joint distribution of Experiment 1. However, some have been judiciously altered so as to conform to the above constraints.

Experiment 3 joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0382	0.0277	0.0177	0.0177	0.0277	0.0377	**0.1667**
Orangutan	0.0277	0.0282	0.0277	0.0277	0.0277	0.0277	**0.1667**
Lion	0.0177	0.0277	0.0382	0.0377	0.0277	0.0177	**0.1667**
Panda	0.0177	0.0277	0.0377	0.0382	0.0277	0.0177	**0.1667**
Monkey	0.0277	0.0277	0.0277	0.0277	0.0282	0.0277	**0.1667**
Cheetah	0.0377	0.0277	0.0177	0.0177	0.0277	0.0382	**0.1667**
Total Tree	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**0.1667**	**1.0000**

And the graph of these joint probabilities is:



The careful reader will notice a number of things about the values that we just placed in the joint probability cells. Conspicuous among these things is that some of the joint probabilities are different from the joint probabilities of Experiment 1. This is necessary in order to make Experiment 3 stochastically independent.

The fact that at least some of these joint probabilities is different from Experiment 1 also ensures that not all of these joint probabilities are the product of their component probabilities. For example, the probability of the joint event (Grizzly, Ponderosa) is 0.0382. But the product of its component probabilities is 0.1667*0.1667 = 0.0277.

Also, the joint distribution is no longer uniform, because not all the joint probabilities are 0.0277 (which is 1/36 rounded to 4 decimal places) as in our first experiment. In fact, some of the joint events have probability 0.0177, some 0.0277, and some 0.0377.

However, notice that the individual component distributions remain individually uniform. (These can be seen here in the “Total Mammal” column and the “Total Tree” row.) This means that if you were to throw each die by itself, it would prove to be “fair”. But when you throw them together, the joint probabilities will not be equally likely as expected by the other players, and therefore the game will have been rigged in a clever way that is known to us (the “cheaters”), unknown to the other players, and difficult to detect.

The Dependence of the Mammal and the Tree Chance Variables

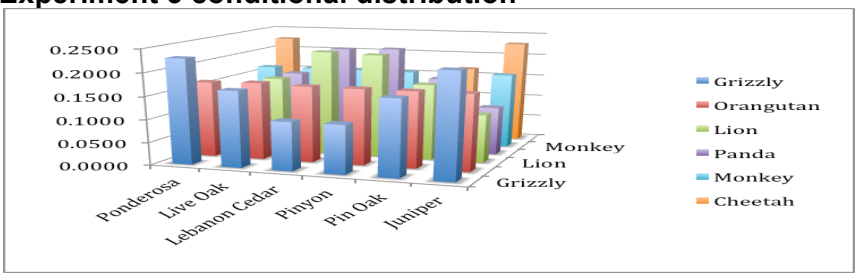
Of course, our goal is to determine whether or not Experiment 3 is stochastically dependent or not. As we learned with Experiments 1 and 2, this is determined by deriving the conditional probability distribution from the above joint probability distribution. The reader will recall that this is accomplished by dividing every joint probability by its row sum. For Experiment 3 this yields:

Experiment 3 conditional probability distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.2292	0.1662	0.1062	0.1062	0.1662	0.2260	**1.0000**
Orangutan	0.1662	0.1692	0.1662	0.1662	0.1662	0.1660	**1.0000**
Lion	0.1062	0.1662	0.2292	0.2262	0.1662	0.1060	**1.0000**
Panda	0.1062	0.1662	0.2262	0.2292	0.1662	0.1060	**1.0000**
Monkey	0.1662	0.1662	0.1662	0.1662	0.1680	0.1672	**1.0000**
Cheetah	0.2262	0.1662	0.1062	0.1062	0.1662	0.2290	**1.0000**

The graph of this conditional distribution is

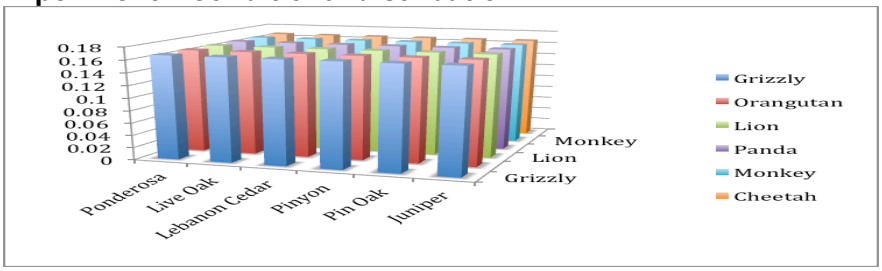
Experiment 3 conditional distribution



With this conditional distribution, we are now in a position to compare all of the rows to see if they are the same. If they are, then the two component chance variables associated with this distribution are stochastically independent. If any of the rows are different, then the two component chance variables associated with this distribution are stochastically dependent.

It is clear by inspection that the rows are not all the same as each other, since the first two are different. For example, compare the above graph with that of the conditional distribution for Experiment 1, where all of the rows are the same.

Experiment 1 conditional distribution



Therefore, we can conclude that the two chance variables, X and Y, of Experiment 3 are stochastically dependent.

Another way to see this is to look at the conditional probability of a particular tree die face landing up for two different mammal faces and find that they are different. For example, the probability of the tree die landing with Ponderosa up is 0.1062 whenever the mammal die that landed up was Lion. But, the probability of the tree die landing with Ponderosa up is 0.2262 whenever the mammal die that landed up was Lion.

Another thing to notice is that our algorithm of calculating a joint probability by multiplying its two component probabilities does not work for dependent chance variables. At least it does not work for all of the joint probabilities in the joint distribution. It may work for some (for example, the joint event (Orangutan, Ponderosa)), but it does not work for all joint events in the table.

So, from this example it looks as though whenever the two chance variables involved in a joint distribution are statistically dependent, there will be at least one joint event (often many) whose probability is not the product of its component probabilities.

But, beware; we have yet to prove this. At this point in this reading, it is merely a strong suspicion.

Things to Wonder About

We have visited the question of whether stochastic independence between two chance variables is symmetric without yet resolving it. This time lets visit the question of whether stochastic dependence is symmetric.

We have said many times in describing this experiment that “the tree die depends upon the mammal die” – or something to that effect. But is it not also true that “the mammal die depends upon the tree die”? Is it possible to have a situation where one die depends upon another, but the other does not depend upon the one? Is it logically possible to have any situation where one party depends upon a second party, but the second party does not concomitantly depend upon the first?

If this suspicion is true, then stochastic dependence is a symmetrical relationship between two chance variables.

If it turns out that stochastic dependence is symmetrical for our two dice; then is it “caused by” the physics of the situation – for example the way magnetism works? Or might this symmetry be fundamentally a probability phenomenon? In other words, is stochastic dependence necessarily symmetric? And for probability reasons alone?

This entire line of questioning brings to issue whether causality and stochastic dependence are logically consistent. Are they compatible philosophies that both explain how one phenomenon “depends upon”, or “originates” another? To delve more deeply into these issues, you may want to read Appendix 3 of this primer, entitled “Three Approaches to Critical Thinking”.

Experiment 4: A Dependent Joint Experiment with Non-Equal Probabilities

In this experiment, we are going to continue with the same “tampered” dice that we used in the previous example. This means that the two die will still have magnets secretly embedded in them. And these magnets will have an effect when both dice are thrown together as their respective magnetic fields interact.

As in the previous examples, how the mammal die lands will exert a certain magnetic attraction to a certain other face on the tree die. And this magnetic field will bias which

face lands up on the tree die. Just as in Experiment 3, this effect shows up as modifications to the joint probability distribution of the two die when thrown together that biases it so that we now have a dependent joint probability distribution.

However, in addition, we shall also use the magnets as a weighted load so that the probabilities of either one of the die when thrown in isolation of the other die will be changed as well. They will no longer be equally likely. In other words, not only will we use magnetized weights so as to affect the outcomes of the two dice when tossed together, but also we shall place the weights off-center in each die so that their probabilities when tossed separately will be changed.

The idea of Experiment 4 is to see what happens when we not only have dependencies between the two die because of the hidden magnets; but also each individual die is loaded with an off-center weight, which changes its basic probabilities from equally likely to not equally likely.

What we want to see is how the dependencies and the non-equal probabilities of the distributions of the faces of the both die effect the joint distribution of probabilities for both dice when they are rolled as a pair.

The Component Probability Distributions

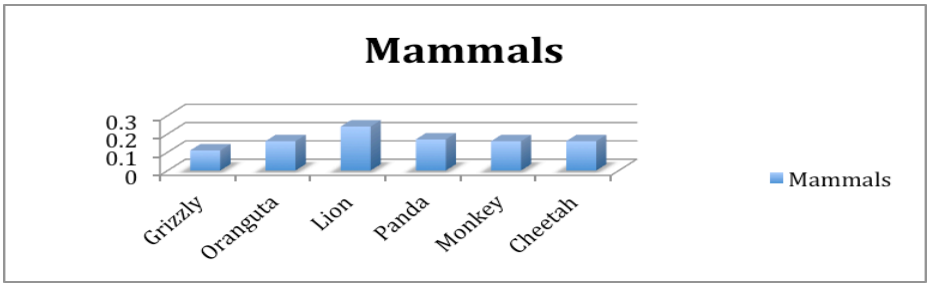
We have already established that for this experiment, the magnets are embedded off-center within both dice with the result that the probabilities of each tie – when tossed in isolation of the other die – will exhibit non-equal probabilities for their faces. The component probabilities for the faces of each die in this experiment are listed in the two tables below.

For the mammal die the probabilities of the six faces, we shall assume that the off-center loading produces the following probabilities:

Mammal Die Component Probability Distribution X

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total
Mammals	0.1100	0.1600	0.2400	0.1700	0.1600	0.1600	1.0000

The probability graph for X is:

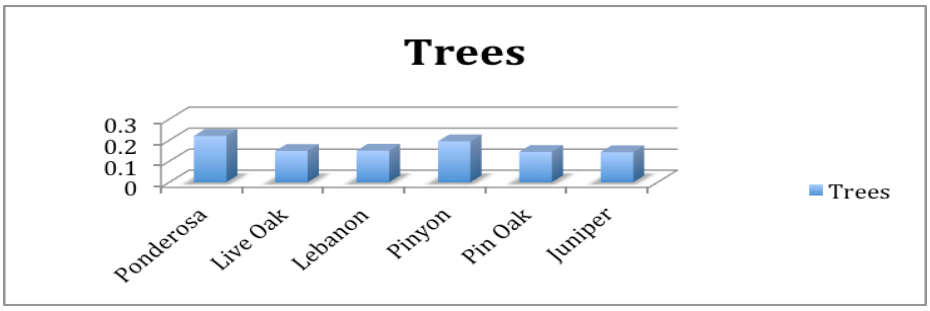


For the tree die the probabilities of the six faces, we shall assume that the off-center loading produces the following probabilities:

Tree Die Component Probability Distribution Y

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total
Trees	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1.0000

The probability graph for Y is:



The Probabilities of the Joint Events

Depending upon how the custom dice vendor orients the polarities of these off-centered embedded magnetized weights, the magnetic fields of these two dice will interact to produce a dependence effect just as in the previous experiment – regardless of the fact that the dice are both "loaded" and the probabilities of the individual die in isolation are non-uniform.

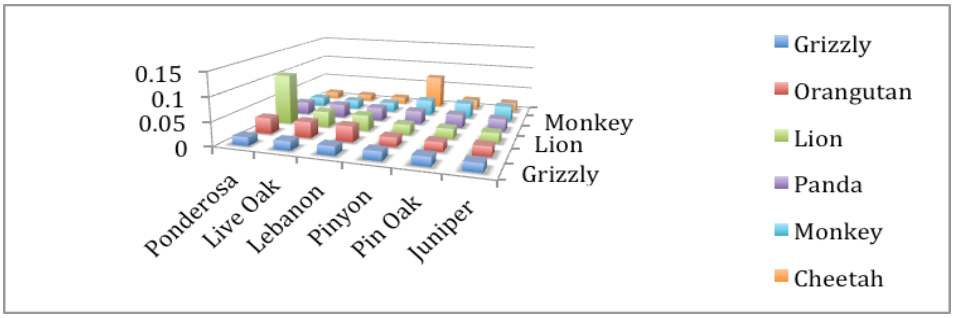
Let us assume that this placement and orientation of the weighted magnets inside of each die results in the following joint probabilities for the two dice when they are thrown together.

Notice that, as with the previous experiments, the total mammal row sums (in the "Total Mammal column) are, in fact, the mammal die component distribution; and that the total tree column sums (in the "Total Tree row) are, in fact, the tree die component distribution. This is always true for any joint distribution.

Experiment 4 Mammal and Tree dice joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

The graph of this joint distribution is:



As with the previous experiment, we want to test to see if these two chance variables are revealed as stochastically independent or stochastically dependent.

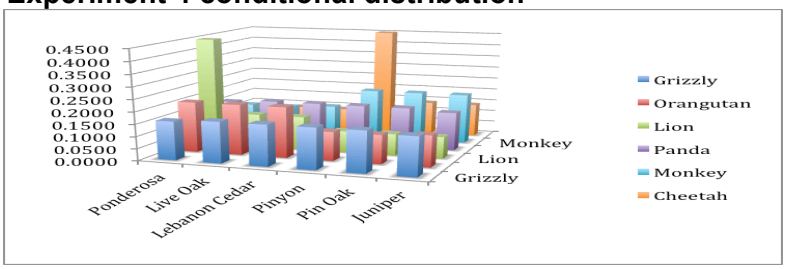
Recall that the first step of our test for stochastic dependency is to derive the conditional probability distribution for these two chance variables from their joint distribution. Recall that this is done by dividing each joint probability (cell) of the above joint distribution table by its row sum. The resulting conditional distribution is:

Experiment 4 conditional probability distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

The graph of this distribution his here:

Experiment 4 conditional distribution



Recall that the next step in our test for stochastic dependency is to compare the rows in the conditional distribution. If all the rows are the same as each other, then the two chance variables are stochastically independent. But, if at least one of them is different, then the two chance variables are stochastically dependent.

Upon inspection, we see that the rows are not all the same. Therefore, the mammal die and the tree die of Experiment 4 are stochastically dependent.

Things to Wonder About

We have seen that the two dice in Experiments 3 are stochastically dependent. The same can be said for Experiment 4.

But, is there some sense in which one of these experiments if “more stochastically dependent” than the other? In other words, is there some meaningful notion of “degree of stochastic dependency”?

If so, then how would it be measured?

If one experiment (joint distribution) were, in some sense, “more stochastically dependent” than another, then how would this fact show up in their respective joint distributions? How would it show up in their conditional distributions? How would it show up in the graphs of their conditional distributions?

How could we go about devising a measure of the degree of stochastic dependence between the two chance variables of a joint probability space?

The answer to this question is the subject of the culminating section of Part II below entitled “Information Theory’s Mathematics of Stochastic Dependence”. This is the section that presents the mathematics of information theory that codifies the formal mathematics of the issues we have been dealing with in these five examples.

So, if you have difficulty answering this question, fear not. It will be answered in due course.

Experiment 5: Another Dependent Joint Experiment with Non-Equal Probabilities

We shall now present our fifth and final example - Experiment 5. The purpose of Experiment 5 is to provide an example of a very highly dependent joint probability space. In fact, Experiment 5 is designed to be very stochastically dependent. It is so dependent that it is not too far away from being deterministic – but not quite.

Experiment 5 is the same as Experiment 4, except that the embedded hidden magnets have been placed into an extreme position inside of their respective dice, and have been given a high degree of magnification, so as to assure a high probability of certain pairs of the two dice landing up. Also, the orientation of the magnets may also have been altered to affect which die faces of the mammal die attract which die faces of the tree die. The effect of all of this is to alter the probabilities of the joint distribution in a manner that makes a few combinations highly probable, while others are rather low.

However, the positioning of the centers of gravity of those two magnets within the dice is essentially the same as it was in Experiment 4. This results in the individual composite probability distributions remaining close to what they are in Experiment 4.

The result of all of this dice loading, magnification and magnet reorientation is the composite probability distributions, joint probability distribution and conditional probability distribution presented below.

One of the ideas of presenting Experiments 3, 4 and 5 is to show a sequence of probability spaces that are gradually increasing in their degrees of stochastic dependence. Experiment 5 is exhibits the highest degree of dependency of the three.

In the next subsection, we shall provide a graphical comparison of these experiments to encourage an intuitive acceptance that these three are, indeed, increasing in their degrees for stochastic dependency.

So, let us now present Experiment 5.

Lets first look at the Mammal Die component distribution:

The Mammal Die component distribution X

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total
Mammals	0.1100	0.1600	0.2400	0.1700	0.1600	0.1600	1.0000

Next we view the Tree Die component distributions:

The Tree Die component distribution Y

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total
Trees	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1.0000

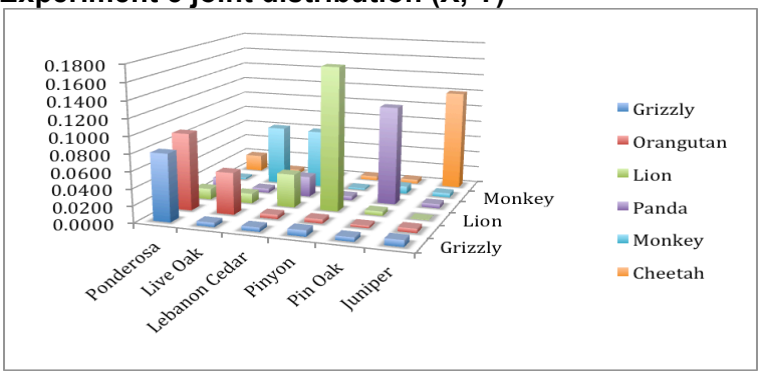
Then, we have the joint distribution for these two loaded and magnetized dice, which of course is the result of the magnifications and orientations of the embedded weights in the two dice:

The (Mammal, Tree) dice joint distribution (X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0800	0.0050	0.0050	0.0073	0.0050	0.0077	0.1100
Orangutan	0.0920	0.0500	0.0050	0.0050	0.0030	0.0050	0.1600
Lion	0.0130	0.0118	0.0400	0.1697	0.0050	0.0005	0.2400
Panda	0.0130	0.0050	0.0246	0.0050	0.1174	0.0050	0.1700
Monkey	0.0020	0.0716	0.0704	0.0020	0.0090	0.0050	0.1600
Cheetah	0.0200	0.0050	0.0050	0.0050	0.0050	0.1200	0.1600
Total Tree	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1.0000

This joint distribution has this graph:

Experiment 5 joint distribution (X, Y)



The reader should notice that the probabilities near the diagonal of this joint distribution are significantly higher than most of the other probabilities, with a few exception. Notice that there is a more conspicuous pattern of the taller towers clustering near the diagonal than in the other two experiments 2 and 4 that have the same two component chance variables X and Y. This pattern of relationships should show up as a higher degree of stochastic dependence (as measured by *mutual information* value) than either Experiment 3 or Experiment 4. So, we shall be looking to compare and contrast these mutual information values.

From this joint distribution, we can calculate the conditional probability distribution for these two chance variables in the usual way by dividing each cell (joint probability) of the joint distribution by its row sum. This action has the effect of "normalizing" all of the "mammal rows" – or the individual probability distributions for the Tree die having been given the outcome of the mammal die for the joint event.

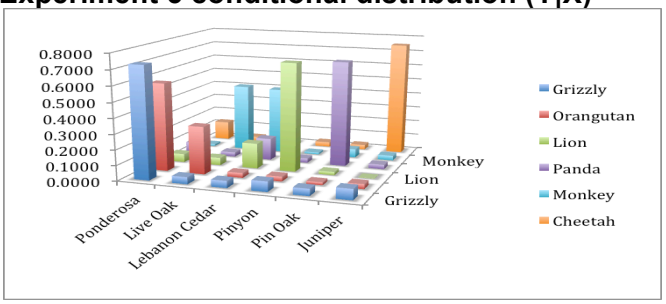
The result is the following:

Experiment 5 conditional distribution (Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.7273	0.0455	0.0455	0.0664	0.0455	0.0700	**1.0000**
Orangutan	0.5750	0.3125	0.0313	0.0313	0.0187	0.0313	**1.0000**
Lion	0.0542	0.0492	0.1667	0.7071	0.0208	0.0021	**1.0000**
Panda	0.0765	0.0294	0.1447	0.0294	0.6906	0.0294	**1.0000**
Monkey	0.0125	0.4475	0.4400	0.0125	0.0563	0.0313	**1.0000**
Cheetah	0.1250	0.0313	0.0313	0.0313	0.0313	0.7500	**1.0000**

This conditional distribution has this graph:

Experiment 5 conditional distribution (Y|X)



Notice also that the probabilities near the diagonal are mostly much larger than the other probabilities in the conditional distribution. There are a few exceptions to this: (Orangutan, Ponderosa), (Monkey, Live Oak) and (Monkey, Lebanon Cedar). But overall, there is a general trend of the higher probabilities clustering around the diagonal of the table. This graphs reveals this fact.

Recall also that the conditional distribution is a "normalized" version of the joint distribution, where all of the rows have been given "equal weight". In a sense, the conditional distribution allows the comparison of "apples to apples" when comparing the rows; whereas the joint distribution does not.

As with the previous experiments, the way we shall currently approach ascertaining the degree of dependency of these experiments is to

1. Develop their conditional distribution from their joint distribution. Dividing each entry of the joint distribution by its row sum does this. We already accomplished this above.
2. Inspect the resulting conditional distribution to see if all the rows are equal to each other. If all the rows are the same, then the two component chance variables are stochastically independent. Otherwise, they are stochastically dependent.

So, what is left for us to do is #2 above. From inspecting the conditional distribution above, we can see that not all the rows are the same as each other. Consequently, we can surmise that the two chance variables involved are stochastically dependent.

Furthermore, we can also see by inspection that several of the cells of the conditional probability table are disproportionately larger than the other values in their same column. Take for example the probability value of 0.7273 in the cell at the intersection of Grizzly and Ponderosa. This means that whenever Grizzly lands up on the mammal die, that there is a 72% probability that Ponderosa will have landed up on the mammal die! (There must be a pretty strong magnetic field between the two magnets to make this likelihood be so large.)

In the next major section of Part II, we shall present the formal mathematical equipment to ascertain the degree of stochastic dependency of any joint probability space. However, for the time being we shall be content to develop an intuitive understanding of the concept – that is by inspection.

Before that, though, we shall continue a little further with our intuitive approach to analyzing the degrees of dependency across our five example dice experiments. In the next section below, we shall use the graphs that we have developed of the joint probability distributions and the conditional distribution to surmise a general sense and intuition of the degrees of dependencies of these five example experiments.

Intuitive Look at the Relative Degrees of Dependency of the Five Experiments

We shall now summarize the developments of the five example experiments by comparing the graphs that we have developed for all five of them. We shall present the graphs of these experiments in their order of stochastic dependency. You will recall that the two chance variables of Experiments 1 and 2 had zero degrees of dependency, because they were both stochastically independent. Then, Experiments 3, 4 and 5 exhibited increasing degrees of dependency.

In this subsection, we are going to compare these graphs and see if we can discern any “visual trends” as we increase the degree of stochastic dependency from experiment to experiment. We are looking for a key to help us figure out how we could possibly develop a function that could calculate the measure of dependency between these two chance variables. We would expect that such a function would utilize the joint distribution and the conditional distribution of a two chance variables - and perhaps even their component distributions.

In order to gain insight into how such a function might be defined, this section will explore this question in an intuitive manner by concentrating on the graphs of the conditional distributions of each of these five experiments. We use the conditional distributions because that’s what we ultimately used in each section in order to assess the stochastic dependency of each of the example experiments.

To perform this analysis, we shall organize all of these graphs into a single table. The table will have five rows – one row for each of the example experiments. Each row will contain two items: the conditional distribution graph and some analysis commentary. This commentary, over all, will attempt to ferret out the salient trends that are significant concerning the increase in stochastic dependency across the 5 experiments.

Finally, after the table, we shall draw conclusions from these comments regarding our intuitive circumspection of stochastic dependency.

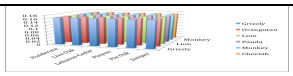
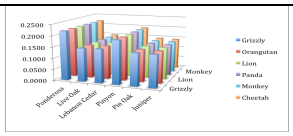
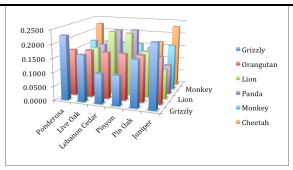
The vertical scales of the following graphs have been sized so that they are consistent with each other – so that the heights of all of the bars in all of the graphs measure approximately the same probability units.

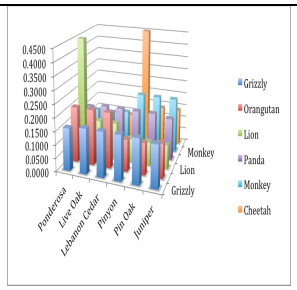
Also, keep in mind that:

1. The value of each bar in both graph types is the probability of a joint sample point.
2. The heights of all of the bars in each single joint distribution graph sum to 1.
3. The sum of all of the bars in each row of each conditional distribution graph sums to 1.

Finally a comment about all the conditional distribution: The rows of a conditional distribution are the result of “normalizing” the rows of its corresponding joint distribution so that each row has the same “weight” as each other.

Stochastic Dependency Increase across the Five Example Experiments

Conditional Distribution (Y|X)	Stochastic Dependency Comment
	Experiment 1. Recall that X and Y are independent of each other in Experiment 1. This can be visually verified by looking at the conditional distribution. The probabilities for any tree face are the same no matter which row (mammal) face it is in. In addition, since this distribution is independent, it has zero amount of dependency. It therefore becomes the basis for comparison for all other joint distributions that have the same two component distributions as this independent one.
	Experiment 2. Recall that X and Y are independent of each other in Experiment 2. This can be visually verified by looking at the conditional distribution. The probabilities for any tree face are the same no matter which row (mammal) face it is in. Another way to say this is: In the conditional distribution, consider any tree face ("column" from front to back). Do all of the bars in that column have the same probability as each other? If the answer to this question is "yes" for every column (tree face) in the graph, then the two chance variables are independent. One can see that this is true for Experiment 2. (It was also true of Experiment 1.) In addition, since this distribution is independent, it has zero amount of dependency. It therefore becomes the basis for comparison for all other joint distributions that have the same two component distributions as this independent one.
	Experiment 3. Consider the conditional distribution. Consider any of its tree faces. Does it have the same probability value no matter which mammal face row it is in? In other words, is it true for every "column" (front to back of the graph), that all of the probability values in that column are the same as each other? The answer is "no". Therefore, the two chance variables in Experiment 3 are *dependent*. In fact, for this distribution, none of the columns has the property that all of its probability values are the same as each other. Since this Experiment is dependent, and because it has the same two component distributions as Experiment 1, then we can ask "How different is it from Experiment 1's conditional distribution?" The answer to this question can give us a sense of its degree of dependence.

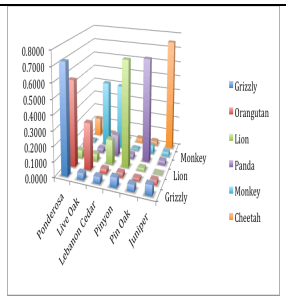


Experiment 4. Consider the conditional distribution. Consider any of its tree faces. Does it have the same probability value no matter which mammal face row it is in? In other words, is it true for every "column" (front to back of the graph), that all of the probability values in that column are the same as each other?

The answer is "no". Therefore, the two chance variables in Experiment 3 are dependent.

In fact, for this distribution, none of the columns has the property that all of its probability values are the same as each other. Moreover, there are a couple of "columns" (Ponderosa and Pinyon) in which there is one value that is conspicuously different from all of the other values in its column.

Since this Experiment is dependent, and because it has the same two component distributions as Experiment 2, then we can ask "How different is it from Experiment 2's conditional distribution?" The answer to this question can give us a sense of its degree of dependence.



Experiment 5. Consider the conditional distribution. Consider any of its tree faces. Does it have the same probability value no matter which mammal face row it is in? In other words, is it true for every "column" (front to back of the graph), that all of the probability values in that column are the same as each other?

The answer is a conspicuous "no". Therefore, the two chance variables in Experiment 3 are dependent.

In fact, for this distribution, none of the columns has the property that all of its probability values are the same as each other. Moreover, for all six "columns", there are one or two values that are conspicuously taller then all of the other values in its column.

Since this Experiment is stochastically dependent, and because it has the same two component distributions as Experiment 2, then we can ask "How different is it from Experiment 2's conditional distribution?" The answer to this question can give us a sense of its degree of dependence.

Intuitively, all six of the columns of this graph are conspicuously different than Experiment 2's columns. However, only two of Experiment 4's columns are conspicuously different from Example 2's. While all of Experiment 3's column's are different from Experiment 1' columns, they are not conspicuously so. This line of reasoning leads one to expect that Experiment 3 is a "little bit dependent", that Experiment 4 is "dependent"; and that Experiment 5 is "very dependent".

Summary Observations Concerning the Five Example Experiments

The Meaning of Stochastic Independence and Dependence

Let's summarize what we have been saying about stochastic independence and stochastic dependence.

To say that two chance variables are statistically independent means that, whenever the two are occur together in a joint situation, knowing the outcome of the first provides no useful information to help you surmise the outcome of the second.

In terms of probabilities, stochastic independence means that the probabilities of the possible outcomes of the second chance variable are the same regardless of the outcome of the first chance variable.

On the other hand, to say that two chance variables are statistically dependent means that, whenever the two are occur together in a joint situation, knowing the outcome of the first *does* provides useful information to help you surmise the outcome of the second.

In terms of probabilities, stochastic dependence means that the probabilities of the possible outcomes of the second chance variable may be different for different outcomes of the first chance variable. In fact, there will be at least one outcome of the first chance variable that will have distinctive outcomes for the second chance variable.

Some Intuitive Characterizations of Stochastic Independence and Dependence

We can draws some general characterizations from the previous section in which we looked at the graphs of the conditional distributions of all five experiments and made some observations about them based upon a visual inspection of these graphs.

We saw a number of visual trends, and now we shall attempt to draw some general intuitive conclusions from them.

First, we saw that each row of a conditional distribution is a probability distribution in its own right; and that the conditional distribution is the result of "normalizing" all of the rows of it associated joint distribution so that they all have the same "weight" as each other.

Second, we saw that if the two chance variables of a joint probability space are stochastically independent, then its conditional distribution has the property that all of its rows are the same "shape" as each other. And, if the chance variables are stochastically dependent, then at least one of the rows of its conditional distribution has a "shape" that is different from the others.

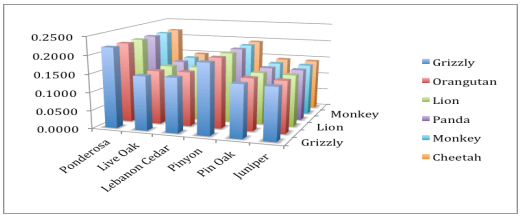
Third, we saw that some dependent joint probability spaces are "more stochastically dependent" than others. In other words, there is a meaningful idea of *degree of dependence*. In fact - given two specific component distributions – there appears to be exactly one stochastically independent joint probability space for those two component distributions (chance variables). And, from what we said above, all of the rows of its conditional distributions are the same "shape" as each other, because of the fact that it is stochastically independent.

On the other hand, there appear to be several possible dependent joint probability spaces for these two component distributions – all exhibiting varying degrees of dependency. And, the rows of their conditional probability distributions are internally different from each other. Apparently, the more different are these rows from each other, the more dependency there may be between the two chance variables.

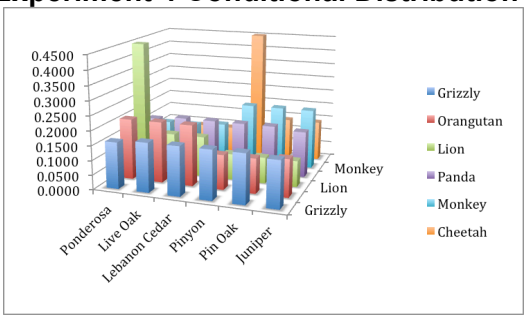
Fourth, we also saw that we can gain an intuitive feeling for "just how stochastically dependent a particular joint probability space is". We can do this by visually comparing the graph of its conditional distribution to that of the stochastically independent distribution based on the same two component distributions.

For example, Experiments 2 and 4 have the same two component distributions, but Experiment 2 is stochastically independent; while Experiment 4 is stochastically dependent. So, we can get some idea as to the degree of stochastic dependence of Experiment 4 by visually comparing its conditional distribution's graph to that of Experiment 2. These are shown here:

Experiment 2 Conditional Distribution



Experiment 4 Conditional Distribution



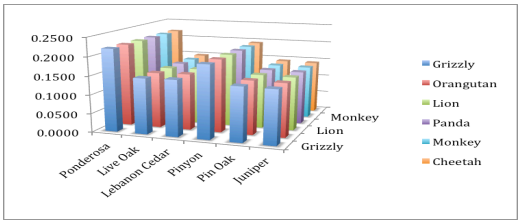
By inspection, we can see that there is “some amount of difference” between these two graphs. Of course, Experiment 2 should have zero amount of stochastic dependence because of the fact that it is stochastically independent. Therefore, any difference that the graph of Experiment 4 exhibits from the graph of Experiment 2 would represent the degree of stochastic dependence of Experiment 4.

If we could find some consitent way of measuring that “amount of difference” of Experiment 4 from Experimenrt 2, then we would have a measure of stochastic dependence of Experiment 4. This is precisely the purpose of the mathematics of information theory, as we shall see in the next major section of Part II. That mathematics, as we shall see, successfully captures numerically this difference between these two graphs in a measuring function – a measuring function named mutual information.

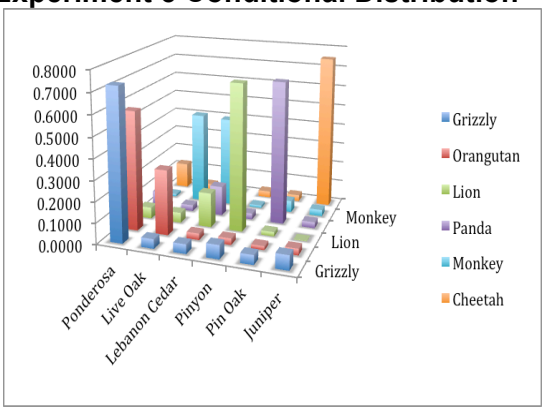
Before we go on to the formalities, though, lets look at Experiment 5 and see how it compares the consideration we just made for Experiment 4.

Recall that our analysis above of Experiment 4’s degree of dependence consisted of comparing its conditional probability graph against that of stochasticaslly independent Experiment 2. So, lets compare the conditional distribution of Experiment 5 against the conditional distribution of Experiment 2 also. This is shown here:

Experiment 2 Conditional Distribution



Experiment 5 Conditional Distribution



Comparing these two graphs with the above two graphs, I shall make the intuitive assertion that Experiment 5 is more different from Experiment 2 than is Experiment 4.

This means that – as best as I can tell by inspecting the graphs - the degree of stochastic dependence of Experiment 5 is greater than the degree of stochastic dependence of Experiment 4.

Lets summarize what we just intuitively concluded regarding the degree of stochastic dependency of Experiment 2, 4 and 5.

First, since Experiments 2, 4 and 5 have the same underlying two component probability distributions X and Y; it is easy to compare their relative degrees of stochastic dependence by comparing their conditional distribution graphs.

Second, the degree of dependency of Experiment 2 must be 0 (zero), since, being stochastically independent, it should have no degree of dependence. This can be seen in the graph of Experiment 2 above because no matter which mammal row you select, the bars are all the same height as they would have been had you selected any other mammal row of bars. In other words, it does not matter which mammal row was actually realized; the probability distribution would be the same. In other words, the probability of the Tree die faces is completely independent of the Mammal die faces.

Third, since Experiment 2 has zero degrees of stochastic dependency, then any "difference" between the graph of the conditional distributions of Experiment 4 and Experiment 2 should show the degree of dependence of Experiment 4. By the same token, any "difference" between the graphs of the conditional distribution of Experiment 5 and Experiment 2 should show the degree of dependence of Experiment 5.

Moreover, the "difference" between those two pairs of graphs is visually greater for Experiment 5 than it is for Experiment 4.

Therefore, whenever we develop the formal information theory mathematics to measure the degree of dependence of two chance variables, it should produce a larger measure for Experiment 5 than it does for Experiment 4. Moreover, such a measure should produce a value of zero (0) for Experiment 2.

And there is one more thing to consider. In what we just discussed, we had to make sure that we were comparing experiments (joint probability spaces) all of which had the same two component probability distributions X and Y. However, the measure of stochastic dependence that we develop in the formal section below should not have that restriction. Such a measure should be able to compare the degrees of dependence between two joint probability spaces even though their underlying component chance variables are not the same.

The Need for A Formal Apparatus to Measure the Degree of Stochastic Dependence

At this point in this primer, we have presented a strong intuitive presentation that has hopefully given the reader a certain level of comfort, readiness and curiosity about stochastic dependence, portent and meaningfulness between two chance variables.

Armed with this, the reader is hopefully ready for a critical treatment of the mathematical constructs that formally describe the mechanisms at work for these issues. These constructs are the essential teaching of the second part of information theory. We shall take these up in the next section.

Information Theory's Mathematics of Stochastic Dependence

Now that you have thoroughly considered the issues and problems raised by the above five example experiments, you are in a position to understand how the mathematics of information theory addresses these issues.

In this section, we shall present the mathematical constructs of information theory that represent these ideas.

The Goal

In the beginning of Part II, we expressed an interest in the meaningfulness of one phenomenon to another. We said that the degree of such meaningfulness is a function of the extent to which one of those phenomena portends something about the other. We also concluded that if any such portent was at work, then one of the phenomena enjoyed a dependence on the other.

In this way, we took the three ideas of *portent*, *dependence* and *meaningfulness* between two phenomena as logically equivalent. Thus, we have been using these three terms interchangeably.

And we want to be able to detect not only when one phenomenon depends upon another, but also to what degree or amount. We want to measure the degree to which one phenomenon depends upon another.

But we also most often are dealing with situations that involve some degree of randomness – *chance variation*. We are dealing with sample spaces and phenomena that behave slightly differently with every instance. And we are dealing with many samples, many instances. So, we have to contend with statistical variation and uncertainty in the in these observations and measurements.

So our goal in Part II is to find a function that *measures the degree of dependency between two chance variables*. And, from our experience of thoroughly delving into the five joint experiments above, we know that such a measuring function will have to take into consideration the joint probability distribution of the two chance variables as well as the their two component probability distributions.

Obviously, in this section on the formal development of the mathematical apparatus of the interdependency of chance variables, we must formally develop the notions of component probability distribution as well as joint probability distribution. Of course, a component distribution is simply an ordinary probability distribution of the kind we developed in Part I. So there is no need to develop that again here.

However, the notion of joint distribution is new in Part II. Therefore, our formal development of Part II will begin with the joint sample space, after which we shall immediately define joint probability distribution.

Before arriving at our final destination for Part II – the measure of stochastic dependence between two chance variables that we shall name *mutual information* – there are a number of preliminary constructs that must be gradually built up in order to finally reach the ability to define *mutual information* in terms of them. Beginning with joint sample space, these constructs are:

- Joints sample space
- Joint probability distribution
- Joint entropy
- Conditional probability

- Conditional distribution
- Conditional entropy
- Stochastic independence
- Stochastic dependence
- Relative entropy
- Mutual Information

In the remainder of this section we formally present these ideas and constructs, and provide further insight to connect the intuition that was developed with the example experiments earlier to these formal ideas.

Joint Sample Spaces

Given two chance variables, X and Y, we can form a third chance variable whose sample points are all possible pairs whose first entry is a sample point from X and whose second entry is a sample point from Y. This is referred to as the joint chance variable between X and Y, and it is symbolized by "(X, Y)".

If n(X) is the number of sample points in X and n(Y) is the number of sample points in Y, then the number of sample points in (X, Y) is n(X)*n(Y).

In relationship to the joint chance variable (X, Y), the chance variables X and Y from which it is derived are referred to as the *component chance variables* of (X, Y).

For example, the same joint sample space was used for all five experiments using the two dice above. This joint sample space consists of 36 pairs of dice, each of whose first die face is from the mammal die, and each of whose second die face is from the tree die.

Joint Distributions

As chance variables, not only do X and Y have their own probability distributions, but so also does the joint chance variable (X, Y). This probability distribution is also called the *joint distribution* of X and Y, and is also symbolized by "p(X, Y)".

Of course, the two component chance variables X and Y may each have many different probability distributions – usually at different times. For example, Perhaps chance variable X represents a coin that is tossed. If the coin is "fair", then each of its faces has a probability of ½. However, after tossing the coin for awhile, a player who want to cheat could do something to the coin to alter its balance, and therefore to change its probability distribution so that the probabilities for heads and tails are no longer the same. Thus, the altered coin would have a different probability distribution after the cheater had altered it than it did before.

In this case, it we need some symbol other than "p(X)" to distinguish between the two distributions. In this primer, we shall use subscript for this: e.g. "$p_1(X)$" for the first distribution, and "$p_2(X)$" for the new distribution after the coin had been altered. These subscripts can be any symbol we wish.

In addition, as time goes by, joint distributions can also change. Therefore, any two chance variables X and Y can have multiple joint distributions p(X,Y) as well. These, too, can be distinguished via subscripts: e.g. "$p_1(X,Y)$", "$p_2(X,Y)$", or "$p_3(X,Y)$".

However, if there is no need to distinguish between multiple distributions on either simple or joint distributions, then the subscripts can be omitted.

Multiple Joint Distributions on the Same Two Component Distributions

For a given component sample spaces X and Y with their respective component distributions p(X) and p(Y), there can be many possible joint distributions p(X, Y). In fact, for any given component probability distributions p(X) and p(Y), there are an infinite number of joint probability distribution $p_i(X, Y)$.

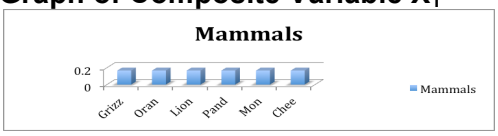
And, as we shall see in the next section, each of these many joint probability distributions p(X, Y) on the same two component distributions p(X) and p(Y) has its own distinct degree of stochastic independence. In other words, for any given pair of component probability distributions, there are different joint distributions, each of which exhibit an range of *degrees of dependency* – ranging from no dependency at all (an independent joint distributions) all the way to some maximum degree of dependency.

Therefore, there is a very good reason to distinguish between all of these joint distributions on the same two component distributions with their own distinct names or symbols.

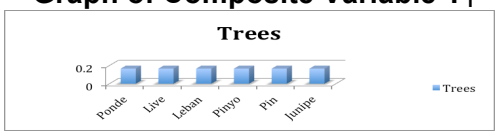
For example, our example Experiments 1 and 3 share the same two component distributions p(X) and p(Y) – the mammal and tree dice with the uniform probability distributions. Lets rename them "$p(X_1)$" and $p("Y_1)$", because we want to distinguish them from the other two component distributions "p(X)" and "p(Y)" – the mammal and tree dice with the non-uniform distributions that are used in Experiments 2, 4 and 5.

These two composite distribution for Experiments 1 and 3 are depicted by the following two graphs.

Graph of Composite Variable X_1



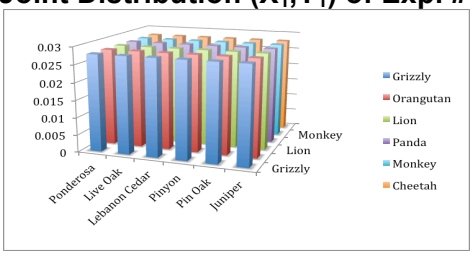
Graph of Composite Variable Y_1



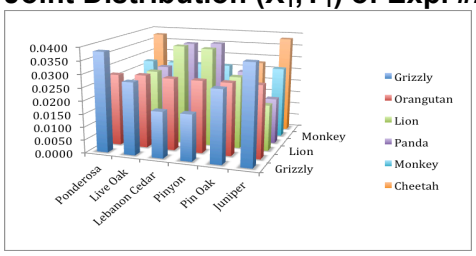
But we have already defined two different joint distributions on these two specific component distributions. One of them is the joint distribution that we used in Experiment 1 and the other is the joint distribution that we used in Experiment 3.

Lets portray the graphs of these two joint distributions again here:

Joint Distribution (X_1,Y_1) of Exp. #1



Joint Distribution (X_1,Y_1) of Exp. #2



Clearly we have here two distinct joint probability distributions on the same two composite probability distributions X_1 and Y_1. We can do this simply by assigning whatever joint probabilities we desire to each joint distribution – as long as they are non-negative and sum to 1.

But, so far, we are using the same symbol "" for both – "(X_1,Y_1)".

In fact, we could create an infinite number of distinct probability distributions on these same two composite probability distributions X_1 and Y_1. And, from the symbolism we have defined so far, they would all be represented by the same symbol: "(X_1,Y_1)".

Clearly, we need to add something to this symbolism to distinguish all of these many possible joint probability distributions on the same two composite probability distributions X_1 and Y_1.

What we shall do to solve this problem is to add a subscript to the end of the last parenthesis. And, we shall allow any subscript that is desired – anything that will help to distinguish one joint distribution on a pair of component distributions from another joint distribution on that same pair of component distributions.

For example, we may choose to name the Joint Distribution (X_1,Y_1) of Experiment #1 above with the symbol "$(X_1,Y_1)_{Exp1}$"; and to name the Joint Distribution (X_1,Y_1) of Experiment #2 above with the symbol "$(X_1,Y_1)_{Exp2}$".

We are going to permit great liberty with these symbols and their subscripts. For example, one may desire to attempt to inject more meaningfulness into the symbols chosen. For example, instead of "X" and "Y", one may choose to use the symbols "M" and "T" for "mammal" and "tree" dice. And, one may choose to use the subscript "u" rather than "1" to indicate that the component distributions that we are dealing with are uniform. As well, one may choose to use the symbols "1" and "3" instead of "Exp1" and "Exp3" to indicate "Experiment 1" and "Experiment 3". With such choices, then, we would have the following two symbols to represent these the joint distributions for Experiment 1 and Experiment 3 above: "$(M_u,T_u)_1$" and "$(M_u,T_u)_2$".

Of course, these same consideration also apply to Experiment 2, 4 and 5 – all three of which use the same two non-uniform component distributions – call them $M_{\sim u}$ and $T_{\sim u}$ instead of "X" and "Y". Here the "M" means "Mammal die" and "T" means "Tree die", and the symbol "$_{\sim u}$" means "non-uniform" distributions. Accordingly, then the following nomenclature would symbolize the three joint distributions for Experiments 2, 4 and 5 respectively: "$(M_{\sim u},T_{\sim u})_2$", "$(M_{\sim u},T_{\sim u})_3$", and "$(M_{\sim u},T_{\sim u})_3$".

The point of all this is to realize these three facts regarding joint probability distributions on the same two component distributions:

1. Any two component distributions can have an infinite number of distinct joint probability distributions created for them.
2. Each of these distinct joint distributions will exhibit its own degree of stochastic dependence.
3. Each of these distinct joint probability distributions needs to be distinguished by having its own distinct name or symbol – the concept that we have introduced in the present subsection.

We want to point out that only occasionally is it necessary to go to the lengths of using this complicated symbolism to distinguish between distinct joint distributions of the same two component distributions. Whenever the context is clear, and there is no mistaking which joint distribution and which component distributions are being referred to, then the simple (X,Y) notation is sufficient. However, whenever we are comparing and contrasting multiple distinct joint distributions on the same two component distribution, or even on different pairs of component distributions, then the more elaborate nomenclature may be needed.

We should also point out that no other known text has seen the need to make all of these distinctions. Thus the complex nomenclature described above is peculiar to this primer – as far as this author knows.

In any event, the failure to make these distinctions may lead the reader to, consciously or unconsciously, assume that the choice of the two component distributions determines whether the joint distribution is stochastically dependent or independent. In this case, the student fails to realize that the same two component distributions can have an infinite number of joint distributions – each of which determine its own degree of stochastic dependence. This is an important point. And these distinctions have been belabored in this section in order to drive home this point.

Degrees of Stochastic Dependence and Joint Distributions

It turns out that the assignment of the joint probabilities of a joint distribution on two component distributions determines the degree of dependence (or portent or meaningfulness) between the components chance variables of the joint distribution. And, as we have just pointed out in the previous subsection, there are many ways to assign joint probabilities to these joint sample points. This results in a large (infinite) number of different possible joint distributions for a given joint sample space.

And, some of these possible joint distributions over this joint sample space will exhibit large degrees of dependence; others will exhibit small degrees of dependence; while still will exhibit intermediate degrees of dependence; and so on.

However, as we shall see, *exactly one* of these joint distributions, on a particular pair of component distributions, will have the property that it is stochastically independent. All of the remainder of the infinite number of joint probability distributions on that same pair of component distributions will be stochastically dependent – and will exhibit varying degrees of stochastic dependence.

In fact, select any two component distributions. This pair will have an infinite number of joint probability distributions that can be defined on it. And, exactly one of those joint distributions on this pair will be stochastically independent, while all the rest will be stochastically dependent. And each of the rest will exhibit varying degrees of stochastic dependency. The ability to measure this degree of stochastic dependence is the final goal of Part II of this primer.

A principle goal of this section is to ascertain the criteria for this difference in degrees of dependence, and to measure the degree of dependence between the two component chance variables based upon this assignment of joint probabilities to the various joint sample points (also called events).

However, there are some constraints that any joint distribution of p(X, Y) must adhere to. We shall refer to the joint sample points of (X, Y) by "(x, y)" where x is a sample point of X and y is a sample point of Y.

The first constraint is that all of the joint probabilities in the joint distribution must sum to 1. Of course, this is true of any probability distribution, and applies also to joint distributions. In our joint probability table in the above examples, this fact translates into saying that all of the joint probabilities in the body of the table – in all of the rows and columns – must sum to one.

The second constraint is that the component probabilities of chance variable X must be the sums of respective rows of the joint probability table (matrix) of p(X, Y). In our joint probability tables in the above examples, this fact translate to saying that all of the joint probabilities in the same row must sum to the value in the "Total Mammal" cell of that row.

The third constraint is that the component probabilities of chance variable Y must be the sums of respective columns of the joint probability table (matrix) of p(X, Y). In our joint probability tables in the above examples, this fact translate to saying that all of the

joint probabilities in the same column must sum to the value in the “Total Tree” cell of that column.

Example of p(X, Y)

In the next subsection, we shall present a formal definition of the joint distribution (X, Y) of chance variables X and Y – as well as some formal terminology and symbolism for some of its internal entries. In order to better understand that symbolism, as well s the interrelationships among some of its internal entries, it will be useful to have an example joint distribution nearby. Therefore, we shall repeat the joint distribution of one of our example experiments here.

Experiment 4 joint distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Formal Definition of p(X, Y)

The table below, p(X, Y), depicts the joint probabilities "p(x,y)", where the y's are sample points of the chance variable X and the y's are sample points of chance variable. X has n sample points and Y has m sample points.

The body of this table contains the actual probabilities of the joint distribution p(X, Y). These joint probabilities are of labeled with the format "$p(x_j,y_k)$". Mathematically, these

The component distributions X and Y are also depicted in this table

Joint Distribution p(X, Y)

X \ Y	y_1	y_2	y_3	y_4	y_5	y_m	P(X)
x_1	$p(x_1,y_1)$	$p(x_1,y_2)$	$p(x_1,y_3)$	$p(x_1,y_4)$	…	$p(x_1,y_m)$	$p(x_1)$
x_2	$p(x_2,y_1)$	$p(x_2,y_2)$	$p(x_2,y_3)$	$p(x_2,y_4)$	…	$p(x_2,y_m)$	$p(x_2)$
x_3	$p(x_3,y_1)$	$p(x_3,y_2)$	$p(x_3,y_3)$	$p(x_3,y_4)$	…	$p(x_3,y_m)$	$p(x_3)$
x_4	$p(x_4,y_1)$	$p(x_4,y_2)$	$p(x_4,y_3)$	$p(x_4,y_4)$	…	$p(x_4,y_m)$	$p(x_4)$
…	…	…	…	…	…	…	…
x_n	$p(x_n,y_1)$	$p(x_n,y_2)$	$p(x_n,y_3)$	$p(x_n,y_4)$	…	$p(x_n,y_m)$	$p(x_n)$
P(Y)	$p(y_1)$	$p(y_1)$	$p(y_1)$	$p(y_1)$	…	$p(y_1)$	1

The body of this table (not in bold) contains the actual probabilities of the joint distribution (X, Y). These joint probabilities are of labeled with the format "$p(x_j,y_k)$". Mathematically, these entries are collectively referred to as the joint probability matrix.

The rows and columns on the four sides of this matrix are headers – but with mathematical significance.

The first component sample space, labeled "X", whose sample points are (x_1, x_2,…, x_n), is listed along the first column of this table. It's entries, of the form "$\mathbf{x_i}$", are the sample points of X. Their probabilities, the component probability distribution p(X) is listed as the last column of the table. It's entries, of the form "p($\mathbf{x_i}$)", are the probabilities of the respective sample points of X.

Notice that the entries in the p(X) column have two meanings. Not only are they the probabilities of the sample points of X, but they are also the row sums of all of the joint probabilities p(x,y) in their row.

And, the component sample space Y is similarly represented. However, its sample points (y_1, y_2,…, y_m) are listed in the top row. Their component probability distribution p(Y) is listed as the bottom row of the table. It's entries, of the form "p($\mathbf{y_j}$)", are the probabilities of the respective sample points of Y.

Notice that the entries in the p(Y) row have two meanings. Not only are they the probabilities of the sample points of Y, but they are also the column sums of all of the joint probabilities p(x,y) in their column.

Armed with this understanding of the p(X, Y) table, its table entries, their symbols and their interrelationships, we are now ready to provide a more formal definition of the joint probability distribution p(X, Y).

Definition: Joint probability distribution p(X, Y):

Let p(X, Y) be the joint distribution of chance variables X and Y. Then define the distribution p(X, Y), called the joint distribution of X and Y, as follows,

$$p(X, Y) = \{ (p(x_i, y_j)) = p(x_i \wedge y_j) \text{ for all } x_i \in X, y_j \in Y \}.$$

This formality says that the joint distribution p(X, Y) consists of the set of all probabilities that are represented by symbols of the form "$p(x_i, y_j)$". However, the definition of each of these symbols is consists of the calculation "$x_i \wedge y_j$", or "x_i and y_j".

Summary

The distribution p(Y|X) is the conditional probability distribution for chance variable Y, given chance variable X.

Its calculation can be described in three ways. The first way says that it is the matrix that is formed by placing all of the consists of all of the conditional distributions p(Y|X=x) into a single matrix – where each possible p(Y|X=x) occupies its own row.

The second way of describing p(Y|X) describes the most direct way of calculating it from it joint distribution. This way says: p(Y|X) has the same number of rows and columns as p(X, Y) - except for the last row p(Y). Each of the entries in the p(Y|X) matrix is calculated by dividing its corresponding entry in the p(X, Y) table by its row sum in the p(X, Y) table.

A third way of describing p(Y|X) is that each of its entries p(y|x) is calculated by:

p(x^y)/p(x).

The Meaningfulness inherent in a Joint Distribution

The question then arises, "How can we look at the joint probabilities of a joint distribution p(X, Y) of chance variables X and Y and ascertain the degree to which X is meaningful to Y?"

A triumph of information theory is that it has worked out how to answer that question.

The first step toward this is to equate the notion of "meaningfulness between two chance variables" as "the degree of stochastic dependence between those two variables". We have successfully linked these two ideas earlier in Part II. So, the issue of how much meaningfulness is there between two chance variables is equivalent to the issue of how much dependence is there between these two chance variables.

We saw when we delved into this issue intuitively with the five experiments above that the first step was to derive a conditional distribution from the joint distribution. But after that, we had to proceed intuitively.

In this section, however, we shall develop the mathematics to proceed formally from the conditional distribution, and to develop a formal function that can produce a number that consistently measures the degree of stochastic dependence between two chance variables.

In order to develop that measure, we must first define a number of other concepts and mathematical constructs that we shall use to define our final formulation of the measure. The first of these is joint entropy.

Joint entropy

Being a probability distribution in its own right, the joint distribution, of course, has entropy. Joint entropy is simply the entropy of a joint distribution - calculated as you would the entropy of any other distribution.

Joint Entropy Definition

Joint entropy is simply the entropy of a joint distribution. To calculate it, simply treat the joint distribution as an ordinary distribution, and calculate its entropy. This means that you can ignore the fact that the joint distribution is normally represented as a 2-dimensional matrix with rows and columns. For example, all five of our dice experiments have 6 rows and 6 columns. Rather you can treat this joint distribution as though it were simply a one-dimensional distribution with 36 sample points. Its joint entropy, then, is calculated against these 36 sample points as you would any ordinary distribution that has 36 sample points.

Lets state this more formally. The sample points of a joint distribution take the form (x, y), and its probabilities are symbolized by "p(x, y)". Therefore, we can substitute "p(x, y)" in place of "p(x)" in our definition of entropy in order to obtain the function of the entropy of the joint distribution of (X, Y), or H(X, Y).

Recall that our definition of entropy, from Part I, is

$$H(X) = \sum_{i \in S} p(x_i) * \log(1/p(x_i))$$

Therefore, to obtain the joint entropy, by substituting "p(x, y)" in place of "p(x)", we get:

$$H(X, Y) = \sum_{i \in S} p(x_i, y_i) * \log(1/p(x_i, y_i))$$

Or, equivalently,

$$H(X, Y) = -\sum_{i \in S} p(x_i, y_i) * \log(p(x_i, y_i))$$

The astute reader will recognize by inspecting the definition above that joint entropy is another one of those *entropic measures* that were discussed at the end of Part I.

Joint Entropy Example

For example, lets consider the joint distribution of Experiment 4. We shall repeat it joint distribution table here.

Mammal and Tree dice joint distribution p(X, Y) – Experiment 4

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

So, to calculate the joint entropy, we can interpret the above definition as requiring the following 3 steps. (All logarithms use a base of 2.)

Step 1: Each entry in matrix p(X, Y) is of the form p(x, y). For each such entry, calculate log(1/(x, y)). For example, the entry at Grizzly and Ponderosa has the value of 0.0176. Thus, calculate log (1/.0176). This, rounded to 4 decimal places is equal to 5.8283.

Step 2: Now, multiply this calculation by the probability of the joint event for this (x, y) taken from the joint distribution p(X, Y). For the joint sample point (Grizzly, Ponderosa), this joint probability is .0176. The product of .0176 and 5.8283 is .1026, rounded to 4 decimal places.

Step 3: Once this product has been calculated for all 36 joint sample points in the joint space, then form their sum. This sum is the value of H(Y|X) for this space.

The author has performed this calculation and obtained the following result:

For Experiment 4,

H(X, Y) = 4.9832.

All 36 results of step 2, and their summation (4.9832) is presented in the following table. The reader is invited to verify their results (accurate to 4 decimal places).

Experiment 4 Calculation of H(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Row Sum
Grizzly	0.1026	0.1074	0.1074	0.1074	0.1074	0.1026	**0.6346**
Orangutan	0.1645	0.1645	0.1645	0.1095	0.1095	0.1162	**0.8287**
Lion	0.3468	0.1645	0.1645	0.1195	0.1195	0.1195	**1.0343**
Panda	0.1414	0.1478	0.1478	0.1478	0.1478	0.1414	**0.8739**
Monkey	0.1095	0.1095	0.1162	0.1645	0.1645	0.1645	**0.8287**
Cheetah	0.0881	0.0881	0.0881	0.2733	0.1228	0.1228	**0.7831**

Joint Entropy H(X, Y) =							**4.9832**

Conditional probability

Suppose you are dealing with a chance variable for which you have a probability distribution. It is possible for you to suddenly receive some new information about that chance variable that would change your previous assessment of the probabilities.

In conditional probability, the way in which the new information is allowed to "change your previous assessment" is by eliminating some of the sample points of the previous distribution.

For example, if your experiment consists of rolling a typical gaming die one, the new information might be that "the die landed with more than 2 dots up". This information effectively changes your sample space from {1, 2, 3, 4, 5, 6} to {3, 4, 5, 6}.

The question that conditional probability helps you answer is "What are the probabilities of the new (revised) sample space if they keep the same proportion to each other as in the initial probability space?"

This "new information" is referred to as a "condition", and the new probabilities are called "conditional probabilities" because they are "conditioned" on the new information.

We have already seen conditional probability at work in our five examples involving the mammal die and the tree die. All of those examples involved two chance variables (the rolling of the mammal die and the rolling of the tree die) that combined to make a joint chance variable (the rolling of both dice together).

In those four examples, the "new information" always pertained to the outcome of the mammal die, and we were interested in whether or not that new information might change our assessment of the probabilities for the tree die.

Sometimes the new information about the mammal die will change our assessment of the probabilities of the tree die, and sometimes it will not.

When the new information about the mammal die did not change our assessment of any of the probabilities of the tree die, we said that the two chance variables are stochastically independent. However, when the new information about the mammal die actually did change our initial assessment of the probabilities of the tree die, then we said that the two chance variables are stochastically dependent.

This whole idea of the fact that new information about a probability distribution can change some of its probabilities is called *conditional probability*.

In this section, we are going to develop a *mathematical definition of conditional probability*. In the next sections following this one, we shall use our mathematical definition of conditional probability to further provide a mathematical definition of stochastic independence as well a mathematical definition of stochastic dependence. And these will form the mathematical basis for all that follows in Part II.

Essentially, conditional probability concerns a situation where you start out with one probability space, and then you receive new information about that space. The new information has the result of eliminating some of the sample points of initial probability distribution. What you want to know is "How does the elimination of some of the initial sample points by this new information change the probabilities of the remaining

sample points?" The assumption is that you want the remaining probabilities to retain their same proportionality to each other as in the initial distribution.

In our five example experiments, we were always working with joint distributions – because we need joint distributions in Part II in order to consider the interrelationships of two chance variables.

However, the concept of conditional probability works also on simple probability distribution as well as on joint probability distributions. In the subsection that follows, we shall use simple probability distributions to further explain the idea of conditional probability. Subsequently, we shall provide examples of conditional probability with both simple and joint distributions.

And as we proceed through the remainder of Part II, we shall see that the idea of conditional probability lies behind any notions of portent, meaningfulness and stochastic dependence; and that conditional probability is at the root of all of the principle ideas of Part II of this primer.

An Intuitive View of Conditional Probability

Lets now see how we give a mathematical expression to these ideas about conditional probability.

Lets begin by taking an intuitive, visual approach using a simple probability space (no joint events). Lets first look at a graphical representation of the probability of an event – we'll call it event B.

(In probability theory, *event* means a set of sample points.)

Event B is represented in Figure 1 below as the circle. Notice that B resides within the entire sample space S, represented by the rectangle. In this picture, event B contains some of the sample points in S, and S of course contains all of the sample points in S.

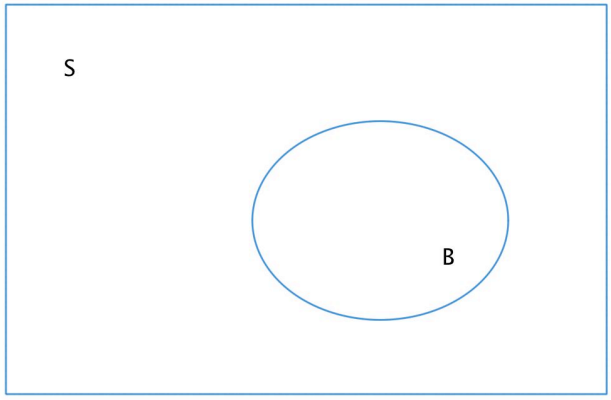


Figure 1

The probability of event B, symbolized "p(B)", is the fraction obtained by dividing the size of B by the size of S. That is,

p(B) = size(B) / size(S)

Now, consider that we have a second event, A. It could very well be that event A has some sample points in common with event B. Figure 2 represents this situation.

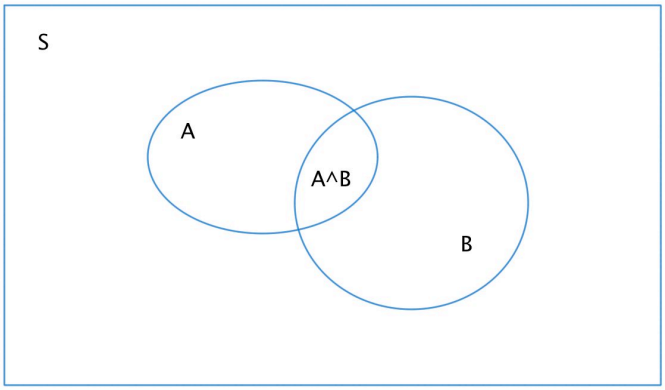


Figure 2

Notice the area where A and B overlap. This area is marked “A^B”, which means “A and B”, because it contains sample points that are in both A and B.

Now, suppose that we are given the following new information:

New information: The sample point that is going to be realized in this trial is definitely in A – which means that it is definitely NOT outside of A.

This new information effectively changes the sample space from S to A! This is because this new information has effectively eliminated anything in S that is outside of A. That is, we can rule out the possibility of anything outside of A being realized during the realization phase of the trial.

The question is: “How does this new information alter our original measurement of the probability of B?

Now, this new information means that we can change the above Venn diagram by “getting rid of” any sample points that are not in A. This we show in Figure 3. In this figure, we use blue shading to indicate that everything in S that is outside of A is no longer a possible outcome and is eliminated from the picture. It has been eliminated by our new information. Thus, the new sample space is now A, and is no longer all of S.

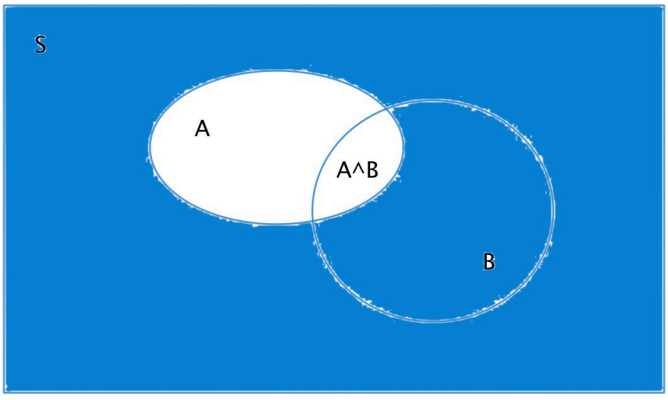


Figure 3

This diagram shows what is left of B. It is the part of B that is also in A! The part of B that is not in A must also be eliminated – because it is also not in A, and nothing that is not in A can be in the “new sample space of possibilities”.

But the part of B that is in A is the region of the diagram marked “A^B”, which means “all sample points that are in both A and B simultaneously”.

Let us then summarize how the diagram has been changed by the new information that the realized event is in A. It means that S has been reduced to A, and that B has been reduced to the region name “A^B”. Figure 4 below shows this.

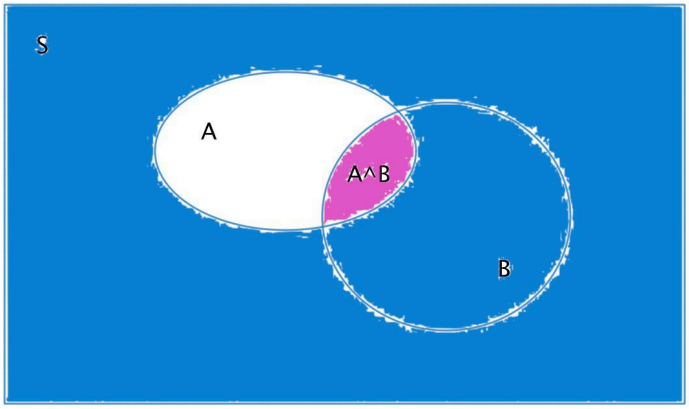


Figure 4

So, the initial probability of B was:

Initial p(B) = size(B)/size(S)

But the new information has effectively replaced “size(B)” in the above with “size(A^B),

And the new information has effectively replaced “size(S)” in the above with “size(A)

So, replacing “size(B)” in the numerator with the new “size(A^B); and likewise replacing “size(S) in the denominator with the new “size(A)”, we have:

New p(B) = size(A^B)/size(A)

Of course, we also need to know what “size(A^B)” means and what “size(A)” means. The “size” of any event is the sum of the probabilities of its sample points. Since sample points are mutually excusive, then this means that:

New p(B) = p(A^B)/p(A)

We need a more descriptive name for the “new p(B)”. We shall call it “the probability of B given A”, and symbolize it with “p(B|A)”, where “|” is the vertical bar, or so-called “pipe” symbol. Therefore:

p(B|A) = p(B) = p(A^B)/p(A)

The Definition of Conditional Probability

We have therefore succeeded in deriving the definition of

Conditional probability of event B given event A, symbolized by “p(B|A)”:

p(B|A) = p(A^B)/p(A)

This concept of conditional probability forms the basis of everything that follows in Part II – and even in Part III – of this primer. The same is true of the mathematical relationship expressed as it definition above.

Some Examples of Conditional Probability

This section will present four examples. The first two will involve simple probability space, while the second two will use joint probability spaces – two of our five example experiments.

A Simple Example of Conditional Probability

As our first example, we shall use the probability space that we introduced this section on conditional probability with. This example involves an ordinary gaming die with six faces, each of which has some number of dots. The numbers of dots range from 1 to 6.

Since the die is "fair", then we know that their probability distribution is the uniform distribution, and that each of the faces has a probability of 1/6. This distribution is represented in the following table:

Ordinary Gaming Die Probabilities

	1 dot	2 dots	3 dots	4 dots	5 dots	6 dots	Total
Probabilities	1/6	1/6	1/6	1/6	1/6	1/6	1

Initially, we are interested in the probability that when the die is rolled that a face lands up that has either 5 dots or 6 dots. Lets call this event – 5 or 6 dots – event B.

We know from above that

P(B) = p(B)/p(S)

Of course,

Initial p(B) = p(5 dots) + p(6 dots) = 1/6 + 1/6 = 1/3

Now, suppose that we obtain the new information that, when the die was rolled, that the number of dots that landed up was greater that 2.

Lets name this new information event A. Event A contains 4 constituent sample points: 3 dots, 4 dots, 5 dots and 6 dots – because all of these events has greater than 2 dots. Therefore,

p(A) = p(3 dots) + p(4 dots) + p(5 dots) + p(6 dots)
= 1/6 + 1/6 + 1/6 + 1/6 = 2/3.

What we want to find out is the new probability of B, given the new information A. This new probability of B is called "the conditional probability of B given A", and is symbolized "p(B|A)". And, from above, its formula is

p(B|A) = p(A^B)/p(A)

At this point, we know the value of one of its inputs, p(A), which is 2/3. But we do not yet know the value of its only other input, p(A^B). So we must calculate that.

Now, "A^B" means all of the events that are in both A and B simultaneously. These are all the die faces that are both "greater than 2" (event A) and "is either a 5 or a 6" (event B) at the same time. The logic of this statement dictates that the sample points that satisfy both of these events at the same time are the set

Event (A^B) = {5 dots, 6 dots}

Therefore,

P(A^B) = p(5 dots) + p(6 dots) = 1/6 + 1/6 = 1/3.

Now we have both of the inputs to calculate p(B|A) = p(A^B)/p(A), which is what we are trying to calculate. Thus

p(B|A) = p(A^B)/p(A) = (1/3)/(2/3) = ½

The conclusion is, given that the die face that lands up is greater than two, the probability that it is either a 5 or a 6 is ½.

This example is simple enough that this answer of ½ should meet with intuitive expectations.

A Second Simple Example of Conditional Probability

This example also involves a simple probability distribution, and will feature a component distribution from our example Experiments 2 and 4.

Suppose we toss the mammal die that we used in experiments 2 and 4. This die was loaded so that the probabilities are not equally likely. You may recall that the probability distribution for that die when thrown alone is:

Loaded Mammal Die Probabilities

	Grizzly	Orangutan	Lion	Panda	Monkey	Cheetah	Total
Mammals	0.1100	0.1600	0.2400	0.1700	0.1600	0.1600	1

From this table, we can look up that the probability that the face that will land with Panda up is 0.1700 – correct to 4 decimal places. We can also see that the probability that the face that lands up is Lion is 0.2400.

Now, suppose that, after the die is thrown, a friend tells us that the face that landed up is a bear. This new information might have an effect on our assessment of the probabilities in the above table. For example, the probabilities that the face that has landed up is a Panda may need to be adjusted based upon this new information.

The same is true for our initial assessment of the probability that it is a Lion, as it is for every other face on this die. However, in the interest of time, we are going to look at only two of the faces right now to see how they are affected.

Lets first consider the revised probability for the Panda. For this question, let B be the event that “Panda lands up”. And we want to factor in the new information we have been given – which is that some bear has landed up. Lets call this event A: “A bear landed up”.

The revised probability that we want to calculate is a conditional probability: p(B|A), the probability of B given A.

We know from above that

p(B|A) = p(A^B)/p(A)

So, we must calculate p(A^B) and divide it by our calculation of p(A):

p(A^B) = p(sample points that are both “a Panda” and “a Bear”) = p(“a Panda”) = .17.
p(A) = p(sample points that qualify as “a Bear”) = p(“a Grizzly”)+p(“a Panda”) = .11+.17= .28.

Thus, p(B|A) = p(A^B)/p(A) = .17/.28 = .6071.

Our conclusion is that: the probability that our mammal die lands with a Panda up – given that a bear lands up – is .6071.

Our second question asks for the conditional probability that a Lion lands up, given that a bear lands up. Intuitively, we know the answer should be zero, since knowing that a bear has landed up should have eliminated any possibility of a Lion landing up.

But lets go through the calculations anyway to verify that our formula for conditional probability gives the answer we expect – that of 0.

So we shall again use the definition for conditional probability of B given A:

p(B|A) = p(A^B)/p(A)

So, we must calculate p(A^B) and divide it by our calculation of p(A). This time, though, events B and A^B are different than before, as follows:

p(A^B) = p(sample points that are both "a Lion" and "a Bear") = p(<empty set>) = 0.0.
p(A) = p(sample points that qualify as "a Bear") = p("a Grizzly") + p("a Panda") = .11+.17= .28.

Thus, p(B|A) = p(A^B)/p(A) = (0.0)/(0.28) = 0.0.

So, the probability that the mammal die lands with Lion up – given that a bear has landed up – is zero, as expected.

An Example of Conditional Probability involving an Independent Joint Distribution

Joint distributions are already naturally organized to consider conditional probabilities – especially when the "given condition" (the "new information") is the first chance variable (such as the mammal die), and when the sample point whose probability we are interested in is of the second chance variable (such as the tree die).

For example, suppose we are working with our Experiment 2 – which you will recall involves rolling both the mammal and the tree dice together. Also suppose that we are interested in the probability that the tree die lands with Lebanon Cedar up, given that the mammal die landed with Lion up.

Notice that with this, our "given condition" pertains to the first chance variable (the mammal die), and the condition that whose probability we are interested in is the second chance variable (the tree die). This is the conditional probability situation that joint probability distribution are naturally set up to work with.

We can represent this conditional probability as

p(Lebanon Cedar | Lion)

- the probability that Lebanon Cedar will land up on the tree die, given that Lion landed up on the mammal die.

We have duplicated the joint probability table of Experiment 2 below so that it will be easy for us to work with here.

Experiment 2 joint distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0242	0.0163	0.0165	0.0213	0.0159	0.0158	**0.1100**
Orangutan	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Lion	0.0528	0.0356	0.0360	0.0466	0.0347	0.0344	**0.2400**

Panda	0.0374	0.0252	0.0255	0.0330	0.0245	0.0243	**0.1700**
Monkey	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Cheetah	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

It is easy to use this table to calculate conditional probabilities like p(Pinyon | Lion), because the element of its formula are readily read from this table. Applying the formula for conditional probability to this we get,

p(B|A) = p(A^B)/p(A)

Here, p(B|A) is p(Lebanon Cedar|Lion), so:

p(Lebanon Cedar|Lion) = p(Lebanon Cedar^Lion)/p(Lion)

Now, p(Lebanon Cedar^Lion) is given by the cell at the intersection of Lebanon Cedar and Lion, which is 0.0360.
And p(Lion) is given by the row sum of the Lion row, which is also the cell in the "Total Mammal" column at the Lion row, which is 0.2400.
So, p(Lebanon Cedar|Lion) = p(Lebanon Cedar^Lion)/p(Lion) = .0360/.24 = .15.

The interpretation of this is: Suppose we rolled both dice together, but did not look at the results. However, a friend tells us that the mammal die landed Lion. Then the question is, how do we use the new information about mammal landing Lion to calculate a revised probability for the tree die having landed with Lebanon Cedar up?

The answer is: by calculating the conditional probability p(Lebanon Cedar|Lion). Which we just did, and we found out that it is .1500.

A very interesting question is "Did this extra information about the mammal die landing with Lion up actually help us refine our estimate of the probability of Lebanon Cedar?"

The answer to this question will be "Yes" if our revised probability of Lebanon Cedar is different from the probability of Lebanon Cedar that we already had before being given this new information. So we want to compare p(Lebanon Cedar|Lion) – our revised probability – against p(Lebanon Cedar) – our initial probability estimate before we had the new information.

We just calculated p(Lebanon Cedar|Lion) = .15.

However, looking at the above table, p(Lebanon Cedar) is found in the cell immediately below the cell containing "Lebanon Cedar". This value is 0.1500.

But so was the p(Lebanon Cedar|Lion). Therefore they are the same probability. We would have to conclude, then, that for this example, the "new information" about mammal landing with Lion up did not end up revising our initial estimate of 0.1500.

If this is true for p(Lebanon Cedar|Lion) for this joint distribution, it is worth asking if it is also true of other joint events in the same distribution.

For example, is it also true for the joint event p(Live Oak|Monkey)? The calculation for this conditional probability yields 0.148125, whereas p(Live Oak) = .1484 according to the table above. So, the two values are very close, but not equal, according to this calculation.

However, in actuality the two really are equal at 0.148125. The value that I calculated for p(Live Oak) is slight off due to the choice that I made to round off the calculations to 4 decimal points. In actuality, to achieve consistent result, I should not have chosen to

round off the results of these tables when displaying them in my spreadsheet computer application. The lesson here is to watch for these kinds of errors when using a computer, as I was when developing the joint probability tables for this primer.

So, in fact, it actually does turn out – at least for this joint distribution – that no matter which joint event is considered, knowing which face lands up on the mammal die does not provide any useful information that would help you refine any of the probabilities of any of the faces of the tree die.

But, remember, we are dealing here with a joint distribution of two chance variables in which the two chance variables are statistically independent. Thus, it should not be surprising that no new information about which mammal fact turned up can give us a different, or refined, probability about which tree face will turn up. That is, no new information regarding which mammal face turned up can be helpful to us in predicting which tree face also turned up. But this is the very meaning of stochastic independence.

Our next example below will deal with the same issues, but we will work with a stochastically dependent joint distribution.

An Example of Conditional Probability involving a Dependent Joint Distribution

Lets look at one more example of conditional probability involving a joint distribution. But this time the two chance variables involved will be statistically dependent. The example we shall use is our fourth experiment.

As with the example above, we shall be interested in conditional probabilities that do revise our initial probability estimates so that any new information about the first chance variable is helpful to us in making better predictions about the second chance variable.

In the previous example, we looked to see if new information about how the mammal die landed could help us better predict the outcome of the tree die. And we found that such new information never helped us in the previous example.

Recall that in our fourth experiment, both dice have been loaded with off-center weights to make the probabilities non-equal. As well, the weights have been magnetized, so that when both dice are rolled together, their magnetic fields will affect which faces of the two die are attractive to each other. The result of all of this is that the joint probability distribution results in the two chance variables being statistically dependent.

We present below the joint probability table for this experiment as it appeared above when we first introduced it.

Experiment 4 joint distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

What we are trying to find out is whether any new information that tells us how which face of the mammal die lands up (when the two dice are thrown together) ever helps us refine our estimates about which face of the tree die lands up.

To find this out, we are going to look to see if there is any tree die face that has a conditional probability that is different from its "unconditional probability". By "unconditional probability" is meant it probability when its die is thrown by itself.

Conditional probability is calculated, as before, by the formula

p(b|a) = p(a^b)/p(a)

where "a" is the "given" mammal die face and "b" is a tree die face of interest.

In order to calculate this we obviously need "p(a^b)" and "p(a)" – both of which we can read from the above joint distribution table as follows:

p(a^b) is found in the cell at the intersection of the "b" column and the "a" row.
p(a) is found at the intersection of the Total Mammal column and the "a" row.

So, we are trying to see if there is a joint event (a, b) in this table such that the conditional probability of p(b|a) is different from "unconditioned" probability of "b" – p(b) - without having any extra information about "a".

And, in order to declare this joint probability space, and its two chance variables, "stochastically dependent", all we have to find is one joint event where there is this difference.

I claim that the joint event (Lion, Pinyon) has this property. That is, I claim that p(b|a) is not the same as p(b). Now, since a = "Lion" and b = "Pinyon", then we are actually trying to see if p(Pinyon|Lion) is different from p(Pinyon). Lets calculate both of these to see if they are, in fact, different.

First, the value of p(Pinyon) is given in the cell immediately below the word "Pinyon" in the table. This probability is, then,

p(Pinyon) = 0.1940.

Second, lets calculate p(Pinyon|Lion). From the formula for conditional probability we know that

p(Pinyon|Lion) = p(Pinyon^Lion)/p(Lion) = 0.0216/0.2400 = .09

Thus, p(Pinyon|Lion) is not equal to p(Pinyon) because one is .1940 and the other is .09.

So, we have shown for this particular joint distribution that it is not always the case that p(y|x) = p(y).

Conditional Probability Distribution of Y Given a Sample Point of X: (Y|X=x)

In this section we are going to look at a situation where we currently have a joint distribution for two chance variables p(X, Y). But subsequently realization occurs, and exactly one of the joint sample points (x_k, y_k) is determined. Suppose that we find out the exact sample point x_k. Thus this new information help us to make a more informed prediction as to which sample point y_k is?

Discussion

Suppose that we have a joint distribution of our two dice, (X, Y), such as Experiment 4. We shall redraw it matrix here for convenience.

Experiment 4 joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Suppose the two dice are tossed, and we find out that the mammal die landed with Monkey facing up.

That new information essentially altered the tree die probability distribution that we initially had. Before we received this new information, our knowledge of the tree die led us to using the tree die component distribution. This tree die component distribution happens to be the same distribution that is featured in the last row of the above matrix – in the row entitled Total Tree.

However, this new information that “the mammal die landed with Monkey up” has the effect of telling us to no longer use the Total Tree row as the distribution of the tree die faces. Instead, this new information is telling us to use the “Monkey” row’s entries as the revised Tree Die probability distribution.

In other words, before we received the information that “the mammal die landed with Monkey up” our assessment of the tree die probability distribution was:

Initial Assessment of the tree die probability distribution p(Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Total Tree	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1.0000**

However, now that we have received the new information that "the mammal die landed with Monkey up", then our new assessment of the tree die probability distribution is:

Revised Assessment of the tree die probability distribution p(Y|X=x)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**

This revised distribution is the result of having eliminated all of the joint probabilities in the joint distribution matrix except for those that are in the Monkey row.

However, there is something wrong with the Revised assessment distribution above. Its probabilities sum to 0.1600 – rather than 1.0000!

And, this is a problem because all probability distributions must be normalized so that they sum to 1. Of course, this problem is the result of having eliminated all of those other probabilities. Therefore, we must normalize this revised probability distribution of the tree die.

But we must also revise it in such a way that its probabilities retain their current proportionality to each other. Dividing every probability in the revised distribution by 0.1600 can do this. This is the value that is currently in the "Total Mammal" column, and that is also the row sum of the Monkey row. If we do this division, we obtain a resulting row that actually is a legitimate probability distribution.

We call this distribution "the conditional tree die probability distribution given that the mammal die is Monkey". We symbolize it (Y|X=Monkey)

This result is:

Conditional tree die probability distribution given mammal=Monkey (Y|X=Monkey)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**

Thus, we have "normalized" the probabilities of the Monkey row of the joint distribution by dividing all of the probabilities of the Monkey row by its row sum. This can be seen by the fact that the new row sum (Total Mammal column) is now 1.0000.

We just did this for the case that the mammal die landed with Monkey up. But, of course, we can do this for every other possible sample point of the mammal die as well. This would develop a total of 6 (Y|X=x) conditional distributions. We call all of these distributions of type "conditional distribution given a sample point".

We shall calculate two more of these here – one for Lion and the other for Grizzly:

Conditional tree die probability distribution given mammal=Lion (Y|X= Lion)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**

Conditional tree die probability distribution given mammal= Grizzly (Y|X= Grizzly)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**

We leave the calculation of the other three conditional distributions given a sample point for Experiment 4 to the reader.

Notice that the symbolism that we use for this type of conditional distribution is "(Y|X=x)". As in all of our symbolism, capital letters such as X and Y represent a component probability spaces; while lower case letters such as x and y represent sample points of their respective probability spaces.

Therefore, "(Y|X=x)" means "the space Y given that X has the value x".

Formal Definition of (Y|X=x)

In table form, (Y|X=x) looks like this:

$(Y|X=x_k)$

	y_1	y_2	y_3	y_4	y_5	y_6	Total Y
x_k	$p(y_1\|x_k)$	$p(y_2\|x_k)$	$p(y_3\|x_k)$	$p(y_4\|x_k)$	$p(y_5\|x_k)$	$p(y_6\|x_k)$	1.0000

Definition: Conditional Distribution (Y|X=x)

Let (X, Y) be the joint distribution of chance variables X and Y. Then define the distribution (Y|X=x), called the conditional distribution of Y given X=x, as follows:

$$(Y|X=x_k) = \{ (p(y|x_k) = p(x_k \wedge y)/p(x_k) \text{ for all } y \in Y \}.$$

Summary

The distribution (Y|X=x) is the conditional probability distribution for Y, given a specific sample point of X. It is calculated by taking the row of the joint distribution (X, Y) that belongs to the X sample point "x", and dividing all of its joint entry points by their row sum p(x).

Conditional Distributions of Y Given distributions for X: (Y|X)

In this section, we are going to develop the conditional distribution for chance variable Y given chance variable X.

Discussion

In the previous section we developed the idea of a "conditional probability distribution of the chance variable Y given a specific sample point of chance variables X". We symbolized this as (Y|X='specific sample point).

For example, in the previous section, we looked at (Y|X=Monkey) for Experiment 4. This is the probability distribution for chance variable Y given that chance variable X had a value of Monkey. In words, this distribution has the probabilities of all of the tree die faces whenever it is also known that the mammal die landed with Monkey up.

Obviously, there are six total such conditional probability distributions for the tree die outcomes given various possible values of the mammal die. We also looked at the case when the mammal die was Grizzly and when it was Lion. We also asked the reader to develop the other three for Cheetah, Panda and Orangutan.

These six conditional probability distribution are to be compared against the initial component probability distribution Y for the tree die faces. If any of them are different from this component distribution, then the conditional distribution could be considered as providing a more accurate refinement of the initial component distribution. As a result, it should replace the initial component distribution in terms of using it to predict the various tree die faces.

In the previous, section, we developed all of these conditional distributions separately.

However, it would be convenient to have them all in one place – one matrix.

Such a matrix is called "the conditional probability distribution of chance variable Y given chance variable X"; and is symbolized by (Y|X).

(Y|X) is a matrix whose rows are the set of all of the (Y|X='specific X value') conditional distributions that were developed in the previous section.

The simplest way to define (Y|X) is to start with the joint distribution matrix (X, Y), and replace all of its joint probabilities by that joint probability divided by its row sum.

For example, consider Experiment 4 again. Its joint probability distribution is repeated below.

Experiment 4 joint distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

To calculate the associated conditional distribution p(Y|X), divide all entries in the body of the above table (all joint probabilities) by their row sum (the entry in the Total Mammal column in the same row. The result is the following conditional probability matrix:

Experiment 4 conditional distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

The astute reader will recognize this conditional distribution as the construct that we used repeatedly in our exposition of the five Experiments for the purpose of ascertaining whether the two chance variables X and Y are stochastically independent.

As one can see, this matrix simply contains all if the rows if all of the “conditional distribution of Y given a specific value of X” that considered in the previous subsection. A purpose of this “distribution” is simply to consolidate all of these conditional distributions in the same table for convenience.

In practice, exactly one of these rows would obtain – the one whose X value was realized. Once that is known, then all of the other rows can be ignored for the specific trial in question. And only the row for the X value that was actually realized is used.

In fact, the realization of a specific X value acts as a selector to eliminate all of the other rows – for that trial.

The astute reader will notice that the above matrix, in its entirety, is not a probability distribution. This is clear from the fact that the sum of all if its joint probabilities is greater than 1. Rather, this matrix contains several probability distributions – one for each row.

Even so, this “conditional distribution” p(Y|X) has mathematical significance in its own right. We shall see toward the end of Part II that there are a number of important mathematical relationships in which p(Y|X) figures prominently.

In fact, p(Y|X) will turn out to be an essential basis for all that follows in Parts II and III – as we shall see.

Formal Definition of p(Y|X)

In table form, p(Y|X), where X has n sample points and Y has m sample points, looks like this:

p(Y|X)

	y_1	y_2	y_3	y_4	y_5	y_m	Total Y
x_1	$p(y_1\|x_1)$	$p(y_2\|x_1)$	$p(y_3\|x_1)$	$p(y_4\|x_1)$	…	$p(y_m\|x_1)$	1.0000
x_2	$p(y_1\|x_2)$	$p(y_2\|x_2)$	$p(y_3\|x_2)$	$p(y_4\|x_2)$	…	$p(y_m\|x_2)$	1.0000
x_3	$p(y_1\|x_3)$	$p(y_2\|x_3)$	$p(y_3\|x_3)$	$p(y_4\|x_3)$	…	$p(y_m\|x_3)$	1.0000
x_4	$p(y_1\|x_4)$	$p(y_2\|x_4)$	$p(y_3\|x_4)$	$p(y_4\|x_4)$	…	$p(y_m\|x_4)$	1.0000
	…	…	…	…	…	…	1.0000
x_n	$p(y_1\|x_n)$	$p(y_2\|x_n)$	$p(y_3\|x_n)$	$p(y_4\|x_n)$	…	$p(y_m\|x_n)$	1.0000

Definition: Conditional Distribution p(Y|X)

Let p(X, Y) be the joint distribution of chance variables X and Y. Then define the distribution p(Y|X=x), called the conditional distribution of Y given X=x, as follows:

$$p(Y|X) = \{ (p(y_k|x_j), \text{ where } p(x_j \wedge y_k)/p(x_k) \text{ for all } x_j \in X, y_k \in Y \}.$$

Conditional Entropy H(Y|X)

Conditional entropy is simply the entropy of a conditional distribution (Y|X).

Like all other distributions that we have encountered so far, this type of entropy is calculated by taking each <entry> in its matrix and calculating the log(1/<entry>). After that, we then want to multiply this result by its probability. Finally, we want to sum all of those products.

Of course, each entry in the conditional distribution p(Y|X) is of the form p(y|x). For example, lets look again at the conditional probability matrix for Experiment 4:

Experiment 4 conditional distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

Take for example, the entry at the intersection of Orangutan and Pinyon. Its value is 0.1200. It means that the probability that the tree die lands with Pinyon up, given that the mammal die landed with Orangutan up, is .12. So, this is of the form “p(y|x)” where “y” is the sample point “Pinyon” and “x” is the sample point “Orangutan”.

So we can use this expression “p(y|x)” to develop the definition (formula) for the conditional entropy H(Y|X). In fact, “p(y|x)” is the “<entry>” in the p(Y|X) matrix that we talked about above.

We said above that for each entry “p(y|x)” in the p(Y|X) matrix, we want to perform log(1/(y|x)).

Next, we want to multiply log(1/(y|x)) by its probability. Now, this probability is obtained from the joint probability table for this entry – which is the value in the table (X, Y) in the same position as the entry that we are currently working on.

Finally, once we have multiplied the value log(1/(y|x)) by its probability for every entry in the p(Y|X) matrix, then we want to add all of these products together. And that sum is the conditional entropy, H(Y|X).

Lets first give an example of this using Experiment 4. Then we shall articulate the definition of H(Y|X) more formally.

In order to calculate H(Y|X) for Experiment 4 (or any other joint probability space) we shall need both its joint distribution and its conditional distribution. We shall repeat both tables here for convenience.

Experiment 4 Joint Distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Experiment 4 Conditional Distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

So, to calculate the conditional entropy, take the following steps. (All logarithms use a base of 2.)

Step 1: Each entry in matrix p(Y|X) is of the form p(y|x). For each such entry, calculate log(1/(y|x)). For example, the entry at Grizzly and Ponderosa has the value of 0.1600. Thus, calculate log (1/.16). This, rounded to 4 decimal places is equal to 2.6439.

Step 2: Now, multiply this calculation by the probability of the joint event for this (x, y) taken from the joint distribution p(X, Y). for (Grizzly, Ponderosa), this joint probability is .0176. The product of .0176 and 2.6439 is .0465, rounded to 4 decimal places.

Step 3: Once this product has been calculated for all 36 joint sample points in the joint space, then form their sum. This sum is the value of H(Y|X) for this space.

All 36 results of step 2, and their summation (4.9832) is presented in the following table. The reader is invited to verify their results (accurate to 4 decimal places).

Experiment 4 Calculation of H(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Row Sum
Grizzly	0.0465	0.0478	0.0478	0.0478	0.0478	0.0465	**0.2843**
Orangutan	0.0757	0.0757	0.0757	0.0587	0.0587	0.0612	**0.4056**
Lion	0.1244	0.0953	0.0953	0.0750	0.0750	0.0750	**0.5401**
Panda	0.0719	0.0739	0.0739	0.0739	0.0739	0.0719	**0.4393**
Monkey	0.0587	0.0587	0.0612	0.0757	0.0757	0.0757	**0.4056**
Cheetah	0.0500	0.0500	0.0500	0.0829	0.0635	0.0635	**0.3601**
Conditional Entropy H(Y|X) =							**2.4351**

Having calculated an example conditional entropy H(Y|X) for Experiment 4, lets codify what we did into a definition of conditional entropy.

Definition: Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} \sum_{y \in Y} p(x,y) * \log(1/p(y|x))$$

Or, equivalently,

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) * \log(p(y|x))$$

The astute reader will recognize by inspecting the definition above that conditional entropy is another one of those entropic measures that were discussed at the end of Part I.

Conditional Distributions of X Given distributions for Y: p(X|Y)

This section develops p(X|Y), the conditional probability distribution of X given Y.

Discussion

In the previous two sections, we have developed p(Y|X). This is the conditional probability distribution of chance variable Y, given that a specific sample point of X was realized.

More accurately, given a joint distribution p(X, Y), given that a realized sample point (x,y) has an "x" value of x_j, then what is the probability distribution that describes its y-value probabilities. p(Y|X) contains a set of probability distributions for the y-values of (x,y).

In actuality, then, p(Y|X) is a set of probability distributions – one for each possible outcome (sample point) of X. Each of these is the probability distribution that one would use for all of the Y sample points for a given value of X.

Think of p(Y|X) as a matrix of rows, where each row contains the probability distribution of Y that one would use for a given value of X. If the X value is x_j, then the j-th row of the matrix contains the probability distribution that one uses for the y values when one knows that the X value of (x,y) is x_j.

In this section, we are going to develop p(X|Y): the conditional distribution for chance variable X given chance variable Y. This is precisely the transverse of the issue we faced with p(Y|X).

This time it will be the columns of the matrix p(X|Y) that we are now interested in.

So, given any Y-value of (x,y), we want to develop a set of probability distributions for the X's. So, given a particular Y value, we want to know the probabilities of all the X values. It will turn out that the y-th column of the p(X|Y) matrix contains the probability distribution for all of the X's, given that specific Y value.

Calculation of p(X|Y)

Recall that we calculated (Y|X) by starting with the joint distribution (X, Y). Then we developed the matrix (Y|X) by dividing each entry (joint probability) in (X, Y) by its row sum.

For example, consider Experiment 4 again. Its joint probability distribution, as you will recall is:

Experiment 4 joint distribution p(X, Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Notice that we also included all of its row sums in the "Total Mammal" column.

To calculate the associated conditional distribution p(Y|X), we divided all entries in the body of p(X, Y) - all joint probabilities - by their row sum (the entry in the Total Mammal column in the same row). The result is the following conditional probability matrix p(Y|X):

Experiment 4 conditional distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

We have already pointed out that p(Y|X) above is not really a "probability distribution", even though we call it a "conditional distribution". This is obvious because the sum of its cells is not 1. Rather, each of its rows is a probability distribution because it sums to 1.

Having reviewed the meaning of p(Y|X), lets now turn our attention to defining p(X|Y).

We are going to define p(X|Y) in a very similar way to p(Y|X) – except we are going to use column sums rather than row sums.

In fact, p(X|Y) will be given by also starting with the joint distribution p(X, Y). This time, however, we shall use its column sums. These are in the bottom row of the above p(X,Y) matrix.

Then, to calculate the p(X|Y) matrix, we shall divide all of the joint probabilities of p(X,Y) by their column sums. The result is p(X|Y), the conditional probability distribution of X given distribution Y.

Experiment 4 conditional distribution p(X|Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper
Grizzly	0.2292	0.1662	0.1062	0.1062	0.1664	0.2260
Orangutan	0.1662	0.1692	0.1662	0.1662	0.1664	0.1660
Lion	0.1062	0.1662	0.2292	0.2262	0.1664	0.1060
Panda	0.1062	0.1662	0.2262	0.2292	0.1664	0.1060
Monkey	0.1662	0.1662	0.1662	0.1662	0.1682	0.1672
Cheetah	0.2262	0.1662	0.1062	0.1062	0.1664	0.2290
Total Tree	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

In this p(X|Y) matrix, it is the columns that contain the conditional distribution for the X's, having been given a specific value for Y.

For example, suppose we find out that, after the two dice are rolled, that the Y die came up "Pinyon". This means that the "Pinyon" column contains the probabilities for the X's. We know then, for example, that the probability that X=Panda is 0.2292 in this case. Whereas, if Y=Live Oak, then the probability of X=Panda is 0.1662.

Formal Definition of p(X|Y)

In table form, p(X|Y), where X has n sample points and Y has m sample points, looks like this:

p(Y|X)

	y_1	y_2	y_3	y_4	y_5	y_m	Total Y
x_1	$p(x_1\|y_1)$	$p(x_1\|y_2)$	$p(x_1\|y_3)$	$p(x_1\|y_4)$	…	$p(x_1\|y_m)$	1.0000
x_2	$p(x_2\|y_1)$	$p(x_2\|y_2)$	$p(x_2\|y_3)$	$p(x_2\|y_4)$	…	$p(x_2\|y_m)$	1.0000
x_3	$p(x_3\|y_1)$	$p(x_3\|y_2)$	$p(x_3\|y_3)$	$p(x_3\|y_4)$	…	$p(x_3\|y_m)$	1.0000
x_4	$p(x_4\|y_1)$	$p(x_4\|y_2)$	$p(x_4\|y_3)$	$p(x_4\|y_4)$	…	$p(x_4\|y_m)$	1.0000
	…	…	…	…	…	…	1.0000
x_n	$p(x_n\|y_1)$	$p(x_n\|y_2)$	$p(x_n\|y_3)$	$p(x_n\|y_4)$	…	$p(x_n\|y_m)$	1.0000

Definition: Conditional Distribution p(X|Y)

Let p(X, Y) be the joint distribution of chance variables X and Y. Then define the distribution p(Y|X=x), called the conditional distribution of Y given X=x, as follows:

$$p(X|Y) = \{ (p(x_j|y_k), \text{ where } p(x_j \wedge y_k)/p(y_k) \text{ for all } x_j \in X, y_k \in Y \}.$$

Conditional Entropy H(X|Y)

Conditional entropy of X given Y, H(X|Y), is simply the entropy of a conditional distribution p(X|Y).

Like all other distributions that we have encountered so far, this type of entropy is calculated by taking each entry in its matrix and calculating the log(1/<entry>). After that, we then want to multiply this result by its probability (<entry>). Finally, we want to sum all of those products.

Therefore, H(X|Y) as the following formula:

Definition: Conditional Entropy:

$$H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) * \log(1/p(x|y))$$

Or, equivalently,

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) * \log(p(x|y))$$

So, how do we interpret either of these formulas as far as applying it to the (X|Y) matrix? Lets work with the second formula since it is slightly easier to calculate.

First notice that both “p(x,y)” and “p(x|y)” appear in this expression. This means that we will be using both the joint distribution p(X,Y) and the conditional distribution p(X|Y).

Lets use Experiment 4 again as our example.

Using the second formula above, for each entry in the p(X|Y) matrix, we calculate the $\log_2$ of the entry. (Log_2 of a probability will always be as non-positive value.) Then we multiply that result by the corresponding p(x,y) entry from the (X, Y) matrix. Finally, we multiply that result by -1, which makes this a non-negative value. This is how we calculate the term of each of the 36 entries in the p(X|Y) table.

Finally, to obtain the conditional entropy for the entire matrix p(X|Y), we sum all 36 of these terms. For Experiment 4, we obtain the following results. Each entry of this matrix is the product of the $\log_2$ of each of the entries in p(X|Y) multiplied by its corresponding entry in p(X, Y), multiplied by -1. When we sum all 36 of these values, we obtain 2.5414, which is H(X|Y) that we seek.

Experiment 4 – The calculation of H(X|Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Row Sum
Grizzly	0.0812	0.0717	0.0573	0.0573	0.0717	0.0808	0.4200
Orangutan	0.0717	0.0723	0.0717	0.0717	0.0717	0.0717	0.4308
Lion	0.0573	0.0717	0.0812	0.0809	0.0717	0.0572	0.4199
Panda	0.0573	0.0717	0.0809	0.0812	0.0717	0.0572	0.4199
Monkey	0.0717	0.0717	0.0717	0.0717	0.0720	0.0719	0.4308
Cheetah	0.0809	0.0717	0.0573	0.0573	0.0717	0.0812	0.4200
Conditional Entropy H(X|Y)) =							2.5414

The Probabilistic Definition of Statistical Independence

We are now ready to give a formal definition of stochastic independence.

In the first sections of Part II, we pursued a learn-by-example approach, and our main focus has been to encourage the reader to consider what the meaning of stochastic independence and stochastic dependence might mean probabilistically. It is now time to articulate these ideas into more formal language.

In words, we have insinuated that:

> For chance variable Y to be statistically independent of chance variable X, it means that any new information regarding the outcome of X does not help us to estimate the outcome of chance variable Y.

In Terms of the Conditional Probability Distribution

But, what does this mean probabilistically?

What we have here is a "before and after" situation concerning chance variable Y and its probability distribution p(Y) – and whether or not any new information about X changes what we think the probability distribution of Y is.

If this new information changes our conception of the probability distribution for Y, then the changed probability distribution is a correction or refinement of our initial probability distribution of Y. In that case, the new information about X is therefore helpful to us in correctly assessing the probabilities of the Y sample points.

On the other hand, if this new information about X did not change our conception of the probability distribution of Y, then the new information about X would not be helpful to us - and we can ignore it.

In the first case, where the new information about X is helpful, the chance variable Y is said to be stochastically dependent of X. If the new information about X is not helpful to predicting Y, then chance variable Y is said to be stochastically independent of X.

In other words, Y is dependent of X if knowing the outcome of X changes Y's probability distribution; and Y is independent of X if knowing the outcome of X does not change Y's probability distribution.

The best way to tell whether one chance variable is independent or dependent of another is to look at their conditional distribution.

Test for stochastic independence:

> If all of the rows p(Y|X=n) of the conditional distribution (Y|X) are equal to each other, and also to the component probability distribution p(Y), the Y is stochastically dependent of X.

Test for stochastic dependence:

> But, if at least one of the rows of the conditional distribution is not equal to the others, or is not equal to the component distribution p(Y), then Y is stochastically dependent of X.

Let's review what this means to an example – say Experiment 2. Following is the component probability distribution p(Y) for this experiment.

Experiment 2 component distribution p(Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total
Trees	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1

And, following is the conditional probability distribution p(Y|X) for this experiment.

Experiment 2 conditional distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**
Orangutan	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**
Lion	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**
Panda	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**
Monkey	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**
Cheetah	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	**1**

Notice that all of the conditional probabilities p(y|x) within the p(Y|X) distribution are the same as their corresponding component distribution Y probabilities p(y). This is how we know that Y is stochastically independent x of X.

Let's take another example. This time, Y will be stochastically dependent on X. This time we'll look at Experiment 5. Following is the component probability distribution Y for this experiment.

Experiment 5 component distribution p(Y)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total
Trees	0.1107	0.1601	0.2391	0.1700	0.1601	0.1601	1.0000

And, following is the conditional probability distribution (Y|X) for this experiment.

Experiment 5 conditional distribution p(Y|X)

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.9900	0.0020	0.0020	0.0020	0.0020	0.0020	1.0000
Orangutan	0.0020	0.9900	0.0020	0.0020	0.0020	0.0020	1.0000
Lion	0.0020	0.0020	0.9900	0.0020	0.0020	0.0020	1.0000
Panda	0.0020	0.0020	0.0020	0.9900	0.0020	0.0020	1.0000
Monkey	0.0020	0.0020	0.0020	0.0020	0.9900	0.0020	1.0000
Cheetah	0.0020	0.0020	0.0020	0.0020	0.0020	0.9900	1.0000

Notice that at least one of the conditional probabilities p(y|x) within the p(Y|X) distribution is not the same as their corresponding component distribution p(Y) probabilities p(y). Take for example, the Grizzly row of p(Y|X). (In fact, all of them are different in this example.) This is how we know that Y is stochastically dependent x of X.

In Terms of a Probabilistic Formulation

The test above for stochastic independence says that all of the probabilities in the rows of the conditional distribution p(Y|X) will be the same as all of the probabilities of the initial "unconditional" probability distribution of p(Y) – which we have been calling the component distribution of Y.

Symbolically, the criteria for stochastic independence is written:

$p(y|x) = p(y)$, for all $x \in X$, $y \in Y$.

In other words, this says that you have stochastic independence whenever: for any Y sample point y, the conditional probability of y – p(y|x) – is the same as the "unconditional" probability of y = p(y) – no matter which sample point x of X is given as "new information".

We now have an articulation of stochastic independence that is worthy of a formal definition. This is one of two different, but equivalent definitions of stochastic independence that we shall present. This particular definition accommodates semantics. The second one that we shall present below is more accommodating of calculation.

First Definition: stochastic independence: Chance variable Y is statistically independent of chance variable X if and only if:

$p(y|x) = p(y)$, for all sample points x in X and y in Y.

From this, we also have a formal definition of stochastic dependence.

First Definition: stochastic dependence: Chance variable Y is statistically dependent of chance variable X if and only if:

$p(y|x) \neq p(y)$, for at least one sample points x in X and y in Y.

A Simpler Equivalent Definition of Stochastic Independence

We have just said that, by definition, chance variable Y is stochastically independent of chance variable X if and only if

$p(y|x) = p(y)$, of all x and y.

But, we also know that, by the definition of conditional probability – for both stochastic independence and stochastic dependence, that

$p(y|x) = p(x \wedge y)/p(x)$.

Therefore, for the special case of stochastic independence, we can equate the right sides of the above two equations to each other. This gives

$p(y) = p(x \wedge y)/p(x)$.

But we can simplify this equation by multiplying both sides by p(x), giving

$p(x)*p(y) = p(x \wedge y)$.

This proves that,

If $p(y|x) = p(y)$, then $p(x)*p(y) = p(x \wedge y)$

Of course, the meaning of "$p(y|x) = p(y)$" is stochastic independence by definition. This is the same as saying that

If Y is stochastically independent of X, then $p(x)*p(y) = p(x \wedge y)$.

If we run the above argument backwards, we get

$p(x)*p(y) = p(x \wedge y)$
$p(y) = p(x \wedge y)/p(x)$
$p(y) = p(y|x)$

The conclusion of this is that

Y is stochastically independent of X if and only if $p(x)*p(y) = p(x \wedge y)$ for all $x \in X$ and $y \in Y$.

So, this gives us our second definition of stochastic independence:

Second Definition: stochastic independence: Chance variable Y is statistically independent of chance variable X if and only if:

$p(x)*p(y) = p(x \wedge y)$ for all $x \in X$ and $y \in Y$.

From this, we also have a formal definition of stochastic dependence.

Second Definition: stochastic independence: Chance variable Y is statistically dependent of chance variable X if and only if:

$p(x)*p(y) = p(x \wedge y)$ for all $x \in X$ and $y \in Y$, for at least one $x \in X$ and $y \in Y$.

Recall that when we were constructing example Experiment 1 and Experiment 2 in the earlier sections above, we came up with a "rule of thumb" for our joint probabilities. This "rule of thumb" was to multiply the component probabilities. In other words, our "rule of thumb" was that, if we wanted a joint probability to be stochastically

independent, we would assign it to be the product of its two component probabilities. But, this is exactly what the above Second Definition of stochastic independence says! Thus, our earlier “rule of thumb” for what the probabilities of independent joint events must be is vindicated.

As an example of this new definition of stochastic independence at work, lets take another look at the joint distribution for our second example experiment. This time we shall not even have to calculate the conditional distribution in order to determine whether Y is stochastically independent of X.

Experiment 2 joint distribution

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0242	0.0163	0.0165	0.0213	0.0159	0.0158	**0.1100**
Orangutan	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Lion	0.0528	0.0356	0.0360	0.0466	0.0347	0.0344	**0.2400**
Panda	0.0374	0.0252	0.0255	0.0330	0.0245	0.0243	**0.1700**
Monkey	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Cheetah	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Remember that Experiment 2 is stochastically independent. With the old definition of stochastic independence, we essentially have to perform all of the calculations that compute the conditional distribution of p(Y|X). However, with the new definition, all we have to do for each of these 36 entries is to multiply the row sum of that entry by the column sum of that entry. If this product is the same as the entry – for all 36 entries, then we have shown, according to our new definition – that Y is stochastically independent of X.

Without, calculating all 36 products, lets just calculate a few – for the sake of the example. Consider one of the cells we used before: (Orangutan, Live Oak). Our new test becomes this: we simply multiply the row sum of (Orangutan, Live Oak) by its column sum. This is 0.1600 x 0.1484 which is 0.0237 (rounded to 4 decimal places). Now compare that answer with the value in (Orangutan, Live Oak) – which is also 0.0237.

Thus, the cell (Orangutan, Live Oak) passes our new criterion for stochastic independence. Of course, to prove that Y is stochastically dependent of X, we have to perform this new test for the remaining 35 cells. However, by quick inspection, this can be seen to be true for all of the cells of this joint distribution. Obviously, this new calculation was quite simpler than the earlier one, which involved the formula

$p(x^y)/p(x)$.

Of course, in order to prove a joint distribution as stochastically dependent, one only has to find a single entry that does not conform to the relationship $p(x^y) = p(x)*p(y)$. But this is simpler than the initial definition also.

A Simpler Equivalent Definition of Stochastic Dependence

We have just said that, by definition, chance variable Y is stochastically independent of chance variable X if and only if

$p(x^y) = p(x)*p(y)$, of all x and y.

Conversely, then, we can state that chance variables X and Y are *stochastically dependent* if and only if

$p(x \wedge y) \neq p(x)*p(y)$, for at least one x and y.

Pause for Reflection

Perhaps surprisingly, this simple criterion for stochastic dependence is all that is necessary to provide, at least, the possibility that there is some degree of predictability between one chance variable and another.

A little later in Part II, we shall consider the further potentiality that some notion of the "degree of strength" of this predictability exists, and we shall explore a way of determining such strength.

In part III, we shall then exploit this phenomenon to continue to construct the mathematical apparatus to determine when this predictability exists, its strength and what the predictions are.

And - don't forget – all of these ideas began with the root concept of conditional probability. Obviously, we have now hidden that fact with our simpler criterion using the product of probabilities. Nevertheless, conditional probability lies at the base of what we are doing in parts II and II.

The Symmetry of Stochastic Independence

The question arises "If Y is independent of X, then must X be independent of Y?"

We shall show that the answer is "Yes". Once we show that this is true, then we can stop saying that "Y is independent of X" or the "X is independent of Y" and start saying that "X and Y are independent".

But, what does it mean to say, "Y is independent of X"? And what does it mean to say, "X is independent of Y"? To show that they are logically equivalent (mean the same thing), we must construct the mathematical equation for each, and then show that if one of those equations is true, then the other one must be true also, and conversely.

The first question we asked is "What does it mean to say that Y is independent of X"? Above, we have defined that "Y is independent of X" means that:

$p(y|x) = p(y)$

The second question we asked is "What does it mean to say that X is independent of Y"? From above this means that:

$p(x|y) = p(x)$

So, a) we have to show that

If "$p(y|x) = p(y)$" then "$p(x|y) = p(x)$".

And, b) we also have to show that

If "$p(x|y) = p(x)$" then "$p(y|x) = p(y)$".

Let's prove a) first:

$p(y|x) = p(y)$ implies

$p(x \wedge y)/p(x) = p(y)$ implies

$p(x \wedge y) = p(x)*p(y)$ implies

$p(y \wedge x) = p(y)*p(x)$ implies

p(x^y)/p(y) = p(x) implies
p(x|y) = p(x)

This proves a), that if p(y|x) = p(y), then p(x|y) = p(x).

To prove b), merely reverse the statements of the proof for a).

Thus, we have proven that chance variable Y is independent of chance variable X if and only if chance variable X is independent of chance variable Y. Consequently, it is reasonable to stop saying both statements and, instead, simply say

"Chance variables X and Y are independent".

Another way of saying what we have proven is that stochastic independence is a symmetric relationship between two chance variables.

The Symmetry of Stochastic Dependence

Also clearly, by a similar argument, *stochastic dependence* is symmetric. (It is left to the reader to prove this somewhat trivial assertion.) So it also makes sense to say "X and Y are stochastically dependent".

This would mean "whenever Y is dependent on X then X is dependent on Y".

The implication is that it is not possible for Y to be dependent on X without concomitantly having X be dependent on Y!

(Admittedly, even though X and Y are stochastically dependent, it does not mean that they are both "predictors" of each other. This is because time is involved in prediction. Even if two chance variables are stochastically dependent, both of them cannot precede the other in time. This issue of prediction is taken up in Part III of this primer.)

Conditional Distributions of Statistically Independent Chance Variables

We have seen that we can calculate a conditional probability distribution "p(Y|X)" from a given joint distribution p(X,Y) of two chance variables. Thus, there is clearly a relationship between the two. In fact, we could say that these two distributions are two different ways of looking at the statistical relationship between two chance variables X and Y.

From this observation, we can say that an interesting question is "Can we look at the conditional distribution and more easily see some things about the joint distribution that would "stand out" more obviously in the conditional distributions view?

In particular, if the joint distribution were statistically independent, might the conditional distribution reveal that fact more obviously than does its corresponding joint distribution?

Lets look at an example to see if anything stands out. We'll take our example experiment number 2, which we know to be statistically independent. Then we'll calculate its conditional distribution and see if anything interesting and obvious shows up.

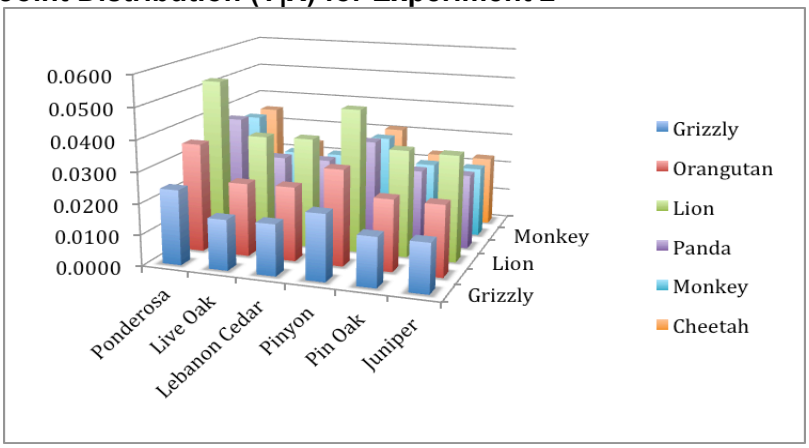
First, here is the joint distribution (again) of our second example experiment.

Mammal and Tree dice joint distribution p(X,Y) – second experiment

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0242	0.0163	0.0165	0.0213	0.0159	0.0158	**0.1100**
Orangutan	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Lion	0.0528	0.0356	0.0360	0.0466	0.0347	0.0344	**0.2400**
Panda	0.0374	0.0252	0.0255	0.0330	0.0245	0.0243	**0.1700**
Monkey	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Cheetah	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

The graph of this joint distribution is:

Joint Distribution (Y|X) for Experiment 2



Now, lets calculate the corresponding conditional distribution for this experiment. Recall how this is done:

Divide each cell of the joint distribution by its row sum. That is, divide every cell in the table by the entry for its row in the “Total Mammals” column. The resulting table is the p(Y|X) table – the conditional distribution for Y given X.

So, we shall calculate the first row of the conditional probabilities table by multiplying all of the joint probabilities of the first row in the joint probabilities table by 2/9. Apply this rule to all six rows, and the resulting conditional probabilities table for this experiment is:

Mammal and Tree dice conditional distribution p(Y|X) – second experiment

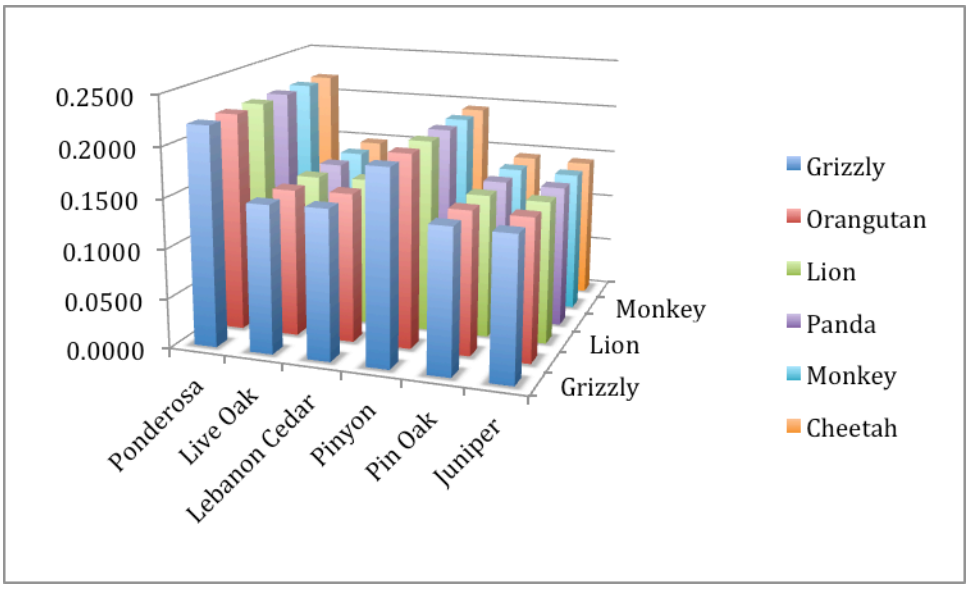
	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	Total Mammal
Grizzly	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Orangutan	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Lion	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Panda	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Monkey	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1
Cheetah	0.2200	0.1484	0.1500	0.1940	0.1444	0.1432	1

Notice that a couple of things stand out. First, all six of the rows are the same as each other! That is, all of the conditional probability distributions for each of the mammal die are the same as each other!

The second thing that jumps out is that all six of the rows of the conditional distribution p(Y|X) are the same as the “Total Tree Probabilities” row of the joint distribution p(X,Y)!

A picture shows this even better. Consider a bar graph of this conditional probability table:

Conditional Distribution (Y|X) for Experiment 2



In this graph, each row of the table is represented by a “series” in the graph. So, “series 1” is a graph of the first row, etc. Notice that the individual graphs for each “series” (table row) is the same as all the other individual graphs.

In other words, the main implication here is that the conditional distribution for a pair of statistically independent chance variables has the property that all if its row are the same as each other.

Question: Does this relationship hold for all statistically independent pairs of chance variables?

Hint: Simply look at the definition of stochastic independence – that p(Y|X) = p(Y) and apply that relationship to any table.

Conditional Distributions of Statistically Dependent Chance Variables

From the previous section, we would also expect that the conditional distribution p(Y|X) for two chance variables that are statistically dependent would have the property that at least one of its rows would have a different set of probabilities in it than the other rows.

So, lets look for corroboration of this suspicion. We shall do so by taking the joint distribution (X,Y) of our fourth example experiment – which we have shown to be statistically dependent, and develop its conditional distribution p(Y|X).

First, lets present once again its joint distribution p(X,Y):

Mammal and Tree dice joint distribution p(X,Y) – Experiment 4

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

From this, lets calculate its corresponding conditional distribution as before:

Mammal and Tree dice conditional distribution p(Y|X) – Fourth Example Experiment

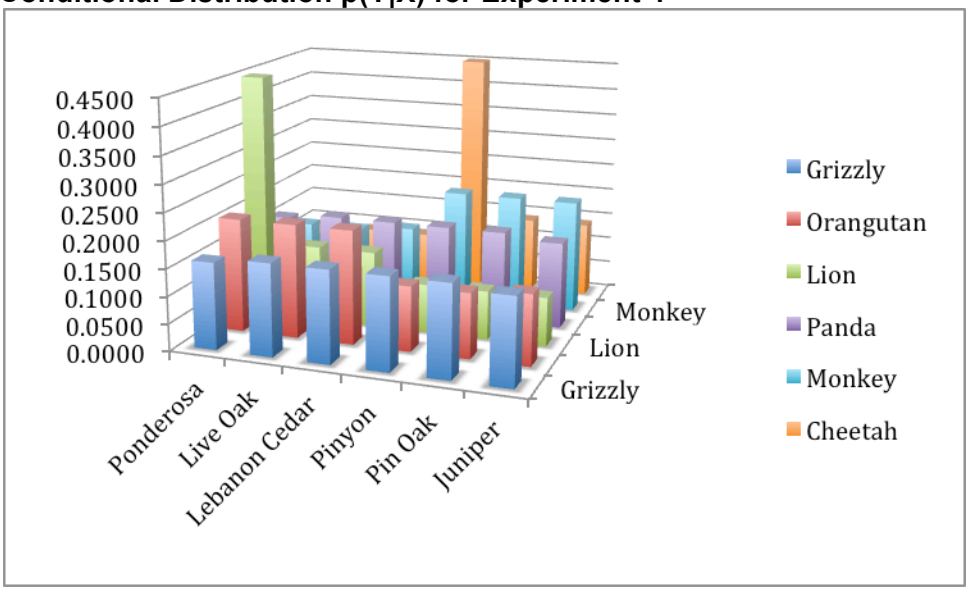
	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Orangutan	0.2100	0.2100	0.2100	0.1200	0.1200	0.1300	**1.0000**
Lion	0.4500	0.1400	0.1400	0.0900	0.0900	0.0900	**1.0000**
Panda	0.1600	0.1700	0.1700	0.1700	0.1700	0.1600	**1.0000**
Monkey	0.1200	0.1200	0.1300	0.2100	0.2100	0.2100	**1.0000**
Cheetah	0.0900	0.0900	0.0900	0.4500	0.1400	0.1400	**1.0000**

Recall that our suspicion is that not all of the rows of this conditional distribution table will be the same as each other. This is clearly the case. In fact, for this example, it appears that no two rows are the same.

Not only that, but there is at least one row of p(Y|X) is different from the “Total Tree” row of p(X,Y). In fact, in this example, they all are!

This fact stands out very well when we create a bar graph for the above table, as we do here:

Conditional Distribution p(Y|X) for Experiment 4



Again, each “series” in this graph represents a single row of our conditional probability table above. We can see that, in fact, it is not the case that all six of these “series” graphs are the same. In fact, for this particular example, no two are alike.

However, we know logically, that we do not require that all six graphs (rows in the table) be different from all of the other five. We merely require that at least one be different from the other five in order to have stochastic dependence.

In any event, we can conclude from this section and the previous one a strong suspicion that we are developing - that it is easy to discern whether a joint distribution p(X,Y) represents two chance variables that are statistically independent or statistically dependent by generating and inspecting the corresponding conditional distribution p(Y|X).

It appears that if all of the rows of the conditional distribution p(Y|X) are the same as each other, then the two chance variables X and Y are statistically independent. Otherwise, they are statistically dependent.

It is left to the reader to prove these suspicions. Hint: They are both true.

The Symmetry of Joint Distributions

It turns out, because of the probability relationships that we have shown above in Part II to be true for statistically independent chance variables X and Y, that there is a number of other interesting and helpful conclusions that we can also draw pertaining to them.

And also, because of the complementary relationship between stochastic independence and stochastic dependence, we can also draw similarly interesting conclusions about statistically dependent chance variables as well.

We shall present without proof some of these. The reader is invited to consider what it would take to prove these relationships.

Proportionality of the Rows of p(X, Y)

Lets turn first to joint distributions of independent chance variables. We showed in the previous section that the corresponding *conditional distributions* for statistically independent chance variables have the property that all of their rows are equal to each other.

The question is, "Can we say something similar for the *joint distribution* of statistically independent chance variables?" The answer is "yes".

The rows of the *joint distribution* p(X,Y) of two statistically independent chance variables X and Y are *proportional* to each other.

Graphically, this means that the graph of such a joint distribution looks like a loaf of bread where, if you slice it in the "across" direction, then all slices of the bread are shaped the same as each other. But, some of the bread slices may be taller than others.

This also has some implication for statistically dependent p(X, Y). For *statistically dependent* chance variables, at least one of the "bread slices" must be of a different shape than the rest.

The Dual Proportionality of p(X, Y)

Recall that we have already proved the symmetry of stochastic independence – and of stochastic dependence. This symmetry has implications for our "loaf of bread" analogy.

Given a statistically independent p(X, Y), the symmetry property says that slicing the loaf of bread in either direction (across or up and down) necessarily results in the slices being in the same "shape" as each other.

More precisely, if X and Y are independent, then slicing the loaf "across" results in all the slices having the same shape as each other. And, slicing the loaf "up and down" also results in the slices all having the same shape as each other.

However, the "across" slices will not, in general, have the same shape as the "up and down" slices.

Of course the corresponding "loaf of bread" analogy when applied to statistically dependent p(X, Y) means 1) for the "across" slices, at least one will be a different shape than any of the others, and 2) for the "up and down" slices, at least one will be a different shape than any of the others.

Mutual Information

We have said from the beginning that the goal of Part II is to develop a function that measures the degree to which one chance variable is meaningful to, or portends something about, another. Such a measuring function should be able to determine whether two such chance variables are meaningful to each other at all – or not. And if they are meaningful to each other, then the function should also be able to tell you how much meaningfulness there is between the two chance variables.

The purpose of this section is to develop such a measure of the degree of portent, or degree of stochastic dependence, between two chance variables. This measure is called the *mutual information* between two chance variables.

As such, mutual information will be shown to be the conclusion of information theory's mathematics of stochastic dependence. In fact, *mutual information* is the essential measure of the degree of stochastic dependence, or of degree of portent, or degree of meaningfulness between two chance variables that participate in a joint distribution.

But there can be many possible joint distributions that two chance variables X and Y can have. This suggests that the value of the mutual information of two chance variables is going to be determined by which specific joint distribution on those two chance variables is involved.

Therefore, when we define *mutual information of two chance variables*, we must indicate which joint distribution we are measuring with respect to.

What We Are Trying to Measure

To summarize what we said in the previous section, we are intending to develop a measure of the *amount of dependence between two chance variables X and Y*.

We have just shown that such an amount, or degree, of dependence is determined by the set of probabilities that are specified a joint distribution on joint sample space (X,Y). But, (X,Y) can have many possible joint probability distributions – each of which would determine its own degree of dependence. Therefore, when measuring the *degree of dependence between X and Y,* we must first specify *which* joint distribution on (X,Y) we are talking about. In other words, our measuring function (and its name) must specify which joint distribution on (X,Y) we are measuring with respect to.

We have said that we have a particular name that we want to give this measure of dependency on two chance variables X and Y that we are about to define. This name is *mutual information*. But, before we can define the mutual information on X and Y, we must first specify a particular joint distribution on X and Y that we shall use to make our definition and calculation of mutual entropy. This is because the calculation of degree of dependency is determined by those joint probabilities, as we discussed in the previous paragraph. This also means that two different joint distributions on X and Y will yield, in general, two different amounts of dependency between X and Y.

In any event, this all means that we must extend the name “mutual information” to include all of the things that must be specified in order to define the concept of stochastic dependence unambiguously. These include the names of the pair of chance variables being measured as well as the specific joint probability distribution being used in the calculation.

Thus, the full name of mutual information must be the mutual information of chance variables X and Y with respect to joint probability distribution $p_k(X,Y)$.

Below, in the section where we develop the actual definition (formula) for mutual information, we shall also use a new symbol for it. Obviously, this new symbol must include all of the elements of the “whole name” of mutual information. Thus, the symbol needs to have an indication that we are talking about *mutual information*, as well as the names of the two chance variables involved, as well as which joint probability distribution on X and Y is being used. The symbol that we shall develop below will be:

$I_k(X;Y)$

Where:

X and Y are the chance variables whose dependence is being measured
k is the index of the joint probability distribution $p_k(X,Y)$ involved

Admittedly, most texts on information theory do not bother to deal with the fact that more than one joint probability distribution on (X,Y) may be involved – but rather that exactly one of them has been established and used. In such a case, there is no need

to make a fuss over the fact that distinct joint distributions in (X,Y) yield two distinct measure of mutual information on X and Y. For those cases, then, there is no fro the subscript "k" in the symbolism. However, we want to be more thorough and include the case that multiple joint distributions on (X,Y) have been defined and are in play – thus we shall include the subscript "k" in the symbolism.

Approach to Mutual Information

Given any two chance variables X and Y, we have shown that there is exactly one joint probability distribution on (X,Y) for which X and Y are *stochastically independent*. This is the joint distribution for which each $p(x,y) = p(x)*p(y)$.

We also want for this particular joint distribution to have dependence measure – a mutual information value – of zero (0), since it has no amount od stochastic dependency. So, since this one and only stochastically independent joint distribution on X and Y has zero amount do dependence between X and Y, lets call this particular joint distribution "$p_0(X,Y)$".

We also know that there are an infinite number of other joint probability distributions all of which are *stochastically dependent* – and therefore have ought to have some positive degree of stochastic dependence when measured by the mutual information function that we are about to define below.

In conclusion, then, any definition of *mutual information* that we invent should give a measure of zero when $p_0(X,Y)$ is used as the joint probability space; but should give a positive measure whenever any other joint probability distribution $p_k(X,Y)$ is used, where $k \neq 0$.

But since the mutual information when $p_0(X,Y)$ is used is zero, and the mutual information when $p_k(X,Y)$ is used is positive, then the measure when p_k is used should be the same as the difference in the measures between when p_k is used and when p_0 is used.

In other words, the measure of dependency, mutual information, when using p_k should be calculable as the difference between the measure of p_k and the measure of p_0.

While it may not seem that substituting "$\underline{p_k - p_0}$" should be simpler than merely using "$\underline{p_k}$", we shall see below that we can use our knowledge of what p_0 is to develop our definition of (formula for) *mutual information*.

In fact, there is something specific that we know about p_0 that we shall be able to use to develop our definition of mutual information that we shall mention in advance right now. It is this fact: since p_0 is *stochastically independent*, then we know by definition that $p_0(x,y) = p(x)*p(y)$, where p(x) uses the component probability distribution on X and p(y) uses the component probability distribution on Y. Thus, whenever we need to, we can substitute "$p(x)*p(y)$" in place of "$p_0(x,y)$". This substitution will be useful below when we develop our definition of *mutual information*.

Reviewing Relative Entropy

Since we have established that we are going to define mutual information as the relative entropy of two specific joint distributions, $p_k(X, Y)$ and $p_0(X, Y)$, then we should review briefly the definition of relative entropy from Part I.

In Part I, we developed a measuring function named *relative entropy*. Its purpose is to contrast the "uncertainty stories" that two probability distributions tell about the same sample space. Recalling that the essence of a probability distribution is to give a picture of the uncertainty inherent in a sample space, it is natural to wonder what the

difference is between the two "uncertainty pictures" put forth by two different probability distributions on the same sample space. Such is the purpose of the measure named relative entropy.

We use the symbol D(p || q) to mean the relative entropy of distributions p and q with respect to p – where both p and q share the same sample space. In words, D(p || q) is the expected value (using the probabilities of p) of the difference between the uncertainties of p from the uncertainties of q.

In a sense, then, we are measuring "how far off" our "initial guess" of the *a priori* distribution q is from, what later turned out to be, the *actual distribution* p. So, now that we know the "actual distribution" q, we shall use it as a baseline, and subtract its "degree of uncertainty" $u_p(x)$ from the "degree of uncertainty $u_q(x)$ of the *a priori* ("initial guess") distribution q.

Recall that the uncertainty of any sample point "x" according to "p" is log(1/p(x)), and the uncertainty of any sample point "x" according to "q" is log(1/q(x)). Consequently, their difference is log(1/q(x)) - log(1/p(x)). To calculate the expected value of this expression using the probabilities of p, we have

$$D(p \,||\, q) = \sum p(x) \, [\log(1/q(x)) - \log(1/p(x))]$$

Which is the definition of relative entropy that we presented in Part I.

However, we showed that this expression can be simplified to:

$$D(p \,||\, q) = \sum p(x) \log (p(x)/q(x))$$

This second expression is the usual articulation of the definition of the relative entropy of two probability distributions p and q with respect to p.

Formal Definition of Mutual Information

We have said that we want to define the *mutual information of chance variables X and Y with respect to joint distribution p* as the *relative* entropy of two particular joint distributions of the joint sample space (X, Y). This means that we are going to apply the concept of *relative entropy* to a joint distribution, rather than to a simpler distribution – as the definition of relative entropy suggests.

It also means that we are selecting two particular joint distributions on the joint sample space (X, Y) to use with the relative entropy formula. These two particular joint distributions are: 1) the joint distribution whose degree of stochastic dependency we are intending to measure, called here $p_k(X, Y)$, and 2) the unique joint distribution on (X, Y) that is *stochastically independent*, called here $p_0(X, Y)$.

Thus, to define mutual information of X and Y, we shall substitute $p_k(X, Y)$ for p and for $p_0(X, Y)$ for q in the final formula above for simplicity of calculation.

This gives:

The mutual information of X and Y:

$$D(p_k(X, Y) \,||\, p_0(X, Y)) = \sum p(x,y) * \log (p(x,y)/q(x,y))$$

We can simplify this expression further. But before we do, lets make a few remarks about it.

First, the notion of relative entropy makes a distinction between the two probability distributions, p and q, that it uses. The probability distribution p is called the *a posteriori* distribution and the distribution q is called the *a priori* distribution.

In our case, though, the "p(x)"s are shorthand for the joint probabilities of joint distribution $p_k(X, Y)$. And, the "q(x)"s are shorthand for the joint probabilities of joint distribution $p_0(X, Y)$. The joint probabilities of $p_k(X, Y)$ are generically called "p(x,y)". While the joint probabilities of $p_0(X, Y)$ are generically called "q(x,y)".

In our case, the joint distribution $p_k(X, Y)$ is the *a posteriori* distribution and the distribution $p_0(X, Y)$ is the a priori distribution. This is true because $p_k(X, Y)$ is the actual distribution, while $p_0(X, Y)$ is the "benchmark" that we are comparing it against. Of course it is the *a posteriori* distribution whose probabilities are mentioned in the formulation immediately after the $\sum$ sign.

Also, we can make a simplifying substitution into the formula. It is for the expression "q(x,y)". It turns out that "q" is the *a priori* distribution $p_0(X, Y)$. Now, this distribution has the property that any of its joint probabilities q(x,y) = q(x)*q(y), since q, or $q(X, Y)_0$, is *stochastically independent* by definition. Therefore, we can substitute "q(x)*q(y)" anytime we have "q(x,y)" in the above formula.

However, q(x) = p(x) and q(y)=p(y) for any sample points x and y. This is true because the component distributions for X and Y are the same for both $p_k(X, Y)$ and $p_0(X, Y)$. Only their joint probabilities are different. Thus, rather than substituting "q(x)*q(y)" for "q(x,y)" in the above formula, we may as well go ahead and substitute "p(x)*p(y)" for "q(x,y)". This point is essential, and is the key to simplifying the initial definition above of mutual information.

Additionally, for the "$\sum_{x \in X}$", we should substitute "$\sum_{x \in X} \sum_{y \in Y}$", since we must sum over both chance variables in the joint distribution. Making these substitutions, then, we get the following

Formal definition for the *mutual information* of X and Y with respect to joint probability distribution $p_k(X, Y)$:

$$I_k(X; Y) = \sum_{x \in X} \sum_{y \in Y} p_k(x, y) * \log (p_k(x, y)/(p(x)*p(y)))$$

Where
- $p_k(X, Y)$ is the *a posteriori* joint distribution
- p(x) is the component probability distribution on X
- p(y) is the component probability distribution on Y

Notice that we have been able to get rid of any references to the $p_0(X, Y)$ distribution in the above simplification. This means that all probabilities mentioned in the formula are either of the joint distribution $p_k(X, Y)$, whose stochastic dependence we are measuring, or of the two component distributions. We could do this because – since $p_0(X, Y)$ is by definition stochastically independent, and since the component probability distributions are the same for both $p_k(X, Y)$ and $p_0(X, Y)$, we were able to express any "$p_0(x,y)$" as "p(x)*p(y)".

The Mutual Information of Two Stochastically Independent Chance Variables

Suppose we have a stochastically independent joint distribution. We have said that we want to construct mutual information, our "measure of stochastic dependence", in such a way that the mutual information of a stochastically independent joint distribution is zero (0).

So, lets see if our above definition of mutual information provides this. To do this, let's look at our formulation to see if, in fact, it produces a value of zero whenever our a posteriori distribution is also stochastically independent. Note that this would mean that our a posteriori distribution, as well as our a priori distribution, would also be $p_0(X, Y)$.

Let's first repeat our definition from the previous section:

The mutual information of X and Y for joint distribution K:

$$I_0(X; Y) = \sum_{x \in X} \sum_{y \in Y} p_0(x, y) * \log (p_0(x, y)/(p(x)*p(y)))$$

Since, in this case, our a posteriori distribution $p_0(X, Y)$ is also stochastically independent, then it would also have the property that $p_0(x, y) = p(x)*p(y)$.

Thus, in the argument to the log function in this definition, we can substitute "p(x)*p(y)" in place of "$p_0(x, y)$". This give the following replacement for $I_0(X; Y)$:

$$I_0(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) * \log ((p(x)*p(y))/(p(x)*p(y)))$$

Notice that now the argument to the log function is "p(x)*p(y)" divided by itself – which is 1. So we can rewrite the above as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) * \log (1) = 0.$$

But, the log(1) = 0. Therefore:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) * 0 = 0.$$

Thus, we have shown what we desired to do: that the mutual information of a stochastically independent joint distribution is zero.

Some Examples of the Mutual Information of Two Chance Variables

In this section we are going to describe three joint distributions on the same two chance variables with the same two composite distributions. Then we are going to compare the values of the *mutual information* of each.

We shall visit these three in the order of their degrees of stochastic dependency. Therefore, we should expect the mutual information value to get larger as we progress through these three examples.

The three examples that we shall use are example Experiments 2, 4 and 5. You will recall that all three of these experiments use the loaded pair of mammal and tree dice – where the *component distributions* for both die are the same two non-uniform probability distributions. Even so, the *mutual information* of these three examples will different for each.

For Experiment 2, the two dice are stochastically independent. Therefore, we should expect its mutual information to have a value of zero.

Experiments 4 and 5, however, are stochastically dependent, with embedded magnets in both dice forming a magnetic field that results in the outcome of each die being influenced by the outcome of the other die. However, the magnetic fields used in Experiment 5 are stronger than those used in Experiment 4. Therefore, the two dice in Experiment 5 is even more stochastically dependent that the two dice used in Experiment 4.

This fact should result in Experiment 5 having a larger (positive) mutual information value than Experiment 4 (also positive). And both should have a larger mutual information value than Experiment 2, whose joint distribution is stochastically independent, and therefore has a mutual information value of zero (0).

Let us now proceed to calculate the mutual information value of all three of these experiments.

Recap of the Mutual Information Definition for These Three Examples

Perhaps the most intuitive way to understand mutual information is through performing the following procedure:

1. You perform a particular calculation for each sample point of the joint sample space involved. In this case, the joint sample space is the set of all 36 possible outcomes (x,y) of the two dice – the mammal die X and the tree die Y. This joint sample space (X,Y) has many joint probability distribution. But exactly one of them must be selected for the purpose of calculating the *mutual information of X and Y*. The select one shall be referred to as $p_k(X, Y)$.
2. This particular calculation subtracts the uncertainty of each sample point based on its actual joint probability $p_k(X, Y)$ from the uncertainty of the same sample point using the stochastically independent joint probability distribution $p_0(X, Y)$. That is, for each joint sample point (x,y), perform the following subtraction $u_0(x,y) – u_k(x,y)$. In other words, this procedure calculates, for each sample point, the difference in the amount of uncertainty as determined by $p_k(x,y)$ and $p_k(x,y)$ for sample point (x,y).
3. Next, you calculate the mean of all of these 36 differences in the two uncertainty values. The result is the mutual information of the actual joint distribution in question. More precisely, this is the mutual information of the two chance variables of the joint distribution in question.

Of course, in order to calculate the mean in step 3, you multiply each of these differences in uncertainty by its probability as specified by the *selected probability distribution* $p_k(X, Y)$. We then add all of these products together to obtain the mean – which is the *mutual information*.

For each of the three example experiments, we shall represent all of this calculation in a 6x6 matrix – one entry for each of the 36 joint sample points. And, the sum of all 36 cells will be the value of the mutual information that we seek.

We shall name this matrix the “Mutual Information Calculation Matrix” for the experiment. The sum of all of its cells will be the mutual information value for the Experiment.

Let summarize what we have just said. To calculate the mutual information for each Experiment, we shall present a 6x6 matrix. Each cell of this matrix will represent one of the 36 joint sample points, (x,y). And it will contain the following calculation:

$$p_k(x,y)^* [\log(1/p_0(x,y)) – \log(1/p_k(x,y))]$$

where

p_k is the actual, or a posteriori, or "selected" joint probability of the point (x,y), and
p_0 is the stochastically independent, or a priori, joint probability of the point (x,y).

Of course, we have shown alternative and equivalent expressions for this calculation that are easier to calculate, which we present here again:

$$I_k(X; Y) = \sum_{x \in X} \sum_{y \in Y} p_k(x, y) * \log (p_k(x, y)/(p(x)*p(y)))$$

This second, alternative, expression has managed to get rid of any mention of the distribution $p_0(x,y)$ – although its presents is still felt in this articulation.

However, the first of the above expressions is more intuitive, and we shall use it for the calculation in these three examples here.

Finally, to calculate the mutual entropy of the actual joint distribution p, we simply sum all 36 of these values.

We shall present this format for all three of the example experiments below.

Note about These Three Examples

Each of these three example use two joint distributions on the joint sample space (X,Y). One of them is the "selected", or "actual", distribution $p_k(X,Y)$ – the one that whose degree of stochastic dependency we are measuring. This will be different for each of the three examples.

The second joint distribution is the unique stochastically independent joint distribution $p_0(x, y)$. Of course, there is only one such distribution for (X,Y). It is the one whose joint probabilities are the products of their component probabilities. So this will be the same distribution for all three examples.

It happens that Experiment 2 is, in fact, $p_0(x, y)$. Therefore, in these three examples, it is Experiment 2 that acts as the $p_0(x, y)$ that we are comparing against the "selected" or "actual" joint distribution of the example.

Therefore,

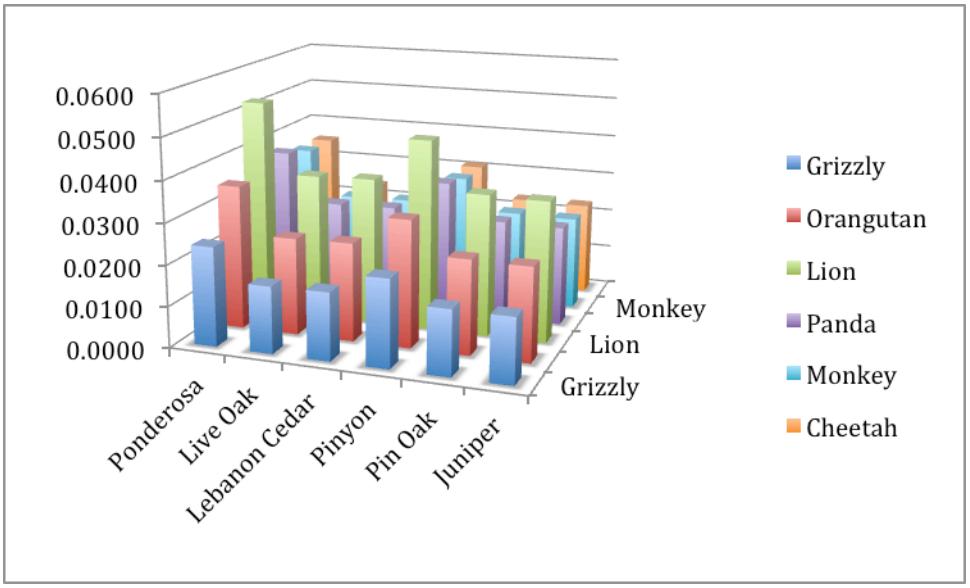
1. In the first example, we shall be comparing Experiment 2 against itself.
2. In the second example, we shall be comparing Experiment 4 against Experiment 2.
3. In the third example, we shall be comparing Experiment 5 against Experiment 2.

Mutual Information of Experiment 2

In this experiment, $p_k(X,Y)$ is Experiment 2. But it turns out that Experiment 2 is, in fact, $p_0(X,Y)$, since Experiment 2 is stochastically independent. Therefore, we are comparing Experiment 2 against itself.

Experiment 2 joint probability distribution $p_k(X,Y)$

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0242	0.0163	0.0165	0.0213	0.0159	0.0158	**0.1100**
Orangutan	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Lion	0.0528	0.0356	0.0360	0.0466	0.0347	0.0344	**0.2400**
Panda	0.0374	0.0252	0.0255	0.0330	0.0245	0.0243	**0.1700**
Monkey	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Cheetah	0.0352	0.0237	0.0240	0.0310	0.0231	0.0229	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**



Note that it is stochastically independent. This can be seen by the fact that every joint probability (in the body of the table) is the product of its "total mammal" and "total tree" composite probabilities.

Since this joint distribution is stochastically independent, then we should expect its mutual information to be zero.

Of course, the calculation of mutual information for this distribution requires that we subtract the actual uncertainty of each sample point from the uncertainty of the joint distribution for these two composite variables if they were stochastically independent. However, they are stochastically independent!

Therefore, you would expect that their difference would be zero for every sample point – because the actual joint distribution and the stochastically independent distribution are the same joint distribution in this case.

Therefore, our Mutual Information Calculation Matrix for Experiment 2 looks like this:

Mutual Information Calculation Matrix for Experiment 2

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper
Grizzly	0	0	0	0	0	0
Orangutan	0	0	0	0	0	0
Lion	0	0	0	0	0	0
Panda	0	0	0	0	0	0
Monkey	0	0	0	0	0	0
Cheetah	0	0	0	0	0	0

Clearly, when we sum all of the cells in this table, the result is:

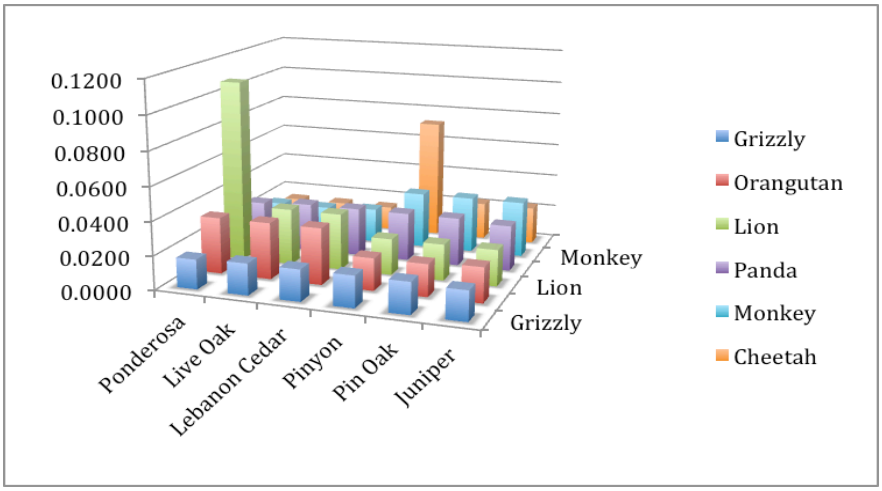
Mutual Information for Experiment 2, $I_k(X;Y) = 0$.

Mutual Information of Experiment 4

In this experiment, the "actual" or "selected" joint distribution $p_k(X,Y)$ is Experiment 4. But we have seen that $p_0(X,Y)$ is in fact Experiment 2, since Experiment 2 is stochastically independent. Therefore, we are comparing Experiment 4 against Experiment 2.

Experiment 4 joint probability distribution $p_k(X,Y)$

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0176	0.0187	0.0187	0.0187	0.0187	0.0176	**0.1100**
Orangutan	0.0336	0.0336	0.0336	0.0192	0.0192	0.0208	**0.1600**
Lion	0.1080	0.0336	0.0336	0.0216	0.0216	0.0216	**0.2400**
Panda	0.0272	0.0289	0.0289	0.0289	0.0289	0.0272	**0.1700**
Monkey	0.0192	0.0192	0.0208	0.0336	0.0336	0.0336	**0.1600**
Cheetah	0.0144	0.0144	0.0144	0.0720	0.0224	0.0224	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**



Note that it is stochastically dependent. This can be seen by the fact that not every joint probability (in the body of the table) is the product of its "total mammal" and "total tree" composite probabilities.

Since this joint distribution is stochastically dependent, then we should expect its mutual information to be zero. The larger it is, the more stochastically dependent the two dice are with respect to each other. And the more the outcome of one portends something about the outcome of the other in a joint outcome.

Of course, the calculation of mutual information for this distribution requires that we subtract the actual uncertainty of each sample point from the uncertainty of the joint distribution for these two composite variables if they were stochastically independent. However, they are stochastically dependent.

Therefore, you would expect that their difference would be non-zero for at least some sample points.

Our Mutual Information Calculation Matrix for Experiment 4 looks like this:

Mutual Information Calculation Matrix for Experiment 4

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper
Grizzly	0.01755	-0.00012	-0.01151	-0.01151	-0.00007	0.01654
Orangutan	-0.00012	0.00061	-0.00012	-0.00012	-0.00007	-0.00017
Lion	-0.01151	-0.00012	0.01755	0.0166	-0.00007	-0.01154
Panda	-0.01151	-0.00012	0.0166	0.01755	-0.00007	-0.01154
Monkey	-0.00012	-0.00012	-0.00012	-0.00012	0.00036	0.00012
Cheetah	0.0166	-0.00012	-0.01151	-0.01151	-0.00007	0.01748

Clearly, when we sum all of the cells in this table, the result is:

Mutual Information for Experiment 4, $I_k(X;Y)$ = **0.04355**

Notice that this value is larger than the mutual information of Experiment 2, which is 0. This is as we expect, since the Experiment 2 is stochastically independent and should have a degree of stochastic dependence of 0. However, Experiment 2 is stochastically dependent, and should have a positive degree of stochastic dependence – which it does.

.

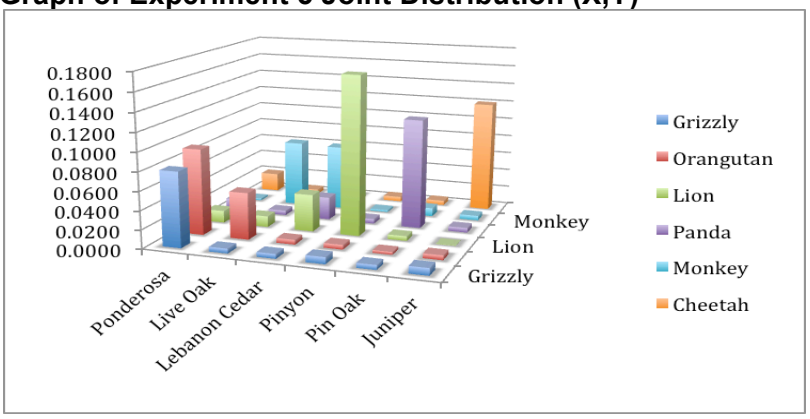
Mutual Information of Experiment 5

In this experiment, the "actual" or "selected" joint distribution $p_k(X,Y)$ is Experiment 5. But we have seen that $p_0(X,Y)$ is in fact Experiment 2, since Experiment 2 is stochastically independent. Therefore, we are comparing Experiment 5 against Experiment 2.

Experiment 5 joint probability distribution $p_k(X,Y)$

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper	**Total Mammal**
Grizzly	0.0800	0.0050	0.0050	0.0073	0.0050	0.0077	**0.1100**
Orangutan	0.0920	0.0500	0.0050	0.0050	0.0030	0.0050	**0.1600**
Lion	0.0130	0.0118	0.0400	0.1697	0.0050	0.0005	**0.2400**
Panda	0.0130	0.0050	0.0246	0.0050	0.1174	0.0050	**0.1700**
Monkey	0.0020	0.0716	0.0704	0.0020	0.0090	0.0050	**0.1600**
Cheetah	0.0200	0.0050	0.0050	0.0050	0.0050	0.1200	**0.1600**
Total Tree	**0.2200**	**0.1484**	**0.1500**	**0.1940**	**0.1444**	**0.1432**	**1.0000**

Graph of Experiment 5 Joint Distribution (X,Y)



Note that it is stochastically dependent. This can be seen by the fact that not every joint probability (in the body of the table) is the product of its "total mammal" and "total tree" composite probabilities.

Since this joint distribution is stochastically dependent, then we should expect its mutual information to be zero. The larger it is, the more stochastically dependent the two dice are with respect to each other. And the more the outcome of one portends something about the outcome of the other in a joint outcome.

Of course, the calculation of mutual information for this distribution requires that we subtract the actual uncertainty of each sample point from the uncertainty of the joint distribution for these two composite variables if they were stochastically independent. However, they are stochastically dependent.

Therefore, you would expect that their difference would be non-zero for at least some sample points.

Our Mutual Information Calculation Matrix for Experiment 4 looks like this:

Mutual Information Calculation Matrix for Experiment 5

	Ponderosa	Live Oak	Lebanon Cedar	Pinyon	Pin Oak	Juniper
Grizzly	0.13800	-0.00853	-0.00861	-0.01130	-0.00834	-0.00795
Orangutan	0.12752	0.05372	-0.01132	-0.01317	-0.00884	-0.01098
Lion	-0.02629	-0.01881	0.00608	0.31663	-0.01397	-0.00305
Panda	-0.01982	-0.01168	-0.00128	-0.01361	0.26506	-0.01142
Monkey	-0.00828	0.11402	0.10930	-0.00791	-0.01224	-0.01098
Cheetah	-0.01631	-0.01124	-0.01132	-0.01317	-0.01104	0.28666

Clearly, when we sum all of the cells in this table, the result is:

Mutual Information for Experiment 5, $I_k(X;Y)$ = **1.1056**

Notice that this value is larger than the mutual information of Experiment 2, which is **0**. This is as we expect, since the Experiment 2 is stochastically independent and should have a degree of stochastic dependence of 0.

Notice also that this value is larger than the mutual information of Experiment 4, which is **0.04355**.

This suggests that Experiment 5 is "substantially more" stochastically dependent than is Experiment 4.

Does this meet with your expectation? Is there something about the two graphs of Experiments 4 and 5 that suggest that Experiment 5 is substantially more stochastically dependent than Experiment 4?

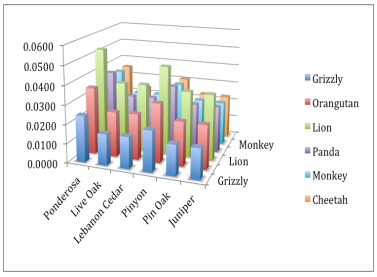
Things to Wonder About

Is there something about the three graphs that suggest that Experiment 2 has *no stochastic dependence*, that Experiment 4 has *some stochastic dependence*, and that Experiment 5 has *a great deal of stochastic dependence*?

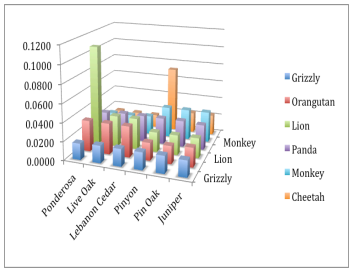
We shall present again the graphs from above of all three joint distributions in the same place below so that the reader can more easily compare and contrast them while pondering this question.

Joint Distributions for:

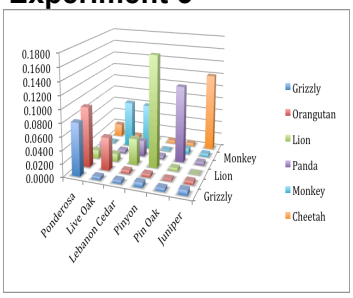
Experiment 2



Experiment 4



Experiment 5



Here's another question. We know that Experiment 2 is stochastically independent. This probably means that we should expect its graph to exhibit some "regularities", or "symmetries". By "symmetries", we mean that some things are "the same *shape*" as other things, even though they might not be the same height. What symmetries can you find in the graph of Experiment 2 that you do not find in the graphs of Experiments 4 and 5? Hint: Look at the "front-to-back slices" of the graph of Experiment 2. Do you find any symmetry there? Also, look at the "left-to-right slices" of the graph of Experiment 2. Do you find any symmetry there?

My answer is this. Look at any of the six "front-to-back" slices of Experiment 2's graph. All of these slices have exactly the same shape as each other – even though they are of different heights. Moreover, they all have the same shape as the Mammal composite graph!

A similar thing is true of you slice Experiment 2's graph from left-to-right. Look at any of the six "left-to-right" slices of Experiment 2's graph. All of these slices have exactly the same shape as each other – even though they are of different heights. Moreover, they all have the same shape as the Tree composite graph!

However, none of these symmetries are true for the graphs of Experiments 4 or 5. This is because they are stochastically dependent. The implication is that "the more stochastically dependent a joint distribution is, the "less symmetrical" are its "slices".

Look again at the graphs of Experiments 4 and 5. Can you see that it is reasonable to say that Experiment 5 is "very far away" from exhibiting these symmetries, but that Experiment 4 is somewhat closer, but doesn't really exhibit these symmetries either? On the other hand, Experiment 2 exhibits these symmetries perfectly.

Mutual information can be seen as a measure of the degree to which there is an absence of these symmetries.

Relationships Among Entropic Measures on Joint Distributions

We have defined a number of *entropic measures* associated with joint probability distributions p(X, Y)[18]. All of these measures are "variations on the theme of *entropy*". That is, they are all generalized forms of entropy.

Formally, what all of these *entropic measures* have in common is that all of them are the expected value of some expression involving "u(x)" – the uncertainty of event. We say that all of these measures are the "expected value of some 'function of u(x)'" – or, symbolically, E(F(u(x)), where F(u(x)) is some function of u(x) and "E" is the *expected value* functional – also know as the *mean*.

Among these *entropic measures* are:

H(X), the entropy of chance variable X.
H(Y), the entropy of chance variable Y.
D(p||q), the relative entropy of probability distributions p and q.
H(X,Y), the joint entropy of chance variables X and Y.
H(X|Y), the conditional entropy of X given Y.
H(Y|X), the conditional entropy of Y given X, and
I(X;Y), the mutual information of chance variables X and Y.

[18] In this section, we are only dealing with one joint distribution at a time on joint space (X,Y). Therefore, we can dispense with the necessity to use the subscript (e.g. "$p_k(X,Y)$") when discussing these joint distributions. This will simplify our symbolism for this section. This is very often the case in probability discussions, and simplifies the symbolism when it is.

All of these are functions provide ways of measuring the mean uncertainty in various relationships among these chance variables. A second interpretation is that these functions measure the degree of randomness of each of these relationships. A third interpretation is that these functions measure the degree to which these relationships are widespread, or "widely adopted". This third interpretation allows us to consider how this "spread" changes in time; in which case we are looking at the dynamics of "how something randomly spreads".

Below, for conciseness, we shall repeatedly mention the first of these interpretations and not the other two. But the reader should remember that the other two meanings are also reasonable interpretations. For any particular application of these entropy measures, any of these interpretations may be used as appropriate.

Review of Initial Definitions of Certain Entropic Measures

Before we begin to discuss various numeric relationships among the above entropic measures, it may be useful to review their direct meanings as specified by their mathematical definitions.

Subsequently, when we look at the numerical relationships among these measures, we shall see how they imply deeper meaningfulness than only these initial definitions.

Initial Definitions of H(X) and H(Y)

H(X) is defined as the *mean uncertainty* of all sample points x of a probability distribution. Recall that the "uncertainty" of a sample point x is defined as the log of the inverse of the probability of x:

$u(x) = \log(1/p(x)) = -\log(p(x))$.

Thus, H(X) is the mean uncertainty of x, or the "mean u(x)" for the entire probability distribution p. To be more specific, we should indicate which probability distribution "p" of the sample space X is being used to define u(x) and H(X). To make this specification, we write "$u_p(x)$" and "$H_p(X)$". However, unless we are comparing two distinct probability distributions p and q on X, we usually don't bother with this delineation.

The important thing to remember is that H(X) is the "mean u(x)", or "mean uncertainty", of a sample space X, based on a specific probability distribution on X.

Everything we just said of H(X) is also true of H(Y).

Below, we are going to expand our understanding of these two by considering certain numeric relationships that they enjoy with other entropic measures.

Initial Definition of H(X,Y)

"p(X, Y)" is a particular joint probability distribution on the joint sample space (X, Y).

Then, "$H_p(X, Y)$" is simply the entropy of all of the joint sample points (x,y) in the joint sample space (X,Y) using joint probability distribution p. If it is clear which joint probability distribution "p" is being used to calculate the entropy, then we can dispense with using the subscript and simply write H(X,Y).

In other words, H(X,Y) is the "mean uncertainty of joint sample space (X,Y)". Or H(X,Y) is the mean u(x,y).

Think of "u(x,y)" as a "fancier", or "generalized" version of "u(x)", where "x" has been replaced by "(x,y)" in the expression "u(x)". In other words, "u(x)" is being replaced by

"u(x,y)". Instead of being interested simply in the "mean u(x)", we are now interested in the "mean u(x,y)".

Think of it this way. The definition of H(X) is

$H_p(X) = \sum_{i \in S} p(x_i)*u(x_i)$, which is the mean value of u(x).

Whereas, the The definition of H(X,Y) is:

$H_p(X,Y) = \sum_{i \in S} p(x_i, y_i)*u(x_i, y_i)$, which is the mean value of some F(u(x)) = u(x,y).

So, we can think of "$u(x_i, y_i)$" as a function of "$u(x_i)$": F(u(x)) = u(x,y). In other words, when we define H(X,Y), we "expand" "u(x)" to a function of u(x), namely u(x,y).

So, what all of these entropic measures have in common is that they are the ma values of some function of u(x).

Initial Definitions of H(Y|X) and H(X|Y)

H(Y|X) is the "mean uncertainty" of all of the (y|x) for all x and y in (X, Y).

That is,

$H_p(Y|X) = \sum_{i \in S} p(x_i, y_i)*u(y_i|x_i)$

Similarly,

$H_p(X|Y) = \sum_{i \in S} p(x_i, y_i)*u(x_i|y_i)$

Initial Definition of Mutual Information I(X;Y)

The mutual information of joint distribution $p_K(X, Y)$ is the "mean difference of uncertainties" between $u_K(x,y)$ and $u_0(x,y)$ for all joint sample points (x,y).

Here, $u_K(x,y)$ is the uncertainty of (X,Y) using the joint probability distribution $p_K(x,y)$ to calculate u(x,y). Whereas, $u_0(x,y)$ is the uncertainty of (X,Y) using the unique stochastically independent joint probability distribution $p_0(x,y)$ to calculate u(x,y).

That is, the mutual information $I_K(X;Y)$ based on joint probability distribution $p_K(x,y)$ is the "mean difference between $u_K(x,y)$ and $u_0(x,y)$" for all joint sample points (x,y) of the joint sample space (X,Y). Or, the mean of $[u_0(x,y) - u_K(x,y)]$.

However, it is often implied that the joint probability distribution to be used in calculating the mutual information of (X,Y) is $p_K(X;Y)$. Therefore, the "k" subscript can be dropped, and "$I_K(X;Y)$" become simply "I(X;Y)".

This puts us into the position to be able to articulate I(X;Y) as the "mean of a function of u(x)", as we have been discussing. In this case the "function of u(x)" is actually a function of u(x,y), the uncertainty of a joint distribution:

$F(u(x,y)) = u_0(x,y) - u_K(x,y)$

Thus,

$I(X;Y) = \sum_{i \in S} p_K(x_i, y_i)*[u_0 (x_i,y_i) - u_K(x_i,y_i)]$

In words, the mutual information measures the "average uncertainty distance" that the sample points of a joint distribution of interest are from the themselves if their uncertainty had been calculated using the unique stochastically independent distribution for X and Y.

In other words, the *mutual information of X and Y with respect to joint distribution* p_K, I(X;Y) measures "how far away from stochastic independence" is joint distribution (X,Y), assuming that (X,Y) has joint distribution p_K.

Or, I(X;Y) measures the degree of stochastic dependence of X and Y in joint distribution p_K.

Some Numerical Relationships Among Entropic Measures

There are some specific numerical relationships that are enjoyed among these measuring functions, regardless of the joint distribution. We shall discuss a number of these relationships in this section – and also their implications to applications of information theory to probability spaces.

We shall take a "discovery" approach to these relationships by looking at the results from calculating these measures for all five of our example Experiments. Here, then, is a summary of those results. We have also included some of these numerical relationships results for each of the five experiments.

Entropic Measure Relationships for Example Experiments 1- 5

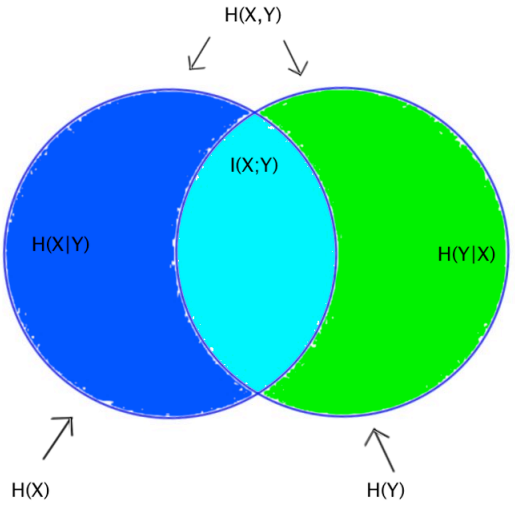
		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
1	**H(X)**	2.5850	2.5481	2.5850	2.5481	2.5481
2	**H(Y)**	2.5850	2.5632	2.5850	2.5632	2.5632
3	**H(X, Y)**	5.1699	5.1113	5.1264	4.9832	4.0057
4	**H(X|Y)**	2.5850	2.5481	2.5414	2.4200	1.4425
5	**H(Y|X)**	2.5850	2.5632	2.5414	2.4351	1.4577
6	**I(X:Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
7	**H(X) - H(X|Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
8	**H(Y) - H(Y|X)**	-0.0000	0.0000	0.0436	0.1281	1.1056
9	**H(X) + H(Y) - H(X,Y)**	0.0000	-0.0000	0.0436	0.1281	1.1056
10	**H(X) + H(Y|X)**	5.1699	5.1113	5.1264	4.9832	4.0057
11	**H(Y) + H(X|Y)**	5.1699	5.1113	5.1264	4.9832	4.0057

We shall now point out a number of interrelationships that can be inferred by the above table. Afterwards, we shall attempt to draw some general conclusions about these relationships. We shall leave proofs of these relationships to the reader, since they mostly follow immediately from the definitions. As well, we shall also point the reader to proofs in published literature.

We shall refer to the first six *entropic measures* in the table above as "the six entropies". The final five entities are arithmetic expressions involving some of these entropic measures that add up to the value of some others of these entropies. These constitute the numerical relationships of interest to this section.

Diagram for Entropic Relationships

All of the entropic measures and their relationships that are discussed in this section can be represented diagrammatically in a Venn-like diagram. This is depicted below. Also see [Cover and Thomas 1991; p. 16].



In this diagram, the “entropy” (entropic measure) of the entire joint probability space H(X,Y) is depicted as all regions in the diagram that are colored either blue, green or turquoise. The entropy H(X) is depicted as all regions that are either blue or turquoise. And the entropy H(Y) is depicted as all regions that are either green or turquoise.

Breaking these entropies down further, we see that the entropy H(X) is equivalent to the *entropic sum* I(X;Y) + H(X|Y). And that the entropy H(Y) is equivalent to the *entropic sum* I(X;Y) + H(Y|X). While these particular equivalences are not listed in the above table of relationships, they nevertheless can be calculated from rows 6 and 7 and rows 6 and 8 of the table.

Finally, recall that the table above shows that the expressions in rows 7, 8 and 9 are equivalent expressions for row 6 - the *mutual information of X and Y*. Lets see how this diagram attests to those relationships.

Obviously, the turquoise area - intersection of H(X) and H(Y) in the diagram - is marked as “I(X;Y)”. So this is the area that we are saying is equivalent to the three expressions in rows 6, 7 and 8. Lets look at each of these three expressions and see of it makes sense graphically to claim that it is equivalent to the turquoise area (the intersection of H(X) and H(Y)).

Lets first look at the expression in row 7: H(X) – H(X|Y). Now H(X) is represented by the blue plus the turquoise regions. And if we “take away” H(X|Y), which is the blue region, you are left with the turquoise region – which represents I(X;Y).

The diagram similarly represents that row 8 is graphically equivalent to I(X;Y). The reader can work through the colored regions on this one.

Finally, lets see how the diagram supports the claim that the expression in row 9 – "H(X) + H(Y) – H(X,Y)" – is equivalent to I(X;Y). To see this, lets break the expression in row 9 into two parts: the "H(X) + H(Y)" and the "H(X,Y)" part. We want to show that, together, they are equivalent to I(X;Y), which is the turquoise region of the diagram.

The first part, "H(X) + H(Y)" is constituted in the diagram by 1 blue, 1 green and 2 turquoise regions. The reason that 2 turquoise regions are involved in "H(X) + H(Y)" is because each of both H(X) and H(Y) bring their own turquoise region amount with them. Now lets look at the second part: "H(X, Y)". This is constituted in the diagram by 1 blue, 1 green and 1 turquoise region. This time only 1 turquoise region is represented.

But, the expression in row 9 says to subtract the second part from the first part. This means subtracting 1 blue, 1 green and 1 turquoise region from 1 blue, 1 green and 2 turquoise regions. After this subtraction, the blue and the green regions go away, and we are left with 1 turquoise region. This is the answer we are looking for, which represents I(X;Y).

This discussion suggests a certain interpretation of I(X;Y). The turquoise region can be considered either as a part of H(X) or as a part of H(Y). When considered as a part of H(X), the diagram tell us that I(X:Y) accounts for all of the remaining uncertainty of H(X) that is not accounted for by the uncertainty of H(X|Y) – which is the uncertainty of X after it has been reduced by some knowledge (certainty) of Y. But this means that I(X;Y) must be the amount of uncertainty of X that is not reduced by any knowledge of Y.

By the same argument, taking the turquoise region as a part of H(Y), I(X;Y) must also be the amount of uncertainty of Y that is not reduced by any knowledge of X. Thus, this turquoise region must represent the amount of uncertainty of X and Y that is not reduced by either some knowledge of X or by some knowledge of Y. Therefore, the turquoise region, I(X:Y), is the "mutual, pure uncertainty" of X and Y. But, from Part I, we know that our measure of "information" is equivalent to the measure of uncertainty whose removal resulted in that information. Therefore, this "mutual uncertainty" can just as well be called mutual information.

Thus, we can see that this diagram is a good graphical representation of the interrelationships amount these various entropic measures.

The Reduced Uncertainty of Conditional Probability

In this section we shall suggest that the idea of conditional probability provides additional information (the conditions) to a joint distribution. In so doing, it should reduce the amount of uncertainty that is inherent in the initial joint distribution to which it provided this "new information". Or, at the very least, this new information should not increase the amount of uncertainty.

That is, H(Y|X) should be less than or equal to H(Y). The same is true for H(X). That is, H(Y|X) ≤ H(Y) and H(X|Y) ≤ H(X). So, lets see if our table bears out this relationship – for all of the example Experiments.

We shall repeat the above table for proximity of reference.

		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
1	**H(X)**	2.5850	2.5481	2.5850	2.5481	2.5481
2	**H(Y)**	2.5850	2.5632	2.5850	2.5632	2.5632
3	**H(X, Y)**	5.1699	5.1113	5.1264	4.9832	4.0057
4	**H(X|Y)**	2.5850	2.5481	2.5414	2.4200	1.4425
5	**H(Y|X)**	2.5850	2.5632	2.5414	2.4351	1.4577
6	**I(X:Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
7	**H(X) - H(X|Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
8	**H(Y) - H(Y|X)**	-0.0000	0.0000	0.0436	0.1281	1.1056
9	**H(X) + H(Y) - H(X,Y)**	0.0000	-0.0000	0.0436	0.1281	1.1056
10	**H(X) + H(Y|X)**	5.1699	5.1113	5.1264	4.9832	4.0057
11	**H(Y) + H(X|Y)**	5.1699	5.1113	5.1264	4.9832	4.0057

We first point out that the joint entropy H(X,Y) is larger than both H(X) and H(Y) for all five experiments. This often happens, if for no other reason than that the joint distribution has more sample points than either of the component distributions. And, one of the forces that make entropy larger is the number of sample points in the distribution. However, joint entropy is not always larger than both of the two component entropies. For example, regardless of the component entropies, it is possible for the joint entropy to be zero (0). This can happen if exactly one of the joint sample points has probability of 1, while the probability of all of the other joint sample points is zero.

Notice that the conditional entropy H(Y|X) is never greater than the entropy H(Y). This is because the uncertainty of Y is always reduced (or unchanged) by having been given "extra information" about the other chance variable X. In other words, "finding out something extra about X" "can't hurt". It can only reduce – or leave unchanged - the amount of uncertainty about Y. At the worst, it will leave the uncertainty about Y the same as before. And the same is true the other way around: The conditional entropy H(X|Y) is never greater than the entropy H(X).

You can see this by looking at the above table for all five experiments. Notice, for all five experiments, that H(Y|X) is always the same or less than to the value of H(Y) for any given experiment. In fact, H(Y|X) is the same as H(Y) for Experiments 1 and 2; and it is less than H(Y) for Experiments 3, 4 and 5.

The same is true for H(X|Y) and H(X). This reflects the fact that conditional probability has to do with being given “extra information” about “the other chance variable”. And that extra information can never hurt. In fact, it can sometimes help.

But, when can it help? That is, under what conditions can the “extra information” provided by conditional probability actually reduce the amount of uncertainty? We can tell this by whether the conditional entropy H(Y|X) ends up being actually less than the entropy H(Y) – rather than being the same as H(Y). (The same goes for H(X|Y) versus H(X).)

To find this out, we need to inspect the cases where H(Y|X) is the same as H(Y) and where they are different. We need to see if there is consistently some condition under which H(Y|X) is the same as H(Y) and some opposite condition under which they are different.

We also need to do the same for H(X|Y) and H(Y). What conditions exist consistently when these two are the same, and what opposite conditions exist consistently when they are different.

Let start with H(Y|X) and H(Y). When are they the same and when are they different? Well, H(Y|X) is the same as H(Y) for Experiment 1 and Experiment 2. Looking at the table, we can see that they both have a value of 2.5850 for Experiment 1 and 2.5632 for Experiment 2. So too for the values of H(X|Y) and H(X). They both have the value 2.5850 for Experiment 1 and 2.5481 for Experiment 2.

But, we have to ask, what “condition” does both Experiment 1 and Experiment 2 exhibit that Experiments 3, 4 and 5 do not? The answer, you will recall, is they are *stochastically independent*.

Now, lets look at Experiments 3, 4 and 5 – which are stochastically dependent – and see if all three of them have the property that 1) their H(Y|X) values are less than their H(Y) values; and 2) their H(X|Y) values are less than their H(X) values.

For Experiment 3, H(Y|X) = 2.5414 and H(Y) = 2.5850; while H(X|Y) = 2.5414 and H(X) = 2.5850. So on both cases conditional probability in Experiment 3, the conditional probability has reduced the entropy, and thus the uncertainty, of target component distribution. Proving that the same is the case for Experiments 4 and 5 is left for the reader.

Nevertheless, we have shown consistent evidence that whenever two chance variables are stochastically dependent, then conditional probability of one chance variable given the outcomes of the second reduces the uncertainty of the first when there is no extra information about the second.

And, we have also provided evidence that whenever the two chance variables are stochastically independent, then extra knowledge regarding the second chance variable does nothing to reduce the uncertainty of the first.

Mutual Information and Stochastic Dependency

Look at the row labeled “I(X:Y)” in the above table. This is the mutual information of the two chance variables X and Y. We have advertised mutual information as a “measure of the degree of stochastic dependency between two chance variables”. Lets see if the values in the above table regarding our five experiments support this claim.

Notice that the values of Experiments 1 and 2 in the above table are both zero (0). This would suggest that both of these experiments have zero amounts of stochastic dependency.

But they should! Because both of these joint distributions are stochastically independent – and therefore should exhibit zero stochastic dependency!

Now, look at the values of I(X;Y) for Experiments 3, 4 and 5. Notice that all three values are greater than zero: **0.0436**, **0.1281** and **1.1056** respectively. This would suggest that all three of these experiments exhibit some degree of stochastic dependency. And, we have claimed all along that all three of these experiments are, indeed, stochastically dependent.

However, notice that these values are increasingly positive. This fact suggests that there are varying degrees of stochastic dependency – with some joint distributions being more stochastically dependent than others. The implication is that stochastic dependency is not merely a binary condition. Rather, stochastic dependency exhibits a continuum of degrees. And therefore it makes sense to ask, "How much stochastic dependency does a joint distribution exhibit?"

On the other hand, stochastic independence only has one mutual information value – that of zero (0). So, stochastic independence is a binary condition. If a joint distribution has a mutual information value of zero, it is stochastically independent. If it does not have a value of zero, then it has a positive value and it is stochastically dependent.

Thus, the phenomenon of stochastic dependency can be measured, by mutual information, on a continuous scale ranging from zero to some upper limiting positive value (where the upper limit is distinct for each size of joint sample space.) If the information value is zero, then the joint distribution is said to be stochastically independent. If the mutual information value is positive, then the joint distribution is said to be stochastically dependent.

Moreover, if a joint distribution is stochastically independent, then $H(Y|X) = H(Y)$ and $H(X|Y) = H(X)$. Otherwise, if a joint distribution is stochastically dependent, then $H(Y|X) < H(Y)$ and $H(X|Y) < H(X)$.

Entropic Measures of the Five Experiments

First, we note that Experiments 1 and 3 both use the same two component probability distributions X and Y. So, we shall be comparing these two experiments with each other. In fact, distributions X and Y are both the uniform distribution for a sample space with six items.

Also, Experiments 2, 4 and 5 use the same two component distributions as each other. These two distributions are different from the two that are used by Experiments 1 and 3. These two distributions are different from each other, as well as being different from the distributions used in Experiments 1 and 3. Specifically, both of these distributions are non-uniform.

Regarding Experiments 1 and 3, the joint distribution for Experiment 1 is stochastically independent, while the joint distribution for Experiment 3 is stochastically dependent. Therefore, we should expect the mutual information of Experiment 1 to be zero (0); and we should expect the mutual information of Experiment 3 to be positive.

Regarding Experiments 2, 4 and 5, the joint distribution for Experiment 2 is stochastically independent, while the joint distributions for Experiments 4 and 5 are stochastically dependent. Therefore, we should expect the mutual information of Experiment 2 to be zero (0); and we should expect the mutual information of Experiments 4 and 5 to be positive. However, the joint distribution of experiment 5 is "more different" from the joint distribution of Experiment 2 than is the joint distribution

of Experiment 4. Therefore, we should expect the mutual information of Experiment 5 to be greater than that of Experiment 4.

We shall repeat the above table for proximity of reference.

		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
1	**H(X)**	2.5850	2.5481	2.5850	2.5481	2.5481
2	**H(Y)**	2.5850	2.5632	2.5850	2.5632	2.5632
3	**H(X, Y)**	5.1699	5.1113	5.1264	4.9832	4.0057
4	**H(X|Y)**	2.5850	2.5481	2.5414	2.4200	1.4425
5	**H(Y|X)**	2.5850	2.5632	2.5414	2.4351	1.4577
6	**I(X:Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
7	**H(X) - H(X|Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
8	**H(Y) - H(Y|X)**	-0.0000	0.0000	0.0436	0.1281	1.1056
9	**H(X) + H(Y) - H(X,Y)**	0.0000	-0.0000	0.0436	0.1281	1.1056
10	**H(X) + H(Y|X)**	5.1699	5.1113	5.1264	4.9832	4.0057
11	**H(Y) + H(X|Y)**	5.1699	5.1113	5.1264	4.9832	4.0057

To summarize the situation with mutual information for Experiments 2, 4 and 5: Experiment 2 has I(X;Y) = **0.0000**; Experiment has I(X;Y) = **0.1281**; and Experiment 5 has I(X;Y) = **1.1056**.

Graphically, this means that: for Experiment 5, the H(X) and the H(Y) ovals have an intersecting area of the size 1.1056; for Experiment 4, the H(X) and the H(Y) ovals have a somewhat smaller intersecting area of the size 0.1281; and for Experiment 2, the H(X) and the H(Y) ovals do not intersect at all.

Mutual Information and the Interdependency Relationships

Mutual information is equivalent to a number of alternate interpretations. We shall explore these in this section.

We shall repeat the above table for proximity of reference.

		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
1	**H(X)**	2.5850	2.5481	2.5850	2.5481	2.5481
2	**H(Y)**	2.5850	2.5632	2.5850	2.5632	2.5632
3	**H(X, Y)**	5.1699	5.1113	5.1264	4.9832	4.0057
4	**H(X|Y)**	2.5850	2.5481	2.5414	2.4200	1.4425
5	**H(Y|X)**	2.5850	2.5632	2.5414	2.4351	1.4577
6	**I(X:Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
7	**H(X) - H(X|Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
8	**H(Y) - H(Y|X)**	-0.0000	0.0000	0.0436	0.1281	1.1056
9	**H(X) + H(Y) - H(X,Y)**	0.0000	-0.0000	0.0436	0.1281	1.1056
10	**H(X) + H(Y|X)**	5.1699	5.1113	5.1264	4.9832	4.0057
11	**H(Y) + H(X|Y)**	5.1699	5.1113	5.1264	4.9832	4.0057

It may not have escaped your notice that – for all five experiments – the values in rows 7, 8 and 9 are equal to the I(X:Y), or the mutual information, of that experiment in row 6.

This fact has the effect of providing three alternative ways to understand the meaning of, or to interpret, mutual information. Since mutual information is one of most important ideas of information theory, any increased understanding of it is welcomed. We are going to explore this further in this section.

We are going to declare, without proof for now, that each of the expressions in the last three rows are in fact always equal to the mutual information I(X;Y) of X and Y – and to each other - for any joint distribution.

From a practical standpoint, these equivalences also provide three alternate methods of calculating I(X;Y). For example, instead of calculating I(X;Y) as defined above, one could calculate, instead, H(Y) – H(Y|X).

In any event, lets now discuss the implications of the fact that I(X;Y) as defined above also has the same meaning as “H(Y) – H(Y|X)”, as “H(X) – H(X|Y)” and as H(X) + H(Y) – H(X,Y). We’ll also discuss the implications of the fact that these three have the same meaning as each other.

Lets first scrutinize the expression in row 8: “H(Y) – H(Y|X)”. “H(Y)” depicts the amount of uncertainty, or randomness, or spread of probability associated with chance variable Y’s probability distribution. On the other hand, “H(Y|X)” depicts the amount of uncertainty of Y once the outcome of X is given. H(Y|X), then should represent a reduction in the uncertainty of Y that is obtained by finding out some information about X. In other words, H(Y|X) accounts for part of the uncertainty inherent in Y, but not all of it.

So, “H(Y) – H(Y|X)” must measure the amount of uncertainty that remains in Y once you have removed (subtracted) the amount of uncertainty in Y that has been informed by the certain knowledge of X.

But we just shown that this amount of uncertainty is equal to I(X:Y) – the mutual information of X and Y. So, this is another way to understand I(X;Y). So I(X;Y) is the amount of uncertainty left in Y once you have removed the amount of uncertainty in Y that has been informed by knowledge of X. In one sense, it is the “pure uncertainty” of Y.

Row 7 considers the opposite of row 8: “H(X) – H(X|Y)”. We have said that expression too has the same value as I(X;Y), and therefore as “H(Y) – H(Y|X)”. All three of these expressions have the same value.

Of course this says that it is also true that the mutual information I(X;Y) is the amount of uncertainty left in X once you have removed the amount of uncertainty in X that has been informed by knowledge of Y. In one sense, it is the “pure uncertainty” of X.

What seems newly surprising here is that the “amount of pure uncertainty” left over in Y is the same as the “amount of pure uncertainty” left over in X just discussed are exactly the same as each other! And that they are both the same as I(X;Y) – the mutual information of X and Y.

In fact, this sameness may be why these values are called “mutual information”. (Maybe a better term would have been “mutual uncertainty”.)

This relationship – that H(Y) – H(Y|X) = H(X) – H(X|Y) – says something else, too. It suggest that, in a joint distribution, providing extra information about Y has as much influence on X as providing extra information about X has on Y. There is a necessary symmetry between two chance variables in a joint distribution. We have already seen hints of this symmetry when we showed that X depends on Y if and only if Y depends on X.

Lets now turn our attention to row 9. It’s value is also the same as the value mutual information of X and Y. This means that:

$$H(X) + H(Y) – H(X, Y) = I(X;Y)$$

That is, another way to understand the mutual information of X and Y is that it is the sum of the entropies of both X and Y, reduced by the entropy of the joint distribution (X,Y). This fact also seems to lend credence to the choice of the term “mutual information” for this value.

Now we know that H(X,Y) is often greater than H(X) and also often greater than H(Y). But this relationship implies that H(X,Y) cannot be greater than H(X)+H(Y) together. (Otherwise, I(X;Y) could be negative sometimes.)

Another thing to notice is the fact that whenever H(X,Y) = H(X) + H(Y), then H(X) + H(Y) – H(X, Y) would have to be zero (0). In which case I(X;Y) would also = 0. But, whenever I(X;Y) = 0, we have already shown that X and Y are statistically independent. In other words, whenever the sum of the two composite entropies is equal to the joint entropy, then X and Y are stochastically independent.

Of course, all of these relationships are interesting and probably very useful. However, we have not really proved any of them. The proofs of all of the assertions made in this section, however, are simple exercises in algebra after applying the definitions of the various forms of entropy mentioned. Thus, we leave these proofs as an exercise for the reader. Also see [Cover and Thomas 1991; pp. 19 - 20].

Chain Rule for Joint Entropy

Joint entropy is equivalent to a number of alternate interpretations. We shall explore these in this section.

Again, we display the same table as before here for the purpose of reference proximity.

		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
1	**H(X)**	2.5850	2.5481	2.5850	2.5481	2.5481
2	**H(Y)**	2.5850	2.5632	2.5850	2.5632	2.5632
3	**H(X, Y)**	5.1699	5.1113	5.1264	4.9832	4.0057
4	**H(X|Y)**	2.5850	2.5481	2.5414	2.4200	1.4425
5	**H(Y|X)**	2.5850	2.5632	2.5414	2.4351	1.4577
6	**I(X:Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
7	**H(X) - H(X|Y)**	-0.0000	0.0000	0.0436	0.1281	1.1056
8	**H(Y) - H(Y|X)**	-0.0000	0.0000	0.0436	0.1281	1.1056
9	**H(X) + H(Y) - H(X,Y)**	0.0000	-0.0000	0.0436	0.1281	1.1056
10	**H(X) + H(Y|X)**	5.1699	5.1113	5.1264	4.9832	4.0057
11	**H(Y) + H(X|Y)**	5.1699	5.1113	5.1264	4.9832	4.0057

In this subsection we are going to focus on the fact that, for all five experiments, the values in rows 10 and 11 of the table are equal to each other as well as to row 3. For example, for Experiment 1 all three values are equal to 5.1699. For experiment 3 they are all equal to the value 5.1264. And for Experiment 4 they are all three equal the value 4.9832.

This fact suggests a general conclusion that, like in the previous subsection, we shall assert as being true for all joint probability distributions, but whose proof we shall leave to the reader. Also see [Cover and Thomas 1991; p. 20].

The assertions then are that:

$H(X) + H(Y|X) = H(X, Y)$, and that
$H(Y) + H(X|Y) = H(X, Y)$

The first relationship (row 10) implies that, together, H(X) and H(Y|X) “add up” to constitute H(X, Y). Another way to look at this is that the entire joint space (X,Y) can somehow be “partitioned” into two mutually exclusive parts as measured by: H(X) and H(Y|X).

The second relationship (row 11) implies that H(X, Y) can also be “partitioned” in a second way: H(Y) and H(X|Y). Together, rows 10 and 11 seem to be saying that the entire joint probability space can be “broken down” into two parts in two different ways. One way involves Y and the part of X that is informed by Y. The other way involves X and the parts of Y that are informed by X.

The Venn-like diagram in the following section will shed light on these relationships as well as those discussed in the two preceding subsections.

Multi-dimensional Joint Probability Spaces

Suppose that our experiments involving the two dice used 3 dice instead. What if they used 4 dice, or 10 dice, or 120 dice, or more?

The question that we ask in this section is

Can we still use all of the machinery that we have developed so far in Part II to manipulate these probability spaces of large dimensionality?

What We Have Achieved So Far

In our five example experiments, we have always been dealing with exactly two dice – the mammal die and the tree die. With these two dice, we were able to weave a quite rich theory of a joint probability space with two chance variables. We developed ideas such as (X,Y) the joint distribution, (Y|X) the conditional probability distribution of Y given X, (X|Y) the conditional probability of X given Y.

In addition, we built upon the notion of H(X) and H(Y), the entropies of the component distributions X and Y respectively, in order to develop generalizations of this entropy to help us to measure the degrees of uncertainty of these ideas related to joint distributions. These new "entropies" are H(X,Y) joint entropy, H(Y|X) and H(X|Y) conditional entropies, and I(X;Y). I(X;Y), the mutual information of X and Y, is an entropic functional that essentially measures the degree of stochastic dependence between the two chance variables of a joint probability space.

Mutual information between chance variables in joint distribution is a very important measure because it characterizes the degree of portent, or meaningfulness, between the two chance variables. Mutual information also measures the degree to which the two chance variables are indicators, or predictors, of each other – and is therefore of immense practical value.

Finally, we have seen that we can use various numerical relationships among these generalized entropies to characterize these entropies and their interrelationships.

This has been a pretty good start at investigating how the concept of entropy can be used to characterize probability spaces – which is the essence of information theory. In Part I, we concentrated on probability spaces that has exactly 1 chance variable, which we usually named "X". Then in Part II we have looked at probability spaces involving exactly two chance variables. In this scenario, the questions of 1) whether they are stochastically dependent or not, and 2) the extent to which they are stochastically dependent. And, we have also been able to leverage the concept of entropy to define a function that measures that degree of dependence between two chance variables – mutual information.

Joint Probability Spaces with Any Number of Chance Variables

We began this section by asking the following two questions:

- Suppose that our experiments involving the two dice used 3 dice instead. What if they used 4 dice, or 10 dice, or 120 dice, or more?
- Can we still use all of the machinery that we have developed so far in Part II to manipulate these probability spaces of large dimensionality?

We shall demonstrate here that the answer to the last question is yes.

Admittedly, things start to get pretty complex pretty fast as the number of chance variables (dimensions) involved gets larger. So, we shall concentrate on the simplest case – three chance variables. This will give us an idea as to what is involved increasing the number of dimensions past two.

However, there is no limit on the number of dimensions that can be used. In fact, we can even contemplate joint probability spaces with a countably infinite number of dimensions. And, in fact, we shall do just that in Part III of this primer.

But for this section, we shall stick with a finite number of chance variables (dimensions). In fact, we shall focus on a space with exactly 3 chance variables.

Our primary goal will be to describe how all of the concepts that we have so far contemplated for 2 dimensional probability spaces can be generalized (in natural ways) to apply to probability spaces of larger dimensions. Specifically, we shall see how we can generalize to larger numbers of dimensions the concepts joint sample pace, joint probability distribution, conditional probability distributions, joint entropy, conditional entropies, and mutual information.

We shall provide definitions for these concepts for any finite number of dimensions. However, it will be left to the reader should imagine how these concepts generalize to higher dimensions. However, we regard it as beyond the scope of this primer to exemplify probability spaces of higher numbers of chance variables than 3. These concepts are developed in more advanced texts such as [Cover and Thomas 1991] and the graduate seminar syllabus [Kleeman 2012].

Joint Distributions for 3 or More Chance Variables

In this section we are going to briefly describe joint probability spaces that use 3 chance variables instead of 2. We shall suggest the need for such species of joint distribution by the discussing chance games that use 3 or more dice.

However, due to the complexity of these constructs, we shall only describe, and not provide examples to the level of detail of our previous five 2-dice experiments. We are getting to the point where we shall have to resort to abstraction rather than specific, detailed examples.

Our focus will be on defining and describing joint and conditional distributions and on joint and conditional entropies as well as mutual information of information spaces involving 3 and more chance variables instead of merely 2.

We shall not take the time or space to present a thorough investigation of these 3-D joint sample spaces and probability distributions, nor their Venn diagrams. What we shall do instead is generalize from the 2-D joint information spaces, suggest what they portend regarding the 3-D and higher dimensional information spaces, and leave further considerations regarding the other relationships to a more advanced course in information theory.

We have been using capital letters near the end of the alphabet - X and Y – to identify the chance variables in a joint distribution. However, we shall need to identify perhaps a large number of chance variables for some of these joint spaces. So, it will be better if we begin to use a single indexed variable. So we shall use begin to use X_1, X_2, X_3, etc., to name our chance variables.

Of course, we shall continue to surround a comma-separated list of these indexed variables to name the joint probability space. Thus, instead of using “(X,Y)” to identify the 2-D space whose chance variables are X and Y, we shall use “(X_1, X_2)” instead. Of course, a joint space with 3 chance variables will be (X_1, X_2, X_3), of 4 will be (X_1, X_2, X_3, X_4), etc.

Likewise, for a single joint sample point of such a space, we shall use lower-case variables, for example, (x_1, x_2, x_3, x_4). Of course, for a specific member sample point of one of these spaces, constants will replace variables in this ordered-tuple naming scheme, as we have been doing with the joint probability space involving two dice.

For example, suppose we enhance our dice game with a third die – the "bird" die. Lets assume that the bird die has the following six faces: {sparrow, dove, lark, robin, raptor, eagle}. Of course, we are still using the mammal and the tree dice.

So, we shall rename the mammal die chance variable from X to X_1; the tree die chance variable from Y to X_2; and we shall name the bird die chance variable as X_3. Thus, the new 3-D joint probability space is named (X_1, X_2, X_3). Also, an example sample point from this space would be (lion, Lebanon Cedar, sparrow).

And, of course, we are interested in the following foundational measuring functions p(lion, Lebanon Cedar, sparrow), the probability of this sample point, and u(lion, Lebanon Cedar, sparrow), the uncertainty of this sample point.

Conditional Distributions for 3 or More Chance Variables

We have seen that, in the case of a joint probability space with two chance variables, say (X_1, X_2), that we can factor in certain "additional information" that we find out about one of the chance variables, say X_1, in order to reduce our level of uncertainty about the outcomes of the other chance variable X_2. This is called "conditional probability". Specifically in the example just described, this is called "the conditional probability of X_2 given X_1; and symbolized as $p(X_2|X_1)$. Similarly, we also defined $p(X_1|X_2)$, and thus had more than one way of applying "additional information" to reduce the degree of uncertainty of one of the chance variables.

In this section, we would like to extend the idea of conditional probability to 3 and more dimensional joint information spaces. Lets begin with the case of exactly 3 dimensions – 3 chance variables.

Kinds of Conditional Distributions in 3-D Information Space

Just as with the case of a 2-D information space, we have many ways that conditional probability can be defined. However, whereas we only had two ways to view conditional probability in a 2-D information space, we have many more ways than that to view conditional probability in a 3-D information space.

There are three of these ways that we shall be most interested in. All three of these describe the probability of one of the chance variables given two of the others. More accurately, all three of these consider the probability of one of the chance variables given the joint 2-D space involving the other two. For the space (X_1, X_2, X_3), these three kinds of conditional probability are:

$p(X_1,|(X_2, X_3))$
$p(X_2,|(X_1, X_3))$
$p(X_3,|(X_1, X_2))$

We can characterize these three as "the conditional probability of one of the chance variables given values of the joint space of all the others".

In addition, we shall also be interested in the conditional probabilities of all of the ways that we can consider the probability of one of these 3 chance variables given the outcome of another. There are six of these, which are:

$p(X_3|X_2)$
$p(X_3|X_1)$
$p(X_2|X_1)$
$p(X_2|X_3)$
$p(X_1|X_3)$
$p(X_1|X_2)$

All six of these actually ignore one of the three chance variables and work with the other two. For higher numbers of dimensions, this idea generalizes to selecting two of the chance variables at a time and ignoring all if the remainder for that case. Every possible way of taking the chance variables two at a time constitute this generalization.

A specialization of this way of forming conditional probabilities in information spaces with more than 2 chance variables is to form the conditional probability of any selected chance variable having been given the other chance variable whose index number is one less than the first. In 4-D space, for example, this type includes:

$p(X_4|X_3)$,
$p(X_3|X_2)$, and
$p(X_2|X_1)$

In general this translates to:

$p(X_i|X_{i-1})$

The main application of this latter form of conditional probability is to time series. A time series is a sequence of chance variables that occur one after the other at distinct point in time. A more frequently used term for a time series is *stochastic process*. More formally, a stochastic process is defined[19] as a "sequence of chance variables".

In other words, our definitions of conditional probabilities do not require that all of the chance variables involved have to take place simultaneously. As well, they may occur at different points in time. When they do, however, associate an ordering to these chance variables, and stipulate that the "earlier" variables in the order occur prior, in time, to the chance variables in "later" variables. For example, in the 4-D stochastic process (X_1, X_2, X_3, X_4), X_1 occurs first in time, followed by X_2, X_3, and X_1, in that time order. So, a stochastic process is represented by a joint probability space, where the chance variables involved are ordered in time.

Often, when dealing with stochastic processes, we have need to consider what will happen in some time step "n" when we know what happened in time step n-1. These kinds of questions are modeled well by using the general form of conditional probability, "$p(X_i|X_{i-1})$", that we just discussed - where only two chance variables are involved, the two chance variables represent contiguous time steps, and the second (X_i) is conditioned by the first (X_{i-1}).

This situation is useful if the outcome of an event occurring in one time step portends the probabilities of the event that occurs at the very next time step. For example, suppose we are playing poker, and a card is dealt to each player in such a way that all other players can see the face of the card. Then this information tells the players something about which cards are no longer in the card deck to be dealt on the next dealing. That is, the information about the outcome of one time step has an effect on,

[19] In most probability theory texts, the term *stochastic process* is defined as "a sequence of random variables". The reader will recall that we have defined "random variable" as a special case of "chance variables" in which each sample point is associated with a real number *value* – as well as being associated with a probability. We are defining stochastic processes here in terms of *chance variables* instead of *random variables* because we want to be more general. That is, our definition includes both *chance variables* and their special case *random variables*. There is no need to require stochastic processes to only allow random variables. That is, there is no requirement for the chance variables to be mapped to a real number value, but it is acceptable if they are.

alters, the probabilities of the next time step. This is the kind of situation described by the conditional probability "$p(X_i|X_{i-1})$".

We shall concentrate on this kind of situation in Part III, which investigates the possibility of predicting the outcomes of chance variables in time. We can characterize these two as "the conditional probability of the chance variable with index i given a value of the chance variable with index i-1".

The reader can see that there are many other possible conditional probabilities that can be defined on the joint space (X_1, X_2, X_3). However, we shall confine our interests to these three.

Defining Conditional Distributions in 3-D

Let us now look at mathematically defining these five examples of conditional probability on space (X_1, X_2, X_3).

We can define any of these conditional distributions by applying the basic definition of conditional probability, which, as we defined it earlier in the section on conditional probability, is:

$p(B|A) = p(a \wedge b)/p(a)$ for all $a \in A$, $b \in B$.

In order to define any of these 3-D version of conditional probability, we apply this general formula, with the expression prior to the "|" symbol substituting for B and the expression after the "|" symbol substituting for A.

Applying this to $p(X_3,|(X_1, X_2))$, for example, we get:

$$p(X_3,|(X_1, X_2)) = p(X_3 \mid (X_1, X_2)\ ) = p(\ X_3 \wedge (X_1, X_2))/p(X_1, X_2)$$

Similarly, we also obtain:

$$p(X_1,|(X_2, X_3)) = p(\ X_1 \wedge (X_2, X_3))/p(X_2, X_3), \text{ and}$$
$$p(X_2,|(X_1, X_3)) = p(\ X_2 \wedge (X_1, X_3))/p(X_1, X_3)$$

For any of the conditional probabilities identified above that involve on two chance variables, including

$p(X_4|X_3)$,
$p(X_3|X_2)$, and
$p(X_2|X_1)$

the definition is clear. For example, consider

$p(X_4|X_3) = p(X_4 \wedge X_3)/p(X_3)$

Stochastic Independence for 3 or More Chance Variables

Recall that for two chance variables X_1 and X_2, we say that X_1 and X_2, are stochastically independent if and only if $p(x_1 \wedge x_2) = p(x_1\ )*p(x_2,)$ for all $x_1 \in X_1$, $x_2 \in X_2$.

This express was derived from the fact that saying that X_1 and X_1 are stochastically independent by definition means that $p(X_2|X_1)$ is the same value as $p(X_2)$ – since the "new information" given about X_1 makes does not change the value of $p(X_2)$.

But, if $p(X_2|X_1) = p(X_2)$, then one can derive the fact that $p(x_1 \wedge x_2) = p(x_1\ )*p(x_2,)$ for all $x_1 \in X_1$, $x_2 \in X_2$. We also showed that whenever X_2 is stochastically independent of X_1,

that it necessarily follows that X_1 is stochastically independent of X_2. Thus, its is reasonable to say, then, that X_1 and X_2 are stochastically independent of each other.

Consequently, chance variables X_1 and X_2 are stochastically independent of each other if and only if $p(x_1 \wedge x_2) = p(x_1)*p(x_2,)$ for all $x_1 \in X_1$, $x_2 \in X_2$.

This articulation of stochastic independence is easy to generalize to information spaces that involve any number of chance variables in the following manner:

Definition: Stochastic independence of two or more chance variables: Let $X_1, X_2,\ldots,X_n$, be n chance variables involved in joint probability space $(X_1, X_2,\ldots,X_n)$. Then, $X_1, X_2,\ldots,X_n$ are said to be stochastically independent if and only if

$$p(x_1 \wedge x_2 \wedge\ldots\wedge x_2) = p(x_1)*p(x_2)*\ldots* p(x_n) \text{ for all } x_1 \in X_1, x_2 \in X_2,\ldots, x_{2n} \in X_n.$$

Stochastic Dependence for 3 or More Chance Variables

Definition: Stochastic dependence of two or more chance variables: Let $X_1, X_2,\ldots,X_n$, be n chance variables involved in joint probability space $(X_1, X_2,\ldots,X_n)$. Then, $X_1, X_2,\ldots,X_n$ are said to be stochastically dependent if and only if they are not stochastically independent.

In other words, $X_1, X_2,\ldots,X_n$ are stochastically dependent if and only if there exists at least one $x_1 \wedge x_2 \wedge\ldots\wedge x_2$ in $(X_1, X_2,\ldots,X_n)$ where

$p(x_1 \wedge x_2 \wedge\ldots\wedge x_2) \neq p(x_1)*p(x_2)*\ldots* p(x_n)$.

As with joint probability spaces with exactly two chance variables, stochastic dependence is ultimately the “backbone” of information theory. We have seen that stochastic dependency comes in varying degrees, from none to much. Moreover, the degree of stochastic dependency is measured by an *entropic functional* of the probability distribution named *mutual information* – which we shall extend to joint information spaces of any number of dimensions below.

Joint Entropy in for 3 or More Chance Variables

We shall now generalize the notion of joint entropy to information spaces whose dimensionality (number of chance variables) is greater than two.

Recall that we defined the joint entropy of a 2-D joint probability space (with a renaming of chance variables) as

$$H(X_1, X_2) = -\sum_{i,j\in S} p(x_{1i}, y_{2j})*\log(p(x_{1i}, y_{2j})).$$

Expressed more generally as a function of u(x), this can be expressed as

$$H(X_1, X_2) = \sum_{s\in S} p(s)*u(s), \text{ where s is a sample point in joint space } (X_1, X_2).$$

Preferring the later form, we shall now generalize joint entropy for n chance variables as:

$$H(X_1, X_2,\ldots, X_n) = \sum_{s\in S} p(s)*u(s), \text{ where s is a sample point in joint space } (X_1, X_2,\ldots, X_n).$$

Conditional Entropy for 3 or More Chance Variables

We discussed a number of possible ways to define conditional probability for an n-dimensional information space above. Each of these has its corresponding version of conditional entropy for an n-dimensional information space. We shall not present all of these here. Rather we shall choose a version of conditional probability for a quite general case – a case that we shall find useful in Part III.

Consider the following conditional probability distribution of an n-dimensional information space:

$p(X_n|(X_1, X_2,\ldots, X_{n-1}))$

We shall define the conditional entropy of this conditional distribution as follows:

$$H(X_n|(X_1, X_2,\ldots, X_{n-1})) = -\sum_{i,j\in S} p(x_{1i}, y_{2j})*\log((p(x_n) \wedge p(x_{n-1}, x_{n-2},\ldots, x_1))/ p(x_{n-1}, x_{n-2},\ldots, x_1))$$

We can express this as a more general function of u as follows:

$$H(X_n|(X_1, X_2,\ldots, X_{n-1})) = \sum p(x_n, x_{n-1},\ldots, x_1)*u(x_n| x_{n-1}, x_{n-2},\ldots, x_1)$$

Mutual Information in for 3 or More Chance Variables

The mutual information of two chance variables X and Y in a 2-dimensional joint distribution measures the degree of stochastic dependence of the two chance variables.

This idea can be generalized to a measure of the mutual degree of stochastic independence of n chance variables in an n-dimensional joint distribution. We shall define this generalization in this section.

We have defined the mutual information of the two chance variables of a joint probability space in a couple of different, but mathematically equivalent, ways. The first way accommodated a direct understanding behind what is being measured. The second way provided a simpler calculation, but obscured the rationale behind it.

We shall present both of these again in this section, and then go on to generalize both to information spaces of n-dimensions.

The Intuitive Definition

The intuitive definition directly measures the mean, or expected value, of the difference in uncertainty calculations for every sample point in the sample space according to two distinct joint probability distributions. The first of these probability distributions is the one in that is being measured – called “p”. The second of these probability distributions is the only one on the same sample space that is stochastically independent.

Consequently, we can characterize the mutual information of (X, Y) with joint distribution p as the mean difference in uncertainty between $u_0(x,y)$ and $u_p(x,y)$ over all sample points (x,y) of the joint sample space.

This means that we have defined $I(X,Y) = \sum_{(x,y)\in S} p(x,y)*[u_0(x,y) - u_p(x,y)]$

This is easily generalizable to a joint space of n dimensions, as follows:

$$I(X_1; X_2;\ldots; X_n) = \sum_{(x1, x2,\ldots, xn)\in S} p(x_1, x_2,\ldots, x_n)*[u_0(x_1, x_2,\ldots, x_n) - u_p(x_1, x_2,\ldots, x_n)]$$

The Practical Definition

The practical definition of I(X;Y) that we gave above for an information space involving exactly two chance variables X and Y, and that is given in most text books on information theory is:

$$I(X;Y) = \sum_{(x,y)\in S} p(x,y)*\log(p(x,y)/(p(x)*p(y)))$$

This is easily generalized to an n-dimensional information space as follows:

$$I(X_1; X_2;\ldots; X_n) = \sum_{(x,y)\in S} p(x_1, x_2,\ldots, x_n)*\log(p(x_1, x_2,\ldots, x_n)/(p(x_1)*p(x_2)*\ldots*p(x_n))$$

Summary and Conclusions of Part II

Part II of this information theory primer has been concerned with how to characterize the extent two which two chance variables operating together jointly portend something meaningful about each other's outcomes.

We have shown that this degree of portent or meaningfulness between these two jointly operating chance variables is due to the extent to which the outcome of one of these chance variables is dependent upon the outcome of the other chance variable within the joint outcome.

The test for whether this dependence exits between two such chance variables is the extent to which a specific occurrence of one of these variables specifies the probabilities of the other variable within the joint occurrence.

In other words, if the probability distribution of the outcomes of one of the variables in the joint event depends upon (is peculiar to), or is conditioned upon, what the outcome of the other variable is, then the second chance variable is stochastically dependent on the first.

In such a situation, the probability distributions for the outcomes of the second variable are in general different – given any of the outcomes of the first chance variable. Each of these specific probability distributions is called the conditional probability distribution for the second chance variable, given the specified outcome of the first chance variable. This distribution is symbolized as p(Y|X=x).

If the conditional distributions p(Y|X=x) for the all of the outcomes of the first chance variable are place in within the same matrix, with each row being one of the p(Y|X=x) conditional distributions, then the resulting matrix containing all of these distribution is called the conditional probability distribution for Y given X, and is symbolized (Y|X).

If all of the p(Y|X=x) conditional distributions are exactly the same as each other, then the probabilities for the outcomes of Y do not depend upon the which one of the X sample points was the outcome for X. In such a case, X and Y are stochastically independent. And neither X nor Y portends anything about the other.

But things are more interesting if at least some (and maybe all) of the p(Y|X=x) conditional distributions are distinct. This says that which probability distribution to use for Y depends upon which outcome of X was realized. This situation is the foundation for most of what is interesting in information theory. However, we shall treat the stochastic independence case as the case of their being "zero" amount of stochastic dependence. So, everything in information theory can be treated as stochastically dependent. It just that the stochastically independent case is treated as being "zero" amount of stochastic dependence. Every other joint distribution has some positive amount of stochastic dependence.

It turns out that this conditional distribution (Y|X), and its related distribution (X|Y), is all that is necessary to define anything we want to know concerning the stochastic dependence of the two chance variables X and Y. We have suggested that there are two other interpretations of this stochastic dependence between X and Y. The first is whether X portends something about Y, or vice versa. The second is whether knowledge of the outcome of X is meaningful to the outcome of Y, or vice versa.

We have then gone further to suggest that there is a way to use the conditional probability distributions (Y|X) and (X|Y) to measure the degree to which X and Y are stochastically dependent (or meaningful to each other).

We have answered this question in the affirmative, and then proceeded to develop a measure of the degree of stochastic dependence between two jointly related chance variables. The approach that we took to develop such a measure was to once again apply the concept of *entropic measures*, or *entropic functionals*, that we developed in Part I.

This time, we developed a specific entropic measure named the *mutual information between chance variables X and Y with respect to joint probability distribution p*. This is symbolized by $I_p(X;Y)$. Mutual information essentially looks at the actual joint distribution between X and Y – which we called $p_K(X,Y)$ – and compares it to the joint distribution between X and Y if the two variables were stochastically independent - which we called $p_0(X,Y)$. Mutual information then measures the difference between these two situations. The "further away" $p_K(X,Y)$ is from $p_0(X,Y)$, the "more stochastically dependent" is $p_K(X,Y)$ measured to be.

To be more precise, what the formula for mutual information actually does, for each sample point, is measure the "amount of uncertainty with respect to $p_K(X,Y)$" and subtracts it from the "amount of uncertainty with respect to $p_0(X,Y)$". It then takes the mean of all of these "differences in uncertainties" across all of the sample points in the sample space. This mean difference in uncertainties is, then, the mutual information of the joint distribution X and Y with respect to joint distribution $p_K(X,Y)$ - and is symbolized by $I_K(X;Y)$.

Part III: Prediction: The Meaningfulness of Uncertainty in Time

<>

Epilogue: The Mathematics of the Unknown

Information theory is the branch of mathematics that can breach the boundary between the known and the unknown.

Standing upon the shoulders of probability theory and the theory of stochastic processes, information theory can pierce the established limits of certainty where determinism is stopped at the gate and advance hesitantly into the vast reaches of the unknown.

But information theory cannot roam the intellectually inaccessible reaches of knowing that only the mystics dare to tread. Its itinerary in this space is restrained. It must carry with it the gear and detritus of, at least, some certainty.

There are two particulars that information theory must know. First, it must know some mutually exclusive, exhaustive and mathematically measurable set of possibilities – a sample space. And secondly, it must have at least one, and possibly two, assignments of likelihood of its members – their probabilities.

Thirdly, it must also assign a measure of the uncertainty of those members. But it can derive that measure on its own – after passing through the boundary to the territory of the uncertain.

But these two passengers should not be considered lightly. It is quite a lot to ask of a system of knowledge to pack both a sample space and a probability distribution – and a machine to convert probabilities into uncertainties as well. It is muscular machinery, sturdy intellectual equipment, that information theory must bear as it traverses the bounds of the certain into the regions of the unknown.

And, although it does not have a lexicon for the elements that it finds, it can nevertheless ascertain the portent of one for another – whether the portent exists, and also its degree if it does. And, this ability enables it to sometimes assert predictions in time – which can be handy indeed in the territory of the unknown.

And the vast regions beyond the horizons to the west into which it cannot enter must humble it, indeed.

Appendix 1: The Uncertainty of an Event

In Part I, we presented a function, "u(x)", that measures the amount of uncertainty inherent in any single event of a probability distribution.

The development of this measure lies at the very heart of information theory. Most every consideration that information theory makes after the introduction of u(x) is defined in terms of it.

Even entropy itself – which is the essential measure of information theory and which measures the degree of uncertainty of an entire probability distribution - is defined as the average of u(x) over all sample points of the distribution. Thus, pretty much every significant idea in information theory is defined in terms of entropy. So, u(x) is the most fundamental concept that information theory adds to probability theory.

In Part I, we introduced the definition of u(x) and described it as "the uncertainty of an *event* of a probability space". But, we have also been talking about the "uncertainty of a sample point". The same thing is true of the very notion of probability as well. Sometimes we talk about "the probability of an *event*", and at other times, we speak of "the probability of a *sample point*".

You will recall that there *is* a difference between a *sample point* and an *event*. An *event* is a set of sample points of the same sample space. And therefore a sample point by itself is *not* an *event*. So, there is an inconsistency in this language, and we need to clear up at this time. So, which is it – the probability of a *sample point* or of an *event*? And, which is it – the uncertainty of a *sample point* or the uncertainty of an *event*? We shall clear this issue up for both probabilities and for uncertainties right now.

Recall that we formally defined *probability* as a measure of "events" – rather than of sample points. However, if instead of a single sample point "x", we are willing to consider the set whose only element is x, then such a singleton set would, in fact, be an event because it is a set of sample points. This set would be symbolized as {x} – meaning the set whose only element is the sample point x.

Therefore, if x is a sample point, then "p(x)" really means "p({x})"; and "u(x)" really means "u({x})". This means that both *probability* and *uncertainty* are functions that are defined on *events*. They both act on an event and map it to some real number. That is, they *measure* the event. But they each measure different *aspects* of the event. *Probability* measures the *likelihood* of the event; while *uncertainty* measures the uncertainty of the event.

Anyway, in Part I, we showed how to calculate the *uncertainty of an event* u(x). But, lacking in Part I was any explanation of why this measure is defined the way it is, and how that definition fits with an intuition of what a *measure of uncertainty* ought to be. For any reader who wants to understand information theory at a "gut level", this kind of intuitive understanding is important. This appendix has been provided for such readers.

Most treatises on information theory begin with *entropy* as their starting point. However, this author finds it much simpler and more intuitive to begin with the simpler idea of "the measure of the uncertainty of a single sample point or event", and then define entropy in terms of it. The main reason for this is that it is easy to appeal to intuition when defining a measure of uncertainty for a single sample point.

Subsequently defining entropy as the average (mean) of that measure follows intuitively also. Until one understands that entropy is simply the mean of a simpler

idea, it can be difficult to comprehend what the definition of entropy is proposing to measure.

How can Uncertainty be defined?

We shall first consider the question of what is a reasonable approach to defining such a function, and attempt to find a definition that fits in some way with our intuition.

This question is very old, and has been asked as early as the time of the ancient Greeks. Aristotle, for example, proposed that *uncertainty depends on likelihood*. In fact, it is reasonable, according to Aristotle [Vedral 2010], to say "the more unlikely, the more uncertain." Stating this idea in terms of "likelihood" (rather than unlikelihood), it follows that

"Uncertainty increases as likelihood decreases".

In other words, as *likelihood* gets smaller, *uncertainty* gets larger. And, conversely, as *likelihood* gets larger, *uncertainty* gets smaller. So, likelihood and uncertainty enjoy an *inverse relationship*. When one goes up the other goes down, and conversely.

Another issue is, "What are we describing as exhibiting these various degrees of uncertainty?" Essentially it is some happening, or event whose outcome we do not yet know.

Adding this idea of "what" we are describing as "uncertain" to Aristotle's meaning we obtain:

As the likelihood of a happening decreases, the uncertainty of that happening increases.

Fortunately, all of the ideas just discussed – "likelihood" and "happenings" have correlates in probability theory, where "likelihood" is equivalent to *probability*, and "happenings" are equivalent to *events*.

Substituting "event" for "happening" we get:

As the likelihood of an event decreases, the uncertainty of that event increases.

In fact, it seems that probability theory is tailor made to address the question of "how to define uncertainty".

Of course, in probability theory, we defined the notion of events in terms of a more primitive concept: that of the sample points and sample space of a probability distribution. When we defined probability space above, we said that the set of events E were defined as the set of logical combinations of subsets of the sample space, where the singleton set of every sample point is an event in E.

In other words, we can start out by assigning probabilities to the sample points and obtain probabilities of all of the events in E from there.

In conclusion, we can mathematically develop a measure of the amount of uncertainty in a sample point by embellishing Aristotle's above statement. In order to use probability theory, we shall substitute the word "probability" for the word "likelihood" and the word "event" for the word "happening". We also use the fact that by "event" we mean a set of sample points of the same population, or sample space.

As the probability of an event decreases, the amount of uncertainty of that enent increases.

So, we have articulated Aristotle's idea of the inverse relationship between likelihood and uncertainty into a statement that is almost ready to compose into a mathematical definition.

In probability theory, it is true that we deal with events. But these events are defined as collections of sample points. And, as we have already discussed, we can treat a single sample point as a set whose only member is that sample point. And in treating a sample point this way, we are able to be consistent with the fact that both *probability* and *uncertainty* are functions that act upon *events*, rather than single sample points.

We now have the basic elements that probability theory can work with: events (and sample points) and probabilities. And we know the basic relationship between them: When probability of a sample points goes up, the uncertainty of that sample point goes down, and conversely.

But what we have not yet establishes is "by how much?" That is, we still need to know "How much does uncertainty go up when probability goes down", and "How much does uncertainty go down when probability goes up?"

We shall consider this issue in the next subsection.

How Fast should Uncertainty Rise as Probability Falls?

So far, we have established that defining our uncertainty function u(x) as varying inversely as probability seems to capture our intuition about how these two measures change.

But we have not yet answered the question of "how much" does uncertainty go up as probability goes down, and vice versa. We have to answer this question of "how much" in order to precisely define our u(x) function.

And, of course, it is possible to define u(x) in a number of different ways, all of which preserve the idea the uncertainty is inversely related to probability. Some of these ways will have uncertainty rising and falling relatively fast as compared with probability; while others of these ways will have uncertainty rising and falling relatively slowly as compared with probability.

We need to look at some candidate functions to get a feeling for what is involved, and to get a better idea of which of these function will suit our needs the best. What we are going to do in this section is to define some candidate function definitions, and see which, if any, we like for the job of defining our uncertainty function. All three of these candidate functions will provide an inverse relationship between uncertainty and probability. Where they all differ is in "how fast" uncertainty goes up and down as probability goes down and up.

Lets try as our first candidate functional definition of *uncertainty* the simple relationship:

$$u'(x) = 1/p(x)$$

In other words, we shall temporarily define the *uncertainty of sample point x* is defined as the inverse of p(x) – and see how we like the results of that attempt.

To test out this candidate definition of u'(x), lets look at some of the values that it would produce. We could look at any set of probabilities. But, for this test, lets look at fractions of the powers of ten as they get smaller and smaller, and then compare them to their inverses:

x:	1	1/10	1/100	1/1000	1/10000	1/100000	1/1000000…

u'(x):	1	10	100	1000	10000	100000	1000000…

The problem with this attempt is that the u'(x) values rise too fast as the x values drop. In fact, the x values drop much slower. That is, the distance between the rising u'(x) values seems excessive as compared with the dropping speed of the x values.

Lets look at another, similar set of values and apply them also to the same relationship u'(x) = 1/p(x). This time, lets look at fractions of the powers of 2 as they get smaller and smaller and see what happens to their inverses.

x:	1	1/2	1/4	1/8	1/16	1/32	1/64…
u'(x):	1	2	4	8	16	32	64

Admittedly, the fractions of the powers of two do not rise as "wildly out of control" as the fractions of the powers of ten. However, they still rise too fast for some applications. It might be more convenient if we could "slow down" these u(x)'s as the x's increase.

Here's an idea that would accomplish just that. Instead of using 1/p(x), suppose we used "log(1/p(x))"? That is, take the logarithm of 1/p(x) instead. That should "slow down" u(x) as x gets smaller. And at the same time, we are still preserving the inverse relationship between uncertainty and probability. That is, with this new candidate function, it is still true that as probability goes down, uncertainty goes up, and vice versa.

In fact using a log base of 10 with our first example will be easy – so lets do that. Therefore, our second candidate definition of u(x) will be:

$u_{10}(x) = \log_{10}(1/p(x))$

If we re-apply this definition to our first example involving the fractions of the powers of 10, we obtain:

x:	1	1/10	1/100	1/1000	1/10000	1/100000	1/1000000…
$u_{10}(x)$:	0	1	2	3	4	5	6…

Of course, we can use this same definition against the second example involving the fractions of the powers of 2, but the answers would not come out as whole positive integers. Rather, we would have to use a logarithm table or calculator to derive the results. The results would be positive numbers, but not integers.

We would achieve the same type of result if we used a log base other than 10 – but the result would rise at a different "speed". For example, if we used a log base of 2. Of course, using a log base of 2 against the second example would be easy to calculate, as follows, and would result in the following definition:

$u_2(x) = \log_2(1/p(x))$

If we re-apply this definition to our second example involving the fractions of the powers of 2 (because it is so easy to calculate, we obtain:

x:	1	1/2	1/4	1/8	1/16	1/32	1/64…
$u_2(x)$:	0	1	2	3	4	5	6…

Of course, we can use this same definition against the first example involving the fractions of the powers of 10, but the answers would not come out as whole positive integers. Rather, we would have to use a logarithm table or calculator to derive the results. The results would be positive numbers, but not integers.

In fact, we could use the logarithm to any positive base to define u(x) and we would achieve the same kind of effect that we did by using log bases of 2 and 10. The only difference is the "speed" with which u(x) would rise for a given set of x probabilities.

Anyway, it seems that we have arrived at one approach to defining u(x) so that the value of u(x) would "slow down" the way that we want it to as p(x) gets smaller: taking the logarithm (to some log base) of the inverse of the probability of x.

Of course, there are many other approaches that we could take to "slow down" the result of u(x). Taking the logarithm of the inverse of the probability is only one way. However, as we shall see in the next section, there is another very good reason why this might be the best approach.

What Do We Require of Our Measuring Function?

We have said that we want our measure u(x) of sample point x to vary inversely as its probability.

But we want more than this. We want u(x) to exhibit any other properties that we would expect of any kind of measuring function.

In this section, we shall lay out what the attributes of any measuring function ought to be, and then consider how u(x) must be defined so that it satisfies those attributes.

Subsequently, in a following section, we shall again propose some candidate definitions for u(x) and inspect each candidate to see if it satisfies all of these attributes. From this inspection we shall select the definition of u(x) used by information theory.

There are three attributes that we want any measuring function to exhibit:

1. Measures attribute of interest: The measure should exhibit the right behavior for what we are using it to measure.
2. A measure of zero: The measure should assign a value of zero to the "things" whose "size" should be zero.
3. Additivity for Certain Events: For certain pairs of things being measured, the sum of the measures of the pair should be equal to the measure of a thing that is created by joining the pair. This is not true for "things that overlap". But it should be true for "things that are disjoint".

We shall look at these three attributes more closely now, and decide how they apply to probability spaces.

Measures attribute of interest

We have already established that we want this measure of uncertainty u(x) to "behave" like Aristotle suggested. That is, when the probability of a sample point is relatively large, we want its measure of uncertainty to be relatively small. And when the probability of a sample point is relatively small, we want its measure of uncertainty to be relatively large.

In other words, we want the measure of the uncertainty of events to vary inversely as the probability of those events. Note that this implies that if $p(x) < p(y)$, then $u(x) > u(y)$.

We have not yet decided what the scale of increase of decrease should be for our definition of uncertainty. We shall get to that issue in the next subsection.

However, we do know that since u(x) should vary inversely as probability p(x), then "p(x)" will be appear as an input variable somewhere in the expression that ultimately use to define u(x). And, it may well be that p(x) is the only input variable in that expression.

A Measure of Zero

There are a couple of other features that we want any measuring function to have, including u(x). The first is that a "nothing thing" should have a measure of zero (0). In our case, the "things" that we are measuring are sample points and events consisting of sets of those sample points.

Now, if a sample point or event has no chance of happening, then we would like to make sure that its measure of uncertainty is zero. This is true because if a sample point is impossible and has no chance of happening, then its *certainty* is assured! That is, we are *absolutely certain* that it will not happen! Since its occurance has absolute certainty, then it has no uncertainty, and the measure of that uncertainty should be zero.

Also, if the probability of a sample point is one (1), then it is certain to happen! Thus, its measure of uncertainty should also be zero.

In conclusion, the function that we are defining for measuring the uncertainty of a sample point should produce a zero (0) value in two cases: 1) when the probability of the sample point in question is zero, and 2) when the probability of the sample point in question is one (1). In all other cases, $0 < p(x) < 1$, the uncertainty of x, u(x), must be positive, because some amount of uncertainty exists regarding x in those situations.

Additivity

There is usually some notion of joining two of the "things to be measured" in such a way that this jointing results in a single "thing" that can also be measured. As long as the two things that are joined do not "overlap" to begin with, then we would like our measuring function to work in a manner that produces a measure for the "joint thing" that is the sum of the measures of the two "things" before they were joined.

For example, if we measure the length of two boards – using a meter stick - and then join them by gluing them together end-to-end, then the measure of the resulting board should be the sum of the measures of the initial two board. However, if instead of gluing them end-to-end, we overlap them and nail them together in the overlap; then the new joined boards entity would have a length that is less than the sum of the two initial lengths.

In other words, "additivity" should apply in the case that we join the two initial boards in a way that does *not* involve overlapping them. However, additivity is not expected to apply if we joint the boards in a way that involves overlapping them. This is additivity of our measure of "board length" using a meter stick.

From this example, we can see that we do not expect our measures to be additive whenever they are "overlapped" in the process of joining them. But we do expect them to be additive when they are not overlapped.

This attribute of a measuring function is often stated, "The whole is equal to the sum of its parts". While this relationship is not true of everything, we usually do like for it to be

true of our measuring functions when no “overlapping” is involved. This feature is also given a name. Sometimes it is called *additivity* and sometimes it is called *linearity*.

The question is, then, in our realm of probability spaces, is there some way that we can “join” two “non-overlapping probabilistic entities” in such a way that the sum of the uncertainties of those initial two non-overlapping entities is equal to the measure of uncertainty of the jointed probabilistic entity? And if so, then what *are* those entities?

The answer to the first question is “Yes”. There is some way that we can “join” two non-overlapping probability space entities that results in the uncertainty measure of the joint probability entity being equal to the sum of the uncertainty measures of the two probability space entities.

The second question is “What are those probability entities?” Lets preface the answer by pointing out that there are three entities involved in this answer. The first two are the “non-overlapping” entities that are being joined. The third is the resulting joint entity.

But what kind of entities are they? The answer is that the first entity is an event from one sample space named X – lets call it event x. The second is an event from another (possibly the same) sample space Y – lets call it event y. The third entity, the *joint entity*, is the pair (x, y) – which is from a third sample space that is called the *joint sample space (X, Y)*.

Since X, Y and (X, Y) are three sample spaces from three probability spaces then the all have probability distributions, which we shall call p(x), p(y) and p(x, y).

Think of this as being similar to our situation of joining boards. With the boards, we have two different supplies of boards – supply X and supply Y. We can join a pair of boards by taking board x from supply X and board y from supply Y. The result is a space of joined board, (X, Y).

To keep the analogy like the one from the joint probability spaces, we will have to say that every possible pair of boards x and y already have a “joining plan” that is predetermined. Some of the pairs are predefined to overlap by varying degrees and others are predefined to not overlap at all when they are joined. This predetermined joining plan is analogous to a *probability distribution of the joint probability space (X, Y)*. It is this joint probability distribution that determines which pairs of events (x, y) are “overlapped” and by how much.

Of course, each of the initial supplies of boards, X and Y, has its own probability distribution. Moreover, the probability assignments of the joint distribution p(x, y) are somewhat determined by the probability distributions of X and of Y – but not totally so. There are many possible joint probability distributions for (X, Y) all of which comply with the constraints imposed by the probabilities of X and the probabilities of Y.

Now lets look at how the “additivity” property translates to this analogy. In the “board analogy” we said that additivity means that the sum of the lengths of the two boards is the same as the length of the joined board whenever the two boards do not overlap. On the other hand, the sum of the length of the two boards is less than the length of the joined board whenever the two boards *do* overlap.

For the joint probability space situation, the results will be analogous. Additivity means that the sum of the measures of the two events x and y is the same as the measure of the joined event whenever the two initial events do not overlap. On the other hand, the sum of the measures of the two events x and y is less than the measure of the joined event whenever the two initial events *do* overlap.

The only question that remains for this analogy is “What does it mean for two events x and y to ‘overlap’?”

We shall spend all of Part II of this primer answering and exploring the consequences of this question. But for the time being, let me “relieve the suspense” and spill the answer. Whenever the two initial events x and y are *stochastically independent* (also known as *statistically independent*) do we require that *additivity* be preserved by u(x).

Summary of What We Require in Our Measure of Uncertainty

So, in the process of defining our u(x) measure, we shall have to make sure that it exhibits all of the following properties.

First, in keeping with Aristotle’s observation about “uncertainty varying inversely with probability”, the u(x) must be high when p(x) is low, and conversely. Whatever mathematical expression that we invent to define u(x) must ensure this relationship.

Second, we need to investigate whether probability spaces have “things to be measured” – namely sample points and collections of sample points (events) - that should have an uncertainty measure of zero. Any such “things” would be ones that “have no uncertainty”.

In other words, these sample points or events must be absolutely certain. We have already decided that there are two cases that are in fact absolutely certain. One of them is when the probability of the sample point is 1. This occurs when p(x) = 1 for sample point x. the other case is when p(x) = 0. In this case it is absolutely certain that sample point x will not occur. Therefore, the measure of sample points whose probabilities are either 0 or 1 should have a measure of uncertainty of zero. For these, wee require that u(x) = 0.

Thirdly, we must find some way of “joining” sample points (or events) so that the measure of the joint event is equal to the sum of the measures of the constituent events that are joined – at least for pairs of events that, in some sense, do not overlap. However, if the two constituent events *are* in some manner “overlapped”, then we would not expect equality here. Thus, in addition to identifying some form of “joining” of events, we must also identify some notion of “overlap” of pairs of joint events.

Moreover, these ideas of “joining” and of “overlap” must be useful and intuitive in the realm of probability spaces. And finally, we must define the “joining” as a mathematical relationship that provides all of these features. We have given a preview of what this “way of joining” is and of what this “manner of overlapping” is.

Considerations for Defining the Uncertainty of a Sample Point

We have shown that any measuring function should have the following three traits if possible: 1) measures attribute of interest, 2) possesses a zero measure and 3) exhibits *additivity* for “non-overlapping” joined (or *joint*) events.

In the three sections that follow, we shall consider three potential candidate definitions for this measure of uncertainty. Then we shall select the best one – which of course is the one used by information theory.

The criteria that determines which one “the best” will be the extent to which it satisfies all three of these desired properties. At the outset, there seems to be no reason why it is not possible to find (or invent) several such measuring functions, all of which exhibit all three of these desired properties. However, we shall soon see that finding such a measure of uncertainty is not so easy, and that there are not so many of them as one might expect.

Measures attribute of interest

The essential attribute that interests us in any measure of uncertainty – as pointed out by Aristotle – is that of "the inverse relationship between probability and uncertainty". This means that our definition should produce larger measures of uncertainty for sample points with low probability, and smaller measures of uncertainty for sample points with high probability. In other words, the measure that we finally select should be monotonically decreasing with respect to probability.

We are now going to conduct a "discovery" exercise to find such a measuring function. To save time, we are going to offer three candidate functions, each of which satisfies this first property. One of them will, in fact, be the measuring function for uncertainty that is actually used by information theory. By the end of this section, we shall have revealed which one it is. But for the time being, lets continue with our discovery exercise.

To identify them uniquely, we should give names to our three candidate functions. For now, let's call them "$u_1(x)$", "$u_2(x)$" and "$u_3(x)$", respectively. In all three cases, "x" represents the "thing" that we are measuring – which in this case is some sample point (or collection of sample points – event) of our probability distribution.

Once we have narrowed our selection to exactly one of these measures, we shall then rename it to "u(x)" for the remainder of this primer.

For all three candidate functions, let "p(x)" mean "the probability of x" according to probability distribution "p".

The three candidate functions that we have chosen to work with are:

$u_1(x) = 1/p(x)$

$u_2(x) = 1/p(x)^2$

$u_3(x) = \log(1/p(x))$

Without proving it, we shall assert that all three of these functions measure "uncertainty", which of course is our "attribute of interest" – or what we are trying to measure. This attribute, or behavior, is: As p(x) get larger, the value of the measuring function - $u_1(x)$, $u_2(x)$ or $u_3(x)$ – should get smaller. It is easy to show that all three candidate functions satisfy attribute of interest of "u(x) going up as p(x) goes down, and vice versa." We shall leave it to the readers to satisfy their understandings that this is so. (There is one exception to this attribute. It turns out, as we shall see, that u(x) needs to have a value of 0 whenever p(x)=0, and this breaks the monotonicity for this one case. However, monotonicity needs to be maintained for all other probability values.)

But, since all three of these functions satisfy this first condition, then so far we have not seen any reason to select one of these functions over the other. So we shall have to look at the other two desired properties to see if any of these three measures gets "culled out" by them.

Actually, the first function $u_1(x)$ is simpler than the other two. And, admittedly, that is a good reason to choose it. However, it is going to turn out that $u_1(x)$ fails to satisfy one of the other two desired properties. And that fact will be a show stopper for $u_1(x)$. But we shall delay that conversation until we discuss that property.

A Zero Measure

As discussed above, the "zero measure" property says that the measure u(x) of any sample point "x" that "has no degree of uncertainty" should have a measure of zero. Of course, events that "have no uncertainty" are of two types. The first type is events whose probability is 1. That is, if p(x) = 1, then u(x) should be equal to zero. In symbols, if p(x) =1, then u(x) = 0.

However, there is also another type of (finite) event that "has no uncertainty", or that is "absolutely certain". This is an event with probability of zero. The reason for this is that, if a finite event has probability of zero, then it is *absolutely certain* that it will *not* happen.

In summary, if p(x) = 1 then we are absolutely certain that "x" *will* happen; and if p(x) = 0, then we are absolutely certain that "x" *will not* happen. In either case, there is *absolute certainty*, and absolutely *no*, or zero, *uncertainty*. Therefore, whenever either p(x) = 0 or p(x) = 1, then u(x) = 0. Otherwise, p(x) > 0.

Thus, we must test each of our three candidates to make sure that they assign a result of zero whenever p(x) = 0 or p(x) = 1.

Lets first test $u_1(x)$ in this regard. If we substitute the value of 1 for p(x) in $u_1(x)$, the result is 1 – not 0. Therefore, $u_1(1)$ fails for this property.

And, if you substitute the value of 0 for x in $u_1(x)$, the result infinity, which is undefined. So, $u_1(x)$ fails to produce a zero value for both a 0 and a 1 probability. So, $u_1(x)$ fails to yield the desired result for both p(x) = 0 and p(x) = 1. This is not good for candidate $u_1(x)$.

Lets now consider the second candidate $u_2(x)$. It actually has exactly the same problems as $u_1(x)$. Moreover, candidate $u_2(x)$ is a little more complicated than $u_1(x)$, which makes it an even worse choice than candidate $u_1(x)$.

Lets now test $u_3(x)$. As with the other two, we desire that $u_3(0)$ and $u_3(1)$ both produce a value of zero. Clearly, $u_3(0)$ is not defined mathematically because it involves division by zero. This was the same problem we had with both $u_1(0)$ and $u_2(0)$. So, this fact makes candidate $u_3(x)$ no better than the first two candidates.

However, we still need to check $u_3(x)$ for the case when p(x) = 1. This time, candidate $u_3(x)$ passes the test, because $u_3(x) = \log(1/1) = \log(1) = 0$. And this is the desired value!

Therefore, $u_3(x) = \log(1/pr(x))$ scores better than the other two functions for the first criteria – the "zero measure". It is the only one that produces a zero when x is 1.

Admittedly, none of these three candidates works for both cases of p(x) = 0. But $u_3(x)$ fairs better than the other two candidates, because it does work when p(x) = 1.

Of course, we still have the problem that $u_3(x)$ also fails when x = 0. But we can handle this by defining $u_3(0)$ arbitrarily to equal 0 whenever p(x) = 0.

Of course, we could have forced either of the other two functions to behave correctly at both 0 and 1 also. But only forcing $u_3(x)$ at one point is much better than having to force either of the other two functions at both points.

In any event, the candidate $u_3(x)$ is a slight winner in the second criteria: "a zero measure".

So far, all three candidates fair well for the first criteria: "measures attribute of interest". And, we have just seen that the third candidate, $u_3(x)$, pulled ahead in the second criteria: "a zero measure".

Finally, we must test all three candidates against the third criteria: "additivity for 'non-overlapping' events". We shall do this next. But lets relieve the suspense. We are about to see that the third candidate, $u_3(x)$, is the only one of the three to pass the "additivity" test.

Additivity for Stochastically-Independent Events

So far, all three candidate measures worked satisfactorily for the "measures attribute of interest" property, because the all "go down" whenever the probability "goes up" and conversely. For the "zero measure" property, however, the function $u_3(x)$ scored better than the other two. "$u_3(x)$" wasn't perfect there either, but we were able to "fix it up" so that it did the job.

We now turn to the third property we want for our measure – additivity for "disjoint" events.

What are Non-overlapping Events in a Joint Probability Space?

In order to discuss additivity in a joint probability space, then we must first identify some distinction among these joint events into "overlapping" and "non-overlapping" (or "disjoint") categories. Once we have done that, then we can assess whether each of our candidate uncertainty measures is additive for the joint events that are disjoint (non-overlapping) and is non-additive for the joint events that are "overlapping".

We discussed above a way of joining two events (sample points) x and y (from two sample spaces X and Y) by pairing them to produce the new joint event (x, y) – which resides in a new *joint sample space (X, Y)*.

Therefore, if the concept of "additivity" is going to apply to this type of event joining, we have to figure out what it means in this context for two events to "be overlapping" once they are joined, and conversely, for two events to be "disjoint" once they are joined.

Once we determine what "disjoint" and "overlapping mean for our kind of event joining, then we can require that our uncertainty measure u(x) of the new joint event (x, y) be equal to the uncertainty measure of x plus the uncertainty measure of y whenever x and y are *non-overlapping*. In symbols, we say it like this: If x and y are non-overlapping (disjoint), then we shall require of our measuring function u that u(x) + u(y) = u(x, y). However, if x and y are *overlapping*, then we would *not* expect u(x) + u(y) = (x, y).

In words, what we mean when we say that two events are "overlapping" is that they "share some meaning". Moreover, this "sharing of meaning" the occurrence of one of these events containing some information that lets an observer *infer* something about *whether the other event is going to occur*. For example, if we see dark clouds in the sky, then we can infer that it might rain pretty soon. It may not rain, but it would be reasonable to infer that it might. In other words, if we see dark clouds in the sky we can assert that there is a higher probability that it will rain soon that if there were no clouds in the sky.

This is true because the two events – "dark clouds in the sky" and "raining" – that we can say that these two events are semantically related. Certainty, they are probabilistically related. They are related because they have some meaningfulness, or *information*, in common. It is in this sense that these two events "overlap" when they are "joined" (occur together).

If these two events have no meaningfulness in common, then they do not "overlap" when they are joined.

For example, suppose we flip a coin and it turns up "heads". What does this tell us about how the coin will turn up if we flip the coin a second time? The answer is "absolutely nothing". Knowing that the coin turns up heads the first time *gives us no information at all* about how the coin will turn up the second time. This is because *there is no overlap in meaning* between the first event (flipping the coin the first time) and the second event (flipping the coin the second time). These two events are "disjoint", whereas the two other events (the dark clouds and the rain) are "overlapping" – with respect to meaningful information.

This subject is so important to information theory that we shall devote the entirety of Part II of this primer to it. Therefore, we shall delay further discussion until then. However, we shall go ahead and introduce some of the terminology of Part II at this time to make our discussion more efficient.

Whenever two events are "overlapping" in a probability space (like the dark clouds and the rain), we shall say that they are *stochastically dependent*. And whenever they are non-overlapping, or semantically disjoint, we shall say that they are *stochastically independent*.

It is the *stochastically independent* pairs of events that are *additive* in our joint probability spaces because the uncertainty measures of their joint events (x, y) are equal to the sum of the uncertainties of their component events x and y. That is, we need to make sure that we find a measuring function u(x) such that u(x, y) = u(x)+u(y) for stochastically independent ("semantically disjoint") joint events (x, y). However, if the pair of events (x, y) are *stochastically dependent*, then we would expect that our measure u(x) *will not be* additive.

An Example of Additivity in a Joint Probability Space

I shall now provide a meaningful example of such a way of joining. Suppose we have a coin and we want to toss or flip it to see whether it lands with heads (H) or tails (T) facing up. This experiment is a probability space with sample space {H, T} and probability distribution p = {(H, ½), (T, ½)}.

So, what would be a reasonable way to "join" two sample points to produce a new sample point in this situation of flipping a coin? Suppose we extend this experiment of flipping a single coin and define a new experiment that flips one coin, and then subsequently flips that coin again. So, the new experiment flips two coins in a row; whereas the initial experiment flips one coin.

Notice that, the first experiment is defined by a singleton outcome – T or H. However, a pair of outcomes describes this new experiment. The first member of this pair describes the outcome of the first toss and the second member describes the outcome of the second toss. The possible sample space of this new experiment is:

S' = { (H, H), (H, T), (T, H), (T, T) }.

In fact, these four pairs form the sample space of this new "joint" experiment – formed by two successive trials[20] of the first experiment. This new joint experiment, though related to the first experiment where one coin was flipped, is nevertheless a different experiment. In fact, its sample space has 4 sample points, whereas the sample space of the first experiment has only 2 sample points.

[20] This same joint probability space can also represent the experiment in which two coins are flipped at the same time.

Moreover, the probabilities are different for the joint experiment, where each sample point has a probability of ¼. So, the probability distribution p' for the joint space is:

p' = { ((H, H), 1/4), ((H, T), 1/4), ((T, H), 1/4), ((T, T), 1/4)}

So, we have found a reasonable way to "join" two probability spaces (or, as we say, chance variables) to obtain a third probability space, or chance variable.

In this particular example, all possible pairs of outcomes are *stochastically independent*. This means that knowing whether the first coin flip produced a heads or a tails give us no hint*, no helpful information*, as to what the second flip will turn out to be.

This is unlike the dark clouds and rain situation, where knowing that there are dark clouds in the sky portends that there is a higher probability of rain than not knowing what the clouds situation is. In that case, knowing that there are dark clouds in the ski *is helpful information* for guessing whether it is going to rain. This is because there is *some information in common* between "dark clouds in the sky" and "rain". This "information in common" is called *mutual information*. If two events are *stochastically dependent*, it is because there is some *mutual information* shared by them (some "overlap"). If they are *stochastically independent*, then there is *no mutual information between them*. They are "semantically disjoint".

Our quest right now is to find some measuring function of uncertainty that is *additive for stochastically independent events*, and not additive for *stochastically dependent* events. (For the time being, we shall concentrate on whether our candidate measures are additive for independent events, and shall not bother with inspecting that each measure is also not additive for dependent events.)

Since, all of the joint events of our coin flip situation in the present example are *stochastically independent*, then our measuring function "u" (when we finally find it) will be additive for all joint events in this example.

(Notice that in this example, the probability distribution p is equally-likely; and so is the joint probability distribution p'. However, this need not be so. In fact, the probabilities of distribution p' need not be tied to the probabilities of distribution p. It is possible that either is equally likely while the other is not. It is even possible that neither is equally likely. The probabilities of both distributions may, or may not, in fact, be unrelated to those of the other.)

Testing $u_1(x)$ for Additivity for Independent Events

We shall now test candidate measure $u_1(x)$ to see if it is additive for stochastically-independent x and y. To say that $u_1(x)$ is additive means for an independent joint sample point (x,y) that

$$u_1(x,y)) = u_1(x) + u_1(y)$$

for all stochastically independent x, y.

But, to say that x and y are stochastically independent means that p(x^y) = p(x)*p(y). And, of course, by definition, "(x, y)" means the case of both x and y, or "x^y". Therefore, the statement above that we are testing,

$$u_1((x,y)) = u_1(x) + u_1(y)$$

Is equivalent to

$$u_1((x^y)) = u_1(x) + u_1(y)$$

So, the test we are conducting asks the question of whether the above equality is true for our first candidate measure of uncertainty $u_1(x)$. So, lets check both sides of this equation to see if they have to be the same for all x, y.

Regarding the left side,

$u_1((x^y) = 1/p(x^y) = 1/(p(x)*p(y))$

Now, lets check the right side, to see if it is the same as the left side for all x, y.

$u_1(x) + u_1(y) = 1/p(x) + 1/p(y)$

So, in order to show that is the first candidate measure $u_1(x)$ is additive for stochastically independent x and y, we have to show that

$1/(p(x)*p(y)) = 1/p(x) + 1/p(y)$

for all x and y.

However, it is easy to show by counter-example that this relationship is not always true. For example, suppose p(x) = 1/3 and p(y) = ½. Then the left side of the above relationship is

$1/(p(x)*p(y)) = 1/((1/3)*(1/2)) = 1/(1/6) = 6.$

On the other hand, the right side evaluates to:

$1/p(x) + 1/p(y) = 1/(1/3) + 1/(1/2) = 3 + 2 = 5.$

Consequently, we have found some values for p(x) and p(y) for which the relationship

$u_1((x^y) = 1/p(x^y) = 1/(p(x)*p(y))$

is not true whenever x and y are stochastically independent. And we have therefore proven by counter-example that the measure $u_1(x)$ is not additive for all stochastically independent x and y.

Thus, $u_1(x)$ fails two out of three of our requirements for a measure of uncertainty of an event or sample point.

Testing $u_2(x)$ for Additivity for Independent Events

This time, we shall test candidate measure $u_2(x)$ to see if it is additive for stochastically-independent x and y. To say that $u_2(x)$ is additive means for an independent joint sample point (x,y) that

$u_2(x,y)) = u_2(x) + u_2(y)$

for all stochastically independent x, y. (This is the same test we applied for $u_1(x)$.)

But, to say that x and y are stochastically independent means that p(x^y) = p(x)*p(y). And, of course, by definition, "(x, y)" means the case of both x and y, or "x^y". Therefore, the statement above that we are testing,

$u_2((x,y)) = u_2(x) + u_2(y)$

Is equivalent to

$u_2((x^y)) = u_2(x) + u_2(y)$

So, the test we are conducting asks the question of whether the above equality is true for our first candidate measure of uncertainty $u_2(x)$. So, lets check both sides of this equation to see if they have to be the same for all x, y.

Regarding the left side,

$u_2((x^y) = 1/p(x^y)^2 = 1/(p(x)*p(y))^2$

Now, lets check the right side, to see if it is the same as the left side for all x, y.

$u_2(x) + u_2(y) = 1/p(x)^2 + 1/p(y)^2$

So, in order to show that is the first candidate measure $u_1(x)$ is additive for stochastically independent x and y, we have to show that

$1/(p(x)*p(y))^2 = 1/p(x)^2 + 1/p(y)^2$

for all x and y.

However, it is easy to show by counter-example that this relationship is not always true. For example, suppose p(x) = 1/3 and p(y) = ½. Then the left side of the above relationship is

$1/(p(x)*p(y))^2 = 1/((1/3)*(1/2))^2 = 1/(1/6)^2 = 1/(1/36) = 36.$

On the other hand, the right side evaluates to:

$1/p(x) + 1/p(y) = 1/(1/3)^2 + 1/(1/2)^2 = 1/(1/9) + 1/(1/4) = 9 + 4 = 13.$

Consequently, we have found some values for p(x) and p(y) for which the relationship

$u_1((x^y) = 1/p(x^y) = 1/(p(x)*p(y))$

is not true whenever x and y are stochastically independent. And we have therefore proven by counter-example that the measure $u_1(x)$ is not additive for all stochastically independent x and y.

Thus, $u_2(x)$ fails two out of three of our requirements for a measure of uncertainty of an event or sample point.

Testing $u_3(x)$ for Additivity for Independent Events

This time we shall test candidate measure $u_3(x)$ to see if it is additive for stochastically-independent x and y. To say that $u_3(x)$ is additive means for an independent joint sample point (x,y) that

$u_3((x,y)) = u_3(x) + u_3(y)$

But, to say that x and y are stochastically independent means that p(x^y) = p(x)*p(y). And, of course, by definition, "(x, y)" means the case of both x and y, or "x^y". Therefore, the statement above that we are testing,

$u_3((x,y)) = u_3(x) + u_3(y)$

Is equivalent to

$u_3((x^y) = u_3(x) + u_3(y)$

So, the test we are conducting asks the question of whether the above equality is true for our third candidate measure of uncertainty $u_3(x)$. So, lets check both sides of this equation to see if they have to be the same for all x, y.

Regarding the left side,

u_3((x^y) = log(1/p(x^y)) = -log(p(x^y)) = -log(p(x)*p(y)) = -(log(p(x) + log(p(y))

Now, lets check the right side, to see if it is the same as the left side for all x, y.

$u_3(x) + u_3(y)$ = log(1/p(x)) + log(1/p(y)) = -log(p(x)) – log((p(y) = -(log(p(x)) + log(p(y)).

Consequently, the left side is the same as the right side for all x, y.

And we have therefore proven that the measure $u_3(x)$ is additive for all stochastically independent x and y.

Thus, $u_3(x)$ passes completely the "measures attribute of interest" property and the "additivity for independent events" property. And, it passed the "zero measure" property for x = 1. It did not initially pass the zero measure property for x = 0. But we forced it to by arbitrarily defining $u_3(0) = 0$.

The Measure of the Uncertainty of an Event

Consequently, of our three candidate measures of the uncertainty of a single event (or single sample point), the only one that passed all three tests was $u_3(x)$. Thus, we shall rename "u(x)" to "u(x)", or just "u".

Thus, we finally arrive at the definition of the function that we shall henceforth use to measure uncertainty at the sample point or event level:

Measure of the uncertainty of an event:

u(x) = log(1/p(x))

where x is an event of a probability space, and p(x) is the probability of x with respect to probability distribution p.

We must also consider some other properties of this measure u: 1) the log base used by u and 2) whether u exhibits non-additivity for dependent events.

The Log base of u(x)

It must be understood that the above definition of u(x) is incomplete because we have not yet specified what log base is to be used. The fact of the matter is, that our arguments above will all work no matter which positive log base is used. Therefore, our u(x) actually represents an entire family of functions – one for each possible log base.

It must be noted that in order to use u(x) for calculation purpose, any positive log base will do. One simply selects any log base – as long as one consistently uses the same log base when comparing distinct calculations. Of course, the higher the log base chosen, the faster the results of u(x) will rise. But the choice is strictly arbitrary.

Computer scientists and communications engineers are fond of using a log base of 2, because then the situation is a model of the concept of *bits* used in computers. Mathematicians and engineers have commonly used two other log bases: 10 and e.

One more point is in order. The arguments used in proving that u(x) satisfies all three of these properties does not prove that it is *the only measure* of an event that satisfies all three of these properties.

But, indeed, it is! We address this issue in a subsection later below.

Non-additivity of u(x) for Dependent Events

So far, we have demonstrated that, if events x and y are *stochastically independent*, then u(x, y) = u(x) + u(y). In other words, we have proved that if x and y are independent, then the measure u is additive.

However, we have also made the statement that if x and y are *stochastically dependent*, then u is *not additive*.

This is easily proven simply by "running the earlier proof backwards" – as we shall now prove.

Lemma: The measure u = log(1/p(x)) is *not additive* for stochastically dependent events x and y.

Consider stochastically dependent events x and y.
Then, x and y are not independent.
We shall prove the lemma by contradiction.

Assume that u is additive for x and y.
Then u(x) + u(y) = u(x^y)
And –log(p(x)) – log(p(y)) = -log(p(x^y))
-[log(p(x) + log(p(y))] = log(p(x^y))
-log(p(x)*p(y)) = -log(p(x^y))
Thus, p(x)*p(y) = p(x^y), since log is invertible.
But this contradicts the condition that x and y are dependent.
Thus, the assumption that u is additive for dependent events x and y is erroneous.

Therefore, u is non-additive for stochastically dependent events.

This conclusion, together with the results of the previous subsections, tells us that "u" is additive for independent events, but non-additive for dependent events.

Analogously, this fact is also true for our motivating physical example where we were joining wooden boards. In this analogy, boards that are joined end-to-end are correlated to joint events whose component events are stochastically independent. And boards that are joined by overlapping and nailing are correlated to stochastically dependent events.

In that example, if we joined two boards end-to-end, then the length of the joined boards was exactly equal to the sum of the lengths of the initial boards. Therefore the joint was additive for end-to-end joining. However, if we joined two boards in a manner that overlapped them and then, say, nailed them together in the overlap, then the length of the joined pair was not the same as the sum of the lengths of the two boards separately. (in fact it was less than.)

A similar phenomenon is occurring in our joint probability space example. For stochastically dependent pairs of events, their "overlap" is some "shared information" (mutual information). For stochastically independent events, there is no mutual information – they are "semantically disjoint". Therefore, the measure of two joined independent events, like non-overlapping boards, is equal to the sum of their individual measurements.

Uncertainty Chain Rule

We have shown so far that our uncertainty function u(x) is additive for two stochastically independent event, and non-additive for dependent events.

Because of the additivity of independent events, we can write

u(x, y) = u(x) + u(y)

if x and y are independent. However, that equation does not hold whenever x and y are dependent. The question then arises whether there is some other simple relationship similar to the one above that *can* be made for all pairs of events x and y, whether or not they are independent.

The answer is Yes. It is called the *chain rule* for the events of a joint probability space.

Chain Rule for Events of a Joint Probability Space

u(x, y) = u(x) + u(y|x)

Before we prove this, we shall point out that:

p(y|x) = p(x,y)/p(x) by definition.
Multiplying both sides of this equation by p(x) we get:
p(x,y) = p(x)*p(y|x).

We shall use this last relationship in the following proof by substituting the right side of the above equation for the left.

Proof:

u(x, y) = -log(p(x, y))
= -log(p(x)*p(y|x), by substitution.
=-[log(p(x)) + log(p(y|x))]
= -log(p(x)) – log(p(y|x))
=u(x) + u(y|x)
Thus, u(x, y) = u(x) + u(y|x) QED.

As a matter of interest, we already know that whenever x, y are independent that we have the following additivity relationship for u:

u(x, y) = u(x) + u(y|x)

Obviously, the only difference between the chain rule relationship that we just proved and the additivity relationship for independent events is the last term. It is “u(y|x)” in the chain rule, but “u(y)” in the additivity relationship.

This must mean that “u(y|x)” is equal to “u(y)” in the case that x and y are independent. This fact is easily shown.

Consider u(y|x) in the case that x and y are independent.

By definition, p(y|x) = p(x, y)/p(x).
But, whenever x, y are independent, then p(x, y) = p(x)*p(y) by definition of independence.
So, we can substitute “p(x)*p(y)” for p(x, y) above, giving
p(y|x) = p(x)*p(y)/px).
But the “p(x)”s in the numerator and denominator of the right side cancel out, leaving
p(y|x) = p(y) in the case of x, y independent.

Recall that what we are trying to show is that

u(x, y) = u(x) + u(y|x)

reduces to

u(x, y) = u(x) + u(y)

whenever x and y are independent.

And in order to show this, we have to show that the final terms of each of the above relationships are equal whenever x and y are independent. What we just showed is

that p(y|x) is the same as p(y) whenever x and y are independent. What we are going to do now is to use the fact that p(y|x) is the same as p(y) whenever x and y are independent to show that u(y|x) is the same as u(y) whenever x and y are independent. We shall show this by substitution.

So,

u(y|x) = -log(p(y|x)) = -log(p(y)) whenever x and y are independent. But
u(y) = -log(p(y)).
Thus both u(y|x) and u(y) are equal whenever x and y are independent.

Measure of the Uncertainty of an Event

Therefore, we have found a measuring function that satisfies all three of the properties we desire in our measure of uncertainty of an event – namely, the function u(x).

So we shall anoint the function u(x) as the measure of the uncertainty of a sample point that we seek. Because of this, we shall name it "the uncertainty of event (or sample point) x" of a probability space. Here is a more thorough definition:

Measure of the uncertainty of a sample point:

Given a probability space P = (S, p, E), define the *measure of the uncertainty* of event $x \in E$ as:

$$u(x) = \log(1/p(x)) \quad \text{- for x where } p(x) \neq 0, \text{ and}$$
$$0 \quad \text{- for x where } p(x) = 0.$$

Notice that we had to make a special case for x where p(x) is zero, because the expression "log(1/p(x))" entails division by zero for that case.

This function is foundational in information theory, because it, together with probability theory, is the root of all that follows.

Surprisingly, this function has not been assigned a widely adopted name! The reason for this is that this simple function is immediately used by information theory to define a more complex function that measures the uncertainty of an entire probability space – rather than one of its events. This new function, which we shall define in the next section, is called *entropy* and is ultimately the actual focus of information theory.

Thus, the uncertainty function u(x) that we have developed in the present section has largely been lost in the attention given to entropy – and has not even been named! Although one researcher tried giving it the name "surprisal", but the name has not enjoyed universal adoption. For this text, we shall refer to this measure as the *uncertainty of an event* of a probability space.

Also notice that u(x) depends upon the probability distribution p. Therefore, if there is a sample space on which we have defined two probability distributions, say p and q, then there will be two distinct measures of uncertainty for events "x" of the sample space, namely, $u_p(x)$ and $u_q(x)$. In fact, these actually define two distinct probability and information spaces on the same sample space. This distinction will become necessary in Part II where we will have the occasion to compare the uncertainty measures of two different probability distributions on the same sample space.

Appendix 2: Information Theory and Communications Theory

A look at the literature on information theory raises the question as to whether *information* theory is the same or different from a discipline named *communications theory*.

This primer has already, in the Prologue section above, distinguished between the two disciplines. Specifically, the subject of this primer is information theory – and *not* communications theory. Implicit in this fact is the claim that it is both possible and reasonable to make such a distinction.

I believe that either separating or joining the two disciplines is a reasonable and defensible position, depending on what one desires to accomprlish. However, this primer has many reasons to find it more useful to distinguish them – one of which is that the topics that this primer intends to discuss fall under one of them and not the other (at least according to the separation made by this primer.) Thus, it behooves me to describe why it is possible, reasonable and desirable to make this distinction.

A Look At the Information Theory Literature

Of the four sources on information theory that are most heavily referenced by this primer, three cover both topics to a considerable extent, and all four have the name of one of those two disciplines in their titles.

Lets now take a look at the treatment of these two disciplines on the parts of each of these sources.

Shannon's 1948 Paper

The first source is Shannon's paper entitled "The Mathematical Theory of Communications" [Shannon 1948]. Shannon is largely credited with "inventing" information theory. It is my analysis, as stated in the Prologue of the present article, that Shannon had to invent information theory as a mathematical underpinning upon which he was then able to erect communications theory.

My best guess is that Shannon started out with the intention to develop *communications theory*. After all, he worked for the telephone company. But then found that he needed a mathematical foundation for it that he had to first invent. This mathematical foundation was comprised of a new measure of the uncertainty inherent in a probability distribution, which he named *entropy*. Shannon's idea of entropy can be applied to any situation that has probabilities, regardless of what "application domain" it comes from. My contention is that it is this mathematical foundation that constitutes *information theory*.

Once Shannon had invented that foundation, he was then in a position to move on to his initial intent – the invention of *communication theory*. One can reasonably conclude by looking at Shannon's paper that *communications theory* pertains to "messages" and "channels" through which the messages are sent. On the other hand, it is reasonable to conclude from Shannon's treatment that *information theory* pertains to the measure of the degree of uncertainty inherent in a probability distribution – any probability distribution. In any event, the "information theory" part of Shannon's paper is rather brief – being introduced and discussed in the middle of his first (of five) chapter.

Khinchin's 1957 Book

Khinchin was a mathematician and a member of the famous Russian probability school. His 1957 book, entitled *Mathematical Foundations of Information Theory* [Khinchin 1957] puts information theory on a stronger mathematical foundation than Shannon had been able to do. Essentially, Khinchin describes information theory as the study of entropy.

Khinchin wastes no time positioning information theory as a branch of mathematics – specifically of probability theory. He states, "There is no doubt that in the years to come the study of entropy will become a permanent part of probability theory…." [Khinchin 1957, p 2].

The remainder of Khinchin's book is an exercise in probability theory for the purpose of developing the idea of entropy through probabilistic reasoning. However, later chapters begin to apply these ideas to messages and communications channels. Thus, it can be said that Khinchin's book discusses both disciplines.

Cover and Thomas' 1991 Book

This book, entitled *Elements of Information Theory*, is a thorough and extensive text on information theory. It is most likely intended as a multidisciplinary college textbook on the subject. There have been many advances in information theory since the time of Khinchin and Shannon. In particular, information theory has matured to a general theory of predictability and prediction.

The book appears to be targeted as a textbook for a course in information theory at the level of "advanced undergraduate and graduate students" from a multidisciplinary audience. Specifically, the text is used to supplement the syllabus of a graduate level seminar offered by the Courant Institute of Applied Mathematics at New York University [Kleeman 2012].

One might expect students from mathematics, engineering, economics and any other technically oriented disciplines to use the book. One could expect a course or seminar that used this book to be taught in any number of departments, including mathematics, statistics, economics, engineering, etc.

The mathematical foundations are introduced in the first chapter, and comprise the development of the concept of entropy based upon probability theory and stochastic processes. It is fair to say that the book assumes some knowledge and experience with these mathematical subjects on the part of the reader.

The book then delves into some advance mathematical topics before proceeding to a number of chapters that discuss applications of information theory. These applications certainly include "messages", "information channels", "channel capacity", "Gaussian channels" and other topics that are reasonably labeled as *communications theory*. But, as well, several other applications of the mathematics of entropy (information theory) – that are not treated as "sending messages through channels" - are also covered. These include statistical mechanics, applications to statistics, and stock market analysis.

Kleeman's 2012 NYU Graduate Seminar Syllabus

Richard Kleeman's seminar at the Courant Institute of Applied Mathematics is entitled "Information Theory and Predictability". In addition to its quite readable and informative syllabus, the seminar uses the [Cover and Thomas 1991] book as a principle reference source. And the scope of the course includes the mathematics of entropy and its many

manifestations (entropic functionals), which culminates at its most advanced manifestations (at this point in time) as predictability.

Like [Cover and Thomas 1991], Kleeman's seminar also visits a number of applications of information theory in his later lectures, including data compression, information transfer, statistical mechanics, dynamical systems and weather prediction.

How Sources Define Information Theory and Communications Theory

Generally, then, all four of the principle sources on Information and Communications theories used by this primer include discussions of both theories within their covers, but they treat them differently. This treatment, then, is not conclusive on the subject of whether it is possible or useful to try to distinguish them – as this primer does.

So, lets turn to their respective definitions to see is we can find anything persuasive on the issue.

Communications Theory

[Shannon 1948, p.1] describes communications theory by elaborating the problem that it solves, as follows:

> The fundamental problem of communications is that of reproducing at one point either exactly or approximately a message selected at another point.

It is reasonable, then, to assert that if some arbitrary subject is a part of "communications theory", then that subject must pertain to the reproduction of a message from one point to another.

To explain further, [Shannon 1948, p. 2] identifies the subject of communications theory as being a "communications system", and then goes on to enumerate its components as:

1. An information source
2. A transmitter
3. The channel
4. The receiver
5. The destination

It is reasonable, then, to assert that if some arbitrary subject is a part of "communications theory", then that subject must make considerations that relate to components that correspond to the above five.

To be more historically precise, we shall point out that Shannon was involved in the development of communications theory in his capacity at Bell Labs as an electrical engineer. In fact, being involved in telephone systems, Shannon was precisely interested in communications theory within the context of electrical engineering. After all, he was not dealing with flag semaphores or tin cans with string. The target systems of his interests were all electrical systems.

This likely explains why [Cover and Thomas 2006, p. 1] associate "Communications Theory" under Electrical Engineering. In any event, it is reasonable to categorize modern information theory as a discipline of Electrical Engineering.

Information Theory

[Shannon 1948, p. 10] sets up the problem addressed by information theory as follows:

> Suppose we have a set of possible events whose probabilities of occurrence are p_1, p_2,… p_n. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

Shannon then explains that he is using the term "choice" interchangeably with term "uncertainty".

So, Shannon characterizes information theory as concerned with measuring the amount of uncertainty inherent in a situation when all that is known about the situation is a set of possible events and their probabilities. Of course, a set of possible outcomes and their probabilities is called a *probability distribution*.

In other words, Shannon is implying that *information theory* is concerned with measuring the uncertainty (or choice) that is inherent in a discrete probability distribution.

Moreover, he puts no further constraints on the discrete probability distributions that can be considered by information theory. For example, it does not matter what domain of application any such probability distribution represents. Simply that it has probabilities is sufficient.

Echoing this sentiment, [Kleeman 2012, lecture 1 p. 1] says that

> The central idea of information theory is to measure the uncertainty associated with random variables.

We shall conclude then that information theory is concerned with measuring the degree of uncertainty inherent any probability distribution – regardless of what field of inquiry or concern that probability distribution relates to.

Information Theory vs Communications Theory

[Cover and Thomas 2006, p.4] assert that

> The fundamental quantities of information theory – entropy, relative entropy and mutual information – are defined as functionals of probability distributions. In turn they characterize behaviors of long sequences of random variables and allow us to estimate the probabilities of rare events….

This assertion places information theory squarely within the province of probability theory and therefore of mathematics. Information theory, then, is categorized as a mathematical sub-discipline.

On the other hand, [Cover and Thomas, p. 1] place communications theory as a branch of electrical engineering that is concerned with error rates, channel capacities and other characteristics of Shannon's "information system".

Relationships Between the Two Disciplines

We claim that it is fair to conclude from the above discussion that 1) communications theory is an engineering discipline. 2) Whereas, information theory is a branch of probability theory that studies entropy as a measure of the uncertainty inherent in a probability distribution. As such, information theory is a sub-discipline of abstract mathematics.

Having concluded this, then it is reasonable to assert that information theory has a number of *fields of applications*. One of these fields of application is *communications*

theory. But information theory has other fields of application, including statistical mechanics, economics and computer science.

In fact, any phenomena that has probabilities or any field uses probability distributions can apply information theory.

Therefore, whereas communications theory is a specific engineering discipline, information theory is a general mathematical discipline that has many applications.

And, of course, it is possible for an arbitrary subject of interest to "have probabilities" without pertaining "to the reproduction of a message from one point to another"; or without having components that include "an information source, a transmitter, a channel, a receiver and a destination."

However, as Shannon showed us, if the subject has "an information source, a transmitter, a channel, a receiver and a destination", then it also "has probabilities".

From these observations we can conclude that communications theory is a special case of (is an application of) information theory; but information theory is not a special case of (is not an application of) communications theory.

Analogy

It may be useful to draw an analogy to the relationship between information theory and communications theory. We shall use Newtonian Calculus and Mechanics to analogize the relationship between them.

Whereas, Shannon started out with the intention of developing communications theory, Newton had the initial intention of developing mechanics. However, Newton found that he had to develop the calculus first as an underlying mathematical discipline to mechanics. So too, Shannon found that before he could develop a mathematical theory of communications (an electrical engineering discipline), he first had to develop information theory as its underlying mathematics.

In both of these cases, the underlying disciplines that had to be developed first were eventually understood to be branches of mathematics – that also had applications to other disciplines beyond being applied to the subjects that they were initially intended to provide foundations for.

Also, it is easy to imagine that the calculus was initially attached to mechanics as though the calculus, itself, were a part of physics, rather than being a branch of mathematics. So too was information theory initially attached to communications theory by many – even though it is actually a purely mathematical topic.

It was probably not understood that the calculus was, in fact, pure mathematics and not physics until other applications for it were found outside of physics – for example in economics. So too, it was not initially understood that information theory is, in fact, pure mathematics and not electrical engineering, until other applications were found for it outside of electrical engineering and computer engineering.

In fact, it was the genius of Shannon to point out that, essentially, information theory had already been invented by Gibbs. Of course, Gibbs had not bothered to point out that fact to the world at the time (circa 1902). And, the questions as to whether Gibbs was aware that his formula for entropy had broad application beyond statistical mechanics is only speculated about, even today.

Appendix 3: Three Approaches to Critical Thinking

Abstract

This appendix presents three approaches to critical thinking that are evident in contemporary scientific and other intellectual discourse. In this article, they are referred to as the causality model, the logic model and the statistical inference model.

These three programs of critical thinking are generally presumed to be mostly equivalent and are often freely intertwined within serious discourse. But this article argues that no two of these systems of thought are analogs of the others – and that attempts to show that they are have failed. Specifically, it is argued that all attempts to define "property-preserving transformations" between them (analogs, correlations, equivalence relationships, homomorphisms, symmetries, etc.) ultimately fail.

Nevertheless, all three do have something in common. It is that they all attempt to explain, model or represent some idea of "phenomenal dependence". Phenomenal dependence is the dependency between phenomena on each other for their existence and dynamics. For causality, the pair "causes" and "effects" are the operative phenomenal dependence mechanisms. For the logic model, it is "necessary conditions" and "sufficient conditions". For statistical inference it is whether a certain probability (p-value) is greater or less than a pre-selected confidence threshold named the significance level.

A principle difference between statistical inference and the other two models is that it is non-deterministic. "Non-deterministic" means that the repetition of the same procedure may produce different outcomes for each repetition. This is called random variation. The other two models, causality and the logic model, are deterministic and generally do poorly with random variation.

But because of the complex nature of the natural systems studied empirically by scientific endeavor, random variation dominates the pursuit of science knowledge. That is, multiple samples taken from the same population regularly exhibit different, inconsistent, statistics. (This is the subject of the "Central Limit Theorem".) Consequently, the scientific method must, and does, employ the statistical inference model as its final arbiter of acceptance or rejection of proposed assertions.

However, the scientific method also reserves a place in its earlier steps for both the causality model and the logic model. The application of the causality model – because of its deep intuitive nature – provides inspiration for the articulation of problems observed and assertions that offer potential explanations. The logic model plays the role of translator within the scientific method – rearticulating the assertions of the causality model into testable, refutable hypotheses, which are then acted upon by the statistical inference model for the final act of acceptance or rejection.

In western intellectual discourse, three conspicuous approaches have emerged over the last two thousand years to guide the processes of critical thinking. I shall refer to each of these three approaches as:

- The causality model
- The logic model
- The statistical inference model

We point out that the third of these models, statistical inference, is the subject of Part II of this primer on information theory, and the basis for Part III. We shall show how stochastic dependence is a special case of statistical inference. This fact makes this appendix particularly pertinent to this primer.

Other approaches are also evident, but these three models of critical thought have emerged to dominate intellectual discourse.

Upon inspection, it can be seen that all three share a common trait: each provides its own mechanism to model how the existence – or even origination - of one phenomenon depends in some way on the existences of one or more other phenomena.

It could be said that the logic model and the statistical inference model follow the lead of the causality model in this respect. Whereas, the causality model offers the related notions of cause and effect to model the dynamics of phenomenal dependence, the other two provide their own concepts.

In this appendix, we shall be interested in whether these three models of critical thinking are in some sense analogous to, or analogs of, each other. And if they are, then are there certain elements of each that can be considered as analogs to certain elements of the others?

Since it is common practice to intertwine these three models in intellectual discourse, it would be useful indeed if there can be shown to exists relationships among the three models that renders them as analogs – or in some sense correlated. For example, the scientific method, as we shall see, conspicuously uses all three.

While the causality model deals with phenomena directly, the other two models use the notion of assertions, or indicative sentences, to indirectly represent phenomena. This fact is a subtle unspoken redirection of emphasis from that which is observed (phenomena) to the thinking of one who is doing the observing.

The third of these approaches – statistical inference – is discussed in Part II of this primer, and is shown to be the mathematical root of portent and meaning in information theory. It is in this appendix that this approach to thinking is compared with and contrasted against the two other approaches in contemporary investigative thought.

This appendix presents some informal thoughts on these three approaches to “existential origination”, or “phenomenal dependence” in modern parlance. And it is not to be confused for a formal treatment of these issues.

This appendix suggests that these three views of “phenomenal dependence” all have a similar intention – which is to provide distinct views of how phenomena come into existence through the combined interactions of other phenomena.

I have had to invent some of my own terminology (e.g. “phenomenal dependence”) in order to suggest a commonality among these three models. The intention of this appendix is to provoke thought rather than to present any formal argument. Obviously, considerably more deliberation is needed before a formal investigation into these matters can be undertaken.

By phrases such as “phenomenal dependence” is meant a concern or interest in why some phenomena are originated (or other originating relationship) by other phenomena – and what the origination relationships among phenomena are. All three of these approaches, from their own unique perspectives, provide credible ways of thinking about how one phenomenon “comes out of” another.

The main goal of this appendix is to arouse awareness that all three of these approaches co-exist, and that they are often freely, although perhaps unconsciously, intertwined within the same discourse.

However, people engaged in critical thinking activities - such as mathematics, philosophy, science and engineering – have often run into difficulties because of this

not-too-careful intertwining. And as a result there is the need for more deliberation about how these three approaches are different and what their relationships to each other are. Such is the deliberative pursuit of this appendix.

In what follows, a certain amount of history is discussed, and certain of these three models of thinking have evolved to become more established than others in the realms of investigation involving critical thinking.

What follows are some thoughts regarding each of these three approaches on the part of this author, plus some concluding observations.

The Causality Model

It is a deeply intuitive idea that all phenomena are effects that are caused by other phenomena.

By the time of the ancient Greek philosophers, this notion was broadly adopted. It seemed evident that an expansive hierarchy of events that moved ever forward in time by this cause-and-effect model relates all phenomena.

The Greek philosophers expended considerable argument in attempts to justify this idea. The extent to which they were successful is not clear.

The causality model imposes a sense of time. That is, we know that if x causes y, then y did not occur before x in time. There is a fundamental asymmetry between cause and effect within the causality model.

An intuitive and attractive aspect of causality is its deterministic nature. Determinism makes causality a reliable and dependable model for use by critical thinkers. A specific set of causes results in the same effects repeatedly. This is convenient because it proves a very predictable and therefore trustworthy thought model.

On the other hand, causality's determinism does not do a very good job of representing uncertain or random phenomena. That is, if the same trial is repeated with the same "causes", the causality model requires that the same effects consistently obtain. But this determinism does not faithfully represent contemporary scientific investigations of complex phenomena in which the whole of almost any system of interest is too large and complex to reasonably be able to observe the whole of the system with a single sampling. Rather, multiple samples must be taken in order to try to collectively obtain a description of the whole. And multiple samples of the same population generally exhibit inconsistent results – a situation known as random variation. This inability to account for random variation can be a serious drawback for any deterministic model, and of the causality model specifically.

A second disadvantage of causality as a model for phenomenal dependence is that it offers no internal mechanism for selecting or proving that any phenomenon causes another. This is a serious lacking for any deterministic model of phenomenal dependence. Any model that is used for critical thought should have a test criterion for determining the admissibility or non-admissibility of criteria into the model.

Adoption of the Causality Model

In our own day, the principle of cause and effect is taken, at least, as an article of faith. Little defense is needed by anyone whose uses it a basis of an argument. In this regard, little more needs to be said here about this principle. All of us allude to it every day countless times.

However, the predomination of causality in the sciences has begun to be challenged. Specifically, the causality model has been largely abandoned by arguably the most

successful scientific theory of all times – quantum mechanics. Quantum mechanics has been so successful that virtually all branches of classical physics have been rewritten so as to include a quantum mechanical version.

In quantum mechanics, the same sequence of "causal" events regularly produces different effects – in other words, random variation. Of course, this type of random behavior is the intended domain of the statistical inference model, whose foundations are probability theory. Thus, in quantum mechanics, "causes" have been replaced by "probabilities".

The Appeal of Causality

We must ask, "What is the appeal of the causality model as an explanation of phenomenal dependence?" Admittedly, causality appeals to our deep intuition and offers a persuasive explanation of the interrelationships of all things phenomenal. But what is this appeal?

In the first place, the notion of cause-and-effect directly addresses our deeply held question "Why do things happen the way they do?" It addresses "Why".

Secondly, causality narrows the responsibility of effects to a limited number of "causal" phenomena – or "parties". It narrows the focus of responsibility. This fact serves the controllability of situations – our desire to exert control. If we can limit responsibility for "effects" to a small number of "causes" – preferably just one – then the situation is more likely to be controllable, we know what to "correct", and we can have some hope of being able to control the situation.

A third appeal of causality is that it is deterministic. In the causality model, "causes" determine "effects". Given some specific set of causal conditions, causality demands that the same effect results every time. So causality is repeatable and therefore trustworthy and dependable.

For these reasons, causality is broadly appealing and deeply entrenched. It has enjoyed thousands of years of comfortable adoption across humanity and is unlikely to be dethroned by alternatives.

The only remaining question is how the causality model holds up against the other two, particularly when any two of them are intertwined in discourse, and particularly when the demands of rigorous, critical thinking are evident.

The Logic Model

The "laws of thought", logic, were established in a near-universally acceptable form by Aristotle – who, as it happens, was also a prime contributor to the ideas and principles of the causality model. Aristotelian logic has descended to contemporary times with the same level of near-universal acceptance as has ideas of cause and effect.

However, the two models – causality and logic - are very different. Normally, we take for granted that they are simpatico, and intertwine them at will in our deliberations and conversations.

However, if they were to turn out to be less than equivalent, then we may be in for some trouble – some inconsistencies. And in a strictly logical way they are not equivalent, as we shall see shortly.

The Logic Model in Western Culture

Aristotle's laws of thought have held pretty well for two thousand years. During the nineteenth century George Boole further codified them into a mathematical formulation – algebraic formulation actually.

And after this, Aristotelian logic started being used as the foundation of mathematics by mathematicians such as Peano, Russell, Whitehead and others into the twentieth century. In fact, some mathematicians regard mathematics as an extension to logic. This view of logic is appropriately named mathematical logic.

The Dominance of the Adoption of Logic

By the beginning of the twentieth century logic had been established as the foundational discipline for how proper thought should be conducted within intellectual discourse. It is important to impress upon the reader that since then logic has enjoyed an essential dominance as the final arbiter of thought.

The logic model differs from the causality model in that it shifts the focus from a direct study of phenomena to that of the study of "thinking about phenomena". With this change of focus, there is a subtle shift from the "observed" (phenomena) to the "observer" (one who thinks about phenomena). It is worth wondering why this change of emphasis was undertaken. Perhaps it came to be understood that thinking about phenomena had become sufficiently complex so as to merit more discipline.

Despite this difference, the logic model preserves some of the appeal of the causality model.

Like causality, logic also addresses the philosophical question of "Why?" Also like causality, logic tends to narrow the responsibility for various consequences to a limited number of actors. In so doing, like causality, it offers increased controllability.

And finally, logic – like causality - is deterministic. This means that a logical argument is repeatable. The same argument under the same conditions produces the same conclusion consistently time and again. It can be trusted.

All of this is convenient for anyone who wants to better understand and manage the phenomena of the universe.

How Logic Works

So, how does logic work? Without getting into the obvious and not-so-obvious details of the model, it is reasonable to say that the model consists of three components, all of which work together to produce an instance of a formal system. For example, Euclidean Geometry is an instance of a formal system. And so are classical mechanics, quantum mechanics, biochemistry and market economics. Any system defined by logical deduction is a formal system.

The first component of a formal system is a set of assertions called axioms. The axioms of a system are its starting point. Each formal system has its own unique set of axioms. Once the axioms are chosen, each of them is treated as though it is true.

The choice of axioms is left to the discretion of one who is inventing a particular formal system – as long as they collectively comply with two rules. The first rule, named logical independence, is that none of them can be derived from the others using the rules of logical inference (defined shortly). The second rule, named logical consistency, is that none of them can lead to the negation of any of the others through the use of the rules of logical inference.

There is also a third rule for the axioms that is highly desirable – completeness. A set of axioms is complete if any possible assertion that can be articulated from the elements of the system can be generated, by the rules of logical inference, from the axioms.

The second component of a formal system is the rules of logical inference. This is a set of rules that specify how the assertions of a formal system can be operated upon (combined) to produce new true assertions that can be created and added to the formal system. These rules detail these operations and various specific relations among them. These rules are shared amongst all formal systems.

The third component of a formal system is another set of assertions called theorems. Theorems are assertions that are formed from the rules of logical inference acting upon some combination of axioms and other theorems.

As already stated, together, the axioms and the theorems constitute a formal system.

In reality, a formal system has some other elements also, such as a set of undefined terms. But we shall ignore those details in this appendix.

A Closer Look at logical inference

It is reasonable to say that essential among the operational affects of logical inference are the two ideas of 1) necessary conditions and 2) sufficient conditions.

And, these are the two ideas of logic that concern phenomenal dependence – or the idea that whenever one phenomenon (“condition” or “assertion” in logic) occurs that something is suggested about whether or not some other phenomenon will also occur.

In other words, these two ideas of necessary and sufficient conditions are how logic treats the dependence between two assertions. And, recall that we have already said that assertions in logic are the correlates of phenomena in the causation model.

In this section, we would like to assess how well these two models – causality and logic – can play together. Ideally, it would be desirable to be able to translate either of these two models into the other, since this would enable them to “play together” handily.

In other words, we would like to see if we could show that the two ideas of “cause” and “effect” in the causality model and the two ideas of “necessary conditions” and “sufficient conditions” in the logic model are analogs of each other. We have already established that cause and effect are the two elements within causation that represent phenomenal dependence. And we have already established that necessary conditions and sufficient conditions are the two elements within the logic model that represent phenomenal dependence. However, we have yet to establish whether or not cause and effect are direct analogs of necessary conditions and sufficient conditions in the logic model.

If it is possible to show that the are analogs, then the two models would be essentially equivalent, and any differences between the two would be a matter of semantics. This would be especially convenient; because then we would be able to intertwine the two models at will in discourse and argument.

But, translating between the causality model and the logic model is a matter of being able to translate between the language of cause and effect and the language of necessary conditions and sufficient conditions. So, in order to investigate the “translatability” of the two models, we must first learn more about what necessary conditions are and what sufficient conditions are in logic.

We pointed out earlier that logic does not deal with phenomena directly. Rather it deals with phenomena only indirectly by working with assertions about phenomena. Assertions are indicative sentences that are associated with one of two truth-values: true and false.

Of course, assertions are much more manageable to deal with than phenomena directly. And this facility is part of the appeal of logic. Thus, instead of dealing directly with the phenomena of existence, logic deals with a linguistic artifact – the assertion - that can be taken as representing the phenomena of existence. And instead of being concerned with whether a phenomenon exists or not, logic then becomes concerned with the surrogate situation, the analog, of whether an assertion is true or false.

Having established this, we can now discuss how logic explains the 'truth dependence' (rather than actual "phenomenal existential dependence") between distinct assertions.

The reader should note that when logicians, mathematicians or other practitioners of the logic model are working in the model, their primary focus is on the necessity and/or the sufficiency of various conditions for each other. In many ways, it is fair to say that necessary conditions and sufficient conditions are a primary, if not the primary, tool of their trade.

In other words, if we want the causality model to be "translatable to", or an analog of, the logic model, then the notions of cause and effect within it must be able to be articulated in terms of the ideas of necessary conditions and sufficient conditions within the logic model. We shall attempt to make that translation shortly – after first explaining more precisely what constitutes these two conditions.

Sufficient Conditions

Given two assertions, call them "x" and "y", x is said to be a sufficient condition for y if the statement "if x then y" is a true assertion.

Inspecting this more closely, the statement "if x then y" is saying, "Knowing that x is true is sufficient information to also know that y is also true". Thus, x is a sufficient condition for y whenever the assertion "if x then y" is true.

For example, suppose the following statement is true: "If you have $20, you can purchase this book." Then, if this is true, then it is also true that "having $20" is sufficient to purchase this book. No other condition is required (if this statement is true). Therefore, the condition "having $20" is sufficient.

Necessary Conditions

However, there is something else that we can also say if we know that "if x then y" is a true assertion. It is this: "Knowing that y is true means necessarily means that x must also be true." In other words, x is a necessary condition for y whenever the assertion "if x then y" is true.

For example, consider that it is still true that "If you have $20, you can purchase this book." Then if you actually do purchase this book, then one can conclude also that you must have had $20 – because it is necessary for you to have had $20 in order to be able to purchase the book. Thus, having $20 was a necessary condition for purchasing the book.

The power of viewing phenomena in terms of necessary and sufficient conditions can perhaps be more easily appreciated when it is understood that if both "x is necessary for y" and "x is sufficient for y" are true, then x and y are logically equivalent assertions.

Arguments and Proofs in the Logic Model

Within the logic model a mechanism is provided for establishing proofs of assertions. This mechanism of proof is accepted within logic as concluding that an assertion is true.

Suppose one has a target assertion – that is, one that you would like to prove to be true. Also suppose that one also has a set (the axioms and theorems) of true assertions. If the logical conjunction (one of the operators of the rules of logical inference) of the set of assertions is a sufficient condition for the target assertion, then the target assertion to be proved to be true within the rules of statistical inference. Then this set of assertions is called an argument for the assertion.

What is significant here are that the rules of logical inference is an internal mechanism within the logic model that can prove that an assertion is true or false.

Adoption of the Logical View

The fact that the logic model contains a mechanism – logical inference – to conclude the truth or falsehood of its assertions, gives a powerful advantage to logic over causality – which has no such internal mechanism for proving that one phenomenon "causes" another. For a deterministic system of thinking – which we have shown that both causality and logic are, and which begs for high degrees of certainty - this is a distinct disadvantage of the causality model – and a distinct win for the logic model.

Consequently, since the end of the nineteenth century, mathematical logic has become almost universally adopted and almost totally unassailable in critical discourse. Even with the limitations of logic exposed by both Kurt Gödel and Alan Turing in the 1930s and 1940s, faith in logic's ability to represent the dependency relationships among assertions has been completely unshaken and has remained entrenched.

In the face of this deep penetration of logic as the arbiter of intellectual discourse, it is fair to ask "How have the causality model and the logic model coexisted, and how has their relative adoption rates been affected?"

The answer is that the critical thinkers have mostly and unquestionably accepted both. And furthermore, there is a largely unspoken and unchallenged assumption that the two are completely compatible.

However, in the wave of participation with David Hilbert's audacious program to rewrite all of mathematics in the language of mathematical logic – which occurred starting around the beginning of the twentieth century, attempts were made to rearticulate the cause and effect model into the language of logic – the language of necessary and sufficient conditions.

Lets now look at the extent to which that attempt was, and could be, successful.

Trying to Rectify the Causality Model with the Logic Model

On the face of it, it would seem there should be no difficulty in articulating cause and effect in terms of necessary and sufficient conditions. In fact, identifying a mapping between the pair (cause, effect) in the causality model and the pair (necessary condition, sufficient condition) would be how one would identify phenomenal dependency analogs between these two systems of thinking.

We would like to be able to show that there is a simple mapping, or translation, between "cause" and either necessary conditions or sufficient condition, and "effect"

and either sufficient conditions or necessary conditions (whichever was not mapped to "cause"). Because, such a translation would show that there is a straightforward and intuitive correlation between the two systems.

If we have difficulty in finding, or cannot find, such a simple translation between the two systems, then believing that the two systems are semantically analogical is called into question.

But to see that there is in fact a difficulty, consider the following question:

If, within the language and thinking of the causality model, "x causes y", then which of the following statements from the logical model must necessarily be true for x and y:

1. x is a necessary condition for y.
2. x is a sufficient condition for y.
3. x is both a necessary and sufficient condition for y.

Causes are Not Always Necessary Conditions for Their Effects

The first sentence (that x is a necessary condition for y) does not always hold when "x causes y". This is true because, even if x is a cause of y, it may not be the only condition by which y can be caused – in which case it would not be a necessary condition for y.

For example, in a game of billiards, I may strike the cue ball in such a manner that it then strikes the 6-ball and causes it to go into the side pocket. However, that action does not preclude my taking a different shot – perhaps a bank shot – that also causes the 6-ball to go into the side pocket. Thus in the causality model, "cause" does not imply a necessary condition within the logic model.

Causes are Not Always Sufficient Conditions for Their Effects

The second sentence (that x is a sufficient condition for y) does not always hold either when "x causes y". This is true because multiple causes often occur together, and any one of them by themselves are not sufficient to cause y.

Often, whenever x causes y, it is actually a member of a set of several conditions all of which must be true in order for y to be true. In the billiards example, even if the cue stick strikes the cue ball in such a way that its trajectory subsequently strikes the 6-ball in such a way that its trajectory passes into the side pocket, it is also necessary that a number of other conditions to simultaneously hold in order for the ball to actually go into the side pocket.

For example, none of the other billiard balls currently on the table can lie in the path of either the cue ball trajectory or the 6-ball trajectory. Thus, for each of the balls currently in play, its not lying along that trajectory is also necessary for the ball to go into the pocket. This means that none of them alone can be a sufficient condition. Plus, this also means that our condition "x" – that the cue stick strikes the cue ball in such a way that it strikes the 6-ball in such a way that it goes into the side pocket – also cannot be a sufficient condition for the 6-bll going into the side pocket.

In other words, it often happens that when no scientist would disagree that "x causes y" is a reasonable and true assertion, it cannot then be said that x is a sufficient condition for y.

It cannot be Said that Causes are Necessary and Sufficient Conditions for Their Effects

Thirdly, from what we have just said, whenever x causes y, we can't rely on x being a necessary condition for y, and we can't rely on x being a sufficient condition for y. Therefore we obviously can't rely on x being both a necessary and sufficient condition for y.

Besides, if a cause (causing phenomenon) were required to be both necessary and sufficient for an effect, then in mathematical logic, the two phenomena would have to be equivalent phenomena.

This is true because in the logic model, two assertions that are both necessary and sufficient for each other are equivalent assertions under the model. In fact, the test for equivalence in logic is precisely that: if x is a necessary condition for y, and y is a necessary condition for x, then both x and y are both necessary and sufficient conditions for each other, and are equivalent conditions (assertions).

But requiring a cause to be equivalent to an effect would, then also require the effect to be the cause within the logic model. This is not a property that is shared by a cause and its effect within the causality model.

In fact, with causality, a cause and its effect share a strictly asymmetric relationship. Thus, requiring that cause and effect be both necessary and sufficient conditions for each other disturbs this essential asymmetry of cause and effect.

Thus, there are a number of different ways to see why cause and effect in the causality model cannot be correlated to necessary and sufficient conditions in the logic model. So, while it is true that cause and effect are the two mechanisms within causality that represent phenomenal dependence; and it is also true that necessary conditions and sufficient conditions are the two mechanisms within the logic model that represent phenomenal dependence; we cannot find a mapping between the two models that equates them (or preserves the property of phenomenal dependence).

Therefore, we cannot conclude that causality and logic are equivalent models.

Can The Causality Model be rectified with the Logic Model?

So we must conclude that a cause and effect relationship between phenomenon x and phenomenon y cannot be adequately articulated into the language of necessary and sufficient conditions – that is, into the language of mathematical logic.

Given all of this, it is reasonable to ask, "What have been the relative adoption rates of the causality model versus the logic model?" The answer is that both are still used, but in critical scientific and other intellectual practices, the logic model dominates the causality model.

This dominance of logic should not be surprising. After all, the ability of logic to mitigate between true and false assertions is a powerful capability for a deterministic system of thinking to possess. And the failure of the causality model to provide internally a mechanism for resolving whether one phenomenon enjoys a causative relationship with another is a conspicuous omission.

Still, causality is deeply rooted in our intuitions, and is not easily abandoned. So the two systems of thinking both persist, and are often intertwined. Nevertheless, the dominance of logic over causality as the ultimate arbiter of deterministic thought processes is settled.

To support this observation, consider this. In elementary, high school and undergraduate education, the causality model is used to motivate scientific assertions, such as laws of nature and other principles. This is primarily because the causality

model is rooted in our deep intuition. In order for students to find these assertions plausible, the causality model must be used because of its appeal to their deepest intuitions.

But the logic model is about the rules of thought. Whenever complex thought is involved in scientific deliberation – which it increasingly is with the advent and acceleration of scientific complexity, then concern for accurate, consistent and critical thought begins to outweigh appeal to intuition.

Impact on the Field of Mathematics

Modern mathematics since the end of the 19th century, as we have discussed, is based upon the logic model, and does not refer to the causality model at all. In fact, if causality were ever a part of mathematics, then it was abandoned after Hilbert's program to re-write mathematics in terms of mathematical logic and set theory around the transition to the 20th century. And, as we have shown, the logic model cannot be consistently translated into, or shown to be an analog of, the causality model.

Hilbert's program and other changes in mathematical attitudes about mathematics in the late eighteenth centuries convinced many pure mathematicians that mathematics stands alone from nature and the sciences. To these mathematicians, mathematics began to be understood as an abstract exercise in pure logic. Certainly, mathematics is a system for creating abstract analogs, or models, of "realizable" systems in nature and in science. But that does not make it a part of science.

Understandably, perhaps few physicists or other scientists have ever seen mathematics in this light. Of course, scientist's application of mathematics is to model, or analogize, systems of interest in their own research pursuits – so it would be a mute point to them whether mathematics is "science" or something else. Generally, the claim that mathematics is not science, but "something else", is simply not an interesting discussion for many, if not most, scientists.

But for mathematicians, the observation that mathematics is not natural science was both revolutionary and important. It was revolutionary, because since the time of the ancient Greeks, mathematics was treated as empirical. It was important because such a realization resulted in determining how the field of mathematics would change and grow. If mathematics is not an empirical observation of an external natural world, then what is it? And how would it be defined, develop and advance? All of this was the essence of Hilbert's program and other meta-mathematical exercises that were going on at that time and since.

The Statistical Inference Model

During the latter parts of the 19th century, it began to become evident to critical thinkers, scientists in particular, that systems of interests in nature are usually too complex to wholly observe at once and to wholly understand. Consequently, the best that science can consistently expect for partial degrees of certainty regarding its knowledge of natural systems.

In other words, whether or not nature is deterministic, the human pursuit of the knowledge of it cannot be. So, human endeavors to understand the universe must in general be treated non-deterministically by its pursuants. One of the first scientists that we know of in the modern era to come to this understanding was James Clerk Maxwell, who introduced statistics into the study of the atomic nature of matter.

Until this understanding began to emerge, mathematics was treated strictly deterministically, and it wholly adopted the rules of thinking dictated by the logic model. But with the revelation of partial knowledge and the increasing acceptance of the necessity for science to work within a domain of various degrees of uncertainty, a new mathematics was born – probability theory.

Probability theory was a new way of thinking. Probability theory is interested in identifying all of the alternative possible outcomes of an ensuing event. It assigns a non-negative number (fraction), the probability of the alternative, to all of those alternatives, with each assignment giving a measure of the likelihood that its alternative will be the actual outcome of the event, once the event occurs. In addition, for convenience, all of the probabilities are "normalized" so that they sum to 1 for a given event.

Within its domain of application, probability theory can seldom prove anything. The best one can usually hope for is to infer that it is reasonable to accept an assertion, based upon it probability, or some function thereof.

Of course, some subjective judgment is involved with this type of thinking, because one must first decide what one's "threshold of acceptability" is. In other words, one must first decide how probable the truth of an assertion must be before one "feels satisfied" in "behaving as though" the assertion is true. If the probability of some assertion is found to be lower than that threshold, then it is reasonable to refuse to statistically infer the truth of that assertion.

And, of course, to be intellectually honest one must maintain awareness that such an assertion has not actually been deterministically proven through statistical methods – and that probabilities of assertions can change with time. Rather, one should be vigilant and either 1) open to new evidence that the probability of the assertion has changed, or 2) that one's "acceptability threshold" (or so-called significance level) for the minimum probability of truthfulness has changed.

This new way of thinking is called statistical inference. Notice that the word "inference" is used, rather than "implication", because nothing has actually been "proved".

Obviously, statistical inference, as a way of critical thinking, gives up a lot of comfortable properties of the deterministic models of causality and logic – properties like certainty. But under the circumstance recounted above (random variation), these deterministic models can simply no longer be supported.

Like the logic model, statistical inference is ultimately concerned with dependency relationships between two (or more) assertions. Recall that the logic model works by deducing deterministically that various truth-values of one or more assertions collectively imply a truth-value of another assertion (a theorem).

However, statistical inference operates on the principle that the probability of a concluding assertion being true, given that certain other assertions are true, lies within an acceptable threshold of "believability" (significance level).

Let us now look more closely at the model of statistical inference.

Chance Variation: The Ubiquity of Uncertainty and Randomness

This appendix makes the following working assumption about science as a human endeavor:

> Science is interested in constructing a body of assertions about the natural world that are sufficiently reliable and trustworthy that it becomes reasonable to drive for near-universal adoption of those assertions by critical thinkers.

If this is true about science, then trying to develop this “body of assertions” by collecting the opinions, speculations, philosophies and biases of participants has not proven to be a way to develop unanimity.

Rather, the endeavor of science has long turned to collective observation as a way to rid its pursuits of these troublesome biases. In fact, science has developed an elaborate procedure, collectively honed over time, in order to ensure against such difficulties. We shall visit this procedure, called the scientific method, in some detail below.

But first we must discuss another difficulty that persists – even in the face of this collective observation procedure named the scientific method. This new difficulty is the phenomenon called chance variation. The scientific method cannot banish chance variation, however. Rather, it must accommodate it, as we shall see.

Chance variation is the repetition of the same procedure that results in possibly different outcomes with each repetition. In other words, with chance variation, there is no guarantee that the same outcome will occur consistently in each of multiple repetitions of the same procedure (even using the same inputs).

Of course, some may argue that, if different outcomes are had with each repetition, then the procedures of each repetition were not actually the same. The researcher only thought that they were. And, of course, this is precisely the position of a philosophy that the universe must be deterministic - and only be deterministic.

But we are not here to argue whether or not non-determinists are befooling themselves (or whether determinists are), or to argue philosophy at all.

Rather, we are saying (mathematically) that 1) situations in which repetitions of the same procedure always results in the same outcome are called deterministic. Whereas, 2) situations in which repetitions of the same procedure can result in the different outcome with each repetition are called non-deterministic. Without loss of generality, we shall stipulate that both types of situations may exist.

(Philosophers may argue that either one of these two situations cannot ever exists. However, behaving as mathematicians, we merely identify these two logical possibilities and work with them.)

Moreover, we shall stipulate that whether the repetitions are of the same procedure exactly or not is left to the discretion of the observer/experimenter. Thus, the same experiment can be considered either deterministic or non-deterministic depending on the fineness of the definition of “sameness of procedure”. For example, tossing the same pair of dice multiple times can be considered by an experimenter to be the “same procedure across multiple repetitions”. However, another investigator may insist that the procedure actually changes across these repetitions because the initial conditions vary with each repetition – and that therefore the experiment is not non-deterministic. This appendix considers both of these positions to be viable, and which one is chosen is left to the discretion of the investigator.

Having established this critical distinction between deterministic and non-deterministic, lets return to the predicament that scientists that are attempting to define and codify the scientific method find themselves in.

We mentioned above that, at least by the 19th century, scientists were beginning to encounter difficulties with the fact that natural systems that they were interested in studying were becoming too complex to be able to observe wholly – all at once.

So, they were reduced to observing what they could – which turned out to be relatively small subsets of those natural systems of interest. In other words, in the language of statistics it had become impossible to make observations about the population (the whole of their system of interest). And they were reduced to make observations about samples of the population (subsets of their system of interest).

However, the real trouble came when they discovered inconsistencies among the samples, even though the samples were taken from the same population.

Now, this situation is exactly what we described above when we defined non-determinism. This situation is precisely the repetition of the same procedure with the production of different results.

In this case, the "procedure that was being repeated" was this:

> The deliberate and careful selection of a sample from the population, followed by the observation of the behavior of that sample.

And, the results that were inconsistent (often changed for each repetition) were the various measures that were consistently taken for each repetition. These measures eventually became formalized by the discipline of mathematical statistics and were given names such as mean, median, mode and moments. In these terms, the problem is that the value of these statistics was found to vary across these multiple samples – all of which were taken from the same population.

The fact that these measure were generally different for each sample taken from the same population meant that a non-deterministic thinking process was called for. This fact did not completely rule out either the causality model or the logic model – since they could perhaps be brought to bear somewhere within the scientific method procedure.

However, it did mean that some non-deterministic critical thinking process must of necessity be involved within the scientific method. And, moreover, it meant that such a non-deterministic thinking process would of necessity be the final arbiter within the scientific method for arriving at conclusions pertaining to the acceptance or rejections of proposed assertions.

This thinking process that the scientific method arrived at, that is able to represent the non-deterministic nature of scientific empirical behavior, is the statistical inference model that we discussed at length above. Statistical inference was derived from probability theory and is properly a subject of mathematical statistics. We shall develop it more fully in the following section.

Statistical Inference Described

Suppose that we are observing the behavior, the dynamics, of some complex phenomenon. After some amount of observation, we may begin to suspect that "some causal agent is at work here". We may become suspicious, more specifically, that "agent x is causing sub-phenomenon y".

However, we have a couple of stumbling blocks in our way that are preventing us from proving that "agent x is causing phenomenon y". One of them is that the causality model is of no help here, because it provides no criteria as a part of its model for how to determine that some phenomenon is causing another phenomenon. Causality as a way of thinking asserts that phenomena cause other phenomena. But causality is not a practice, because it offers no mechanism for ascertaining which phenomena actually causes which other phenomena.

But worse than that, our system of interest is too complex to be observed all at once. So we are reduced to taking samples of it – subsets of it – and observing them. And we find that these samples – as sample are wont to do – are not exhibiting the same behavior as each other.

Chance variation has entered the picture and foiled our ability to use either the causality model or the logic model of critical thinking to model sample results. These inconsistencies all by themselves prohibit the use of any deterministic thinking process from being able to "prove" that "agent x is causing sub-phenomenon y".

However, we still strongly suspect a causal relationship is at work. Perhaps what is producing these inconsistent results across these samples is that some other agent, call it z, is also at work alongside agent x, and that the two together are producing the inconsistent results across then sample. If we could only isolate agent x away from agent z, then perhaps these inconsistencies would disappear. Unfortunately, it is not always within our reach to make that separation.

Or, it could be a million other complications that are causing these inconsistencies across these samples. One possibility is that the dynamics of our system of interest is, itself, non-deterministic – as well as our human processes of observing it.

Anyway, what we know is that, given our situation, no deterministic thinking process can sort this out. But, perhaps we could take another approach – one that may enable us to make some assertion, even in the face of the evident chance variation.

> Suppose we could show that that variations we are observing are very unlikely to be the result of chance!

If we could show that "chance is at work" is not a reasonable explanation, then it would be reasonable to conclude that some other mechanism "other than chance" is at work in the phenomenon that we are observing. When we experience such unexpected behavior, we often use the phrase "it's not just a coincidence that…". The use of this phrase signals that we suspect something is at work other than "chance".

We encounter this kind of thing in ordinary living very often. For example, suppose that we experience "ordinary levels of rainfall" over each of several years. And then, for a number of years in a row, we begin to experience "abnormal" levels of rainfall for a number of other years in succession – either more or less rain. At that point we begin to suspect that something is at work "other than chance".

What statistical inference is in a position to do is to assess whether such observed behavior is unlikely to "occur due to chance". And, if so, then it would be reasonable to assert that "something is probably at work here other than chance". Of course, this approach does not prove that something is at work other than chance. But at least one can say, "The observations indicate that the probability that this situation is working by chance is highly unlikely". Therefore, the observations reveal that the "evidence is such that it is reasonable to behave as though something is at work other than chance".

Of course, "occur due to chance" is not a very precise statement. In order to ensure that we are dealing with a critical thinking process, we must be more precise. In fact, we need to provide some kind of criteria – if possible - so that we can unambiguously (if not certainly) decide whether or not it is "reasonable" to conclude that the phenomena occurred by chance.

We shall look to probability theory to hopefully provide us with this kind of precision.

First, though, lets look at a more specific example in order to set the stage for this investigation. Consider a pair of typical six-sided gaming dice. Suppose that we

observe a player who has gone on an unusual winning streak when this pair of dice is involved. The question then arises as to whether the player is cheating. Perhaps he has "loaded" the dice, or performed some other form of tampering, so that they are no longer "fair dice". The question is "How can we go about testing these dice to see if we can get some evidence that they have been unfairly tampered with in some way?"

Before we describe the statistical inference process in some detail, it may be useful to introduce the approach with a kind of "cartoon explanation" – an explanation that sets out the basic idea and intent, without being precise and accurate. So, here is this "cartoon explanation" of the statistical inference process:

> Statistical inference doesn't try to prove anything. Rather, it looks for the actual occurrence of a rare event. Or, at least such an event would be considered rare if one assumed that it was under the sway of chance alone. We then look to see what the probability of that rare event would be if chance alone actually were "controlling" the situation. To do that, we calculate the probability of that rare occurring event – assuming that the actual phenomenon (e.g. tossing a pair of dice) is operating by chance alone. If that probability is sufficiently small (smaller than a previously decided threshold), then the conclusion is reached that the actual observed rare event is simply not behaving like a chance happening – that it is simply has too low of a probability to reasonably happen by chance alone. ("There are no coincidences.") Therefore, it must be more likely that it is behaving like "non-chance" happenings. Therefore, more than likely, "something else is at work other than chance".

That is, statistical inference begins by asking whether the probability that an observed phenomena, called the p-value, using the assumption that the phenomena is operating by chance, is so unlikely as to call into question the assumption that it was, in fact, behaving by chance alone.

Here is the approach that statistical inference uses with our pair of dice. First, one would notice that a certain player has been "just too lucky" in the last several games, and the suspicion arises that he "must be cheating". The "rare occurring event" is the fact that he is winning too often. The observer begins to suspect that the probability of his winning that often just by chance alone is too rare of an event to actually encounter in real life.

But we have to get more specific in order to use statistical inference. The phrase "just by chance alone" translates to something about the probability distribution of a fair pair of dice. More specifically, it translates to saying the probability of a player winning as often as this "cheater" is so low for a fair pair of dice ("a zillion to one") that I find it unbelievable that the probability distribution of fair dice is the one that describes what is going on in this particular game. And since, it is unbelievable to me that the probability distribution of fair dice describes this game, then unfair dice must be involved.

To be mores specific, what has been noticed is that all of the other players have "just been too unlucky" – except for this player. The reason the other players have been "unlucky" is that their first roll of the dice has too often produced a sum of "2" – a configuration known as "snake eyes" in dice parlance. The unluckiness comes from the fact that rolling snake eyes on the first roll is a loosing event. The player loses her bet and the right to continue rolling the dice for the present round if she rolls snake eyes.

The suspicion is that the cheating player is somehow "slipping in" a loaded pair of dice when other players are rolling, and then using a fair pair of dice whenever he is rolling. In other words, we want to really test the hypothesis that "The dice pair being used by the non-cheating players is 'unfair'."

The way that statistical inference precedes in this situation is to 1) select the event of rolling snake eyes, and then 2) calculate the probability of rolling snake eyes (p-value) using the assumption that a fair pair of dice is being used.

After that, one wants to see if this calculated p-value is so low that rolling snake eyes is too improbable to have actually occurred within the probability distribution for a fair pair of dice. If the probability of rolling snake eyes really is that low, then there is more than likely something wrong with our assumption. Of course, the only assumption that we have made is that the dice were fair. So, that assumption must have been wrong. Therefore, it is more likely that the dice that rolled snake eyes were an unfair pair of dice!

Of course, before we calculate the probability of rolling snake eyes, we should determine in advance "Just how low of a probability is too low to satisfy our feelings of believability?" This will be our level if believability, and it is a judgment call based on our comfort level. This is called our "level of significance" for the test. Typically, scientists choose either .05 or .01 for a significance level. A significance level of .05 would mean this: "If the probability of snake eyes (p-value) turns out to be lower than .05, then I'm going to find it hard to believe that the dice were fair.

Other scientist might not want to risk accusing the player of cheating if he really wasn't cheating. This would be an "error of commission", or "Type I error". So they might select a significance level if .01 to reduce the likelihood of such an error. Unfortunately, by playing it safe and selecting a lower significance level, they have run the risk of an "error of omission" – of not accusing a player of cheating when he actually was. This is known as "Type II error". The choice of a significance level is a matter of judgment. You have to decide whether you had rather risk Type I or Type II error, and choose a significance level accordingly.

Anyway, after we select our significance level, then we calculate the probability of rolling snake eyes when using fair dice (the p-value). If this probability this probability turns out to be less than our chosen significance level, then we shall conclude, "It is unlikely that the dice used to roll these snake eyes are fair dice."

We know that fair dice exhibit a particular probability distribution. In this distribution, for example, the probability of rolling a "7" (the sum of the two dice) is 6/36, or 1/6. This is because there are 6 ways that the two die can collectively show a sum of 7, while there are 36 ways in all that they can land.

Using the same thinking, we can enumerate all of the possible sums that the two die could land, and similarly calculate their probabilities. Therefore, we can calculate the probability distribution that describes the chance variation of the two dice if they are fair. Being a probability distribution, we can also calculate both the mean and the standard deviation for this distribution.

Now, we know that even if the two dice are fair, it is still possible for a sample of tosses to exhibit unusual behavior. For example, it is possible to toss a fair die 5 times in a row and to get a sum of 2 ("snake eyes") all five times. However, that eventuality has a very low probability – at least for a fair pair of dice!

And, if someone tossed a pair of dice five times in a row, and "snake eyes" came up all five times, one could reasonably expect to suspect that the dice had been tampered with. This is because the probability of rolling snake eyes five times in a row has a very low probability. In other words, if that actually happened, then it would be reasonable to conclude, "Something is at work other than chance"!

Another way to say the same thing is that: "If snake eyes landed five times in a row, then the likelihood is that the pair of dice used is from some other probability

distribution other than the probability distribution for a fair pair of dice. Therefore, it is reasonable to conclude that this pair of dice has a different probability distribution than a fair pair, and that it must not be a fair pair of dice."

Having set up this example, lets now look at how one can use probability theory to make an unambiguous decision that "something is at work other than pure chance"!

First, if you are suspicious that "something is at work other than pure chance", then you must have observed some behavior that seems "out of the ordinary" for a phenomenon that is suppose be behaving according to chance. At this point, you should identify that particular type of behavior. For our "snake eyes" behavior, it is that snake eyes came up five rolls in a row. This is called your "test event". And if there is some number that you can associate with it, then the number is called your "test statistic".

We are going to assume that we have observed an occasion during the game when a player other than our suspect has, indeed, rolled snake eyes 5 times in a row (across 5 contiguous points of course). This will be our test statistic. It seems to us that the probability of a fair dice coming up snake eyes five times in a row is so small that it is more believable that our assumption that the dice are fair is more likely to be correct, than snake eyes for fair dice actually coming up 5 times in a row.

In other words, the probability that a fair pair of dice would produce five snake eyes in a row seems too small to be believable. What seems more believable is that a pair of dice that produced that result would be an unfair pair of dice!

But you don't really know yet what the probability is for snake eyes to be rolled five times in a row is for a fair pair of dice. So you must first calculate that in order to know whether the probability of that is as small as you suspect.

As a matter of fact, before you calculate the probability of a fair pair of dice rolling snake eyes five time in a row, you should first decide what your criteria is for "too low of a probability to be believable". Any probability below that chosen threshold would then be "unbelievable" to you. This decision is a matter of your own level of comfort. It is subjective. For you, is it a probability of .05? Is it a probability of .01? Or is it some higher level of probability. This "comfort level" of probability is called the significance level.

You are going to first select this significance level based on your own personal comfort. Suppose that you decide that .05 is your significance level. This means that if the probability that you calculated for five snake eyes in a row for a fair pair of dice is lower than .05, then you are finding it difficult to believe that the pair of dice used is actually fair. This disbelief comes from the fact that you think that rolling five snake eyes in a row is too rare of an event for a fair pair of dice – and that therefore you are having difficulty believing that the dice that were just used are actually fair.

So, once you have established the significance level that you are comfortable with, then you actually calculate the probability of a fair pair of dice producing five snake eyes in a row.

So, lets do the calculation. We shall assume that the two dice do not influence each other's outcome. This is called statistical independence. This assumption allows us to calculate the probability of this event easily.

The probability that a fair pair of dice would roll snake eyes five times in a row – assuming that each toss is statistically independent of the other throws – is

$$\text{p-value} = (1/36)^5 = 1.653\text{x}10^{-8} = .00000001653$$

This value is considerably less than a probability of .05.

Our conclusion then is that the probability of a fair pair of dice would produce five snake eyes in a row is 1.653×10^{-8}. And, of course, this probability is very much smaller than our chosen significance level of .05.

Therefore, we can conclude that the probability that a fair pair of dice would produce five snake eyes in a row is just too unlikely for you to believe that the pair of dice actually used could reasonably be assumed to be a fair pair.

In the process of statistical inference, the thought procedure is set up to accommodate this kind of thinking.

If the calculated probability is greater than the significance level

The preceding example worked from the assumption that our calculated probability of rolling snake eyes using a fair pair of dice is less than (or equal to) our chosen significance level.

However, what happens if it turns out that our calculated probability of rolling snake eyes with a fair pair of dice is greater than our chosen significance level. In such a case it would be reasonable to assert, “Snake eyes could actually have been rolled by fair dice”. And therefore we maybe could reasonably conclude that “The dice that were use to roll snake eyes in this observation were, indeed, fair.”

But, wait a minute! Maybe a pair of unfair dice could also have a probability of rolling snake eyes that is greater than our chosen significance level. We don’t know what the unfair dice’s probability distribution actually is. So there is no reason for us to know that the probability of rolling snake eyes in that distribution is not also greater than our significance level. Therefore, for all we know, it might be believable to us (according to our chosen significance level, for the event to have occurred in either the fair dice distribution or the unfair dice distribution – given that our calculate probability is greater than our significance level.

Therefore, finding out that the probability of rolling snake eyes with a fair pair of dice is greater than our significance level does not prove that the pair of dice used must be fair.

In other words, if our calculated probability of rolling snake eyes is greater than our chosen significance level, then the results are inconclusive. We don’t have enough information from this fact alone to conclude that it is the dice used are probably fair.

Thus, if the only probability distribution that we know is that of a fair pair of dice, then we can only make a conclusion if our calculated probability of rolling snake eyes is less than our chosen significance level, On the other hand, if it is greater than our significance level, then we can make no conclusion.

However, if we happen to know the probability distribution of the unfair dice, then we can go further and calculate the probability that snake eyes could have reasonably come from that distribution. That is, we can also calculate the probability of rolling snake eyes using that distribution. And if that probability is less than our chosen significance level, then we can also reject the likelihood that the dice were described by that distribution as well.

Unfortunately, we don’t actually know what the probabilities are for rolling snake eyes are with an unfair pair of dice is. For all we know, it may also be reasonable for snake eyes to be rolled using an unfair pair of dice too! Therefore, even though our calculated probability for rolling snake eyes is greater than our chosen significance level, that does not prove that fair die were actually used!

Consequently, unless we know more about the probability distribution of the “unfair dice”, we can only make a reasonable conclusion whenever the calculated probability is smaller than our chosen significance level. If it is greater than our significance level, we are not in a position to draw any particular conclusion.

Thinking about this situation a little deeper, the more one knows about the potential probability distributions that are available to possibly describe the random behavior of the two dice, then the more opportunities you have of rejecting any hypotheses that the behavior (outcome) that you are calculating for came from each of those possible distributions.

However, this approach does not offer any reason to accept any hypothesis that states that the observed behavior does, in fact, come from any of those distributions.

Thus, statistical inference is set up to reject hypotheses that state that the observed behavior is reasonably described by some particular probability distribution. Use of this approach usually begins by assuming a hypothesis that states that some “chance” distribution would reasonably describe the observed behavior. This is called the null hypothesis. If the calculated probability is less than the chosen significance level, then it is reasonable to reject the null hypothesis. One could then reasonably conclude that the observed phenomenon did not “occur by chance alone”.

If there are other suspected probability distributions that are also candidates for describing the chance behavior of the observed phenomena, then other null hypotheses can also be articulated and subjected to the same test. In this way, statistical inference proceeds to eliminate various hypotheses that the observed behavior is likely to behave according to any particular probability distribution.

However, the practice of statistical inference in science admittedly does not limit itself to the rejection of null hypotheses! Under special circumstance, beyond the scope of this appendix, it is possible to conclude that certain hypotheses can be accepted, rather than having rejection of hypotheses as the only viable possibility. But, in most cases, statistical inference never provides enough evidence to actually accept any hypothesis – only to reject hypotheses.

Nevertheless, whether it is justified or not, scientists often formulate alternative hypotheses to null hypotheses – which they decide to accept if they are able to reject a null hypothesis.

Of course, this is reasonable, for instance in our example, if 1) the distribution of the unfair dice is known, and 2) the probability of the “rare event” (snake eyes five time in a row) were calculated to be high (or higher than our significance level) within that distribution.

In our example, as likely in most cases, there is no reason to assume that the distribution of the unfair dice is any distribution in particular. In fact, there are an infinite number of probability distributions that it could be. What we actually have to work with is the distribution of the “fair dice”. Unfortunately, the fair dice distribution may be enough to decide against the dice being fair. But it generally does not provide us with enough information to conclude that the dice are actually unfair!

We would generally need to be able to narrow down the unfair distribution to a small set of possibilities in order to make such determinations. Such is the nature of statistical inference.

Assessment

The procedure that we just performed is called hypothesis testing.

What we did was to test the behavior of an actual system (the two actual dice) to see if they exhibited behavior that that would be reasonable for a system that was "operating purely by chance" – according to a particular probability distribution for fair dice.

In order to do that, we had to first be more specific about "what behavior would be reasonable for a system that was 'operating purely by chance'" actually means. Our approach to that was notice that "chance behavior" means that the system would be described by some probability distribution.

So, our first task was to determine what the probability distribution for the rolling of two fair dice actually is. From our definition of this probability distribution, we can tell what the probability of any particular kind of behavior is.

Further, we decided that if we performed an experiment with our pair of actual dice, we could ascertain what the probability of its behavior is.

If the actual observed behavior had a low probability for a pair of fair dice (lower than our previously chosen probability, then we would conclude that our pair of dice is probably not fair.

Statistical Inference Process Summary

Our example is a particular type of a more general thought process called statistical inference.

What all statistical inference has in common is that

1. It makes a choice as to what chance behavior means for a particular system of interest. It then specifies that chance behavior by identifying a probability distribution for it. We shall refer to this distribution as the chance distribution.
2. It makes the temporary testable assumption that the system of interest actually does behave according to "chance" – as described by the identified probability distribution. This hypothesis is called the "null hypothesis".
3. It then pre-determines a level of significance. This is a probability threshold such that, if the calculated probability of the "rare event" turns out to be small than the significance level, the it will be judged by the observer that the actual probability of the rare event is simply too low for it to be believable that such an event is described by the chance distribution.
4. It identifies the occurrence of an actual event that the observer suspects is too rare to be explained by the null hypothesis – the assumption that the chance distribution describes the observed behavior
5. It then calculates the probability of the observed ("rare") event by using the chance distribution. Lets call this the chance probability of the observed event.
6. If the chance probability of the observed event is smaller than the significance level, then the conclusion is that the probability of the observed event is too rare to for it to be believable that the its probability distribution is, in fact, the chance distribution. If this occurs, the null hypothesis is rejected.
7. However, if the chance probability of the observed event is greater than the significance level, then we do not have enough information to reject the null hypothesis.
8. If the null hypothesis cannot be rejected, then some versions of statistical inference allow the acceptance of some alternate hypothesis. However, this is not always strictly justified by statistical inference model. An occasion when the prescription of

alternative hypotheses is mathematically justified is beyond the scope of this appendix.

Phenomenal Dependency in Statistical Inference

In this appendix, we have emphasized that we are comparing three models of critical thinking; and we have said that all three of the models, in one way or another, pertain to phenomenal dependence.

By this we mean that the model provides some mechanism for representing dependency between two or more phenomena – or between whatever entity type the model uses to represent phenomena.

In the case of the causality model, phenomena are addressed directly. What is at issue is the existence of these phenomena and whether their existence depends somehow on other phenomena. And the dependency relationships between phenomena are represented by the cause and effect relationship.

On the other hand, the logic model uses assertions, or conditions, to represent phenomena. What is at issue is the truth or falsehood of these assertions. And logical inference is the relationship between assertions – specifically necessary conditions and sufficient conditions, rather than cause and effect as in the causality model.

Essentially, then, in the logic model, assertions are the analogs of phenomena, truth is the analog of existence, and “x implies y” is the analog of “x causes y” in the causality model. However, we tried to show that one couldn’t simply map cause and effect in the causality model to necessary conditions and sufficient conditions in the logic model. So the two models of thought are not strictly analogs, or correlatives, of each other. They both, however, both address the nature of phenomenal dependence.

What we want to do now is to explain how statistical inference fits as an analogous model into this phenomenal dependence scheme if we want to offer it as a third model of critical thought.

In the first place, there is something that is conspicuously missing in the statistical inference model that would prevent it from fitting into this phenomenal dependence scheme – at least so far.

What is missing is an analog of a relationship between two phenomena in the causality model, and of a relationship between two assertions in the logic model. In other words, what is so far missing from the statistical inference model is an analog to cause and effect in the causality model, or to logical inference in the logic model.

So, we shall now show how phenomenal dependence is represented in the statistical inference model.

We shall first comment that the statistical inference model, like the logic model, also uses assertions as its analog of phenomena in the causality model. However, the statistical inference model calls assertions by a different name – “hypotheses”.

But, unlike the logic model, statistical inference does not attempt to show that one hypothesis implies another.

Rather, statistical inference – as we have seen – investigates whether it is reasonable to believe that an observed phenomenon could have actually occurred “by chance”. If the observed phenomenon has a calculated probability that is unbelievably small as calculated by a chance distribution, then concluding that it could have actually occurred by chance is deemed to be unreasonable.

In such a case, then, any assertion that “the observed phenomenon occurred by chance” is unlikely to be true. Such an assertion, within the statistical inference model, is named the “null hypothesis”. If the calculated probability (“p-value”) of the observed phenomenon is, in fact, unacceptably low, then – within the framework of model – one rejects the null hypothesis”.

“Rejecting the null hypothesis” in the statistical inference model is an analog to proving the falsehood of an assertion within the logic model. Of course, “rejecting a null hypothesis” within the statistical inference model is very different from “proving falsehood” in the logic model. The first has a degree of uncertainty about it, whereas the second is brimming with the confidence of certainty.

Statistical inference also provides mechanisms that are analogs of “true” and “false” in the logic model. In statistical inference, these analogs pertain to accepting and rejecting hypotheses. Of course, the analog mapping is not precise here either. The issue of rejecting hypotheses is easy to map to “false” in the logic model. But the issue of an analog to “true” in the in statistical inference is, as we have seen, murkier. Moreover, acceptance and rejection in statistical inference do not share a precisely analogous relationship to each other within statistical inference that “true” and “false” do within the logic model – where they are complementary.

The principle difference between statistical inference on the one hand, and either the causality of the logic model on the other, is that of the degree of certainty involved. Within the logic model, “proving falsehood” is an absolutely certain act. But within the statistical inference model, “rejecting a hypothesis” is not absolutely certain at all. Rather, it is a statement of probability – or degree of certainty. And, this fact characterizes perhaps the principle difference between a deterministic model and a non-deterministic one.

In any event, at this point we have established some analog within the statistical inference model of the notions of “assertion” and of “truth and falsehood” in the logic model. These are the notions of a “null hypothesis”, “rejection of the null hypothesis” and possibly the “acceptance of an alternative hypothesis in statistical inference.

However, we still have not developed within the statistical inference model the analog of “cause and effect” in causality or “logical inference” in logic. Both of these pertain to the truth of a compound statement regarding two assertions, one of which enjoys a phenomenal dependence on the other. For example, in causality, we observe the existence of one phenomenon causing another. In logic, we make the statement “x implies y”, and then we set about to prove it using some set of axioms and the rules of logical inference.

Here is what is done in the statistical inference model to represent phenomenal dependence….

We articulate a hypothesis concerning two assertions to the effect that one of the assertions implies the other.

Subsequently, like any other hypothesis that we investigate in the model, we articulate a null hypothesis that would be false if the initial hypothesis were true. At this point, then, we proceed, as we would do with any null hypothesis within the statistical inference model: We observe phenomena and subject them to the same hypothesis testing procedure that we have been describing.

We then see if we can make a reasonable assertion that the null hypothesis should be rejected and an alternative hypothesis accepted – within the realms of the observed probabilities.

These are the mechanisms at work behind tests of relationship in applied statistics, including correlation coefficients and other such tests. Also, the mathematical foundations behind these techniques are joint distributions.

Thus, the statistical inference model provides for the ability to make acceptable statements concerning the phenomenal dependency of two assertions – at least within the limits of certainty established by the rules of probability.

And this is how the statistical inference model provides a third model of critical thinking to both the causality model and the logic model.

Information Theory

Another type of statistical inference is found in information theory. Information theory emphasizes the measure of the degree of uncertainty of a probability distribution – a measure called entropy. Entropy is defined completely in terms of the probabilities of the distribution. That is, probabilities are the only input parameters to the formula for entropy.

Entropy measures the degree of uncertainty inherent in a probability distribution by directly measuring how "evenly spread" the probabilities are across all of the sample points of the distribution. The more uniform is the spread, the less certain one can be about exactly which sample point will, or will not, be realized. So there is this relationship between uncertainties and the even spread of probabilities to which that the idea of entropy speaks.

So, entropy plays a similar role in information theory as the p-value, or probability value of an observed event, plays in mathematical statistics. Both are arbiters of rarity, uncertainty, believability or acceptability. Statistics asks how the p-value compares to the significance level in order to determine acceptability of an assertion ("hypothesis"). On the other hand, information theory asks, "What is the entropy" to determine the acceptability of assertions.

And, in both cases, complex functions of initial assertions can be transformed into richer assertions, which can then also be tested by the probabilistic methods of both mathematical statistics and of information theory. Mathematical statistics generates these enhanced assertions through "functions of a random variable". Whereas, information theory uses a number of "entropic functionals" that it defines.

Both forms of statistical inference also provide elaborate mechanisms (based on joint probability distributions) to make assertions regarding the relationships between two or more assertions, and then to assess the likelihood of the acceptance of these compound assertions. Regression and correlation analysis represents tests of multiple chance variables in mathematical statistics; whereas stochastic dependency, mutual information and other mechanisms do so in information theory, all of which are the subject of Part II of this information theory primer.

Pros and Cons of the Three Models of Critical Thinking

At this point, we have introduced and discussed in some detail, three broadly adopted models of phenomenal dependence, all of which are used as guides to critical thinking: the causality model, the logic model and the statistical inference model.

In this section we shall summarize, compare and contrast the three models and suggest some advantages and disadvantages of each when used within various situations. We shall find that for each there are domains of application for which they

are advantageous – but that one of the three has emerged historically as the final arbiter within the domain of the scientific method.

The Causality Model

The causality model is a deeply intuitive and broadly appealing model of phenomenal dependence – and of critical thinking. Historically, it provided an apparently reasonable alternative to a reliance on magic and superstition.

On the other hand, causality is deterministic and does not accommodate uncertainty or randomness very well. That is, if the same trial is repeated with the same "causes", the causality model requires that the same effects consistently obtain.

But determinism does not faithfully represent contemporary scientific investigations of complex phenomena in which the whole of almost any system of interest is too large and complex to be able to be observed at once. Rather, samples must be taken. And multiple samples of the same population generally exhibit inconsistent results – a situation describes as random variation. This inability to represent and account for random variation is a serious drawback for any deterministic model, and of the causality model specifically.

A second disadvantage of causality as a model for phenomenal dependence is that it offers no internal mechanism for selecting or proving that any phenomenon causes any other phenomenon. This is a serious lacking for any deterministic model of phenomenal dependence. Consequently, causality may better serve as a philosophy rather than as a program to guide critical thinking.

Third, any deterministic model of phenomenal dependence should be able to be articulated as, or translated into, a corresponding logical argument. This ability would require that the notions of cause and effect being articulated as some kind of necessary conditions or sufficient conditions, or both - since these are the mechanisms within the logic model that accounts for phenomenal dependence. However, attempts over the ages to provide such an articulation has not been forthcoming.

Finally, the causality model has been largely abandoned by arguably the most successful scientific theory of all times – quantum mechanics. In quantum mechanics, deterministic "causes" have been replaced by non-deterministic "probabilities".

Thus, it is difficult to argue that the causality model has not been, in the end, largely abandoned by science as a final arbiter of thought when science's most successful discipline, quantum mechanics, has chosen an alternative.

Certainly, causality still finds a place within science. Because of its powerfully intuitive aspects, causality often serves as the inspiration for the origination of scientific ideas, which, through the use of logical inference, eventually find articulation as the hypotheses upon which the statistical inference model if privileged to act.

The Logic Model

As a model of phenomenal dependence, the logic model works indirectly using assertions (indicative statements) about phenomena, rather than directly with phenomena themselves.

This is actually very useful, because humans deal with phenomena not only through direct perception but also through thinking and talking about phenomena. In fact, the human treatment of phenomena through thinking is often a very complex business that is layered on top of our mere perception of phenomena – and therefore deserves a

deliberate treatment in its own right. The logic model addresses this aspect of phenomenal consideration.

Within the causality model, the dependency relationship between phenomena that treats phenomenal dependence is that of cause and effect. Two phenomena enjoy this dependence relationship whenever one of them is a cause of the other, and the other is the effect of the former.

In the logic model, on the other hand, assertions about phenomena behave as analogs to phenomena in the causality model. And, necessary conditions and sufficient conditions provide the relationships among assertions that account for the analog of phenomenal dependency within the logic model.

We could conveniently claim that the logic model is completely analogous to the causality model if we could show a relationship between the two models overall (a mapping) that consistently correlates assertions to phenomena, necessary conditions to either causes or effect, and sufficient conditions to either effects or causes.

Unfortunately, attempts to show that there exists such a complete analogy between logic and causality have never been successful. It seems that neither cause nor effect can be articulated as sufficient conditions or as necessary conditions. These kinds of perfect analogies between abstract systems are very important to both mathematics (where they are named equivalences, property preserving transformations, homomorphisms and other names) and to physics (where they are generally labeled symmetries).

The logic model has certain advantages over the causality model in some situations. In particular, the logic model provides an internal mechanism for accepting or rejecting assertions. It works by applying the rules of logical inference to a set of pre-given assertions called axioms. (Logical inference is more correctly called logical implication – which is actually a more accurate term. The two are used interchangeably.) This process produces a set of new assertions that are called theorems. The axioms are assumed to be true, and any theorems that can be derived through logical inference from the axioms are specified to be a true theorem by the logic model.

Logical implication provides a test of acceptance or rejection for assertions within a particular logical system. If the assertion can be shown to be a theorem that can be deduced through implication from the system's axioms using the rules of logical inference, then it is a (true) theorem of the system, and is acceptable. Otherwise, if its negation (another assertion) can be proved to be a theorem, then the assertion is rejected.

This ability of the logic model to provide an internal mechanism for the acceptance or rejection of its assertions (surrogates for phenomena) sets it apart from the causality model, which has no such mechanism – and is a strong advantage for the logic model.

Another limitation of the logic model is that it provides no internal mechanism for assessing the truth or falsehood of the axioms of any formal system. Rather, all axioms are arbitrarily chosen are assumed to be true within the logic model. As we have said, the truth of the theorems of a formal system (its ultimate product) is determined by applying the rules of logical inference to the axioms. The logic model by definition trusts the viability of the rules of logical inference. However, the logic model offers no internal test of the truth of the axioms. They must be assumed to be true.

Moreover, any test for the truthfulness of the axioms of a formal system must be found outside of the logic model! So this can be a serious shortcoming of the logic model.

A systemic property that the logic model shares with the causality model is that they are both deterministic. Determinism in a model of critical thinking has both its advantages and its disadvantages.

Determinism can be advantageous because it produces consistent conclusions. This renders the model reliable and trustworthy.

However, determinism can be an inflexibility that causes its thought model to "break" when trying to represent certain natural phenomena or human endeavors.

The pursuit of scientific knowledge on the part of humans is one such endeavor. The essential problem is that the natural systems about which scientist are pursuing knowledge happen to be too complex to be observed completely in a single view. Consequently, scientists are reduced to having to observe subsets (samples) of these systems in a series of single views. The problem is that such a set of samples generally produces differing, inconsistent data concerning the same underlying natural system (the population).

This inconsistency is called random variation. Random variation occurs when the same process (at least as far as the observer can tell they are the same) produces different results when repeated.

The logic model has difficulty modeling phenomena – such as the pursuit of scientific knowledge – that exhibit random variation, because logic is deterministic. This does not mean that logic cannot play a role within the methods used by scientists. But it does mean that the logic model is compromised whenever it attempts to provide the mechanism for acceptance and rejection of proposed assertions (hypotheses) that have been produced during the pursuit of scientific knowledge.

For example, logic can play a role in the steps of the method that involve developing candidate assertions, which will eventually, within the scientific method, be subjected to tests of acceptability. However, logic is lacking in an ability to make such a test of acceptability on its own – owing to the presence of random variation in the sample-taking (observational) steps of the scientific method.

The Statistical Inference Model

Statistical inference involves observing "suspicious events". These are apparently "rare" events that may suggest that something more than "chance" is at work within a phenomenon of interest. Statistical inference provides a way to subject those suspicious events to statistical scrutiny. This scrutiny involves calculating the probability that such an event would occur, using the "chance profile" (probability distribution) of the phenomenon as a chance phenomenon.

If that calculation produces a probability that is so low that it seems unlikely to have produced the "suspicious event" by chance, then it is reasonable to infer that the "suspicious" event did not come from the probability distribution of the chance event. And that it more like came from some other probability distribution instead – a probability distribution that describes that "something else is at work other than chance" (This is often expressed loosely as "there are no coincidences".)

If this occurs, then it is reasonable to reject the idea that the "suspicious phenomenon" is a result of chance, and that, indeed, it is not a "result of chance" and that "something else is at work".

By the way, this "idea that the 'suspicious phenomenon is a result of chance" is called the null hypothesis. If you, in fact, do suspect that "something else is at work", then you

are looking to reject the null hypothesis, since the null hypothesis is a statement at it is chance that actually is at work, rather than “something else”.

Thus, if the calculated probability of the “suspicious phenomenon” is less than the chosen significance level, then you can reject the null hypothesis, and infer (not conclude) that chance alone is not a good explanation of the phenomenon, and that something else is more likely at work.

However, if the calculated probability of the “suspicious phenomenon” is greater than the chosen significance level, then you cannot reject the null hypothesis – but neither can you accept it! This is because such a probability could also be exhibited by any other probability distribution that the “suspicious phenomenon” happens to actually exhibit.

In other words, if the calculated probability of the “suspicious phenomenon” is less than the chosen significance level, then you can reject the null hypothesis. This is your inference. However, if the calculated probability of the “suspicious phenomenon” is greater than the chosen significance level, then you can’t infer anything!

And, statistical significance never proves anything. It only allows you to reasonably infer – based on probabilities – that the observed phenomenon is unlikely to have occurred by “chance”, where “chance” is defined according to some probability distribution.

Statistical inference, like causality and like the logic model, is also a model of phenomenal dependence. As a model of phenomenal dependence, the statistical inference model works indirectly with assertions (indicative statements) about phenomena, rather than directly with phenomena themselves. However, statistical inference calls these assertions hypothesis.

Its approach to working with these assertions, or hypotheses, is that it makes hypotheses that may be able to show are unlikely to be true. Whenever it can do this, then it is in a position to reject the hypothesis because it has shown that it is unlikely to be true. The model proceeds in this way, making and rejecting additional hypotheses, until – by process of elimination; the realm of possibilities is narrowed.

An obvious disadvantage of statistical inference as a model of critical thinking is its reluctance to advance any certain conclusions.

But the “other side of this coin” is the conservative nature of this way of thinking. Ultimately, it can eliminate possible explanations as being highly unlikely. However, it leaves open the possibility of anything being ultimately true.

In the statistical inference model, assertions about phenomena behave as analogs to phenomena in the causality model – just as they do in the logical model. Here, acceptance and rejection of hypotheses provide the relationships among assertions that account for the analog of phenomenal dependency within the statistical dependence model.

The overriding advantage of the statistical inference model, however, is complete accommodation of chance variation. This feature makes it ideally suited to the empirical nature of scientific investigation – the application which motivated it and for which it was initially developed.

Despite the disadvantages of the statistical inference model in the area of its inability to commit to final and certain conclusions – an ability enjoyed by both the causation and the logic models, the statistical inference model exhibits a capability that is absolutely required by the scientific method – it ability to work with chance variation.

While the inability of the other two models of critical thought to reckon with chance variation pretty much eliminates them as essential to the acceptance or rejection of experimental conclusions within the scientific method, the statistical inference model on the other hand is tailor made for the job.

The other two models certainly do have their important places within the scientific method – although they are optional ones. However the statistical inference model, as we shall see, is the only one of the three models of critical thought that has any hope of being in the position to be the final arbiter of the acceptance or rejection of any candidate conclusions that the scientific method hopes to draw.

The Scientific Method and the Three Models of Critical Thinking

It is broadly understood that science is an empirical practice. That is, science involves the observation of natural phenomena, and subsequent interpretations of those observations.

As a culture, it is the deliberate intension of science to articulate its interpretations and findings as assertions that achieve a high degree of irrefutability and persistence; and therefore, hopefully, near-universal acceptance and even adoption. This results in science practice being a very careful and conservative culture.

Doing Science and Random Variation

Working against this conservatism, however, is the fact that its chosen domain of interest – nature – is too large and too complex to be wholly observable. In fact, even the pieces of nature about which science wants to make assertions are also too large and complex to be wholly observable. Only pieces of those pieces can be observed at once.

This fact is problematic for science, because the observation of these different “pieces” often produces inconsistent results. These inconsistencies present a problem for science, because science wants to make assertions about these wholes of nature – which are too big for it to observe at once. This introduces the notion of uncertainty.

Of course, uncertainty is at odds with the conservative nature of science – which as we have already discussed is driving toward near-universal acceptance of its pronouncements.

Fortunately, there exists a set of disciplines whose purview is this kind of uncertainty. These include mathematical statistics, its foundation probability theory, as well as the study of uncertainty itself, information theory. Statistics is the art of working with “pieces” of the “whole” – formally known as “samples” of a “population” within that discipline. Consequently, statistics is made to order for scientific investigation. In this appendix, we are collectively referring to the methodology behind all of these disciplines as statistical inference.

These disciplines may not have been what science has traditionally had in mind as its chosen way of thinking – principally because admitting to the presence of uncertainty only certifies that it exists – which can often work against “driving near-universal adoption” of ideas. Moreover, adopting a way of thinking into science’s investigative processes whose sole purpose is to accommodate uncertainty is nothing less than an admission that uncertainty is present.

Perhaps more friendly to science’s appetite for irrefutability, certainty and persistence are the two other “ways of thinking” that are the subject of this primer: the causality model and the logic model. This “friendliness” to science by these two disciplines of

thinking is so due to the fact that both are deterministic by nature. And this determinism is consistent with science's desire for certainty.

The causality model responds to the deep intuition of most scientists. For this reason alone, it should find a place within any methodology that scientist use to direct their practices. However, the causality model lacks the specificity to be able to articulate refutable and testable assertions.

The logic model does, however, provide the necessary mechanism to either refute or accept propositions that scientists may find themselves asserting. This is the mechanism of logical inference, commonly referred to as logical proof. Unfortunately, the logic model is also deterministic, and does not provide a mechanism for resolving the multiple and possibly conflicting results of multiple observations ("samples") of the same unobservably-large systems of interest ("populations").

So, what science needs is the definition of a method, or process, that instructs practicing empirical scientists on how to proceed with their investigation. This is the scientific method.

Such a method will accommodate the presence of uncertainty through the use of statistical inference. But it should also make way for the use of the causality model and the logic model where possible, since both of these accommodate the bias of science – that of having as much certainty as possible in order to drive near-universal agreement and adoption.

Let us now look at what has evolved as the broadly adopted methodology for conducting scientific investigation – also referred to as "doing science" – the scientific method.

The Procedure of the Scientific Method

The scientific method is the specification of the general procedure for practicing empirical scientific inquiry. It is the accepted method that all scientists are expected to use to conduct investigations. The scientific method draws on all three of our critical thought models at some point in its implementation.

There is general agreement across the scientific community as to how the procedure is specified. However, a search for a delineation of the steps of the procedure reveals a number of different interpretations. Fortunately, all of the ones reviewed by this researcher were in general agreement.

What follows is an abstraction from a number of these sources of the steps of the scientific method. Accordingly, we shall describe the scientific method consisting of these steps:

Steps of the Scientific Method

1. Observe
2. Question
3. Propose Explanations
4. Articulate Testable Assertion
5. Articulate Refutable Hypotheses
6. Test Hypothesis
7. Assess Results
8. Infer Conclusions

My apologies to the reader if I have not faithfully abstracted the steps of the scientific method from the sources I have found.

We shall now briefly describe each of these steps, try to surmise the general flow of their conduct and emphasize the role within the scientific method of the three models of critical thought that we have discussed at length in this appendix.

Every scientific investigation is targeted at a particular system, which we have been referring to here as the "system of interest".

Observe

Make an informal study of the system of interest. Observe in some detail the dynamic and static aspects of the system. Look for the apparent incongruities and other unexplained phenomena.

Question

Questions will arise naturally out of these informal observations. Pay particular attention to any causality related questions. It is in this step that the causality model is introduced.

Look for certain effects whose cause(s) are not readily discernable, but that are of particular interest to the investigators.

Record these questions and consider them for further investigation. Select some interrelated subset from among these as the basis for the pursuant investigation.

Propose Explanations

Propose some potential causes for the observed effects. Articulate these cause and effect phenomenal relationships as proposed explanations.

Articulate Testable Assertion

Rearticulate the proposed explanations as predictable, testable, refutable assertions. Appeal to the logic model can be especially helpful here.

Articulate Refutable Hypotheses

Articulate testable, refutable hypotheses from the testable assertions. These should include null hypotheses and optionally, alternative hypotheses.

Test Hypothesis

Apply the statistical inference model against the refutable hypotheses. Samples are taken, data is collected, and statistics are calculated from the samples.

Assess Results

The scrutinize the statistics that resulted from the hypothesis testing .

Infer Conclusions

Reject as many of your hypotheses as is called for by the results. Also, optionally accept alternative hypotheses after deliberation of the assessment.

Conclusion

Within the scientific method, the final arbiter as to whether assertions (hypotheses) are accepted are rejected is the statistical inference model. These deliberations occur near the end of the method when any proposed hypotheses are tested.

While the other two models may be, and usually are, involved earlier in the scientific method process for the purpose of formulating refutable hypotheses, when it comes to articulating the hypotheses and applying the final tests of acceptability or refutability, the statistical inference is the sole proprietor.

Ultimately, this is true because of the ability of statistical inference to accommodate uncertainty in the form of random variation.

Therefore, statistical inference can refute a hypothesis, but it does not generally require the acceptance of any hypothesis. This is the meaning of the statement “Science can disprove but it cannot prove anything.”

Appendix 4: Example – Cellphone-by-Continent

This appendix will present a complete example of a joint distribution that exemplifies every construct that has been developed in this primer.

For our example, we shall select a new probability space. We are going to use a new example for a number of reasons. First, we want to use an example with a smaller number of joint sample points. Our five dice experiments have 36 sample points, and this is too length for a complete example. Second, we would like to use a more realistic example that could be associated with a more serious application – one from science or business.

The example that we have selected pertains to the adoption of different mobile phone technologies across the world. We are interested in how various technology brands have been adopted across different global regions, or continents. In fact, we would like to investigate whether there is any stochastic dependence between global region and brand adoption. And, if we find that there is a dependency, then how much.

Admittedly, our example will be a "toy" one – mainly because the number of technology brands identified (4) and the number of global regions used (also 4) is not as many as an actual business analysis would use.

However, we have kept these numbers small so as to accommodate the fact that these examples will be quite lengthy. Nevertheless, these 4*4=16 joint sample points will be a much smaller number than the 6*6=36 joint sample points of our other five "dice" Experiments.

Another "toy" aspect of this example is that we have chosen to allow one of the component chance variables (phone brand) to take the uniform distribution. While this choice does not reflect the real world (some phone brands are in fact more popular than others), making it does afford us with the opportunity to see what a mixed pair of component distributions – one non-uniform and the other uniform – looks like.

Our discussion of these Part II mathematical constructs for our "Cellphone Brands by Continent" example will be organized into three sections:

1. Component, joint and conditional distributions of the joint probability space
2. Entropic measures
3. Entropic measure relationships

Component, Joint and Conditional Distributions

In this example, we are interested in investigated how various cellphone brands are chosen across the continents of the world. Specifically, we are interested in whether or not there is any relationship (portent, meaningfulness, dependence) between the brand of cellphone in use and the regions of the world (continents) where they are purchased.

There are three probability distributions for which we have gathered purchasing data. One of these has determined the relative frequency (probability) of phones that are in use by brand. Four brand categories are distinguished by the data. (More than 4 categories would be more interesting, but we are trying to keep the data organization manageable for this section.)

Another probability distribution has determined the relative frequency (probability) of continent where the phones are used. Four continental areas are distinguished by the data.

Finally, data for the specifically observed joint distribution, using the two previous distributions as component distributions, has been gathered to reveal the relative frequency of phones in use in the sixteen categories that are determined by the joint combinations of the 4 categories of phone brand and the 4 categories of continent of the two component distributions. We shall use the symbol $(X, Y)_K$ for this particular joint distribution on X and Y – the subscript K being used to distinguish this particular observed joint distribution from any other joint distributions on the same composite distributions X and Y.

These three empirically observed probability distributions, along with their graphs, are displayed below. Be aware that all three are empirically observed. None has been derived from any of the others.

Also be aware that the joint distribution below is just one of an infinite number of other possible joint distributions on the same two component distributions. Our focus in this example is to ascertain "how much stochastic dependence" is inherent in this particular joint distribution on these two component distributions. (Any of the other possible joint distributions on these two particular component distributions would, in general, exhibit a different amount of stochastic dependence.)

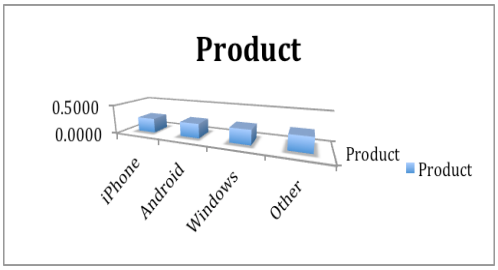
Here then are the three (empirically observed) probability distributions of interest for our Cellphone probability space.

X and Y: The Two Component Distributions – Observed

We now present the actual observed component distributions X and Y.

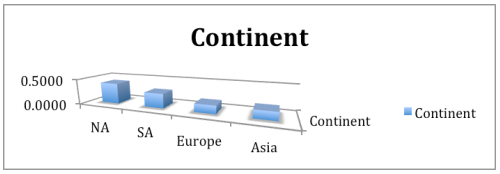
X: Phone Brand Component Distribution – Observed

X	iPhone	Android	Windows	Other	**Total**
Product	0.2500	0.2500	0.2500	0.2500	**1.0000**



Y: Continent Component Distribution – Observed

Y	NA	SA	Europe	Asia	**Total**
Continent	0.4063	0.2813	0.1563	0.1563	**1.0000**

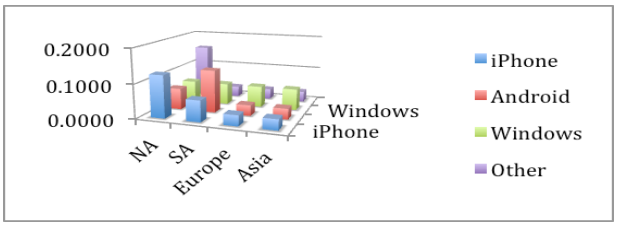


$(X, Y)_K$: (Phone Brand, Continent) Joint Distribution – Observed

We now present the actual observed joint distribution $(X, Y)_K$ of the these two chance variables.

$(X, Y)_K$: (Phone Brand, Continent) Joint Distribution – Observed

	NA	SA	Europe	Asia	**Total Phone**
iPhone	0.1250	0.0625	0.0313	0.0313	**0.2500**
Android	0.0625	0.1250	0.0313	0.0313	**0.2500**
Windows	0.0625	0.0625	0.0625	0.0625	**0.2500**
Other	0.1563	0.0313	0.0313	0.0313	**0.2500**
Total Continent	**0.4063**	**0.2813**	**0.1563**	**0.1563**	**1.0000**



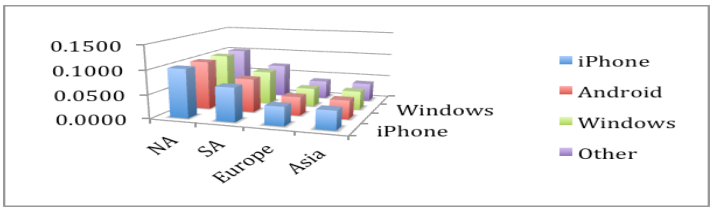
Since we are ultimately interested in "how different" this joint distribution $(X, Y)_K$ is from the statistically independent joint distribution $(X, Y)_0$ on these same two component distributions X and Y, we shall portray $(X, Y)_0$ also.

Later below when we calculate the mutual information of $(X, Y)_K$, we shall see that it is a function of the "difference" between $(X, Y)_K$ and $(X, Y)_0$. So we present both of them now so that the reader can contemplate how they might be different. Essentially, the mutual information of $(X, Y)_K$, is a measure of "how far away" $(X, Y)_K$ is from $(X, Y)_0$.

So, here we depict $(X, Y)_0$.

$(X, Y)_0$: Stochastically Independent Joint Distribution – Calculated

	NA	SA	Europe	Asia	**Total Phone**
iPhone	0.1016	0.0703	0.0391	0.0391	**0.2500**
Android	0.1016	0.0703	0.0391	0.0391	**0.2500**
Windows	0.1016	0.0703	0.0391	0.0391	**0.2500**
Other	0.1016	0.0703	0.0391	0.0391	**0.2500**
Total Continent	**0.4063**	**0.2813**	**0.1563**	**0.1563**	**1.0000**



Besides these empirically observed distributions, we can derive the two conditional distributions: “Y given X” (Y|X) and “X given Y” (X|Y).

(Y|X): Conditional Distribution – Calculated

We derive this conditional distribution by dividing every cell in the joint distribution (X, Y) above by its row sum. We shall use the table below to calculate all 16 cells of the matrix (Y|X).

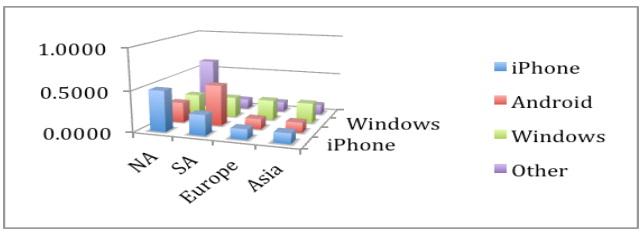
(x,y)	p(x,y)	p(x)	p(y|x) = p(x,y)/p(x)
iPhone, NA	0.12500	0.25000	0.5000
iPhone, SA	0.06250	0.25000	0.2500
iPhone, Europe	0.03125	0.25000	0.1250
iPhone, Asia	0.03125	0.25000	0.1250
Android, NA	0.06250	0.25000	0.2500
Android, SA	0.12500	0.25000	0.5000
Android, Europe	0.03125	0.25000	0.1250
Android, Asia	0.03125	0.25000	0.1250
Windows, NA	0.06250	0.25000	0.2500
Windows, SA	0.06250	0.25000	0.2500
Windows, Europe	0.06250	0.25000	0.2500
Windows, Asia	0.06250	0.25000	0.2500
Other, NA	0.15625	0.25000	0.6250
Other, SA	0.03125	0.25000	0.1250
Other, Europe	0.03125	0.25000	0.1250
Other, Asia	0.03125	0.25000	0.1250

The final column contains the conditional probability p(y|x) of each cell. This value is calculated by dividing p(x,y) by p(x), as indicated in the table. These sixteen results can then be represented in matrix form:

$(Y|X)_K$

	NA	SA	Europe	Asia	Total Phone
iPhone	0.5000	0.2500	0.1250	0.1250	**1.0000**
Android	0.2500	0.5000	0.1250	0.1250	**1.0000**
Windows	0.2500	0.2500	0.2500	0.2500	**1.0000**
Other	0.6250	0.1250	0.1250	0.1250	**1.0000**

With this graph:



(X|Y): Conditional Distribution – Calculated

We derive this conditional distribution by dividing every cell in the joint distribution (X, Y) above by its column sum. We shall use the table below to calculate all 16 cells of the matrix (X|Y).

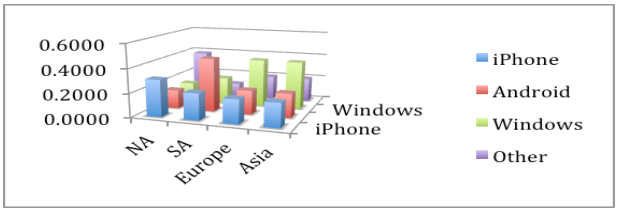
(x,y)	p(x,y)	p(y)	p(x,y)/p(y)
iPhone, NA	0.12500	0.40625	0.3077
iPhone, SA	0.06250	0.28125	0.2222
iPhone, Europe	0.03125	0.15625	0.2000
iPhone, Asia	0.03125	0.15625	0.2000
Android, NA	0.06250	0.40625	0.1538
Android, SA	0.12500	0.28125	0.4444
Android, Europe	0.03125	0.15625	0.2000
Android, Asia	0.03125	0.15625	0.2000
Windows, NA	0.06250	0.40625	0.1538
Windows, SA	0.06250	0.28125	0.2222
Windows, Europe	0.06250	0.15625	0.4000
Windows, Asia	0.06250	0.15625	0.4000
Other, NA	0.15625	0.40625	0.3846
Other, SA	0.03125	0.28125	0.1111
Other, Europe	0.03125	0.15625	0.2000
Other, Asia	0.03125	0.15625	0.2000

The final column contains the conditional probability p(xly) of each cell. This value is calculated by dividing p(x,y) by p(y), as indicated in the table. These sixteen results can then be represented in matrix form:

$(Y|X)_K$

	NA	SA	Europe	Asia
iPhone	0.30769	0.22222	0.20000	0.20000
Android	0.15385	0.44444	0.20000	0.20000
Windows	0.15385	0.22222	0.40000	0.40000
Other	0.38462	0.11111	0.20000	0.20000
Total Continent	**1.00000**	**1.00000**	**1.00000**	**1.00000**

With this graph:



Entropic Measures

Entropy, Joint Entropy, Conditional Entropy and Mutual Information

Information theory presents a number of measuring functions for various phenomena relating to probability distributions. These measuring functions are called entropic measures.

Some of these functions measure aspects of single joint probability distributions. Examples include entropy and relative entropy. Others, such as joint entropy, conditional entropy, and mutual information are measures of various phenomena regarding joint distributions.

In all of these cases, however, there is something in common. What is in common is that all of these entropic measures are a generalization of the concept of entropy.

Recall that entropy is defined (among other equivalent definitions) as:

$$H(p(x)) = -\sum_{i \in S} p(x)*\log(p(x)).$$

We can generalize this formulation by substituting any function of p(x) in place of the second occurrence of "p(x)" in the above definition. And, without loss of generality, we could symbolize any such function of p(x) as "F(p(x))".

We could then write such a "generalized entropy" as:

$$H(F(p(x))) = -\sum_{i \in S} p(x)*\log(F(p(x))).$$

Moreover, we could then create specific instances of these "entropic functionals" by substituting various functions of p(x) in place of "F(p(x))" in the above expression.

For example, entropy itself is a trivial special case of H(F(p(x))) – where F(p(x)) is, in fact, "p(x)". All we have to do is to substitute "p(x)" in place of "F(p(x)))" above, and we have:

$$H(p(x)) = -\sum_{i \in S} p(x)*\log(p(x)),$$

Which of course is the formulation for entropy. Thus entropy is a special case of an entropic measure. Thus, entropic measures are a generalization of entropy.

By substituting other expressions for F(p(x)), we obtain some of the entropic measures that we have already been working with. For example, we could let

F(p(x)) = q(x)/p(x), where q(x) is another probability distribution on x.

With this assignment, then we have

$H(F(p(x))) =$

$-\sum_{i \in S} p(x)*\log(F(p(x))) =$

$-\sum_{i \in S} p(x)*\log(q(x)/p(x)) =$

$\sum_{i \in S} p(x)*\log(p(x)/q(x)) = D(p||q)$

Thus, this special case of F(p(x)) = q(x)/p(x) results in the definition of relative entropy. Consequently, relative entropy is an entropic measure.

We can also get some interesting extensions to the notion of entropic measures by including joint probability distributions. In this case, we have to extend our notion of probability from "p(x)" to "p(x,y)".

And instead of functions of p(x), “F(p(x))”, we are now dealing with “F(p(x,y))”. And instead of H(F(p(x)))), we are now dealing with H(F(p(x,y))).

So, we could, for example, define F(p(x,y)) = p(x,y)/p(x). This would certainly be a function of p(x,y). So, our H(p(x,y)) is now defined by

H(F(p(x,y))) =

$-\sum_{i \in S} p(x,y)*\log(F(p(x,y))) =$

$-\sum_{i \in S} p(x,y)*\log(p(x,y)/p(x)) =$

$-\sum_{i \in S} p(x,y)*\log(p(y|x) = H(Y|X)$

But this is the definition of the conditional entropy H(Y|X).

Perhaps most interesting is the definition of mutual information. It too is an entropic measure. In this case,

F(p(x,y)) = p(x)*p(y)/p(x,y)

Thus,

$H(F(p(x,y))) = -\sum_{i \in S} p(x,y)*\log [p(x)*p(y)/p(x,y)] =$

$\sum_{i \in S} p(x,y)*\log [p(x,y)/p(x)*p(y)] = I(X;Y)$

H(X): Entropy of the Phone Brand Component Distribution

Here is the calculation of the entropy H(X) using the formulation

$H(p(x)) = -\sum_{i \in S} p(x)*\log(p(x)).$

	p(x)	1/p(x)	log(1/p(x))	p(x)* log(1/p(x))
iPhone	0.25000	4.0000	2.0000	**0.5000**
Android	0.25000	4.0000	2.0000	**0.5000**
Windows	0.25000	4.0000	2.0000	**0.5000**
Other	0.25000	4.0000	2.0000	**0.5000**
Entropy				**2.0000**

H(Y): Entropy of the Continent Component Distribution

Here is the calculation of the entropy H(Y) using the formulation

$H(p(x)) = -\sum_{i \in S} p(x)*\log(p(x))$.

	p(y)	1/p(y)	log(1/p(y))	p(y)*log(1/p(y))
NA	0.40625	2.4615	1.2996	**0.5279**
SA	0.28125	3.5556	1.8301	**0.5147**
Europe	0.15625	6.4000	2.6781	**0.4184**
Asia	0.15625	6.4000	2.6781	**0.4184**
Entropy				**1.8796**

H(X,Y): Joint Entropy of the Cellphone Space Distribution

Here is the calculation of the entropy H(X,Y) using the formulation

$$H(p(x,y)) = -\sum_{i \in S} p(x,y)*\log(p(x,y)) = \sum_{i \in S} p(x,y)*\log(1/p(x,y))$$

	p(x,y))	1/p(x,y)	log(1/p(x,y))	p(x,y)* log(1/p(x,y))
iPhone, NA	0.12500	8.0000	3.0000	**0.3750**
iPhone, SA	0.06250	16.0000	4.0000	**0.2500**
iPhone, Europe	0.03125	32.0000	5.0000	**0.1563**
iPhone, Asia	0.03125	32.0000	5.0000	**0.1563**
Android, NA	0.06250	16.0000	4.0000	**0.2500**
Android, SA	0.12500	8.0000	3.0000	**0.3750**
Android, Europe	0.03125	32.0000	5.0000	**0.1563**
Android, Asia	0.03125	32.0000	5.0000	**0.1563**
Windows, NA	0.06250	16.0000	4.0000	**0.2500**
Windows, SA	0.06250	16.0000	4.0000	**0.2500**
Windows, Europe	0.06250	16.0000	4.0000	**0.2500**
Windows, Asia	0.06250	16.0000	4.0000	**0.2500**
Other, NA	0.15625	6.4000	2.6781	**0.4184**
Other, SA	0.03125	32.0000	5.0000	**0.1563**
Other, Europe	0.03125	32.0000	5.0000	**0.1563**
Other, Asia	0.03125	32.0000	5.0000	**0.1563**
Joint Entropy				**3.7622**

H(Y|X): Conditional Entropy of the Continent, Given the Phone Brand

Here is the calculation of the entropy H(Y|X) using the formulation

$$H(p(y|x)) = -\sum_{i \in S} p(x,y)*\log(p(y|x)) = \sum_{i \in S} p(x,y)*\log(1/p(y|x))$$

	p(x,y))	p(y|x)	1/p(y|x)	log(1/p(y|x))	p(x,y)* log(1/p(y|x))
iPhone, NA	0.12500	0.5000	2.0000	1.0000	**0.1250**
iPhone, SA	0.06250	0.2500	4.0000	2.0000	**0.1250**
iPhone, Europe	0.03125	0.1250	8.0000	3.0000	**0.0938**
iPhone, Asia	0.03125	0.1250	8.0000	3.0000	**0.0938**
Android, NA	0.06250	0.2500	4.0000	2.0000	**0.1250**
Android, SA	0.12500	0.5000	2.0000	1.0000	**0.1250**
Android, Europe	0.03125	0.1250	8.0000	3.0000	**0.0938**
Android, Asia	0.03125	0.1250	8.0000	3.0000	**0.0938**
Windows, NA	0.06250	0.2500	4.0000	2.0000	**0.1250**
Windows, SA	0.06250	0.2500	4.0000	2.0000	**0.1250**
Windows, Europe	0.06250	0.2500	4.0000	2.0000	**0.1250**
Windows, Asia	0.06250	0.2500	4.0000	2.0000	**0.1250**
Other, NA	0.15625	0.6250	1.6000	0.6781	**0.1059**
Other, SA	0.03125	0.1250	8.0000	3.0000	**0.0938**
Other, Europe	0.03125	0.1250	8.0000	3.0000	**0.0938**
Other, Asia	0.03125	0.1250	8.0000	3.0000	**0.0938**
Entropy					**1.7622**

H(X|Y): Conditional Entropy of the Phone Brand, Given the Continent

Here is the calculation of the entropy H(XIY) using the formulation

$$H(p(x|y)) = -\sum_{i \in S} p(x,y)*\log(p(x|y)) = \sum_{i \in S} p(x,y)*\log(1/p(x|y))$$

	p(x,y))	p(x|y)	1/p(x|y)	log(1/p(x|y))	p(x,y)* log(1/p(x|y))
iPhone, NA	0.12500	0.3077	3.2500	1.7004	**0.2126**
iPhone, SA	0.06250	0.2222	4.5000	2.1699	**0.1356**
iPhone, Europe	0.03125	0.2000	5.0000	2.3219	**0.0726**
iPhone, Asia	0.03125	0.2000	5.0000	2.3219	**0.0726**
Android, NA	0.06250	0.1538	6.5000	2.7004	**0.1688**
Android, SA	0.12500	0.4444	2.2500	1.1699	**0.1462**
Android, Europe	0.03125	0.2000	5.0000	2.3219	**0.0726**
Android, Asia	0.03125	0.2000	5.0000	2.3219	**0.0726**
Windows, NA	0.06250	0.1538	6.5000	2.7004	**0.1688**
Windows, SA	0.06250	0.2222	4.5000	2.1699	**0.1356**
Windows, Europe	0.06250	0.4000	2.5000	1.3219	**0.0826**
Windows, Asia	0.06250	0.4000	2.5000	1.3219	**0.0826**
Other, NA	0.15625	0.3846	2.6000	1.3785	**0.2154**
Other, SA	0.03125	0.1111	9.0000	3.1699	**0.0991**
Other, Europe	0.03125	0.2000	5.0000	2.3219	**0.0726**
Other, Asia	0.03125	0.2000	5.0000	2.3219	**0.0726**
Entropy					**1.8826**

I(X;Y): Mutual Information of Cellphone-by-Continent – Intuitive Definition

As we have said many times, deriving the mutual information of a pair of random variables (and their joint distribution), symbolized I(X;Y) is the principle goal of the analysis of the degree of stochastic dependence between those two chance variables. We have finally arrived at the calculation of I(X;Y) for the Cellphone example.

Mutual information is so important that we are going to present two different calculations for it here. The first calculation will accommodate the more intuitive explanation that we have been giving throughout Part II.

We have emphasized that mutual information is a measure of the degree of stochastic dependence between two chance variables in a joint distribution. And as such, mutual information can be interpreted to also be a measure of the degree of portent or meaningfulness between those two chance variables.

However, the textbook formulation for calculating mutual information that is seen in textbooks on information theory does not necessarily reveal how it is that mutual information is, in fact, a measure of these phenomena just mentioned.

Therefore, this primer has "reverse-engineered" this textbook formulation to an equivalent formulation that is more revealing.

Therefore, we shall present both formulations, and show that they produce the same resulting value for the mutual information of this particular Cellphone example.

The present subsection shall present the more intuitive formulation; while the next subsection will present the textbook formulation – which has a few less calculation steps.

Lets now present the more intuitive formulation of the mutual information I(X;Y) as an entropic measure.

In this case,

$F(p(x,y)) = -[u_{0_0}(x,y) - u_K(x,y)]$

Where,

$u_{0_0}(x,y)$ is the uncertainty of (x,y) with respect to the joint distribution $(X,Y)_0$.
$u_K(x,y)$ is the uncertainty of (x,y) with respect to the joint distribution $(X,Y)_K$.

This means that

$u_0(x,y) = \log_2(1/p(x,y)_0)$, and
$u_K(x,y) = \log_2(1/p(x,y)_K)$

Therefore,

$F(p(x,y)) = -[u_{0_0}(x,y) - u_K(x,y)] = -[\log_2(1/p(x,y)_0) - \log_2(1/p(x,y)_K)]$

Thus, by definition,

Definition 1: Mutual Information of two chance variables X and Y:

$I(X;Y) = H(F(p(x,y))) = \sum_{i \in S} p(x,y) * \log [\log_2(1/p(x,y)_0) - \log_2(1/p(x,y)_K)]$

In the two tables below, we shall apply this formulation to our Cellphone investigation.

An intuitive discussion of the values in this table follows them.

Part 1: Derivation of I(X;Y) - Intuitive Definition: Mean of [$u_0(x,y) - u_K(x,y)$]

	$p(x,y)_k$	$p(x,y)_0$	$1/p(x,y)_k$	$1/p(x,y)_0$	$u_K(x,y) = \log_2(1/p(x,y)_K$	$u_0(x,y) = \log_2(1/p(x,y)_0)$
iPhone, NA	0.12500	0.10156	8.00000	9.84615	3.00000	3.29956
iPhone, S	0.06250	0.07031	16.00000	14.22222	4.00000	3.83007
iPhone, Europe	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807
iPhone, Asa	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807
Android, NA	0.06250	0.10156	16.00000	9.84615	4.00000	3.29956
Android, SA	0.12500	0.07031	8.00000	14.22222	3.00000	3.83007
Android, Europe	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807
Android, Asia	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807
Windows, NA	0.06250	0.10156	16.00000	9.84615	4.00000	3.29956
Windows, SA	0.06250	0.07031	16.00000	14.22222	4.00000	3.83007
Windows, Europe	0.06250	0.03906	16.00000	25.60000	4.00000	4.67807
Windows, Asia	0.06250	0.03906	16.00000	25.60000	4.00000	4.67807
Other, NA	0.15625	0.10156	6.40000	9.84615	2.67807	3.29956
Other, SA	0.03125	0.07031	32.00000	14.22222	5.00000	3.83007
Other, Europe	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807
Other, Asia	0.03125	0.03906	32.00000	25.60000	5.00000	4.67807

Part 2: Derivation of I(X;Y) - Intuitive Definition: Mean of [$u_0(x,y) - u_K(x,y)$]

	$p(x,y)_k$	$u_0(x,y) - u_K(x,y)$	$p(x,y)_K * [u_0(x,y) - u_K(x,y)]$
iPhone, NA	0.12500	0.29956	**0.03745**
iPhone, SA	0.06250	-0.16993	**-0.01062**
iPhone, Europe	0.03125	-0.32193	**-0.01006**
iPhone, Asia	0.03125	-0.32193	**-0.01006**
Android, NA	0.06250	-0.70044	**-0.04378**
Android, SA	0.12500	0.83007	**0.10376**
Android, Europe	0.03125	-0.32193	**-0.01006**
Android, Asia	0.03125	-0.32193	**-0.01006**
Windows, NA	0.06250	-0.70044	**-0.04378**
Windows, SA	0.06250	-0.16993	**-0.01062**
Windows, Europe	0.06250	0.67807	**0.04238**
Windows, Asia	0.06250	0.67807	**0.04238**
Other, NA	0.15625	0.62149	**0.09711**
Other, SA	0.03125	-1.16993	**-0.03656**
Other, Europe	0.03125	-0.32193	**-0.01006**
Other, Asia	0.03125	-0.32193	**-0.01006**
Entropy			**0.1174**

Intuitive Discussion - Cellphone-by-Continent Mutual Information Example

We shall now engage in a discussion of the calculations in the above derivation of the mutual information I(X;Y) of these two chance variables X=”Phone Brand” and Y=”Continent”, in joint distribution as a measure of their degree of stochastic dependence.

Relationship Between Probability and Uncertainty

We shall first make some preliminary observations. In information theory, the preferred measure of a chance variable is uncertainty. But, probability is the preferred measure of a chance variable in probability theory. Moreover, the two measures are inversely related. That is, as the probability measure of a sample point gets smaller, its uncertainty measure gets larger. As probability goes up, uncertainty goes down, and vice versa.

Of course, probability is a more fundamental measure than is uncertainty, because uncertainty is mathematically defined in terms of probability.

Not only are the two measures inversely related, but the definition of uncertainty is then "skewed" so that it will not "rise too fast". This skewing comes in handy when we are comparing the degrees of uncertainty of events whose probabilities are vastly different – for example, 1/2 and 1/4096. If we simply inverted these two we would get 2 and 4096. And if we then compared them, they would be "farther apart" than would be mathematically useful for comparative purpose. But if we "skewed" them – for example, on a logarithmic scale, we would get more comparable results. For example if we used a log base 2 scaling factor, then 2 would map to 1; and 4096 would map to 12.

(We could have used other "skewing" formulas besides logarithms, but logarithms do the job quite well.)

As another example, if we had probabilities of, say, ½, ¼, and 1/8, and inverted them, we would get 2, 4 and 8. If we them took the logarithm of these three inversions, we obtain 1, 2 and 3 as a result. That is, a probability of ½ is an uncertainty of 1, a probability of ¼ is an uncertainty of 2, and a probability of 1/8 is an uncertainty of 3.

This combination of both inverting the probability and then taking its logarithm has the dual effect of both "inverting" and of "slowing down" the conversion between probabilities and their associated uncertainty measures.

This is the first point – that in information theory, we most often use uncertainty as our measure of chance variation rather than probability.

The second preliminary point is that we are most often given probabilities and have to convert them to uncertainty before we can get started with using uncertainty. The "conversion" formula, the reader will recall is:

$u_p(x) = \log(1/p(x))$

where
x is a sample point of sample space X.
p(x) is the probability of sample point x according to probability distribution p
log is the logarithm to some arbitrarily selected log base. A log base of 2 is used here.

Note:
There may be other probability distributions defined on X than just "p". Each of them, then has is own definition of u(x). For example, a probability distribution "q" on X will have an uncertainty function $u_q(x) = \log(1/q(x))$.

Strategy to Measure the Degree of Stochastic Dependence

Here is the strategy that information theory uses to measure the "degree of stochastic dependence" between the two chance variables X and Y of a joint probability distribution (X,Y) – call it $(X,Y)_K$.

Information theory reckons that

1. There is exactly one joint distribution on X and Y that is stochastically independent. Call it $(X,Y)_0$. The "degree of stochastic dependence" of $(X,Y)_0$ should be 0.
2. Moreover, the "degree of stochastic dependence" of our joint distribution of interest, $(X,Y)_0$, could be reasonably measured by "the difference between" $(X,Y)_0$ and $(X,Y)_K$.

In other words, it would be a reasonable thing to do to measure the "degree of stochastic independence" of a given joint distribution $(X,Y)_0$ by "measuring how far away" $(X,Y)_K$ is from $(X,Y)_0$.

All we need, then, to accomplish this is to find some function that measures the "difference" between two probability distributions on the same sample space.

But, already have such a measuring function! It is relative entropy, "D(p||q)".

In Part I, we presented a number of interpretations of relative entropy, "D(p||q)". Perhaps the most intuitive was

> Relative entropy between probability distributions p and q, D(p||q), is the mean of $u_q(x)$ - $u_p(x)$ – using p(x) to calculate the mean.

In other words,

> The relative entropy of two probability distributions p and q (on the same sample space) is the expected, or mean, value of the differences between the two uncertainty values - $u_q(x)$ and $u_p(x)$ - of each sample point.

This means that to calculate the relative entropy of two probability distributions, for each sample point calculate $u_q(x)$ and $u_p(x)$ and then subtract them. Next, multiply each of these subtractions by p(x). Finally, add all of these products together. The result is the following formula for relative entropy:

$$D(p||q) = \sum_{i \in S} p(x)*[u_q(x) - u_q(x)]$$

The above expresses exactly the mean, or expected value, of $u_q(x) - u_q(x)$.

Notice that in order to calculate a mean, we have to multiply each of the "$u_q(x) - u_q(x)$" – one for each sample point – by "its probability". But, we are dealing here with two probability distributions! So which one do we use to calculate this mean?

The answer is "p" – not "q". The reason for choosing p for calculating the mean is that it is the a posteriori – or actual – distribution.

Of course, so far we have only seen the "p" and "q" or relative entropy be "normal" (not joint) probability distributions. But there is no reason why we cannot apply the definition of D(p||q) to joint distributions as well. And that is precisely what we shall do in order to define mutual information.

Specifically, for the "p" in "D(p||q)", we shall substitute our joint probability distribution of interest $(X,Y)_K$. And for the "q" in "D(p||q)", we shall substitute our stochastically independent joint probability $(X,Y)_0$.

The result of this substitution will be the beginnings of our definition of the mutual information of $(X,Y)_K$. $I(X;Y)_K$. This is:

$$I(X;Y)_K = D((X,Y)_K || I(X;Y)_0)$$

Thus, our definition of the mutual information between X and Y in the joint distribution $(X,Y)_K$ is the relative entropy $D((X,Y)_K || I(X;Y)_0)$.

All that remains then is to rearticulate this definition into a calculable form.

But, we have already decided above that relative entropy of two probability distributions p and q is the mean of the difference of the uncertainties at every sample point. That is, that

$$D(p||q) = \sum_{i \in S} p(x)*[u_q(x) - u_q(x)]$$

Therefore, I(X;Y), then, must be the mean of the differences between the uncertainties of each sample point based upon the distribution $(X,Y)_0$ and the uncertainties of each sample point based upon the distribution $(X,Y)_K$.

Lets give the name "$u_k(x,y)$" to the uncertainty of sample point x based upon the distribution $(X,Y)_K$. And lets give the name "$u_0(x,y)$" to the uncertainty of sample point x based upon the distribution $(X,Y)_0$.

Intuitive Definition of Mutual Information

We now have enough symbolism to take our definition of mutual information as the relative entropy of $(X,Y)_K || I(X;Y)_0$ make it calculable by the proper substitutions. To say that "mutual information as the relative entropy of $(X,Y)_K || I(X;Y)_0$" means, in symbols – as we already said above:

$$I(X;Y)_K = D((X,Y)_K || (X,Y)_0)$$

But, we have already established that

$$D(p||q) = \sum_{i \in S} p(x)*[u_q(x) - u_q(x)]$$

Thus, by substitution of $(X,Y)_K$ for p and of $(X,Y)_0$ for q in the previous expression, we have our working intuitive definition of mutual information:

Intuitive definition of mutual information:

$$I(X;Y) = D((X,Y)_K || (X,Y)_0) = \sum_{i \in S} p(x,y)*[u_0(x,y) – u_k(x,y)]$$

In words, this says that the mutual information of two chance variables X and Y in joint distribution $(X,Y)_K$ is the relative information of $(X,Y)_K$ and $(X,Y)_0$; which is also expressed as the mean, or expected value, of the difference of the uncertainties $u_q(x) - u_q(x)$ of sample points (x,y).

Cellphone Example of Intuitive Mutual Information

So, lets now apply this intuitive understanding of mutual entropy to our Cellphone by Global Region example. We developed the calculation for the mutual information of this example in parts 1 and 2 of the table above. The result turned out to be I(X;Y) = 0.1174.

What we want to do now is to break down this calculation by looking at our calculation one column at a time. Moreover, we would like to enhance our discussion by graphical representations of these columns.

We first observe that each column is more meaningful organized as the two-dimensional matrix of the joint distribution, where one of the dimensions is the chance

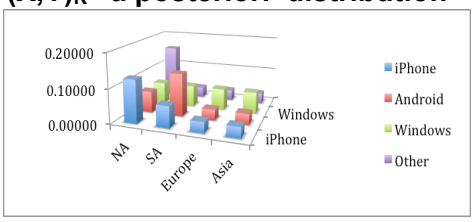
variable X="Phone Brand" and the other is the chance variable Y="Continent". So, we are going to take each of these columns, represent it as 2-dimensions, and then graph it in the usual 3-dimensional graph format that we have been using.

Also, notice that the first six of the columns of the above table are naturally paired – having to do with either the a posteriori joint probability distribution $(X,Y)_K$ whose mutual information we are trying to measure, or the a priori joint distribution $(X,Y)_0$ that we are comparing it against. Specifically, the first two columns depict the two joint distributions $(X,Y)_K$ and $(X,Y)_0$. The second two columns depict the inverse of each of these two joint distributions. And the third pair of columns depict the log of the inverse of these two joint distributions – other wise known as the uncertainties of these two joint distributions $u_0(x,y)$ and $u_k(x,y)$.

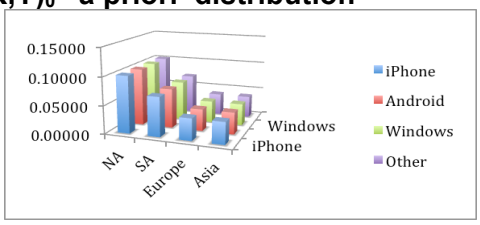
The next to last column of interest contains the calculation of the difference between these two uncertainties, or $u_0(x,y) – u_k(x,y)$ – whose mean, or expected value, is in fact the mutual information value that we seek. And the final column of these calculations obtains the mean by multiplying each result in the previous column by its corresponding probability $p(x,y)_K$ – yielding $p(x,y)_K*[u_0(x,y) – u_k(x,y)]$ for each sample point (x,y). Finally, to obtain the actual mutual information value, we sum all of these $p(x,y)_K*[u_0(x,y) – u_k(x,y)]$ values in this last column. This sum yields the value of 0.1174.

So lets take a graphical look at these calculations. We begin with the first two columns of the above table. These two columns contain the two joint distributions that we are comparing by taking their relative entropy in this calculation: $(X,Y)_K$ and $(X,Y)_0$.

$(X,Y)_K$ - a posteriori distribution



$(X,Y)_0$ - a priori distribution



Our calculation of mutual entropy is very much about detrmining "the difference" between the a priori distribution $(X,Y)_0$, the stochastically independent one, and the a posteriori distribution $(X,Y)_k$. The idea here is that $(X,Y)_0$ should have "zero" amount of stochastic depencendy, since it is stochastically independent. Therefore if we cpould measure the "difference" between $(X,Y)_k$ and $(X,Y)_0$, then that difference should be a measure of the amount of stochastic dependence is exhibited by $(X,Y)_K$.

So, before we go any further with our calculation, look at these two graphs and see if you can "get a feeling" for "how different" they are from each other.

An approach that could be taken by information theory to arrive at a difference between these two distributions – but is not so taken – would be to subtract the values of the above two distributions at each of the 16 sample points, and the take the mean of all 16 of those differences. Those values, of course, are probabilities. Therefore, if this were the approach used by information theory, then it would amount to taking the mean of the differneces of the probabilities of the two differences.

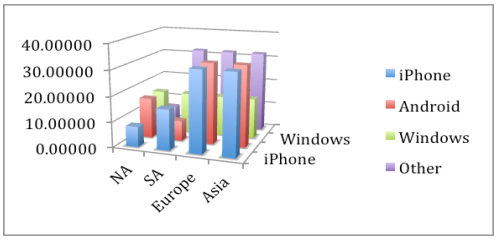
But information theory is almost always more interested in the uncertainty values of sample points than it is directly of the probabilities of those sample points. This is the way information theory. Therefore, information theory is going to want to calculate the

uncertainties of these two distributions first – before taking their differences. Calculating the uncertainties from the probabilities of these two distributions is exacltly what the next 4 columns does.

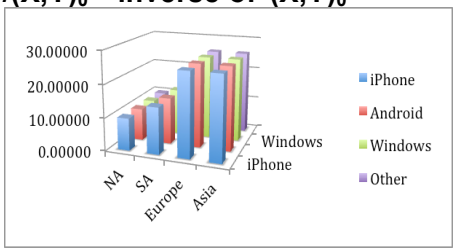
Recall that calculating uncertainties from probabilities is a two-step process. First you invert the probabilities. Second, you take the logarithm of those inverses. The next two columns are dedicated to inverting the probabilities in the first two columns. And the third pair of columns is dedicaated to taking the logarithms (base 2) of the previous two columns. Thus, after the first six columns, we have the individual uncertainties, according to both probability distributions, of each of the 16 sample points involved.

Lets look at both of these steps. First we shall invert the two probability distributions:

$1/(X,Y)_K$ – inverse of $(X,Y)_K$



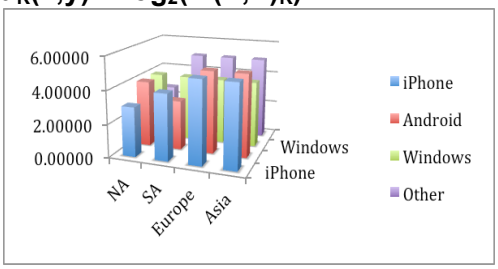
$1/(X,Y)_0$ – inverse of $(X,Y)_0$



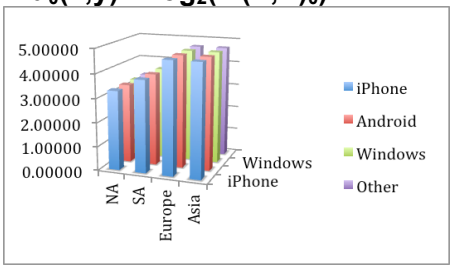
Compare these two graphs to the above two. Notice that sample points that are “taller” in the above two graphs are “shorter” in these two graphs, and vice versa. This “inversion of height” is the result of the inverting operation.

The next step is to obtain the uncertainty values of these two joint distributions, , by taking the logarithms of each of these inverses.

$u_K(x,y) = \log_2(1/(X,Y)_K)$



$u_0(x,y) = \log_2(1/(X,Y)_0)$

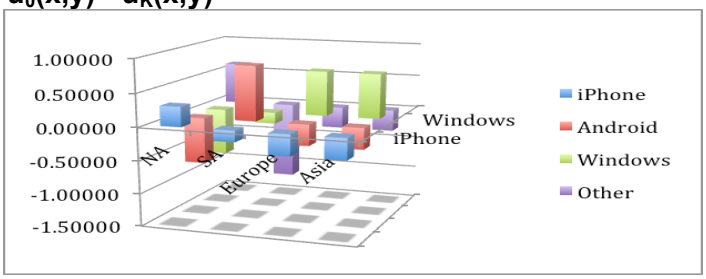


This “taking of the logarithm” of the inverse of the probability has the result of “leveling” those probabilities. This is one of the things that converting from probabilities to uncertainties does: it both inverts the probabilities, but it also “levels” those inverses as well.

Compare these two graphs to the above two. Notice that all of the sample points are the same relative heights as their corresponding sample points in the corresponding graph above. However, all of the sample points have been “squashed” somewhat. And the taller the were, the more they are “squashed” in the immediate graphs above. This higher degree of “squashing” for the taller sample points is the result of taking the logarithm.

The final two steps are the result of “combing” these above two graphs into one graph. This “combining” is actually the subtraction – on a point-by-point basis – of the points of the rightmost graph from the corresponding points of the left graph. The result is the following graph.

$u_0(x,y) - u_K(x,y)$

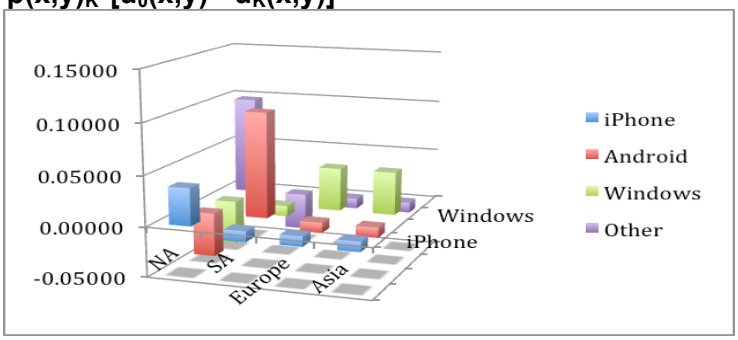


You will notice that some of these values are negative, since their corresponding subtrahends – the values being graphed - are negative.

The next-to-last step in calculating the mutual information of $(X,Y)_K$ is to multiply each of these $u_0(x,y) - u_K(x,y)$ values – one for each of the 16 sample points - by its joint probability $p(x,y)_K$. This yields values of the form $p(x,y)_K*[u_0(x,y) - u_K(x,y)]$ – again, one for each of the sixteen sample points.

This yields the final matrix whose graph is

$p(x,y)_K*[u_0(x,y) - u_K(x,y)]$



Finally, the last step in calculating the mutual information of X and Y in the joint distribution $(X,Y)_K$ is to sum all 16 of the values in the matrix of this graph. We have shown above that that sum is the value 0.1174.

This concludes our graphical look at the calculation of the mutual information of the chance variables X=”Phone Brand” and Y=”Continent” whose joint probabilities are defined by the joint distribution $(X,Y)_K$.

Derivation of I(X;Y) - Textbook Definition: Mean of log2(p(x,y)K/p(x)*p(y))

This section presents a second definition of the mutual information of two chance variables in a joint probability distribution. The definition presented in this section is the one usually seen in textbooks on information theory. An equivalent, but perhaps more intuitive definition of mutual information was presented in the previous subsection.

We have reemphasized that mutual information is a measure of the degree of stochastic dependence between two chance variables in a joint distribution. And as such, mutual information can be interpreted to also be a measure of the degree of portent or meaningfulness between those two chance variables.

Lets now present the more usual definition of the mutual information I(X;Y).

Definition 2: Mutual Information of two chance variables X and Y:

$$I(X;Y) = \sum_{i \in S} p(x,y)*\log (p(x,y)/(p(x)*p(y)))$$

In the two tables below, we shall also apply this formulation to our Cellphone investigation, principally to show that the same result, I(X;Y) = 0.1174, is achieved by the use of this definition as with the other one in the previous subsection.

Part 1: Derivation of I(X;Y) - Textbook Definition: Mean of log(p(x,y)/p(x)*p(y))

	p(x)	p(y)	p(x)*p(y)	p(x,y)	p(x,y)/(p(x)*p(y))	log(p(x,y)/(p(x)*p(y)))
iPhone, NA	0.25000	0.40625	0.10156	0.12500	1.23077	0.29956
iPhone, SA	0.25000	0.28125	0.07031	0.06250	0.88889	-0.16993
iPhone, Europe	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
iPhone, Asia	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
Android, NA	0.25000	0.40625	0.10156	0.06250	0.61538	-0.70044
Android, SA	0.25000	0.28125	0.07031	0.12500	1.77778	0.83007
Android, Europe	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
Android, Asia	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
Windows, NA	0.25000	0.40625	0.10156	0.06250	0.61538	-0.70044
Windows, SA	0.25000	0.28125	0.07031	0.06250	0.88889	-0.16993
Windows, Europe	0.25000	0.15625	0.03906	0.06250	1.60000	0.67807
Windows, Asia	0.25000	0.15625	0.03906	0.06250	1.60000	0.67807
Other, NA	0.25000	0.40625	0.10156	0.15625	1.53846	0.62149
Other, SA	0.25000	0.28125	0.07031	0.03125	0.44444	-1.16993
Other, Europe	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
Other, Asia	0.25000	0.15625	0.03906	0.03125	0.80000	-0.32193
Entropy						

Part 2: Derivation of I(X;Y) - Textbook Definition: Mean of log(p(x,y)/p(x)*p(y))

	p(x,y)	log(p(x,y)/p(x)*p(y))	p(x,y)*[log(p(x,y)/p(x)*p(y))]
iPhone, NA	0.12500	0.29956	**0.03745**
iPhone, SA	0.06250	-0.16993	**-0.01062**
iPhone, Europe	0.03125	-0.32193	**-0.01006**
iPhone, Asia	0.03125	-0.32193	**-0.01006**
Android, NA	0.06250	-0.70044	**-0.04378**
Android, SA	0.12500	0.83007	**0.10376**
Android, Europe	0.03125	-0.32193	**-0.01006**
Android, Asia	0.03125	-0.32193	**-0.01006**
Windows, NA	0.06250	-0.70044	**-0.04378**
Windows, SA	0.06250	-0.16993	**-0.01062**
Windows, Europe	0.06250	0.67807	**0.04238**
Windows, Asia	0.06250	0.67807	**0.04238**
Other, NA	0.15625	0.62149	**0.09711**
Other, SA	0.03125	-1.16993	**-0.03656**
Other, Europe	0.03125	-0.32193	**-0.01006**
Other, Asia	0.03125	-0.32193	**-0.01006**
Entropy			**0.1174**

Entropic Measures Relationships

Our interest in this section is to look at certain forms of entropic measures for the Cellphone-by-Continent example, and ascertain that they exhibit a number of numerical relationships.

The entropic measures in question are:

H(X), H(Y), H(X, Y), H(Y|X), H(X|Y) and I(X;Y).

The relationships that we want to portray for the Cellphone-by-Continent example are:

I(X;Y) = H(X) – H(X|Y)
I(X;Y) = H(Y) – H(Y|X)
I(X;Y) = H(X) + H(Y) – H(X, Y)
H(X, Y) = H(X) + H(Y|X)
H(X, Y) = H(Y) + H(X|Y)

In the following table, the Cellphone-by-Continent example, we have assembled the values of these six entropic measures, as well as an assessments of the above five relationships among these six entropic measures.

H(X)	2.0000	
H(Y)	1.8796	
H(X, Y)	3.7622	
H(X|Y)	1.8826	
H(Y|X)	1.7622	
I(X:Y)	0.1174	
To calculate:		
	I(X:Y)	**H(X) - H(X|Y)**
Does I(X:Y) = H(X) - H(X|Y)?	0.1174	0.1174
	I(X:Y)	**H(Y) - H(Y|X)**
Does I(X:Y) = H(Y) - H(Y|X)?	0.1174	0.1174
	I(X:Y)	**H(X) + H(Y) - H(X,Y)**
Does I(X:Y) = H(X) + H(Y) - H(X,Y)?	0.1174	0.1174
Does H(X) + H(Y|X) = H(X, Y)?	**H(X, Y)**	**H(X) + H(Y|X)**
	3.7622	3.7622
Does H(Y) + H(X|Y) = H(X, Y)?	**H(X, Y)**	**H(Y) + H(X|Y)**
	3.7622	3.7622

Indeed, these five relationships are substantiated by the table, since:

I(X;Y) = 0.1174 = H(X) – H(X|Y)
I(X;Y) = 0.1174 = H(Y) – H(Y|X)
I(X;Y) = 0.1174 = H(X) + H(Y) – H(X, Y)
H(X, Y) = 3.7622 = H(X) + H(Y|X)
H(X, Y) = 3.7622 = H(Y) + H(X|Y).

Appendix 5: Note on Information Theory vs Statistics

It may have occurred to the reader of this primer that *information theory* and *statistics* have a lot in common – as well as a lot that is different. In this situation, it seems useful to entertain a discussion comparing and contrasting the two disciplines.

This note is not intended to be an exhaustive or formal treatise, but rather some thoughts that compare the two disciplines on an informal basis and that discusses some ideas that are germane to information theory.

We shall begin with the observation that both disciplines depend heavily on the concept of the *probability distribution*. This fact argues strongly for both disciplines to be considered as, in some manner, offshoots of probability theory.

Intimacy to Probability Theory

It can be said that the degree of dependency of each of these two disciplines to the *probability distribution* construct determines the closeness of each discipline to the field of probability theory. For each, it is fair and reasonable to ask: "Is the discipline a) *a part of*, b) *an extension to*, or c) *an application of* probability theory?"

Information Theory

Lets first look at information theory in this regard. Information theory can be understood as "the study of uncertainty" [Kleeman 2012, Lecture 1, p. 1]. Information theory first turns its attention to the *constituents* of a probability distribution – its individual sample points or events. So, its first task is to define a measuring function for the degree of uncertainty of these individual sample points. We named such a measuring function, the *uncertainty* of the sample point (or event).

Second, having defined the notion of the uncertainty of individual sample points (and collections of them – events), information theory next turns its attention to defining the notion of the *uncertainty of an entire probability distribution* (i.e., across all of its sample points). And it does this by *averaging the uncertainties of the individual sample points*. The name that it gives to this average uncertainty value is *entropy*. This *entropy of a probability distribution* then becomes the central focus of information theory.

However, the point that we are making is *information theory* begins by *looking inside* of the probability distribution concept in order to begin its quest for defining a measure of uncertainty. Only after that does it construct a measure of the uncertainty of the entire probability distribution[21].

To me, that fact makes information theory very closely related to probability theory.

In fact, it can be argued that the concept of *entropy* should actually be a construct of probability theory, as does [Khinchin 1957, p.1]. It is perhaps only an accident of

[21] Admittedly, Shannon introduces and defines the concept of entropy first [Shannon 1948, p.11], and then subsequently points out that entropy is simply the average of a more fundamental quantity, which he calls the "entropy H_i of each state i" [Shannon 1948, p.13]. Shannon's "entropy H_i of each state i" is what this primer has named the "uncertainty of an event", and symbolized by the functional u(x). Clearly, the "entropy H_i of each state i" is a more fundamental quantity than the entropy of a probability distribution H - since H is the average of all the "H_i"s, as Shannon points out. It is the view of this primer that much of the mystery surrounding *entropy* is dispelled by introducing the more fundamental uncertainty function ("u(x)", or "H_i") first, and then defining H, or "entropy", subsequently as the average of the more fundamental quantity. Some contemporary texts [Vedral 2010] take the same approach in this regard as this primer, while others do not [Cover and Thomas 2006].

history that the entropy concept was invented by two folk in other disciplines (Gibbs and Shannon) who needed it for their extra-mathematical work. And since the probabilists had not gotten around to inventing it, they did. Thus, information theory was born, historically, outside of probability theory.

Because of this intimacy, I prefer to consider information theory as an extension to probability theory.

Statistics

It is fair to regard the discipline of statistics a having a *particular* overall, *practical*, concern about trying to deal with probability distributions in the real world of experimentation.

The particular concern of which I speak pertains to the problem of obtaining a well-defined characterization of the actual probability distribution of a given situation – which we shall refer to as an *experiment*.

In the real world of experimentation, it is either impractical or impossible to "get a handle" on what the actual probability distribution really is. This means, that it is generally impossible to achieve one or both of the following:

1. The exact identity of all of the sample points of the sample space of the distribution, or
2. The exact identity of the probabilities of all of the sample points of the sample space of the distribution.

Rather, the best that an experimenter can expect to achieve on a single observation relating to that distribution is to obtain some *subset* of actual sample space, along with a set of probabilities for each the members of that particular subset.

Moreover, because chance variation is involved in the observation, these probabilities may change each time a subsequent observation of the sample space is taken!

Consequently, in the face of these difficulties, the best that an experimenter can reasonably expect is to obtain a sequence of subsets with their respective sets of probabilities.

The name that is given to each of these successive observations is *trial*. And the name that is given to the successive sets of sample points, and their respective probabilities, is *sample*.

Thus, the discipline of statistics focuses on this difficulty by developing a discipline around the differences that can be expected to obtain between the actual and unknowable probability distribution for the entire sample space versus the individual probability distributions of each of these successive trials – or subsets of the real sample space with their own probability distributions.

Thus, statistics distinguished between the entire probability distribution – which it calls the *population* - and the individual *sample probability distributions*. The *sample distributions* are considered to be mere approximations to the population distribution; and statistics becomes concerned with how one can estimate the population distribution from a collection of sample distributions taken from the same sample space.

Statistics as a discipline, then, starts with the concept of a probability distribution, and then addresses a very real problem that develops in the practical world when trying to apply these concepts from probability theory. The principle problem that it addresses is

the fact that the actual probability distribution is very difficult to determine, and that multiple approximations – the sample distributions - must be dealt with instead.

Thus, it can be fairly stated that statistics is concerned with a particular practical difficulty in attempting to work with probability distributions within the real world of experimentation. This difficulty is that, in practice, only subsets of populations – samples – can realistically be obtained. Moreover, when taking multiple samples from the same population, one encounters chance variation, so that the samples from the same population may all be different. Statistics is concerned with developing probabilistically accurate techniques for estimating the population distribution from these multiple sample distributions.

Unlike information theory, statistics does not start out by delving *inside of the probability distribution*, but rather is concerned with estimating population distributions from sample distributions.

I believe that we can conclude, then, that *statistics is an application* of probability theory, rather than part of it, or an extension to it.

Ways of Characterizing Probability Distributions

We established in the previous section that both statistics and information theory deal with probability distributions, each in its own way. Their dealing with probability distributions is what they have in common and also what makes both of them offshoots of probability theory.

And each of these two disciplines has its own way of characterizing and of placing measurements on probability distributions – according to its own interests. This section explores these differences and their consequences.

Measuring Functionals

One can think of the act of measuring as associating a real number with a "thing". Generally, that "thing" is a complex entity, such as a human being. Thus a "measure" can be thought of as a mathematical function that associates a number to a "thing".

In mathematics, the first functions that we learn about usually associate numbers with other numbers, not "things" with numbers. For example, the function $f(x)=x^2+1$ associates numbers with numbers. For example, it associates a 3 with a 10, a 4 with at 17 and a 5 with a 26.

But functions can be more complicated than this. They can associate numbers to more complicated entities. For example, a function might associate a number to a *person*. Many such functions exist that do. Examples are weight, height, grade-point average and blood pressure. Another example – this time from mathematics – is the definite integral of a polynomial. Such a function associates a single number with a polynomial between an upper and a lower bound.

In mathematics, these functions that associate complicated things with numbers are given a special category. They are called *functionals*.

Both *statistics* and *information theory* use special functionals to characterize – actually, to *measure* – probability distributions. However, the two disciplines most often use completely different measures (functionals) from each other for the purpose of measuring probability distributions. This is because each of these disciplines has different interests than the other regarding probability distributions. Therefore, each discipline defines its own set of functionals to measure probability distributions. And each of these functionals measures different aspects of a probability distribution.

In the two subsections below, we explore the various types of functionals that are defined by each of these two disciplines for the purposes of defining ways of measuring probability distributions.

This might be likened to a human being. Both your medical doctor and track coach have different interests in your body, and therefore have different ways of measuring your body. Some of these measuring functions may be of interest to both of them. But for the most part, they have different interests in your body, and therefore different measuring functions.

For example, your medical doctor is interested in your height, weight, blood pressure, temperature, and several measures that are collected in the laboratory. On the other hand, your soccer coach is interest in your 100-meter dash time, your javelin throw distance, your mile run time and how many hours you run per week. All of these are functionals that associate you with some number. And each of these functional measures a distinct aspect of you.

So too, both statistics and information theory define functionals that measure different aspects of a probability distribution. Both of these disciplines are primarily characterized and distinguished by which of these functionals they are interest in. And, the primary focus of both disciplines is the definition and application of these functionals. The following two subsections delve into the sets of measuring functionals used by these two respective fields.

Measuring Functionals of Statistics

Statistics defines the following set of measuring functionals on probability distributions:

Mean
Median
Variance
Standard deviation
Skewness
Kurtosis
Moments
Central Moments

Categories of Measuring Functionals Defined by Statistics

Each of these measures a different aspect of a probability distribution – just like your weight, your height and your blood pressure measure different aspects of your body. In fact, the above measures can be grouped in to semantic categories.

Specifically, *mean*, *median* and *mode* are called “measures of central tendency”. Each of these three measures some aspect of centrality concerning the sample points of the sample space.

On the other hand, *variance* and *standard deviation* measure, each in its own way, how closely the sample points are grouped about a center (the mean). Therefore, these two functionals are called “measures of variation”.

Skewness and *kurtosis* both measure other aspects of how the sample points of the distribution change.

Statistics also defines two other families of measuring functionals called *moments* and *central moments*. The previous set of measuring functionals just discussed have semantically clear meanings related to intuitive aspects that distributions might exhibit.

However, there is a need for a better mathematical characterization of probability distributions that are more amenable formal theory. Consequently, mathematical statistics defined these two, more formal, categories of measuring functionals. We shall not define these categories in this appendix, nor discuss them further. Suffice it to say that they exist, and are named *moments* and *central moments*.

Parameters versus Statistics

Recall that *statistics* distinguishes itself as a discipline by its interest in a particular problem. The problem is that sample spaces in the experimental world are generally too large to be sampled in their entirety. As a consequence, the practice of experimental statistics requires that one deal with subsets of the actual sample space rather than with its entirety.

These subsets are called *samples*. The entire sample space is called the population. Thus, in experimental statistics, one must deal with *samples* as a proxy for the real sample space. The approach taken to this practical reality is to take multiple samples and use them collectively as an estimate of the entire sample space.

Therefore, statistics as a discipline is constantly interested in "How can the entire sample space – the *population* - be estimated from the current collection of samples?" The approach taken by statistics to making this assessment is to develop the above measuring functionals for each of the samples, and then to use those functionals to estimate the same functionals for the *population*.

Thus, statistical practice develops and maintains two different versions of all of the above-mentioned measuring functionals. One of these versions is for *samples* and the other of these versions is for the *population*. The version these functionals that it keeps for the samples are called *statistics*. The versions of the functionals that it keeps for the population are called *parameters*.

In other words, statistics keeps one version of the mean, median, variance, standard deviation, skewness, kurtosis moments and central moments for the population. These are called *parameters*. And, statistics keeps a separate set of the mean, median, variance, standard deviation, skewness, kurtosis moments and central moments for each of the samples. These are all called *statistics*.

The practice of experimental statistics has a lot to do with 1) developing sets of statistics from samples, and 2) then using them to estimate the parameters of the population. There are a number of techniques based upon theorems from probability theory that experimentalists use in order to make these estimates.

Special Requirements of Statistical Functionals

All of the measuring functionals mentioned above – mean, median, variance, moments, etc. – require certain information as input values. One such information source is the probabilities of the distribution being measured. In fact, all of the functionals require these probabilities as inputs.

However, there is another type of numerical input that also required by all of these functionals from the discipline of statistics. This extra type of numerical information is referred to as the "values" of the sample points. These *values* are a second numerical function, call it "v", on the sample points.

These *values* normally have semantics that are peculiarly germane to the application domain that is being modeled as a statistical application. For example, if the subject of the statistical study is school pupils, the values might be test grades. If the subjects are

competitive athletes, the values might be points-per-game. If the subjects are hospital patients, the values might be various biological vital signs.

In other words, for the statistical functionals (mean, media, standard deviation, etc.), each sample point must have two numbers assigned to it: 1) a probability, and 2) a "value". Thus, we can say that the statistical functionals depend upon two other functionals being defined on the sample space in question: 1) the *probability functional* p(x) and 2) the *value functional* v(x).

These two other functionals (probability and value) are required as inputs to these all of these statistical functionals (*mean, median, variance, moments, etc.*). This fact can be made clear by looking at the algorithm of any one of these statistical functionals.

Lets take the *mean* for example. Like all of the other statistical functionals, *mean* is one way of placing a measurement on a probability distribution. Of course, statistics has two versions of the mean – as it does all of the statistical functionals: 1) a *population mean*, and 2) a *sample mean*. In both case, the formula for calculation is the same. But we are dealing a different set of sample points in each of these two cases: either the sample points of the entire population or the sample points of a subset of the population.

In either case, the formula for the mean says to multiply the *value* of each sample point by the *probability* of that sample point. Obviously, such a product yields a number. The next step in calculating the mean is to sum all of these products. This yields the final value of the mean.

In symbols, then, the mean of a probability distribution p is:

$$\text{mean}(p, v) = \sum_{i \in S} p(x)*v(x)$$

where

$x \in$ sample space S
p(x) is the probability of x with respect to p
v(x) is the *value* of x according to value function "v".

Thus, all statistical functionals are functionals on a sample space S that require the existence of two other simpler functionals p(x) and v(x) on the sample space S. We have used the *mean* as an example of such a functional. However, the same is true of the other statistical functionals that the discipline of statistics embodies (median, variance, standard deviation, moments, etc.).

Not all Probability Spaces have Value Functions

But there is a problem with requiring that all probability distributions have value functionals. The problem is that not all application domains offer natural values functionals for their sample spaces. Examples are:

The fact that statistics requires such value functionals means that some applications domains – the one that do not have value functionals – are out of bounds for the discipline of statistics!

Measuring Functionals of Information Theory

The measuring functionals defined by information theory are all of the same family: the *entropic measures*, which we a have also called the *entropic functionals.* All of these entropic measures have the task of measuring the degree of uncertainty of some aspect of a probability distributions. We have shown in Part I of this primer that an

alternate interpretation of entropic measures is that it measures the *degree of spread of probabilities* of some aspect of a probability distribution.

All of these entropic measures represent different forms of *entropy*. In fact, *entropy* is the simplest form if *entropic measure*. Other entropic measures that we have seen include: conditional entropy, relative entropy and mutual information. Essentially, information theory can be characterized very generally as the "study of uncertainty". More mathematically, however, it can be described as the "study of entropic measures".

What is significant about *entropic measures*, or *entropic functionals*, is that their only input is the probabilities of the probability space.

A significant conclusion of this is that entropic measures do not require that the sample points of their sample space have a value functional – as they do with statistical functionals.

This means that information theory can be applied to any probability space – and to any application domain that "has probabilities". This includes probability spaces that also happen to have value functionals. These value functionals are simply ignored by the entropic measures.

The More General Applicability of Information Theory

This freedom from having to have value functionals is especially significant for the applications whose sample spaces include very complex entities, or sample points. Many of these types of applications concern populations of entities that are so complex that there is no natural association of numbers to the entities.

For example, suppose one is interested in an application to the biological domain. Many of the subject entities of biological studies are sufficiently interesting because of their individual complexity. For these entities, there is no reason to "map them to the real numbers just so that we can take averages, and other statistics".

However, these subjects *do* enjoy substantial chance variation – and are therefore subject to be characterized by the entropic functionals of information theory – specifically *entropy*, *mutual information* and *entropy rate*.

Because of this, information theory can find application in complex domains where the applications of statistical measuring functionals have limitations.

References

[Cover and Thomas 1991] Cover, Thomas M. and Joy A. Thomas; Elements of Information Theory; John Wiley & Sons, Inc.; New York; 1991.

[Jaynes 1957] Jaynes, E. T.; Information Theory and Statistical Mechanics; The Physical Review, Vol. 106, No 4, 620-630, May 15, 1957. Also at http://www.weizmann.ac.il/complex/tlusty/courses/InfoInBio/Papers/JaynesInformationTheory.pdf, 1957.

[Khinchin 1957] Khinchin, A. I; The Mathematical Foundations of Information Theory; Dover Publications, Inc.; 1957; New York.

[Kleeman 2012] Kleeman, Richard; course syllabus: Information Theory and Predictability; Syllabus at http://www.math.nyu.edu/faculty/kleeman/syllabusinfo.html.

[Shannon 1948] Shannon, Claude E.; The Mathematical Theory of Communication; At http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf; 1948.

[Shannon 1963] Shannon, Claude E.; The Mathematical Theory of Communication; First Paperback Edition; University of Illinois Press; 1963.

[Vedral 2010] Vedral, Vlatko; Decoding Reality: The Universe as Quantum Information; Oxford University Press; Oxford, New York; 2010.

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

[Leff 2012] Leff, Harvey S.; Removing the Mystery of Entropy and Thermodynamics – Part I; The Physics Teacher, Vol. 50, p. 28; January 2012.