# UNKNOWNS AND GUESSING
# EXPECTATION AND PROBABILITY

MAURICE J. DUPRÉ

## 1. INTRODUCTION

Mathematics is basically the art of inventing ways to handle information for purposes of knowledge extraction and communication. Statistics is basically an application of mathematics to certain types of real world problems. For elementary statistics and its applications the main branches of mathematics being applied are theories of analysis, probability, and expectation. In order to have the basic tools for handling information, we are going to give an elementary discussion of sets and functions. We will then discuss the general idea of unknowns, as they appear in the real world and see that there are reasonable rules for guessing values of unknowns, based purely on logical consistency. We will find that the constraints of logical consistency and natural consistency with changes of units force a system of rules from which we can easily derive the rules of probability and expectation in complete generality. When reading a given section, if it seems to get too complicated, try skipping ahead to the next section, as there is more in each of the sections on sets and functions than is necessary for understanding the basic rules of probability and expectation. If you then find you need some information from preceding sections, you can go back and look it up. These notes are basically a simplified treatment of the development in [Dupre and Tipler,[4]]. Readers interested in more background should consult [Cox,[1]] and [Jaynes,[5]]. For an elementary treatment which gives more consideration to set theoretical foundations see [2]. For an elementary treatment (using only algebra-no calculus required) of expectation, probability and applications to statistics as developed from these foundations presented here, the interested reader can consult [3].

## 2. SETS AND LOGIC

To begin, let us think of a SET as simply a collection of things. Since you could now ask me what I mean by a collection of things, I would say in response that I am thinking of some kind of aggregation of things. If you ask me what I mean by this last phrase, I might say a set of things, and we would have a circular string of definitions. So, I will take the idea of a set as an undefined term, but I hope you understand what I am talking about from my previous attempts to define the notion of set. We will designate sets with capital letters usually from the first half of the alphabet, such as

$$A, B, C, ..., K, L, ...,$$

and if there are only a finite number of elements in a set, then we just list the elements inside curly brackets, such as

$$A = \{a, b, 5\}.$$

Here we have specified that the set denoted by $A$ has or contains exactly the objects named $a, b, 5$, but no others. The objects in a set are referred to as its MEMBERS or its ELEMENTS, and a convenient and traditional notation here is that $x \in A$ means that $x$ is a member of set $A$, or $x$ is an element of set $A$. In general, a set is completely determined by its members, that is if $G$ and $H$ are both sets and if they have exactly the same members, then they are equal. If every member of $G$ belongs to $H$, then we say $G$ is a SUBSET of $H$, and write this in symbols as $G \subset H$. We can therefore see that $G = H$ is simply the same as $G \subset H$ and $H \subset G$

both being true. When we have too many objects in our set to easily name them all, we will sometimes specify the members with some kind of systematic partial list which alludes to the full list such as

$$B = \{a, b, c, ..., x\}$$

or

$$C = \{3, 4, 5, ..., 435\}.$$

We can even use this device to specify certain types of infinite sets called countably infinite sets such as

$$K = \{3, 4, 5, ...\},$$

which specifies the set of all natural numbers above the number 2, or such as

$$E = \{6, 8, 10, ...\},$$

which specifies all even natural numbers above 5. Notice that actually we could have resorted to using statements to specify sets. For instance, $B = \{x|x$ is a letter of the alphabet preceding y in the ordinary alphabetic order $\}$, and $E = \{x|x$ is an even natural number above 5 $\}$. When we read this last definition of $E$ we should say ”$E$ is the set of all $x$ such that $x$ is an even natural number above 5”. That is, the vertical bar stands for the phrase ”such that”. Sometimes a colon is used in place of the vertical bar so $E = \{x : x$ is an even natural number above 5$\}$. A statement of the form ”x is an even natural number” is technically a whole family of statements if $x$ is allowed to be possibly anything from some large collection of possible values which we would call the UNIVERSE of discourse. Statements of this kind are then more than just simple statements, and are called STATEMENT FUNCTIONS. For instance a SIMPLE STATEMENT would be a statement such as ”3 is an even natural number”. Simple statements state facts that are either true or false. A statement function such as ”x is an even natural number” is neither true nor false until $x$ is replaced by a definite object in our universe. We can write a statement function in general as $p(x)$, so we think of the statement $p(x)$ as stating a property of $x$. For instance in the preceding example, $p(x) =$”$x$ is an even natural number”. Thus, we can think of $\{x|p(x)\}$ as being the set of all things in the universe for which the statement $p(x)$ is actually true. Notice if $P = \{x|p(x)\}$, then the statement functions $x \in P$ and $p(x)$ are logically equivalent, as they are true of exactly the same set of members of the universe, namely the set $P$. So, if $S$ is a set and $p(x)$ is any statement function, then we can form the set $\{x|x \in S$ and $p(x)\}$, and we will henceforth write this as $\{x \in S|p(x)\}$, read ”the set of $x$ belonging to $S$ such that $p(x)$”. If $A$ and $B$ are sets we define their INTERSECTION, $A \cap B$, to be the things common to both sets, that is their overlap. We then define their UNION, $A \cup B$, to be the set of things belonging to at least one of these sets. If we restrict attention to a fixed set $S$, and its subsets, then for any statement functions $p(x)$ and $q(x)$, we can form the subsets $P = \{x \in S|p(x)\}$ and $Q = \{x \in S|q(x)\}$, and notice that

$$P \cap Q = \{x \in S|p(x) \ \&q(x)\}.$$

Here, & is the symbol for ”and”, which in logic is called an example of a logical connective-it is used to connect statements or statement functions together to form more complex statements. Another logical connective is ”or” which in logic is never taken to be the exclusive ”or” of every day speech but rather the inclusive ”or” which allows the possibility of both. That is in every day common speech I might say I am either going to the movies or I will read a book tonight, the implicit implication being that I am not possibly going to do both. But, in logic the possibility of both is always admitted unless explicitly excluded. Thus using the rules of logic strictly, I would have to say that I will go to the movies or read a book but not both, in order to express what I said before in ordinary everyday speech. With this in mind, we see that

$$P \cup Q = \{x \in S|p(x) \ \mathrm{or} q(x)\}.$$

From these two examples, we see that the logical & corresponds to intersection$= \cap$ in set theory whereas logical ”or” corresponds to union$= \cup$ in set theory. Another simple operation on statements is logical negation, "not". If $p(x)$ is any statement function, then not$p(x)$ is the negation of $p(x)$, so not$p(x)$ is true exactly when $p(x)$ is false. If $B, C \subset S$, then the complement of $B$ in $C$ is denoted $C \setminus B$, and is given by $C \setminus B = \{x \in S | x \in C$ but not $x \in B\}$. When set $S$ is clearly understood, it is common to write $B^c$ for $S \setminus B$, that is $B^c = S \setminus B$. Notice that $C \setminus B = C \cap B^c$. Finally we note that

$$P^c = \{x \in S | \text{not} p(x)\},$$

which means that negation in logic corresponds to complementation in set theory. Notice that we have here a complete correspondence between the logical operations and the set operations. Since the set operations can be viewed pictorially using Venn diagrams, we then have a way of viewing logic in terms of pictures. We use the notation $\emptyset$ for the EMPTY SET, which we can define as $\{x \in S | x \neq x\}$, so $\emptyset$ is the set which contains no members at all. Every set contains the empty set as a subset. Here are notations for some of the sets we deal with over and over.

$$\mathbb{R} = \{x |\ x \text{ is a real number}\}.$$

$$\mathbb{Z} = \{x \in \mathbb{R} |\ x \text{ is an integer}\}.$$

$$\mathbb{Z}_+ = \{x \in \mathbb{Z} | x \geq 0\} = \{0, 1, 2, 3, ...\}.$$

$$\mathbb{N} = \{x \in \mathbb{Z} | x > 0\} = \{1, 2, 3, ...\}.$$

## 3. SETS AND FUNCTIONS

We have discussed the basic operations of union, intersection, and complemetation for subsets of a given set $S$, but now we will discuss some operations which take us beyond the set $S$ we start with. To begin, we can form the set of all subsets of $S$ called the POWER SET of $S$, and denoted $2^S$. Thus $2^S = \{A | A \subset S\}$. If $S = \{a, b, c\}$, then

$$2^S = \{\emptyset, \{a\}, \{b\}, \{c\}, \{b, c\}, \{a, c\}, \{a, b\}, \{a, b, c\}\}.$$

We define the cardinality of a finite set as the number of its members, and if $A$ is any finite set, then $card(A)$ denotes its cardinality. Notice that in our example, $card(S) = 3$ and $card(2^S) = 8$, and in fact $2^3 = 8$. This is a special case of the general law that $card(2^S) = 2^{card(S)}$, for any finite set $S$, and this is partly the reason for the notation, although we will soon see there is an even better reason. To see this we must define the concept of a function. Suppose that $S$ and $T$ are any sets. By a FUNCTION $f$ from $S$ to $T$, we mean a specific rule which assigns a member of $T$ to each member of $S$. In this case, we write $f(s)$ for the member of $T$ which rule $f$ assigns to $s \in S$. We generally write

$$f : S \to T$$

or

$$S \xrightarrow{f} T$$

to mean that $f$ is a function from $S$ to $T$. If $f : S \to T$, then we call $S$ the DOMAIN of $f$, so domain$(f) = S$, and we call $T$ the CODOMAIN of $f$, so codomain$(f) = T$. It is also important to keep in mind that if $f : S \to T$ is a function and $s$ belongs to the domain of $f$, then $f(s)$ usually does NOT mean multiplication, it is just a convenient notation for the member of $T$ which $f$ assigns to $s \in S$. It is often useful to think of a function as an input-output device. The members of the domain of the function are the allowable inputs, and $f(s)$ is the output when $s$ is the input to function $f$ and $s$ is in the domain of $f$, that is to say $s$ is an allowable input.

Suppose that $S = \{a, b, c\}$ and $T = \{2, 4, 5, 6\}$. To specify a rule $f$ in this case, it is enough to specify $f(a), f(b)$, and $f(c)$. For instance, we could define $f(a) = 4, f(b) = 5, f(c) = 4$. Notice that the rule can assign the same member of $T$ to more than one member of $S$ and that not every member of $T$ needs to be assigned to a member of $S$. What is necessary is that every member of $S$ has some member of $T$ assigned to it, so if $s$ is any member of $S$ then $f(s)$ is defined by the rule $f$. Such rules are customarily denoted by lower case letters, but not always, and we will have plenty of occasion to use upper case letters for functions. Notice that $f$ is completely specified by giving the list $(f(a), f(b), f(c))$, called a three-tuple of members of $T$. Thus in our example the three-tuple specifying $f$ is simply $(4, 5, 4)$. How many such rules can be made up? Notice that $T$ has 4 members, that is $card(T) = 4$. To make a three-tuple which defines a function here all we have to do is choose a member of $T$ to go in each position of the Three-tuple. If any two three-tuples differ at any location, the rules are different. Since there are 3 slots of the three-tuple to be filled in and for each we have 4 choices, the total number of ways to fill in the three-tuple is 4 times 4 times 4, or $4^3 = 64$. We can denote the set of all functions from $S$ to $T$ by $T^S$, so

$$T^S = \{f|f : S \to T\}.$$

Then we find in general that $card(T^S) = [card(T)]^{card(S)}$. In case that $T_1, T_2, T_3, ..., T_n$ are all sets, then we can let $T$ be their union, set $S = \{1, 2, 3, ..., n\}$, and now a function from $S$ to $T$ is specified by an $n-$tuple. That is if $f$ is a function from $S$ to $T$, then it is completely specified by the $n-$tuple $(f(1), f(2), f(3), ..., f(n))$. Thus we can regard $T^S$ as the set of all such $n-$tuples of members of $T$. Further, if $x \in T^S$, we will simply write $x = (x_1, x_2, x_3, ..., x_n)$, so that if $x$ is an $n-$tuple, then $x_k$ denotes the entry in slot $k$ of the $n-$tuple. Remembering that $T$ was actually the union of the $n$ sets $T_1, T_2, T_3, ..., T_n$, we can form the subset $P$ of $n-$tuples consisting of those with the property that the entry in the first slot comes from $T_1$, the entry in the second slot comes from $T_2$, the entry in the third slot comes from $T_3, ...,$ the entry in the $k^{th}$ slot comes from $T_k, ...,$ and finally, the entry in the last slot comes from $T_n$. This is called the CARTESIAN PRODUCT of the sets and is denoted by

$$P = T_1 \times T_2 \times T_3 \times ... \times T_k \times ... \times T_n = \prod_{k=1}^{n} T_k.$$

Thus we have

$$P = \{x \in T^S | x_k \in T_k, 1 \leq k \leq n\} = \prod_{k=1}^{n} T_k.$$

Maybe not surprisingly, we can build the ordinary counting numbers out of sets. We define $0 = \emptyset, 1 = \{\emptyset\}, 2 = \{0, 1\}, 3 = \{0, 1, 2\}, ..., n + 1 = \{0, 1, 2, 3, ..., n, \}, ...,$ and so on ad infinitum. Notice each natural number is defined as a specific set whose cardinality is that number and that each is defined using only the numbers preceding it. This is called an inductive definition. But now, notice that since 2 is actually a set, that for any set $S$ we should have $2^S$ is the set of functions from $S$ to 2, whereas, we already defined $2^S$ as the set of all subsets of $S$. It appears that we have an inconsistency in our notation, and technically we do. However, we can realize that a function $f$ from $S$ to $2 = \{0, 1\}$ can be used to determine a subset $A$ of $S$, by using the function to specify the subset $A = \{x \in S | f(x) = 1\}$. Notice that since $f$ can only assign the numbers 0,1 to members of $S$, that as soon as we know $A$ we in fact know $f$, since it must be the case that $f(x) = 0$ for each $x \in S \setminus A$. Conversely, given any subset $B \subset S$, we can specify a function $I_B$ from $S$ to $2 = \{0, 1\}$, by the rule requiring $I_B(x) = 1$, if $x \in B$ but $I_B(x) = 0$, if $x \in S \setminus B$. Notice that $B = \{x \in S | I_B(x) = 1\}$. In particular, we have $I_A = f$. We call the function $I_B$ the indicator of the subset $B \subset S$. We have seen that there is a connection between set theory and logic. Now, we will see that there is also a connection between set theory and algebra. In general, if $f, g : S \to \mathbb{R}$, then we can form the functions $f + g$ and $fg$. The function $f + g : S \to \mathbb{R}$ is the rule which assigns to each member $s \in S$ the number $f(s) + g(s)$. Likewise, the rule $fg : S \to \mathbb{R}$ is the rule which assigns to each member $s \in S$ the number $f(s)g(s)$. That is, we add values to form $f + g$ whereas we multiply values to form $fg$. For instance it is easy to see that if $A$ and $B$ are subsets of $S$, then

$$I_{A\cap B} = I_A I_B,$$

and

$$I_{S\setminus B} = 1 - I_B.$$

In particular, this means $I_S = 1$ and $I_\emptyset = 0$, the constant functions with values 1 and 0, respectively. Let's try to make the indicator of $A \cup B$ out of the indicators of $A$ and $B$. Since $I_{A\cap B} = I_A I_B$, it follows that under the correspondence between algebra and set theory we are building, we are finding that set intersection corresponds to multiplication in algebra. Maybe, union corresponds to addition. This means we should try and see if $I_{A\cup B}$ is simply $I_A + I_B$. If we examine $I_A + I_B$, we find that it cannot be an indicator function because if we evaluate this function on a point $s \in A \cap B$, then the resulting value is $1 + 1 = 2$ and an indicator must take all its values in the set $2 = \{0, 1\}$. But, this appears to be the only problem with $I_A + I_B$, for if $s \in S$ is in neither $A$ nor $B$, then the value is 0, just as it is for $I_{A\cup B}$, whereas if $s$ belongs only to $A$ or only to $B$, then $I_A + I_B$ gives the value 1, just as $I_{A\cup B}$ does. In order to fix this problem, we need to subtract 1 from the value of $I_A + I_B$ if $s \in A \cap B$, but do nothing if $s \in S \setminus (A \cap B)$. That is, if we can find a function $g : S \to \{0, 1\}$ such that $g(s) = 1$ when $s \in A \cap B$, and $g(s) = 0$ when $s \in S \setminus (A \cap B)$, then $I_A + I_B - g$ will do the trick. But notice we already have this function, namely $g = I_{A\cap B} = I_A I_B$. We therefore find that

$$I_{A\cup B} = I_A + I_B - I_{A\cap B} = I_A + I_B - I_A I_B.$$

Well, this means that under the correspondence between set theory and algebra, the operation of union of sets is more complicated than simple addition-it involves both addition and multiplication. However, we do see that if $A$ and $B$ are DISJOINT, meaning that they do not overlap, that is to say simply $A \cap B = \emptyset$, then in this special case we do have $I_{A\cup B} = I_A + I_B$. Thus, we should remember that intersection corresponds to multiplication and union corresponds to addition, but that the correspondence for union only works perfectly in the case of a disjoint union. We can also notice that in case that $S$ is finite, we have $card(A\cup B) = card(A)+card(B)$, if $A$ and $B$ are disjoint, whereas this would be generally false if $A \cap B \neq \emptyset$. What would be the general equation for cardinalities that will work even if $A \cap B \neq \emptyset$?

## 4. FUNCTION OPERATIONS

Let us observe how functions in general combine to make new functions. If $f : S \to T$ and $g : T \to U$, then we can form the composition $g \circ f : S \to U$, by using the rule

$$(g \circ f)(x) = g(f(x)), \ x \in S.$$

That is, thinking of a function as an input-output device, we can imagine that $g$ is a function that takes the output from $f$ as input. It is easy to see that if we have a third function $h : U \to V$, then the composition is associative:

$$(h \circ g) \circ f = h \circ (g \circ f),$$

since applying either of these to an $x \in S$ gives the result $h(g(f(x))))$. The simplest function on any set $S$ is the IDENTITY function on $S$ denoted by $id_S$, and whose rule is simply

$$id_S(x) = x, \text{ for each } x \in S.$$

Obviously we have

$$id_T \circ f = f = f \circ id_S.$$

For $A \subset S$ we define the IMAGE of $A$ under $f$, denoted $f(A)$, by

$$f(A) = \{y \in T \mid y = f(x) \text{ for some } x \in A\} = \{f(x) \mid x \in A\}.$$

For $B \subset T$ we define the INVERSE IMAGE of $B$ under $f$, denoted $f^{-1}(B)$, by

$$f^{-1}(B) = \{x \in S \mid f(x) \in B\}.$$

In case $B = \{y\}$ has only one member $y$, then we write $f^{-1}(B) = f^{-1}(y)$. We say that $s : T \to S$ is a section of $f : S \to T$ provided that $f \circ s = id_S$. Notice that this means that for each $t \in T$ we have $f(s(t)) = t$, and therefore that $s(t) \in f^{-1}(t)$, for each $t \in T$. In other words, we have

$$s \in \prod_{t \in T} f^{-1}(t) \text{ if and only if } s : T \to S \text{ is a section of } f.$$

We can observe that if $B$ and $C$ are disjoint subsets of $T$, then $f^{-1}(B)$ and $f^{-1}(C)$ are disjoint subsets of $S$. In particular, if $t_1$ and $t_2$ are different members if $T$, then $f^{-1}(t_1)$ and $f^{-1}(t_2)$ are disjoint subsets of $S$. That is, $\{f^{-1}(t) \subset S \mid t \in T\}$ is a collection of disjoint subsets of $S$ which may contain the empty set. But, if $f$ has a section, say $s$, then since $s(t) \in f^{-1}(t)$ for each $t \in T$, it follows that all the sets $f^{-1}(t)$ for $t \in T$ must be nonempty.

## 5. ALGEBRA, LOGIC, AND UNKNOWNS

We have a correspondence between logic and set theory and we have a correspondence between set theory and algebra. From these two correspondences, we can build a correspondence linking logic directly to algebra. In order to deal with logic at a completely general level, let us develop a notion of unknown number which is completely general. By an UNKNOWN NUMBER, I mean a numerical quantity which has a description so that we know there is a numerical value being described even though we may not know the value exactly. For example, I might say $X$ is the air temperature in degress Fahrenheit at the WWLTV weather station. If you call up this TV station or go to their website, you can find this temperature, but if you have not done this, then you can only guess what the value might be. Let us suppose that we live near that weather station, and we want to dress to go outside. Then we need to make some sort of guess as to the value of the outside temperature in order to know how to dress. In fact, there are many situations in real life where we make decisions based on incomplete information, and where we need to make a guess of some important numerical value in order to proceed. Is that car going to reach my position in the street before I can get across? How fast is it going? Can my mini-van make it under the roof of the parking garage I'm about to drive into? I spot a police car beside the road, I'm going roughly the same speed as everyone else, but maybe I have overlooked a speed limit sign. Am I speeding? You have probably encountered many such situations. That's life. The general notion of an unknown I am trying to describe here could lead to the same paradoxical problems as using arbitrary statement functions to form sets. To get out of this problem, we will simply assume to start that we are dealing with some possibly large set of unknowns, which we will denote by $\mathcal{A}$. I will generally use capital letters near the end of the alphabet for unknowns, but occasionally this convention will be violated. For this reason, I will use the fancy letter $\mathcal{A}$ for the set of all the unknowns I want to consider. I will assume that if $X$ and $Y$ are unknowns, then $X + Y$ is an unknown, whose description is that its value is the result of taking the value of $X$ whatever that is and adding it to the value of $Y$ whatever that is. For instance, if $X$ is the length in inches of a car I see parked across the street and if $Y$ is that car's engine temperature in degrees Celsius, then $X + Y$ is the result of adding those two numbers. Notice that the numbers themselves have no units-the units are part of their descriptions, so any two can be added. Likewise, I can define $XY$ to be the product of the two unknowns $X$ and $Y$. Our first assumption is that if $X, Y \in \mathcal{A}$, then

$$X + Y \in \mathcal{A}$$

and

$$XY \in \mathcal{A}.$$

In mathematics, such a system where we can add and multiply is generally called an ALGEBRA. Clearly the ordinary properties of associativity and commutativity should apply, as well as the distributive law:

$$X(Y + Z) = XY + XZ.$$

You should see that speaking generally of unknown numbers, we have to consider all levels of ignorance of the values from complete ignorance to complete knowledge. For instance, if I say $X$ is the number 5, then there really is no doubt about what $X$ is, whereas, if I say that $X$ is the time in nanoseconds until the next proton in the universe decays, then who knows? At an intermediate level of ignorance, if I know that $(X-5)^2 = 4$, then I can see that $X-5$ is either 2 or $-2$, from which I can then conclude that $X$ is either 3 or 7, but beyond that, I have no way of knowing which it is. Notice this means that every actual number can be considered as a description of itself-that is its is completely known. In other words, $\mathcal{A}$ contains the set of all real numbers, or in symbols, we have $\mathbb{R} \subset \mathcal{A}$. What about logic? Well, here we restrict ourselves to statements of fact which are either true or false. Statements of opinion such as "Liszt's music is more beautiful than Mozart's" are neither true nor false, but rather depend on the person making the statement. Of course a statement like "Joe thinks Liszt's music is more beautiful than Mozart's" where Joe is a definite person, is now a statement of fact which is either true or false-that is, if we ask Joe, then he decides, but until we ask him, we are in the dark. Whether we know a statement to be true or false is another matter of course. Some statements we are pretty sure of and other statements we need to deal with may be a bit "iffy". Again, that's life. If $A$ is any statement, then we can use it to define an unknown in a very simple way reminiscent of the indicator function for a subset, and so we will call it the INDICATOR UNKNOWN of $A$ and denote it by $I_A$ just as for the indicator of a subset. But now, we define $I_A$ to simply be a number which is 1 if $A$ is true and 0 if $A$ is false. If I know $A$ is true, then that is exactly the same as knowing $I_A = 1$, whereas if I know that $A$ is false, then that is the same as knowing that $I_A = 0$. If I do not know whether $A$ is true or false, then I do not know whether $I_A$ is 0 or 1, so $I_A$ is certainly an unknown. Our next assumption is that as far as the unknowns we are allowing in the set $\mathcal{A}$, we will assume that any indicators of any statements we deal with actually belong to $\mathcal{A}$, that is $I_A \in \mathcal{A}$, for every statement $A$ we need to consider. Just as with set indicators, it is easy to see that we have for the connection through indicator unknowns between logic and algebra,

$$I_{A\&B} = I_A I_B, \tag{5.1}$$

and

$$I_{A \text{ or } B} = I_A + I_B - I_{A\&B}, \tag{5.2}$$

so

$$I_{A \text{ or } B} = I_A + I_B - I_A I_B. \tag{5.3}$$

In addition, clearly

$$I_{\text{not } A} = 1 - I_A. \tag{5.4}$$

Thus, as $\mathcal{A}$ is an algebra where we can add and multiply, it follows that if $A, B$ are statements that I can consider, then we can also consider "$A\&B$", "$A$ or $B$", as well as "not $A$.". Let us denote by $\mathcal{I}$ the set of all indicators of statements whose indicators belong to $\mathcal{A}$. Since the equation $x^2 = x$ has only solutions $x = 0$ and $x = 1$, it follows that an unknown $X$ is an indicator unknown if and only if $X^2 = X$. We can therefore say that

$$\mathcal{I} = \{X \in \mathcal{A} \mid X^2 = X\}.$$

We can also say that if $X$ is an indicator unknown, then $X = I_A$ for some statement $A$, and no matter how that statement is worded, it is logically equivalent to the statement $X = 1$, which of course, is true if $A$ is true and false if $A$ is false. Since the English language has many ways of expressing the same thing, there are in general many statements which give rise to the same indicator. If we allow statements to be expressed in any language, then even for a fixed statement, it is doubtful that there is a set of all statements expressing exactly that statement unless I restrict attention to some specific set of languages. But just for simplicity of notation, let us assume we have chosen a set of statements, denoted $\mathcal{S}$, which contains the statement $X = 1$ for each $X \in \mathcal{I}$, and so that $I_A \in \mathcal{I}$, for each $A \in \mathcal{S}$. We can then write

$$\mathcal{I} = \{I_A \mid A \in \mathcal{S}\ \} \subset \mathcal{A}.$$

Moreover, we can reasonably now also assume

$$\text{if } A, B \in \mathcal{S}, \text{ then } A\&B,\ A \text{ or } B, \text{ as well as } \text{not}A \text{ also belong to } \mathcal{S}. \tag{5.5}$$

## 6. GUESSING UNKNOWNS AND EXPECTED VALUE

We are now in a position to begin talking about our objective which is to develop a theory of how to best guess the value of an unknown. To begin here, let us assume that for each unknown $X \in \mathcal{A}$ and each statement $B \in \mathcal{S}$, we have found a best guess denoted $E(X|B)$ which we will think of as the best guess for the value of $X$ based on the information in statement $B$. It may be that if we think of reasonable rules that such choices or guesses should obey, then in fact we can see what should be done. This is a typical mathematical technique for dealing with a problem you do not know how to solve. Just begin with the assumption you have already solved the problem and see if that assumption leads you part way to the answer. First, if $B$ implies that $X = c \in \mathbb{R}$, a definite known number, then we must assume that we picked $E(X|B) = c$ in order to preserve basic logical consistency. That is if the statement $B$ that I am assuming tells me that $X = 5$, then certainly I should guess the value 5 for $X$, which means here that $E(X|B) = 5$. Let us go back to the temperature example, and suppose that I think that my guess for the temperature at the WWLTV weather station reported tomorrow will be 15 degrees Celsius. If you ask me to give my guess in degrees Fahrenheit, in these new units my guess should be determined by my original guess simply by using the conversion formula for converting from degrees Celsius to degrees Fahrenheit. Let's put all this in symbols and see what that tells us. First, let $X$ be the unknown which is the temperature tomorrow at the WWLTV weather station expressed in degrees Celsius, and let $Y$ be the temperature at the WWLTV weather station tomorrow expressed in degrees Fahrenheit. Then I know that for sure, $Y = 32+(9/5)X$. Now, if I guess the value of $X$ to be 15, then that says I think the temperature tomorrow at the WWLTV weather station will be 15 degrees Celsius, and consequently, I would naturally simply convert that guess to Fahrenheit to give my guess for $Y$. That is, my guess for $Y$ must be 59 degrees Fahrenheit. In symbols, know matter what my guess is for $X$, my guess for $Y$ should simply be the result of converting my guess for $X$ using the conversion formula, in symbols, this means $E(Y|B) = 32 + (9/5)E(X|B)$. This same reasoning should apply in any situation of guessing where we decide to change the units in which we express our guess. Since saying $Y$ is the unknown $X$ expressed in some new system of units could in general be given by any equation of the form $Y = a + bX$, where $a, b \in \mathbb{R}$ are definite numbers, it follows that guessing logically consistent with all possible changes of units requires that whenever $B$ implies that $Y = a + bX$, then we must have $E(Y|B) = a + bE(X|B)$, no matter what numbers we might choose for $a$ and $b$. Another way of saying this is that

$$E(a + bX|B) = a + bE(X|B), \text{ for any } X \in \mathcal{A}, a, b \in \mathbb{R}. \tag{6.1}$$

Now notice that this last equation must remain true if we replace $X$ by $0 \in \mathbb{R} \subset \mathcal{A}$. When we make this replacement, it says that $E(a|B) = a$, for any $a \in \mathbb{R}$. Of course, we already had assumed this, but if we hadn't, then we would now see it is necessary just to be consistent with all possible changes of units. More generally, let us suppose that we have two unknowns $X$ and $Y$ which are more loosely related. Let us assume that our information $B$ tells us in particular that whatever $X$ is and whatever $Y$ is it must be the case that $X \leq Y$. Then, it certainly seems that we should guess a value for $Y$ that is at least as much as what we guess for $X$. That is in symbols,

$$E(X|B) \leq E(Y|B), \text{ if } B \text{ implies } X \leq Y. \tag{6.2}$$

Next, I want to ask if receiving new information in the form of a new statement $C$ can change our guess. Certainly it should when the new information is more useful and implies our previous guess was off because of our previously more limited information. On the other hand, suppose that we consider new information which may or may not be true. That is, suppose that we are considering new information which itself could be suspect. First, how can we judge this? Well we know the indicator $I_C$ is 1 if $C$ is true and 0 if $C$ is false, so we know $0 \leq I_C \leq 1$, for sure, and therefore we can definitely now say as far as our guess $E(I_C|B)$ is concerned, by [6.2], we must have

$$0 \leq E(I_C|B) \leq 1. \tag{6.3}$$

How can we interpret this number $E(I_C|B)$? Well, if we have to guess a number between 0 and 1 for the value of $I_C$, and if we think $C$ is very very likely to be true, then we should guess a number very very close to 1. If we think that $C$ is likely to be false, then we should guess a number very close to 0. That is, we should think of $E(I_C|B)$ as a measure of how likely we think $C$ is to be true based on the information we already have from $B$. We define this to be the PROBABILITY of $C$ given $B$, denoted by $P(C|B)$. That is we define probability with the equation

$$P(C|B) = E(I_C|B) \text{ for any } C \in \mathcal{S}. \tag{6.4}$$

In view of [6.3], we then have as an immediate property of probability that

$$0 \leq P(C|B) \leq 1, \tag{6.5}$$

and that if $B$ implies $C$, then $P(C|B) = 1$, whereas if $B$ implies not$C$, that is if $B$ implies $C$ to be false, then $P(C|B) = 0$. It is important to realize that $E(X|B)$ and $P(C|B)$ depend on $B$, and that different people will in general have different information on which to base guesses so will in general arrive at different values for their guesses for values of unknowns and for probabilities of statements. That is to say, that different people use different "$B$'s" in $\mathcal{S}$. In fact, different people might be using different "$\mathcal{S}$'s" and even different "$\mathcal{A}$'s". We should also keep in mind that even though we are thinking of $P(C|B)$ as a measure of how likely that $C$ is to be true given the background information $B$, it is really the guess $E(I_C|B)$, so is subject to the logical restrictions we have so far, and consequently, we have to see where the theory leads in order to evaluate it.

Now, let us return to the problem of seeing if we can formulate a way that new information influences our guesses-that is in some sense, we are formulating a general learning process. So, we are trying to guess a value for unknown $X \in \mathcal{A}$ based on information in statement $B \in \mathcal{S}$, and we are lead to consider new information in statement $C \in \mathcal{S}$, which may or may not be true. As an example, suppose that $X$ is the weight in ounces of a paper weight on your friend's desk which you pick up and hold in your hand. Based on your experience with weights and measures, you might be able to make a reasonable guess. Think of this initial information as statement $B$. But, suppose that you spot an unopened can of mixed nuts on the desk and on the label is the statement "NET WT 21 OZ(595g)". Your friend picks up the can of nuts in one hand and grabs the paper weight in the other hand, and now using comparison by feel as well as the knowledge of what the contents of the can weigh says, "I think the paper weight weighs more than the can of nuts, maybe almost twice as much." Now this provides new information which we can take to be statement $C$. Do you trust your friend's judgement completely, and if you accept this, then certainly you should make your guess now based on $B\&C$. But what if your friend is wrong or trying to mislead you? We somehow have to also use $P(C|B)$. Now, if we form the product $XI_C$, then we have a new unknown which has exactly the same value as $X$ if $C$ is true, but simply has the value 0 if $C$ is false, because 0 multiplied by any number is just 0. Moreover, if $C$ is true, we should use $E(X|C\&B)$ as our guess. But what should we do about the possibility that $C$ may be false? One thing we can notice is that since $XI_C$ is 0 if $C$ is false but equal to $X$ if $C$ is true, it follows that when we guess $E(XI_C|B)$, we have to simultaneously take into account what the value of $X$ is when $C$ is also true as well as how likely it is that $C$ is actually true, and this information should as well be contained in the two numbers $E(X|C\&B)$ and $P(C|B)$. More precisely, at least if you give me the numbers $E(X|C\&B)$ and $P(C|B)$, as well as the statements $B$ and $C$, then from *that information alone* I should be able to determine by some procedure what your guess is for $XI_C$ based on $B$, which of course is $E(XI_C|B)$. This procedure should be the same no matter what the particular unknown $X$ with which we are dealing. In mathematical terms, this means that there should be a rule or function $f_{(B,C)} : \mathbb{R} \to \mathbb{R}$, that is a real valued function whose domain is the set of real numbers, and having the following property:

$$E(XI_C|B) = f_{(B,C)}(E(X|C\&B)), \text{ for all } X \in \mathcal{A}. \tag{6.6}$$

Notice this means that if you tell me $E(X|C\&B) = 8$ and $P(C|B) = .3$, then without knowing what $X$ is, just with the information that your guess is 8, I immediately know that $f_{(B,C)}(8)$ is your guess $E(XI_C|B)$. I am allowing that the procedure or rule may depend on knowing what $B$ and $C$ are, which is why the rule is tagged with the subscript $(B, C)$, and consequently, the rule may also depend on $P(C|B)$. This assumption that some such form of rule exists will turn out to be a very powerful assumption, because we will see shortly what it has to be. But, as $\mathbb{R}^{\mathbb{R}}$ is certainly infinite, it is very impressive that this assumption alone will determine what the rule has to be, and we will see that in fact it does not even depend on the particular statements $B$ and $C$, but rather only depends on the number $P(C|B)$. Now let's work out what the rule has to be. First, recall we have $\mathbb{R} \subset \mathcal{A}$, so if [6.6] holds, then it has to be true if $X$ is replaced by any definite number $r \in \mathbb{R}$. What does the left hand side of [6.6] become when $X$ is replaced by $r$? When we make that replacement, since $I_C \in \mathcal{A}$, we can apply [6.1] with $a = 0$ and $b = r$ to find

$$E(rI_C|B) = rE(I_C|B) = rP(C|B). \tag{6.7}$$

Now let us examine what happens to the right hand side of [6.6] when $X$ is replaced by $r$. Notice we can see $E(r|C\&B) = r$, by using [6.1] for the case that $a = r$ and $b = 0$. This means that when we replace $X$ by $r$ on the right hand side we have simply

$$f_{(B,C)}(E(r|C\&B) = f_{(B,C)}(r). \tag{6.8}$$

Finally, since [6.6] says [6.7] and [6.8] are equal, this means that

$$f_{(B,C)}(r) = rP(C|B), \text{ for all } r \in \mathbb{R}, \text{ and for all } B, C \in \mathcal{S}. \tag{6.9}$$

WHAT COULD BE SIMPLER? The rule simply says that whatever number goes into $f_{(B,C)}$ the number coming out is simply that number multiplied by $P(C|B)$. We can now just go back to [6.6] and use [6.9] on the right hand side of the equation, since $E(X|C\&B)$ is just a number in $\mathbb{R}$, we can use $r = E(X|C\&B)$ in [6.9]:

$$E(XI_C|B) = f_{(B,C)}(E(X|C\&B)) = E(X|C\&B)P(C|B). \tag{6.10}$$

The end result here is the MULTIPLICATION RULE:

$$E(XI_C|B) = E(X|C\&B)P(C|B), \text{ for any } X \in \mathcal{A}, \text{ and for any } B, C \in \mathcal{S}. \tag{6.11}$$

It is customary to call $E(X|B)$ the EXPECTED VALUE or MEAN of $X$ given $B$, and as already indicated, $P(C|B)$ is the probability of $C$ given $B$. Also, when we are dealing with a fixed background statement $B$, it is often left out of the notation, so we would write $E(X)$ in place of $E(X|B)$, write $E(X|C)$ in place of $E(X|C\&B)$, and likewise write $P(D|C)$ in place of $P(D|C\&B)$. When we replace $X$ in [6.11] by the indicator of statement $D \in \mathcal{S}$, we find that $E(I_D I_C|B) = E(I_D|C\&B)P(C|D)$, and then using [5.1] together with the definition of probability, [6.4], we obtain the LAW OF CONDITIONAL PROBABILITY:

$$P(D\&C|B) = P(D|C\&B)P(C|B), \tag{6.12}$$

which of course will usually just be written as

$$P(D\&C) = P(D|C)P(C),$$

when the background $B$ is understood. This law has a lot of intuitive content. For instance if on a certain tropical island we know that it rains 70% of the time, if we know that when it is raining there is a 40% chance that the sun is shining (so we have a sun shower), then translating into probability, we would say that $P$(sunshine|rain)=40%=.4 and $P$(rain)=70%=.7, so by the law of conditional probability, we have $P$(sun shower)=(.4)(.7)=.28. That is, it is raining 70% of the time and 40% of that 70% of the time it's raining in fact the sun is also shining so we have a sun shower, and 40% of 70% is 28%.

Another useful and obvious fact about probability which is an immediate consequence of [5.4], the definition of probability, [6.4], and [6.1] is the LAW OF NEGATION IN PROBABILITY:

$$P(\text{not } A|B) = 1 - P(A|B), \text{ for any } A, B \in \mathcal{S}. \tag{6.13}$$

Indeed, we have

$$P(\text{not } A|B) = E(I_{\text{not } A}|B) = E(1 - I_A|B) = 1 - E(I_A|B) = 1 - P(A|B).$$

You might question pulling out the minus sign in the middle step, but that would just be applying [6.1] to the case $a = 1$ and $b = -1$, as then we have $E(1 - I_A|B) = E(1 + (-1)I_A|B) = 1 + (-1)E(A|B) = 1 - E(A|B)$. Of course the law of negation in probability is very reasonable. If I think that there is a 70% chance that my favorite team will win the Super Bowl, then that means there is a 30% chance they won't win.

Using the law of negation in probability together with the law of conditional probability it is possible to prove the GENERAL LAW OF PROBABILITY:

$$P(A \text{ or } C|B) = P(A|B) + P(C|B) - P(A\&C|B). \text{ for any } A, B, C \in \mathcal{S}. \tag{6.14}$$

I will leave that as a challenge for the reader, but it basically involves DeMorgan's Law of Logic which says that [not $(A\&B)$] is logically the same as [(not $A$) or (not $B$)], from which it immediately follows that [not ($A$ or $B$)] is logically the same as [(not $A$) & (not $B$)]. However, if we accept [6.14] to be true, then when we translate back into expectation notation, it says, in view of [5.2], that

$$E(I_A + I_C - I_{A\&C}|B) = E(I_A|B) + E(I_C|B) - E(I_{A\&C}|B) \tag{6.15}$$

Notice that [6.1] and [6.15] above are both special cases of a more general rule called the LINEAR LAW OF EXPECTATION which says

$$E(aY + bY|B) = aE(Y|B) + bE(X|B), \text{ for any } X, Y \in \mathcal{A},\ B \in \mathcal{S},\ a, b \in \mathbb{R}. \tag{6.16}$$

For, notice that [6.1] is the case where $Y = 1$ in [6.16] and [6.15] is found in two steps with [6.16], the first case where $a = 1$, $X = I_A + I_C$, $b = -1$, and $Y = I_{A\&C}$, followed by the second case applied to the first term using $X = I_A$ and $Y = I_C$. This should lead us to suspect that [6.16] is actually true. At least, [6.1] and [6.14] should lead us to suspect that there is some way to compute the number $E(X + Y|B)$ when we are given $E(X|B)$, $B$, and $Y$. That is, suppose that there is some rule which may depend on both $B$ and $Y$ which allows us to arrive at $E(X + Y|B)$ as soon as we know the number $E(X|B)$. Here, we are saying that maybe, given $B \in \mathcal{S}$ and $Y \in \mathcal{A}$, there is some function $f_{(B,Y)} : \mathbb{R} \to \mathbb{R}$ with the property

$$E(X+Y|B) = f_{(B,Y)}(E(X|B)), \text{ for any } X \in \mathcal{A}. \tag{6.17}$$

Since the function or rule $f_{(B,Y)}$ depends on both $B$ and $Y$, we know that it may also depend on $E(Y|B)$. Keep in mind that there are an infinite number of functions in $\mathbb{R}^{\mathbb{R}}$, so the possibility that there is one that works does not seem so impossible even though actually finding it may, and we are even allowing for the possibility that the rule itself can depend on both $B$ and $Y$. But, here again, since the rule [6.17] must be true for any $X \in \mathcal{A}$, it must also be true for any definite number $r \in \mathbb{R}$. If we replace $X \in \mathcal{A}$ by $r \in \mathbb{R}$, in the right side of [6.17], we have

$$f_{(B,Y)}(E(r|B)) = f_{(B,Y)}(r) \tag{6.18}$$

because $E(r|B) = r$. When we make this replacement on the left side of [6.17] we find by [6.1]

$$E(r+Y|B) = r + E(Y|B). \tag{6.19}$$

Now, when we combine the simplifications by equating the right side of [6.18] to the right side of [6.19], we find that

$$f_{(B,Y)}(r) = r + E(Y|B), \text{ for any } r \in \mathbb{R}. \tag{6.20}$$

But then, we know that $E(X|B)$ is itself simply a number, and we can use $r = E(X|B)$ in [6.20] to find

$$f_{(B,Y)}(E(X|B)) = E(X|B) + E(Y|B). \tag{6.21}$$

Seeing now that the right side of [6.17] and the left side of [6.21] are identical, we can equate the left side of [6.17] to the right side of [6.21] which then gives the ADDITIVE LAW OF EXPECTATION:

$$E(X+Y|B) = E(X|B) + E(Y|B), \text{ for all } X, Y \in \mathcal{A}. \tag{6.22}$$

The GENERAL LINEAR LAW OF EXPECTATION, [6.16], then follows easily from combining [6.22] and [6.1].

At this point, it is useful to notice that we have in particular demonstrated the following five fundamental properties of expectation and the definition of probability:

DEFINITION OF PROBABILITY

$$P(A|B) = E(I_A|B), \text{ for any } A, B \in \mathcal{S}; \tag{6.23}$$

MULTIPLICATION LAW

$$E(XI_C|B) = E(X|C\&B)P(C|B), \text{ for all } X \in \mathcal{A},\ B, C \in \mathcal{S}; \tag{6.24}$$

ADDITIVE LAW

$$E(X+Y|B) = E(X|B) + E(Y|B), \text{ for all} X, Y \in \mathcal{A},\ B \in \mathcal{S}; \tag{6.25}$$

HOMOGENEITY LAW

$$E(rX|B) = rE(X|B), \text{ for all } r \in \mathbb{R},\ X \in \mathcal{A},\ B \in \mathcal{S}; \tag{6.26}$$

POSITIVITY LAW

$$E(X|B) \geq 0, \text{ if } B \text{ implies } X \geq 0,\ X \in \mathcal{A}; \tag{6.27}$$

NORMALIZATION LAW

$$E(1|B) = 1, \text{ for any } B \in \mathcal{S}. \tag{6.28}$$

We can notice that these laws are consequences of the rules we have deduced so far, and at the same time, it is easy to see that from these laws, everything we deduced so far is a consequence. Thus, we can simply take these laws as the starting point for further development. For instance, from [6] combined with [6.26] we see that [6.16] must hold, as

$$E(aX + bY) = E(aX) + E(bY) = aE(X) + bE(Y), \text{ for all } X, Y \in \mathcal{A}.$$

Moreover, applying [6.26] together with [6.28] we have

$$E(r) = E(r1) = rE(1) = r1 = r, \text{ for all } r \in \mathbb{R}.$$

From this and [6] we now see that [6.1] must hold. If we set $a = 1$ and $b = -1$ in [6.16] we see that in fact $E(X - Y) = E(X) - E(Y)$, for any $X, Y \in \mathcal{A}$. From this we see in particular, that if $B \in \mathcal{S}$, and if $B$ implies that $X \geq Y$, then that means $X - Y \geq 0$ so by [6.27] we conclude that $E(X) - E(Y) = E(X - Y) \geq 0$, and this means $E(X) \geq E(Y)$, which is to say that [6.2] must hold. It is also useful to keep in mind that in view of these considerations, we could also say that [6.16] is the same as saying

$$E(aX \pm bY) = aE(X) \pm bE(Y), \text{ for all } a, b \in \mathbb{R},\ X, Y \in \mathcal{A}. \tag{6.29}$$

So far, we have found some general laws for probability and expectation which we deduced by thinking in terms of guessing with logical consistency. These should be thought of as placing restrictions on what we can guess for given unknowns based on our information. We cannot just go wild and guess anything. Now let us see in some simple examples how these laws enable us to figure out exactly what our guess should be.

For our first example, let us consider a box which has a single dice inside which we cannot see, because the box has a lid which is closed. We know it is a standard dice so it is a small cube with six faces, and each face has a number of spots on it. Let us stipulate that the dice is sitting on the floor of the box and there is definitely one of the six faces which is the top face, but we cannot see which one. You now have no information which would allow you to say any one of the six faces is more likely to be on top than any of the other faces. We can let $B$ be the statement of all this background information which is the setup for our situation. Let $X$ be the number of spots on the top face. We do not know what $X$ is, yet we do know that it belongs to the set $\{1, 2, 3, 4, 5, 6\}$. Let $A_1$ be the statement that 1 is the number of spots on the top face, or equivalently, $A_1$ is the statement that $X = 1$. Let $A_2$ be the statement that there are two spots on the top face or equivalently it is the statement that $X = 2$. We can obviously continue in this fashion, so $A_3$ is the statement that $X = 3$, and $A_4$ is the statement that $X = 4$, and $A_5$ is the statement that $X = 5$, and finally, $A_6$ is the statement that $X = 6$. Notice that of the statements $A_1, A_2, A_3, A_4, A_5, A_6$, we know exactly one must be true and all the others false. That is we know that

$$I_{A_1} + I_{A_2} + I_{A_3} + I_{A_4} + I_{A_5} + I_{A_6} = 1. \tag{6.30}$$

If we apply our additive law of expectation with the definition of probability here, we have

$$P(A_1) + P(A_2) + P(A_3) + P(A_4) + P(A_5) + P(A_6) = 1. \tag{6.31}$$

On the other hand we know from [6.5] which is a consequence of [6.2], that $P(A_k)$ must be between 0 and 1 for each $k$ with $1 \leq k \leq 6$. Moreover, we already observed that our background information $B$ does not allow us to choose any of these 6 statements as more likely to be true than the others. The inescapable conclusion here is that all these probabilities must be the same and that means they are all equal to $1/6$. Notice this result is merely a consequence of the state of our information. If I claim that $A_4$ is more likely than $A_3$, I would have to justify that on the basis of our background information, and there is nothing in our background information to justify such a conclusion. Now, what about our guess $E(X)$ for the value of $X$ in this situation? We can multiply both sides of [6.30] by $X$ with the result that

$$X = XI_{A_1} + XI_{A_2} + XI_{A_3} + XI_{A_4} + XI_{A_5} + XI_{A_6} = X. \tag{6.32}$$

Again applying the additive law, we now have

$$E(X) = E(XI_{A_1}) + E(XI_{A_2}) + E(XI_{A_3}) + E(XI_{A_4}) + E(XI_{A_5}) + E(XI_{A_6}). \tag{6.33}$$

Let's examine a typical term from the right hand side of [6.33] and use our multiplication law, for instance, let us examine $E(XI_{A_3})$. When we apply the multiplication law here, we find that

$$E(XI_{A_3}) = E(X|A_3)P(A_3). \tag{6.34}$$

But, when we compute $E(X|A_3)$, we are assuming $A_3$ is true, that is we assume that the face on top has exactly 3 spots on it and therefore we must guess the value 3, which is to say $E(X|A_3) = 3$. Likewise, we see that $E(X|A_5) = 5$, and $E(X|A_2) = 2$, and so on. That is, we see that $E(X|A_k) = k$, for $1 \leq k \leq 6$. On the other hand, $P(A_k) = 1/6$, for $1 \leq k \leq 6$. Thus, for any $k$ with $1 \leq k \leq 6$, we must have

$$E(XI_{A_k}) = E(X|A_k)P(A_k) = k/6. \tag{6.35}$$

This means going back to [6.33] that

(6.36)
$$E(X) = (1/6)+(2/6)+(3/6)+(4/6)+(5/6)+(6/6) = (1+2+3+4+5+6)/6 = 21/6 = 7/2 = 3.5$$

This is sort of an amazing result!! We have found what is the required guess just from applying the laws which restrict what we can guess based on logical consistency alone. Moreover, what we must guess in NOT EVEN A POSSIBLE VALUE for the unknown $X$. What is going on here? Notice that our guessing includes here the guess as to how likely each of the possible values for $X$ is-they must all be equally likely, so our guess for the number of spots has selected a number which is in a very precise sense simultaneously closest to all the possible values. We will see that if any other choice is made for our guess of the value of $X$, then we are more likely going to be FARTHER wrong than we are by guessing 3.5. What I mean by farther wrong is that if we guess 1 and the value of $X$ happens to be 6, then we are WAY OFF the mark, but if we guess 1 and the value happens to be 2, then we are not so far off. In case you guess 2 and the true value is 5, your error is 2-5=-3. Suppose that we decide to give a penalty for being wrong by squaring the error. If you guess 1 and the value is 6, your penalty is $(6-1)^2 = 25$, whereas if you guess 1 and the value is 2, your penalty is only $(2-1)^2 = 1$. You can see that with this kind of penalty in operation, you do not want to take a chance on being far from the true value with your guess. In fact, we can use our theory to see what the expected squared error is, when we guess the value $v$ for the true value of $X$. For now, the error is $(v - X)$, so its square is $(v - X)^2$, and if we replace $X$ by $(v - X)^2$ in the preceding calculation, then we find

$$E((v - X)^2) = [(v - 1)^2 + (v - 2)^2 + (v - 3)^2 + ... + (v - 6)^2]/6, \tag{6.37}$$

and if you try computing this for several values of $v$, you will begin to see that the smallest value happens for $v = 3.5$. In fact, we can use a little algebra and notice that in general here, $(v - k)^2 = v^2 + k^2 - 2vk$, and when we replace the terms with this expanded simplified expression for the square of $v-k$ we will arrive at $v^2+E(X^2)-2vE(X) = v^2-7v+E(X^2)$. If we replace the possible values 1,2,3,4,5,6 with their squares 1,4,9,16,25,36 and redo the calculation of expectation, we find that $E(X^2) = 91/6$. Thus, $v^2 - 7v + E(X^2) = v^2 - 2v(3.5) + (3.5)^2 - (3.5)^2 + 91/6 = (v - 3.5)^2 + (91/6) - (7/2)^2$. Since the smallest the square term can be is zero, we conclude that the smallest we can get is by taking $v = 3.5$ That is, we have

$$E((v - X)^2) \geq (91/6) - (3.5)^2, \text{ for all } v \in \mathbb{R}, \tag{6.38}$$

with equality in[6.38] only when we use $v = 3.5$. That is, the guess $E(X) = 3.5$ represents the guess which minimizes our guess as to the penalty we must pay for being wrong. Remember, the only thing we can go on here is our guess as to what penalty we are dealing with. Let us look at this in more generality. Suppose that $X \in \mathcal{A}$ and $B \in \mathcal{S}$ is our background information. Suppose we guess that $X$ has value $v \in \mathbb{R}$. Then $(X-v)^2$ is the penalty we have to pay and since it is a square of an unknown number in $\mathcal{A}$, we know $E((X-v)^2) \geq 0$ by the positivity law. For simplicity, set $\mu = E(X)$, so $\mu$ is our actual guess dictated by our theory. Notice that $X - \mu$ is the difference between the actual value of $X$ and our ideal guess, and

$$E(X-\mu) = E(X) - \mu = \mu - \mu = 0. \tag{6.39}$$

This means that our best guess for our error when we use the optimal guess is ZERO. However, the same is not true for the squared error. If we guess just any value $v$ for $X$, then we can calculate with algebra

$$(X-v)^2 = [(X-\mu) + (\mu - v)]^2 = (X-\mu)^2 + (\mu-v)^2 + 2(\mu-v)(X-\mu) \tag{6.40}$$

and when we apply the expectation to this equation we see

$$E((X-v)^2) = E((X-\mu)^2) + (\mu-v)^2 + 2(\mu-v)E(X-\mu) = E((X-\mu)^2) + (\mu-v)^2, \tag{6.41}$$

because of [6.39]. Now, we know both terms on the final right hand side of [6.41] are non negative because they are squares and the positivity law applies to the first of those terms. Moreover, the first term, $E((X-\mu)^2)$ is based on the optimal guess according to our theory, so we have no choice here-it is fixed by our theory. The second and last term, $(\mu - v)^2$ will depend on $v$, our supposed alternate guess, and as this term is squared, the smallest it can be is zero which can only happen by choosing $v = \mu = E(X)$. What we have to conclude here is that $E(X)$ is optimal in the sense that as far as we can tell by our guessing method, it seems to lead to the smallest penalty in the sense of the squared error from the true value, and that for this smallest squared error we must guess the value $E((X-\mu)^2)$. As far as this guess of the squared error when we make our optimal guess, $E((X-\mu)^2)$ is concerned, if $E((X-\mu)^2) = 0$, then it turns out be the case that in fact $X = \mu$ has probability 1. That is, if we have a situation where there is more than one possibility for the value of $X$, as in the case of the dice, for instance, then $E((X-\mu)^2) > 0$, meaning our guess for the squared error even when making the optimal guess is a positive number.

Well, now we have seen that essentially when there are a finite number of possible statements where exactly one has to be true and our background information gives no way to select some as being more likely than others, then we should begin by assuming that all of these statements are equally likely to be true, so they all have the same probability based purely on our initial background information. In particular, if $X$ is any unknown in $\mathcal{A}$ and if $C_1, C_2, C_3, ..., C_k, ..., C_n$ are statements in $\mathcal{S}$ of which exactly one is known to be true, then we know that $0 \leq P(C_k) \leq 1$, for $1 \leq k \leq n$, and

$$1 = I_{C_1} + I_{C_2} + I_{C_3} + ... + I_{C_k} + ... + I_{C_n}, \tag{6.42}$$

so applying expectation to both sides of [6.42] gives

$$1 = P(C_1) + P(C_2) + P(C_3) + ... + P(C_k) + ... + P(C_n). \tag{6.43}$$

And then multiplying both sides of [6.42] by $X$ gives

$$X = X1 = XI_{C_1} + XI_{C_2} + XI_{C_3} + ... + XI_{C_k} + ... + XI_{C_n}, \tag{6.44}$$

so again applying expectation we find

$$E(X) = E(XI_{C_1}) + E(XI_{C_2}) + E(XI_{C_3}) + ... + E(XI_{C_k}) + ... + E(XI_{C_n}), \tag{6.45}$$

and applying the multiplication law to each of these terms in [6.45] gives

$$(6.46)\qquad \begin{aligned} &E(X) = E(X|C_1)P(C_1) + E(X|C_2)P(C_2) + E(X|C_3)P(C_3) + ... \\ &+ E(X|C_k)P(C_k) + ...E(X|C_n)P(C_n). \end{aligned}$$

In general then we see that [6.46] allows us to reduce the computation of expectation to the computation of probabilities of the various statements and computations of conditional expectations under the assumption of each of the statements separately. For instance, if $X$ can only have a finite set of values, say $V = \{v_1, v_2, v_3, ..., v_n\}$ is the set of possible values for $X$, then we can take $C_k$ to be the statement that $X = v_k$, for $0 \leq k \leq n$, and [6.46] simplifies to

$$(6.47)\qquad E(X) = v_1P(C_1) + v_2P(C_2) + v_3P(C_3) + ... + v_kP(C_k) + ...v_nP(C_n),$$

and in this situation, we see that the computation of $E(X)$ completely reduces to the computation of the various probabilities of the various alternative values which $X$ can have. Again, the general consideration above on the error in guessing shows that this guess, which is forced on us by logical consistency and its consequences, also has the property of minimizing the squared error of our guess-that is as far as our guessing can tell, our guess for the squared error is also minimized by the guess given by [6.47].

Another useful special case of [6.46] is the case where we take $X$ to be an indicator of a statement, $X = I_A$ with $A \in \mathcal{A}$. Then, as $I_AI_{C_k} = I_{A\&C_k}$, for $0 \leq k \leq n$, we find

$$(6.48)\qquad \begin{aligned} &P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3) + ... \\ &+ P(A|C_k)P(C_k) + ... + P(A|C_n)P(C_n). \end{aligned}$$

Thus, using [6.48] we can break up the computation of probability itself into many separate cases.

Now as our second example, let us consider the case of a box containing 10 blocks of which 4 are red. We are going to draw one block after another until the box is empty, without any replacement of blocks after each draw. Let $R_k$ be the statement that the block drawn on the $k-$th draw is red. Now, to make the state of our information completely clear here, we should assume that this has been already done by an assistant who will tell us truthfully about any question we ask concerning the results of this drawing. Notice, again, we have no way of knowing that any block is any more likely than another to be drawn on any given draw. If we consider that on the first draw there are 10 blocks in the box, then based on our information, we should say $P(R_1) = 4/10$, and then if we are given the result of the first draw was a red block, then the probability of drawing a red block on the second draw is $P(R_2|R_1) = 3/9$. On the other hand, if we ask for $P(R_2)$, then we should realize that all 10 blocks are equally likely to be drawn on the second draw, so $P(R_2) = 4/10$. This may seem confusing at first, but we should keep in mind that all ways of performing this task for which the information is the same at each stage should result in the same value for the probability, since our theory is purely about the state of our information. Now, when we play cards, we do not deal cards by putting them in a box, shaking it up and allowing the players to draw from the box without looking. What we do is shuffle the cards and deal one after another from the top of the deck. The purpose of the shuffle is to keep any player from knowing the position of any specific card in the pile. Notice that once the deck is shuffled, the result of dealing all the cards has been determined, but the players do not know what the result will be. It is the same with the blocks in the box. Instead of drawing the blocks one after another without replacement, just assume that the blocks are stacked and we cannot see how they are stacked but our trusted assistant can see how they are stacked. In that case, if we are asked what the probability is that the second block from the top of the stack is red, then clearly as it is as likely to be any of the blocks, the probability is $P(R_2) = 4/10$. This way of viewing the block drawing problem makes it clear that time is not an element of the problem but rather the state of our information is all that matters. For instance, if we want to compute $P(R_1|R_2)$, then viewing this in terms of time is often confusing, since it appears impossible to consider this even makes sense. But viewed in terms of the stacked blocks, if we know that the second block in the stack is red, then as far as the first block is concerned, it is clear it can be any of the other 9 blocks of which only 3 can be red, so obviously $P(R_1|R_2) = 3/9$. Now if we decide to use [6.48] to compute $P(R_2)$, by breaking it up into cases where $R_1$ is true and where it is false, then it is easy to see the result must be $(3/9)(4/10) + (4/9)(6/10) = (3/9)(4/10) + (6/9)(4/10) = 4/10$. Thus, as should be the case, our rules are consistent and either way the result is the same.

As our final example, let us consider the situation of a basketball player making free throws. Suppose we assume that if he has shot $k$ free throws and made the basket on $r$ of them that when he goes to shoot the next shot his probability of making the basket is $r/k$. Suppose we observed that he misses the first shot and makes the second shot. Let $A$ be this information. Let $M_k$ be the event that he makes the basket on the $k-$th shot. We want to know what $P(M_k|A)$ is, for any $k \geq 2$. Notice that as far as we know, his probability of making a shot is 50%, so we should say $P(M_k|A) = 1/2$, no matter what $k \geq 2$ is. If we carry out the detailed computation using [6.48], then this is what we find. In fact, if we watch the player shoot $n$ shots and actually make the basket on $m$ of the shots we watched, call this statement of information $C$, then the detailed calculation shows that for any shot we did not watch we should calculate $P(R_k|C) = m/n$. This is a very strong indication that the laws of probability are dictating that it is purely about the state of our information and not about so called random events. In reality, the laws of physics are deterministic, so there is no such thing as a random event, there are rather only events which appear to be random to us based on our limited information.

The development of the theory of expectation, covariance, probability, and statistics can now be carried out by simply applying the laws we have deduced above (for instance, see [3] for a treatment using only elementary algebra).

## References

[1] R. T. Cox, Probability, frequency, and reasonable expectation, *Am. J. Phys.*, **14**(1946), 1–13.

[2] M. J. Dupré, Unknowns and Guessing The Laws of Expectation and Probability, on my website.

[3] M. J. Dupré, The Expectation Primer Expectation Covariance Probability, on my website.

[4] M. J. Dupré and F. J. Tipler, Cox's theorem, unknowns and plausible value, preprint posted on LANL arXiv.

[5] E. T. Jaynes, *Probability Theory-The Logic of Science*, Cambridge University Press, Cambridge, U.K., 2003.

TULANE UNIVERSTIY
*E-mail address*: mdupre@tulane.edu