LECTURES FOR BUSINESS STATISTICS

Maurice J. Dupré
Department of Mathematics
New Orleans, LA 70118
email: mdupre@tulane.edu
27 April 2009

1. MONDAY 13 JANUARY 2014

Business Statistics is based on the theory of expectation and probability which is based on mathematics. We therefore begin with a review of some basic ideas at the foundation of mathematics concerning sets and functions. We take the notion of a **set** as an undefined term, even though informally, we think of sets as collections of objects. As such, a set is completely determined by the objects it contains, which is to say if two sets have exactly the same objects, then they are equal as sets. We can specify a set with only a few objects by simply listing them, and the standard mathematical notation here is to enclose the list inside a pair of curly braces: $\{\}$, so for instance we might write

$$A = \{3, \mathrm{Joe}, b, 1, \pi\}$$

to specify the set A which has the five members consiting of the numbers $1, 3, \pi$, the letter b and Joe.

Many of the sets we begin with are sets of numbers. For instance we have the set of all **natural numbers** which we denote by $\mathbb{N}$, so

$$\mathbb{N} = \{1, 2, 3, \cdots\},$$

which of course requires an infinite list. The natural numbers are just the numbers we use for ordinary counting. Obviously $\mathbb{N}$ is an infinite set. If we want to have zero and all negative numbers as well, then we are talking about integers. The standard mathematical notation for the set of all **integers** is $\mathbb{Z}$, so

$$\mathbb{Z} = \{\cdots - 3, -2, -1, 0, 1, 2, 3, \cdots\}$$

which is specified by sort of a doubly infinite list. Unfortunately, there are some infinite sets which are "so infinite" that it is not convenient or even impossible to specify their members with any sort of list. To specify these sets, we use what is called the **set builder** notation, which means we consider the set of all x for which some statement about x is true. For instance, we can specify the set of all **rational numbers** $\mathbb{Q}$ as

$$\mathbb{Q} = \{x \mid x = \frac{p}{q}, \text{ where } p \text{ and } q \text{ are integers and } q \neq 0\}.$$

It is possible to list all the rational numbers in an infinite list, but it is not really convenient. On the other hand, the set of all real numbers cannot be specified by any infinite list. We specify the set $\mathbb{R}$ of all **real numbers** as

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

Here we need to be cautious, as trying to form sets by using any statements about x we like can lead to contradictions, and we will return to this problem shortly.

As a matter of convenient terminolgy and notation, we say that A is a **subset** of B provided that every member of A is a member of B, and we denote this by writing $A \subset B$. Therefore, as sets are completely determined by their members, we have

$$A = B \text{ if and only if both } A \subset B \text{ and } B \subset A.$$

For instance, obviously

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

In addition to the concept of a set we need the more useful concept of a function which allows us to get around from one set to another. A **function** is a **rule** which assigns a member of a set to each member of a set. For instance if D and C are sets and fun is a rule which assigns a member of C to each member of D, then we say that fun is a function from D to C, and we denote this by writing

$$fun : D \longrightarrow C.$$

As a matter of terminolgy, if $fun : D \longrightarrow C$, then we call D the **domain** of fun and we call C the **codomain** of fun. For the member x in D, we denote the member of C which fun assigns to x by the symbol

$$fun(x), \text{ read "fun of ex"}$$

which should not be confused with multiplication, as nothing here may have anything to do with numbers. If D and C are finite sets, then we can specify a function from D to C by simply listing the members of D in the set notation and with a little space allowed list the members of C below, again in the set notation, and simply draw an arrow from each member of D down to the member of C which you would like the rule to specify assignment. The only restriction is that each member of D must have an arrow leading to some member of C. Thus, some members of C may have no arrows leading to them and some may have multiple arrows leading to them. It is easy to see that if D has say 3 members, for instance, if $D = \{5, 2, 3\}$, and if C has 4 members, then to form a function from D to C we have 4 choices for an assignment for 5, we have 4 choices for an assignment for 2, and we have 4 choices for an assignment for 3, so altogether we have 4^3 possible functions from D to C. Because of this simple exponential rule for counting functions, for any sets C, D we denote by C^D the set of all functions with domain D and codomain C. Thus,

$$C^D = \{f \mid f : D \longrightarrow C\}.$$

If we let $n(S)$ denote the number of members in the finite set S, then we have in case C and D are both finite, that C^D is also finite and

$$n(C^D) = [n(C)]^{n(C)}.$$

It is useful to think of a function as a sort of general abstract input-output device. If $f : A \longrightarrow B$, then for x in A we have $f(x)$ is the output in C produced by the function f when the input is x. This leads naturally to the concept of composition of functions, that is using the output of one function as the input of another. If $f : A \longrightarrow B$ and $g : B \longrightarrow C$ then their **composite**, denoted $g \circ f$ is the function from A to C given by the rule

$$[g \circ f](x) = g(f(x)), \text{ for any } x \text{ in } A,$$

so

$$g \circ f : A \longrightarrow C, \text{ if } f : A \longrightarrow B \text{ and } g : B \longrightarrow C.$$

If $f : A \longrightarrow B$, if $g : B \longrightarrow C$, and if $h : C \longrightarrow D$ are all functions, then we can form two functions from A to D, namely $[h \circ g] \circ f$ and $h \circ [g \circ h]$. but in fact, for any x in A we have

$$[[h \circ g] \circ f](x) = h(g(f(x))) = [h \circ [g \circ h]](x)$$

so both these functions have the same rule, the same domain, and the same codomain, and are therefore the same function, that is composition obeys the **associative law:**

$$[h \circ g] \circ f = h \circ [g \circ h].$$

Our set builder notation is really a way of using what are called statement functions to form sets. A statement like "x is a real number" is called a statement function in logic, as it is sort of a function whose outputs are statements which can be either true or false depending on what the input x actually is. Thus, if $P(x)$ is any statement function, then the set

$$S = \{x \mid P(x)\}$$

is the set of all x for which the statement $P(x)$ is true. For instance, if $P(x)$ is the statement the statement that x is a rational number, then

$$\mathbb{Q} = \{x \mid P(x)\}.$$

Unfortunately, when we allow unrestricted use of statement functions to try and define sets we can run into trouble. To see what can go wrong, if I try to specify the finite set A as

$$A = \{7, 4, A\}$$

you should object that I cannot define A using A as part of the definition, as that is a circular definition. In particular, it seems that we should not allow a set to be a member of itself. Of course it is certainly possible to form sets whose members are sets, and in fact when mathematics is founded on set theory, everything is a set, so the members of sets are in fact sets. The problem is the circularity which arises from having a set be a member of itself. To make the point perfectly clear, suppose that we allow the possiblilty that a set can be a member of itself and unrestricted use of statement functions. Certainly some sets are not members of themselves, so following Bertrand Russell and in his honor, we denote by

$$\mathcal{R} = \{A \mid A \text{ is a set which is not a member of itself }\}.$$

Now let us ask whether $\mathcal{R}$ is a member of itself. If it is, then the statement defining $\mathcal{R}$ must be true of $\mathcal{R}$ which means that $\mathcal{R}$ is not a member of itself, whereas if $\mathcal{R}$ is not a member of itself, then the statement defining $\mathcal{R}$ is false for $\mathcal{R}$ itself which means that $\mathcal{R}$ is a member of itself. We have arrived at a contradiction known as **Russell's Paradox.** The way out of Russell's Paradox is an involved mathematical theory called **axiomatic set theory**. In simple terms however, the solution is to never use the set builder process except to form subsets of sets. Thus, if we know we have a set S, and if $P(x)$ is any statement function, then the subset $A \subset S$ of all members of S for which $P(x)$ is true is always a set, so we write

$$A = \{x \text{ in } S \mid P(x)\}$$

and thereby avoid Russell's Paradox. To see why this is true, remember that everything is a set in the axiomatic theory of sets, and in addition, no set is a member of itself, so if $\mathbb{R}_S$ denotes the subset of all member of S which are not members of themselves, then simply $\mathbb{R}_S = S$. No paradox possible. Of course, in the axiomatic set theory one has to assume axiomatically that there is a set to start with. From that we can then proceed to form all the sets we need using various axioms of set formation. For instance, if S is a set whose members are sets, then we can form the **union** of all the members of S which se denote by $\bigcup S$. For instance,

$$\bigcup\{A, B\} = A \cup B$$

is the union of the sets A and B. In axiomatic set theory, one has to assume as an axiom, that if S is a set of sets, then there is a set B with the property that $A \subset B$ for each A in S. Then

$$\bigcup S = \{x \text{ in } B \mid x \text{ in } A \text{ for some } A \text{ in } S\}.$$

To see if you understand what is going on here, for each real number r, consider the set D_r defined by

$$D_r = \{x \text{ in } \mathbb{R} \mid x \geq r\},$$

and let S be the set of sets defined by

$$S = \{D_r \mid r > 0\}.$$

What is a simple description of $\bigcup S$?

2. WEDNESDAY 15 JANUARY 2014

In the previous lecture we defined the union of any set of sets. If $\mathcal{D}$ is any set of sets, meaning that every member of $\mathcal{D}$ is a set, then we defined the **union** of $\mathcal{D}$, denoted $\bigcup \mathcal{D}$, as the set

$$\bigcup \mathcal{D} = \{x \mid x \text{ in } D, \text{ for some set } D \text{ in } \mathcal{D}\}.$$

On the other hand, we can define the **intersection** of $\mathcal{D}$, denoted $\bigcap \mathcal{D}$, as the set

$$\bigcap \mathcal{D} = \{x \mid x \text{ in } D, \text{ for every set } D \text{ in } \mathcal{D}\}.$$

For instance if A and B are any sets, then their union, denoted $A \cup B$, is

$$A \cup B = \bigcup \{A, B\},$$

and their intersection, denoted $A \cap B$, is

$$A \cap B = \bigcap \{A, B\}.$$

Likewise, if A, B, and C are sets, then

$$A \cup B \cup C = \bigcup \{A, B, C\},$$

and

$$A \cap B \cap C = \bigcap \{A, B, C\}.$$

The answer to the question at the end of the previous lecture is simply $\bigcup S$ must be the set of all positive real numbers. If P is the set of all positive real numbers, then

$$\bigcup S = P = \{x \text{ in } \mathbb{R} \mid x > 0\}.$$

To see this, we first observe that for each real number r we have r in fact is a member of D_r. Since S is the set

$$S = \{D_r \mid r > 0\},$$

it follows that every positive real number belongs to $\bigcup S$, which is to say that

$$P \subset \bigcup S.$$

On the other hand, if r is in P, then $D_r \subset P$, and therefore

$$\bigcup S \subset P.$$

As both

$$P \subset \bigcup S \text{ and } \bigcup S \subset P,$$

it follows that

$$\bigcup S = P.$$

Recall that in the previous lecture we defined the set of all functions from A to B as

$$B^A = \{f \mid f : A \longrightarrow B\},$$

because in case A and B are finite sets, if A has exactly a members and if B has exactly b members, then B^A has b^a members. We can use this fact to calclate the number of subsets of a finite set.

Before doing this calculation of the number of subsets of a finite set, we first note that we assume there is a set, that is, some set exists, and call it A. Then, we define the **empty set**, denoted $\emptyset$, by

$$\emptyset = \{x \text{ in } A \mid x \neq x\}.$$

Thus, $\emptyset \subset A$, but $\emptyset$ has no members, which is of course the reason it is called the empty set. If B is any set, then $\emptyset \subset B$, since the only way this could possibly not be true is for the empty set to have a member which is not a member of B, but this cannot possibly be the case, since the empty set has no members. Now, if A is any set, then $\{A\}$ is a new set and is not to be confused with A. In particular, $\{\emptyset\}$ is a non-empty set so is not the empty set. We can use this process to construct the natural numbers out of sets. We define the number zero to be the empty set, so

$$0 = \emptyset$$

and then we define the number one to be the set whose only member is the empty set, so

$$1 = \{\emptyset\} = \{0\} = \emptyset \cup \{0\} = 0 \cup \{0\}.$$

Next, we define the number two as $1 \cup \{1\}$, so

$$2 = 1 \cup \{1\} = \{0, 1\}.$$

Likewise, we define the number three to be $2 \cup \{2\}$, and so on, so having defined the natural number n, the next natural number, $n + 1$ is simply defined as $n \cup \{n\}$.

$$n + 1 = n \cup \{n\}.$$

Notice that each natural number is a specific finite set and the natural number 0 has zero members, the natural number 1 has exactly one member, the natural number 2 has exactly two members and so on, so the natural number n has exactly n members.

Suppose that $f : A \longrightarrow B$ is any function and that $C \subset A$. We define the **image** of C under f, denoted $f(C)$, to be the subset of B given by

$$f(C) = \{y \text{ in } B \mid y = f(x), \text{ for some } x \text{ in } B\} \subset B.$$

On the other hand, if $D \subset B$, then we define the **inverse image** of D under f, denoted $f^{-1}(D)$, to be the subset of A given by

$$f^{-1}(D) = \{x \text{ in } A \mid f(x) \text{ in } D\} \subset A.$$

As an example, suppose that $A = \{a, b, c, d, e\}$ and $B = \{1, 2, 3, 4\}$, and suppose that $f : A \longrightarrow B$ is given by the rule $f(a) = 2, f(b) = 2, f(c) = 3, f(d) = 2, f(e) = 4$. Then, for instance,

$$f(\{a, c, d\}) = \{2, 3\},$$

and

$$f^{-1}(\{1, 3, 4\}) = \{c, e\}$$

and

$$f^{-1}(\{1\}) = \emptyset.$$

Now, to calculate the number of subsets of any finite set, we begin by noticing that as $2 = \{0, 1\}$, if A is any set, finite or not, and if $f : A \longrightarrow 2$, then

$$f^{-1}(\{1\}) \subset A.$$

Therefore each function in 2^A specifies a subset of A. Conversely, if B is any subset of A, then it defines a unique function called the **indicator** of $B \subset A$ and denoted $I_{[B \subset A]}$ which is defined by the rule

$$I_{[B \subset A]}(x) = 1, \text{ if } x \text{ is in } B \text{ and zero otherwise.}$$

We see that there is a one-to-one correspondence between subsets of A and functions in 2^A, so if A is finite and has exactly n members, then 2^A has 2^n members, so there are 2^n subsets of A, and of course the empty set and A it self are members of the set of all subsets of A. For instance, if $A = \{a, b, c\}$, then the set of all subsets of A is

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

We are next going to study the theory of **expectation and probabiity**, and to begin we shall simply think of trying to guess an unknown quantity which we shall denote by X. For instance if I tell you that $4X = 8$, then you know right away that $X = 2$, so there is not really much unknown in this case, but if I tell you that X is George Washington's birthweight in pounds, then you probably do not know what that is, but you certainly know that 300 is not a reasonable guess and you know that 0.003 is not a reasonable guess. When we try to guess an unknown quantity, we need to have some information on which to base our guess. Of course the information we have may or may not be actually true, but that is just a problem we have to live with. We will assume that information is given by factual (true or false) statements. We will use capital letters near the end of the alphabet as symbols for unknowns and capital letters near the beginning of the alphabet for statements. If X is the unknown we are trying to guess and if K is the the statement of our knowledge used to make our guess, then we call our guess for the value of X the **expected value** of X given K and we denote this by writing it as $E(X|K)$, so

$$E(X|K) = \text{ guess for the value of } X \text{ given we assume the statement } K.$$

The first thing we want to do here is see how guessing is limited by considerations of pure logic. The simplest rule dictated purely by logic is that if your information K tells you the value of X, then that value is your guess. If K implies that $X = r$, where r is a specific real number, then r must be your guess for the value of X given K. Now, if K implies that $X = r$, then K and K & $(X = r)$ are logically the same. We can therefore say that pure logic imposes the general rule of guessing called the

Logical Consistency Rule:

$$E(X|K\&[X = r]) = r, \text{ for any } r \text{ in } \mathbb{R}.$$

We next point out that unknowns are quantities which are specified as pure numbers, their units are contained in their specification. This means that if X and Y are any two unknowns, then we can add them forming $X + Y$ and multiply them forming XY. In any situation we are analyzing, we can consider the set of all pertinant unknowns as a set $\boldsymbol{\mathcal{A}}$, called the **algebra of unknowns**. We can consider ordinary numbers as unknowns whose values are actually known no matter what K is. For instance, no matter what K is, we know that 3 is the number three, π is π, and in general, each real number r is a "known unknown". This means that

$$\mathbb{R} \subset \boldsymbol{\mathcal{A}},$$

and therefore the rule of logical consistency dictates

$$E(r|K) = E(r|K\&[r = r]) = r.$$

As another example, suppose that K contains the statement that a dice is in a box where we cannot see it, that the dice is lying on the floor of the box so that one face is on top, our information does not tell us which face is on top, and X is the number of spots on the top face. Also assume that K includes the information as to what a dice is; it is a small cube with 6 faces, each having a number of spots and that number is in the set $S = \{1, 2, 3, 4, 5, 6\}$. Therefore, we know that X is in S, but we do not know which member of S it actually is. We are going to see that there is a very good reason to have $E(X|K) = 3.5$, even though we know this cannot be the value of X.

As another example, suppose that we have a box of envelopes and each contains some money. We choose two envelopes from the box, and let X be the amount of money in the first envelope and let Y be the amount of money in the second envelope. If our guesses are $E(X|K) = 30$ and $E(Y|K) = 40$, then it certainly seems we should have $E(X+Y|K)=70$. Why does this seem so undeniable? Well, we will see that the only possible rule here consistent with the rule of logical consistency is the

Addition Rule:

$$E(X+Y|K)=E(X|K)+E(Y|K), \text{ fo all } X,Y \text{ in } \boldsymbol{\mathcal{A}}.$$

For instance, could it be the case that we should have a rule like

$$E(X+Y|K)=[E(X|K)]^2[E(Y|K)]^3,$$

or maybe some other even stranger rule? The answer will turn out to be NO, as we will be able to show quite generally that

The Addition Rule above is the only possiblity which is allowed by the Logical Consistency Rule.

3. FRIDAY 17 JANUARY 2014

In the last lecture, we began developing the **Theory of Expectation**, and we specified the notation $E(X|K)$ for the **Expected Value** of X **given** K as background information. Another notation for this same quantity is μ_X, called the **Mean** of X. In this second form of notation, the statement K does not appear, so one must be careful with it. The reasons for the two different notations are partly historical and partly a matter of convenience. In general, when dealing with a situation where we have unknowns, there are usually many, so it is useful to denote the set of all unknowns by $\boldsymbol{\mathcal{A}}$, and call it the **Algebra of Unknowns**. So we want to imagine that we need to determine $E(X|K)$ in case X is any member of $\boldsymbol{\mathcal{A}}$ and K is any statement of relevant factual information. Here it is important to realize that K may or may not be actually true, and we generally do not care which is the case, the problem is to determine our guess for the value of X if we assume K is true. Now, we will assume that in our background information we have all the basic facts about real numbers and simple algebra. In particular, each real number is an unknown whose value is actually known no matter the background information K. Thus,

$$\mathbb{R} \subset \boldsymbol{\mathcal{A}}.$$

To begin we will think of $E(X|K)$ as a a real number which is our guess for the value of X bases ONLY on the information K. We want our guessing to be constrained by LOGIC. This means in particular, that if our information K happens to tell us the value of X, then that is our guess for the value of X, no matter whether or not the background information is actually true. This means we have the

Logical Consistency Rule:

$$E(X|K \ \& \ [X = r]) = r, \text{ for any } r \text{ in } \mathbb{R}.$$

As an immediate consequence of the Logical Consistency Rule, we have the

Retraction Rule:

$$E(r|K) = r, \text{ for any } r \text{ in } \mathbb{R}.$$

To see why the Retraction Rule follows from the Logical Consistency Rule, we simply note that as $[r = r]$ for any real number r, we have

$$E(r|K) = E(r|K \ \& \ [r = r]) = r,$$

on simply replacing X by r in the Logical Consistency Rule.

Last time we observed that as the unknowns have any units required as part of their definition, their values are simply pure numbers and can therefore be added and multiplied. Thus,

$$X + Y \text{ and } XY \text{ belong to } \boldsymbol{\mathcal{A}} \text{ whenever } X \text{ and } Y \text{ belong to } \boldsymbol{\mathcal{A}},$$

and this is why we call $\boldsymbol{\mathcal{A}}$ the algebra of unknowns, because it is mathematically a system called an algebra, that is a system where we can add and multiply. This means, we should like to be able to determine $E(X + Y|K)$ from the numbers $E(X|K)$ and $E(Y|K)$. That is, once we have worked out the guess for X and the guess for Y using K, then these two numbers should determine our guess for $X + Y$. For instance, if X is the number of spots on the top face of a dice in a box which I cannot see, and if $E(X|K) = 2$, and if Y is the number of spots on another dice in another box and we have determined $E(Y|K) = 3$, then it seems very reasonable that $E(X + Y|K)$ should have to be $2 + 3 = 5$. We need to see that this is actually forced on us by the Logical Consistency Rule, and in fact it is forced on us by the Retraction Rule which is a restricted version of the Logical Consistency Rule. We call it the

Addition Rule:

$$E(X + Y|K) = E(X|K) + E(Y|K), \text{ for any } X, Y \text{ in } \boldsymbol{\mathcal{A}}.$$

To see why this simple rule is forced on us as soon as we assume that there is some rule for determining $E(X+Y|K)$ from the numbers $E(X|K)$ and $EY|K)$, we can begin with a simple case. Our general rule would in particular allow us to determine the guess for $X+3$ as soon as we know the guess for X, because $E(3|K)=3$. Thus, our general rule would here give us a function

$$f_3 : \mathbb{R} \longrightarrow \mathbb{R}$$

with the property that

$$f_3(E(X|K)) = E(X+3|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

But as $\mathbb{R} \subset \boldsymbol{\mathcal{A}}$, we can apply the preceding formula to the case where X is an actual real number. For instance, if X is replaced by the number 7, then

$$f_3(7) = f_3(E(7|K)) = E(7+3|K) = E(10|K) = 10 = 7+3,$$

and obviously, this would work if 7 were replaced by any other number telling us that f_3 is the function which merely adds 3 to whatever is put into it. That is,

$$f_3(x) = x+3, \text{ for any } x \text{ in } \mathbb{R}.$$

But, tis means that

$$E(X+3|K) = f_3(E(X|K)) = E(X|K)+3, \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

Notice that there is really nothing special about the number three here in this argument, it could have been any real number r. That is, if we assume there is a rule for computing our guess for $E(X+r|K)$ using only our guess $E(X|K)$, then that is really the assumption that there is some function

$$f_r : \mathbb{R} \longrightarrow \mathbb{R}$$

with the property that

$$f_r(E(X|K)) = E(X+r|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

Then for any x in $\mathbb{R}$, we have

$$f_r(x) = f_r(E(x|K)) = E(x+r|K) = x+r,$$

since $x+r$ is simply a real number. This shows that the rule f_r is simply

$$f_r(x) = x+r, \text{ for any } x \text{ in } \mathbb{R},$$

and therefore

$$E(X+r|K) = f_r(E(X|K)) = E(X|K)+r, \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

We have therefore shown that the assumption of some rule for determining the guess for $X+Y$ from the guess for X and the guess for Y all based on K is the simple addition rule, if one of the unknowns is simply a real number, so we will now assume this rule to be always true.

Now, for the case of guessing $E(X+Y|K)$, if we had a general rule for this, it would in particular tell us that for fixed Y we can have a function

$$f_Y : \mathbb{R} \longrightarrow \mathbb{R}$$

with the property that

$$E(X+Y|K) = f_Y(E(X|K)), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

To find out what f_Y has to be, again, we are going to use the same method as above, we simply apply our formula for f_Y here in the case that we take X to simply be the real number x in $\mathbb{R} \subset \mathbb{A}$. we then have, since our addition rule applies already to the case where one of the unknowns is only a real number,

$$f_Y(x) = f_Y(E(x|K)) = E(x+Y|K) = x + E(Y|K),$$

and therefore,

$$E(X+Y|K) = f_Y(E(X|K)) = E(X|K) + E(Y|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

Again, there was nothing assumed about Y other than it belongs to $\boldsymbol{\mathcal{A}}$, so we have therefore demonstrated that the Addition Rule as stated above is the only possible rule for determining the guess for $X+Y$ from the guesses for X and for Y all based on K.

Notice there was nothing special about the fact that we were dealing with addition. If we replace addition everywhere with multiplication in the preceding arguments, it would have given the corresponding multiplication rule that to calculate $E(XY|K)$ we just multiply $E(X|K)$ and $E(Y|K)$. That is, if we assume that there is some rule for calculating $E(XY|K)$ from the numbers $E(X|K)$ and $E(Y|K)$, then the rule would have to be just multiply the two numbers together. However, it can be shown that such a multiplication rule would imply that K tells us the value of every unknown in $\boldsymbol{\mathcal{A}}$. Since this is certainly not true for the example of the dice in a box, it follows we have demonstrated that IN GENERAL, THERE CAN BE NO RULE FOR DETERMINING $E(XY|K)$ FROM THE NUMBERS $E(X|K)$ AND $E(Y|K)$.

Because of this, we need to look at special cases of the multiplication rule. For instance, if X is the number os spots on the top of a dice which we cannot see, and if $Y = 7X$, it certainly seems undeniable that if we have $E(X|K) = 2$, then $E(Y|K) = 14$. As another example, if X is the amount of money in an account expressed in German Marks and if Y is the amount in dollars, and if the exchange rate says each Mark is worth twenty five cents, then guessing that the value of the account is 100 Marks is the same as guessing the value of the account is twenty five dollars.

Therefore the general rule called the

Homogeneity Rule:

$$E(rX|K) = rE(X|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}} \text{ and any real number } r \text{ in } \mathbb{R},$$

seems completely reasonable, and is the only possible rule as our arguments show, for determining the guess for rX from the guess for X. We therefore assume this rule to be true. Notice that this is saying that the general multiplication rule holds if at least one of the unknowns is simply a known real number, that is

$$E(XY|K) = E(X|K)E(Y|K), \text{ for any } X, Y \text{ in } \boldsymbol{\mathcal{A}}, \text{ if at least one of the two unknowns belongs to } \mathbb{R} \subset \boldsymbol{\mathcal{A}}.$$

Notice that if $Y = r$ is in $\mathbb{R}$, then it is an unknown with only one possible value. Let us now consider the case where there are two possible values for Y. To begin, then let us assume that $Y^2 = Y$. This means that the only possible values of Y are zero and one. Notice that if $Y = 1$, then $XY = X$, and if $XY = 0$, then $XY = 0$, so in either case, we should be able to determine the guess for XY. Specifically, if we can determine our guess for X assuming that $Y = 1$ in addition to K, then we should be able to determine the guess for XY based on K. That is, there should be some rule for determining $E(XY|K)$ from $E(X|K \ \& \ [Y=1])$, and it is not now obvious what the rule should be as it was in the previous case of the addition rule. But, the assumption that such a rule exists will by the same method tell us what the rule has to be. We therefore assume there is some function

$$f_Y : \mathbb{R} \longrightarrow \mathbb{R},$$

with the property that

$$f_Y(E(X|K \ \& \ [Y=1])) = E(XY|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}.$$

To determine what f_Y has to be, we do what we did before, namely, we try this formula out in the case where X is merely a number x in $\mathbb{R} \subset \boldsymbol{\mathcal{A}}$. In this case, we have, by the Homogeneity Rule,

$$f_Y(x) = f_Y(E(x|K \text{ \& } [Y = 1])) = E(xY|K) = xE(Y|K), \text{ for any } x \text{ in } \mathbb{R}.$$

It follows that

$$E(XY|K) = f_Y(E(X|K \text{ \& } [Y = 1])) = E(X|K \text{ \& } [Y = 1])E(Y|K).$$

This is the most general form of multiplication rule which we can allow. That is, we have the

Multiplication Rule:

$$\text{If } X, Y \text{ are in } \boldsymbol{\mathcal{A}} \text{ and if } Y^2 = 1, \text{ then } E(XY|K) = E(X|K \text{ \& } [Y = 1])E(Y|K).$$

If B is any statement, then we can form its **indicator unknown**, denoted I_B, which has the value one if B is True and zero if B is False. Thus, the only possible values of an indicator are zero and one. It follows that $I_B^2 = I_B$. In fact, if Y is any unknown equal to its own square, then taking B to be the statement $Y = 1$, we have $Y = I_B$. Thus, we can formulate the multiplication rule as saying that

$$E(XI_B|K) = E(X|K \text{ \& } B)E(I_B|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}}, \text{ and any statement } K.$$

Another useful rule has to do with the ordering of unknowns. If I draw two lines on the blackboard and say that X is the length in feet of the first line and Y is the length in feet of the second line, and if the second line is obviously longer, our information is telling us $X \leq Y$, even though we may not be able to know the exact lengths of either of the lines. Certainly, you would not guess a smaller number for the line that appears longer. We thus have the

Order Rule:

$$\text{If } X, Y \text{ in } \boldsymbol{\mathcal{A}}, \text{ and if } K \text{ implies that } X \leq Y, \text{ then}, E(X|K) \leq E(Y|K).$$

Notice that if B is any statement, then its indicator satisfies

$$0 \leq I_B \leq 1,$$

and therefore

$$0 \leq E(I_B|K) \leq 1.$$

We Define the **Conditional Probability of** B **given** K, denoted $P(B|K)$, by the formula

$$P(B|K) = E(I_B|K), \text{ for any statement } B.$$

Thus, the **First Law of Probabililty** is:

$$0 \leq P(B|K) \leq 1, \text{ for any statement } B.$$

The multiplication rule now becomes:

$$E(XI_B|K) = E(X|K \text{ \& } B)P(B|K), \text{ for any } X \text{ in } \boldsymbol{\mathcal{A}} \text{ and any statements } K, B.$$

4. WEDNESDAY 22 JANUARY 2014

To begin, let us review the rules we have for logically consistent guessing, that is to say, our **Expectation Rules.** In what follows, X and Y are unknowns, K and B are statements, r and s are real numbers.

Logical Consistency Rule:

$$E(X|K \text{ \& } [X = r]) = r.$$

Order Rule:

$$\text{If } K \text{ implies } X \leq Y, \text{ then } E(X|K) \leq E(Y|K).$$

Addition Rule:

$$E(X + Y|K) = E(X|K) + E(Y|K).$$

Homogeneity Rule:

$$E(rX|K) = rE(X|K).$$

Multiplication Rule:

$$\text{If } Y^2 = Y, \text{ then } E(XY|K) = E(X|K \text{ \& } [Y = 1])E(Y|K).$$

We also review our basic definitions.

Indicator Unknown of a Statement:

$$I_B = \text{ One if } B \text{ is True and Zero if } B \text{ is False.}$$

Definition of Probability:

$$P(B|K) = E(I_B|K) = \text{ Conditional Probability of } B \text{ given } K.$$

Next, we have some immediate consequences of these basic rules and definitions.

Equality Rule:

$$\text{If } K \text{ implies } X = Y, \text{ then } E(K|K) = E(Y|K).$$

This is a simple consequence of the Order Rule, since if $X = Y$, then both $X \leq Y$ and $Y \leq X$ are true and therefore both $E(X|K) \leq E(Y|K)$ and $E(Y|K) \leq E(X|K)$ are true so $E(X|K) = E(Y|K)$.

Retraction Rule:

$$E(r|K) = r.$$

This is a consequence of the Logical Consistency Rule and the fact that $r = r$ is always true.

Linearity Rule:

$$E(rX \pm sY|K) = rE(X|K) \pm sE(Y|K).$$

This is simply the result of combining the Addition Rule with the Homogeneity Rule.

First Law of Probability:

$$0 \leq P(B|K) \leq 1.$$

which follows from the fact that $0 \leq I_B \leq 1$, the Order Rule, and the definition of probability.

If B is any statement, then $I_B = 1$ is logically equivalent to B itself. Since

$$I_B = I_B^2 \text{ and } E(I_B|K) = P(B|K),$$

the Multiplication Law can be restated as

Multiplication Rule:

$$E(XI_B|K) = E(X|K \ \& \ B)P(B|K).$$

Notice we can also interpret the Logical Consistency Rule as saying that information cannot be ignored. It is important to keep in mind that the information K need not be true in actuality, but is merely assumed to be true for the purpose of calculating the expected value. It is best to think of this as actually being done by a robot. We give the robot the information and require the robot to **guess every unknown** in the algebra $\boldsymbol{\mathcal{A}}$ of unknowns we are dealing with, in such a way as to obey all the rules. For instance, suppose we have five identical envelopes containing money, which we mark on the outside with numbers one through five, and other than their identification numbers, they all appear to be identical because the robot cannot see what is inside the envelopes. We define X_1 to be the value of the money in the envelope numbered one, in dollars, and likewise, X_k is the value of the money in the envelope numbered k, in dollars. If we tell the robot the total amount of money in the five envelopes is one hundred dollars, the robot knows as a consequence of the information K, that

$$X_1 + X_2 + X_3 + X_4 + X_5 = 100,$$

and therefore by the rules,

$$E(X_1|K) + E(X_2|K) + E(X_3|K) + E(X_4|K) + E(X_5|K) = E(100|K) = 100.$$

Now, the robot is faced with a dilema. He must guess each of these expected values,

$$E(X_1|K), \ E(X_2|K), \ E(X_3|K), \ E(X_4|K), \ E(X_5|K),$$

but his information does not tell him there is any difference between them. As far as the robot is concerned, these five numbers are indistiguishable. Consequently, he **must guess they are all the same.**

We will call this the **Principle of Indifference:**

If the information K does not distinguish the values of $X_1, X_2, X_3, ..., X_n$ in any way but does imply the value of the total is a real number r, then

$$E(X_k|K) = \frac{r}{n}, \ k = 1, 2, 3, ..., n.$$

As an example, suppose that we have ten buckets of jewelry with a total value of ten million dollars. Then based on this information alone, we expect the value of the jewelry in each bucket to be worth one million dollars. Suppose that instead we have ten pounds of jewelry worth ten million dollars. If X is the value in dollars of a pound of this jewelry, and if K is only this information, then $E(X|K)$ =one million. Likewise, two pounds would be expected to be worth two million and so on. Any method of dividing up the jewelry leads to the same conclusion. If the jewelry is put into ten identical boxes, which we cannot see inside, then each box is expected to be worth one million, even if unknown to us some boxes may be empty. In fact, if all the jewelry is put into one of ten identical boxes, the other nine being empty, then as all appear identical, they are all expected to be worth one million.

We now turn to the **Theory of Probability** which is a result of our **Expectation Rules.** First, we must see how indicators work with algebra. If A and B are any statements, then $A \ \& \ B = A \cap B$ is the statement which is true if and only if both A and B are true. We then easily see that

$$I_{A\&B} = I_{A \cap B} = I_A \cdot I_B.$$

On the other hand, the statement A or $B = A \cup B$ is the statement which is true if at least one of the two statements (possibly both, we do not care) is true. Since logical "and" required multiplication of indicators, thinking algebraically, we might suspect addition applies to logical ”or”, but if we add the two indicators and it happens that both statements are true, then the value of the sum would be 2 which is not allowed for an indicator. We must therefore subtract one only in the case both are true, that is we now easily see that

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

If S denotes the statment ”$1 = 1$”, Then S is true no matter what, we call S the **Sure Statement**. Thus,

$$I_S = 1.$$

The negation of A is simply $notA$ and is true exacty if A is false and likewise false exactly if A is true. Thus, as

$$A \cup notA = S,$$

we have

$$I_{notA} = 1 - I_A.$$

As immediate consequences of the Expectation Rules, we then have the

Laws of Probability:

$$P(S|K) = 1,$$
$$P(A \text{ or } B) = P(A \cup B) = P(A|K) + P(B|K) - P(A\&B|K),$$

from which we conclude in particular the

Complement Rule:

$$P(notA|K) = 1 - P(A|K).$$

As a consequence of the Multiplication Rule for Expectation we have the

Conditional Probability Rule:

$$P(A\&B|K) = P(A \cap B|K) = P(A|K\&B)P(B|K).$$

By a **Partition of Unity** we will mean any collection of statements with the property that exactly one is true and all others are false. If P is a finite partition of unity, then

$$\sum_{B \text{ in } P} I_B = 1,$$

and therefore

$$\sum_{\mathcal{B} \text{ in } P} P(B|K) = 1.$$

So, if the information K does not give any distinction among the statements of the partition of unity, then by the principle of indifference, they must all have the same probability given K. We would say in this case we are dealing with the **Model of Equally Likely Outcomes**. If there are exactly n statements in P, then each must have probability $1/n$ given K, if K dictates we have the model of equally likely outcomes. Thus for the dice in the box where the robot cannot see the number of spots on the top face, given only that information the robot would conclude by the Principle of Indifference that each face has probability $1/6$ of being the face on top. Let B_k be the statement that the top face has k spots on top. Then

$$P = \{B_1, B_2, B_3, B_4, B_5, B_6\}$$

is a partition of unity for the dice in the box. Let I_k denote the indicator of B_k. We then see that

$$I_1 + I_2 + I_3 + I_4 + I_5 + I_6 = 1.$$

so by the principle of indifference,

$$P(B_k|K) = \frac{1}{6}.$$

Let X be the number of spots on the top face. Thus B_k says $X = k$, for each $k = 1, 2, 3, 4, 5, 6$. But,

$$X = X \cdot 1 = XI_1 + XI_2 + XI_3 + XI_4 + XI_5 + XI_6,$$

so

$$E(X|K) = E(XI_1|K) + E(XI_2|K) + E(XI_3|K) + E(XI_4|K) + E(XI_5|K) + E(XI_6|K) = \sum_{1 \leq k \leq 6} E(XI_k|K).$$

For each k, we can apply the multiplication rule:

$$E(XI_k|K) = E(X|K \;\&\; [X = k])P(X = k|K) = k \cdot \frac{1}{6} = \frac{k}{6}.$$

We are therefore forced to conclude that

$$E(X|K) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = \frac{7}{2} = 3.5.$$

Notice that our rules have forced the robot to this conclusion that $E(X|K) = 3.5$ Obviously, in any situation where X is an unknown with a finite number of possible values, where our information does not tell us any value is preferred in any way, we must conclude that $E(X|K)$ is simply the **Average** of all the possible values, the total divided by the number of possible values. It is also important to keep in mind that **the expected value may not be one of the possible values**, so if we think in terms of guessing, the robot has decided that it is best to be wrong!! We will soon see why this is so. But more generally, we can observe that if P is any finite partition of unity, and if X is any unknown, the equation

$$\sum_{B \text{ in } P} I_B = 1,$$

can be multiplied on both sides by X to give the equation

$$X = \sum_{B \text{ in } P} XI_B,$$

so by the Addition Rule we have

$$E(X|K) = \sum_{B \text{ in } P} E(XI_B|K).$$

But, by the Multiplication Rule, we have

$$E(XI_B|K) = E(X|K \;\&\; B)P(B|K), \text{ for each } B \text{ in } P,$$

and consequently

$$E(X|K) = \sum_{B \text{ in } P} E(X|K \;\&\; B)P(B|K).$$

This means that whenever we have a partition P with the property that $E(X|K \;\&\; B)$ and $P(B|K)$ is known to us for each B in P, then we calculate $E(X|K)$ as the sum of the products of conditional expected values multiplied by corresponding probabilities of the conditions. In particular, any time we know all the possible values of X and each of their probabilities we just sum the products of values multiplied by probabilities.

As an example of this last formula applied to X, the number of spots on the top of a dice we cannot see, if A is the statement that X is Even and if B is the statement that X is Odd, then clearly $P = \{A, B\}$ is a partition, and our information gives no preference to either of these two statements, so by the Principle of Indifference, each must have probability 1/2. Moreover, if we know X is even, then the only possible values are in the set $\{2, 4, 6\}$ and all of these three possibilities are equally likely, again by the Principle of Indifference, so

$$E(X|K \ \& \ A) = 4,$$

as 4 is the average of these three even numbers. Likewise, if we know that X is odd, then the only possible values are in the set $\{1, 3, 5\}$ and all these three possibilities are equally likely, so again the expected value of X given that it is odd is the average of these three values, which is 3. That is we have

$$E(X|K \ \& \ B) = 3.$$

Our general formula for expectation then gives

$$E(X|K) = E(X|K \ \& \ A)P(A|K) + E(X|K \ \& \ B)P(B|K) = 4 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = \frac{4+3}{2} = \frac{7}{2} = 3.5,$$

again. As another example with the dice, we can take the partition $P = \{A, B, C\}$, where

$$A \text{ says } X \text{ is in } \{1, 2\},$$

$$B \text{ says } X \text{ is in } \{3, 4\},$$

$$A \text{ says } X \text{ is in } \{15, 6\},$$

so again, according to the information given, there is no preference for any of these three statements, so all are equally likely to be true and therefore must each have probability 1/3. Obviously,

$$E(X|K \ \& \ A) = 1.5,$$

$$E(X|K \ \& \ B) = 3.5,$$

$$E(X|K \ \& \ C) = 5.5,$$

so using this partition to calculate $E(X|K)$, we have

$$\begin{aligned} E(X|K) &= E(X|K \ \& \ A)P(A|K) + E(X|K \ \& \ B)P(B|K) + E(X|K \ \& \ C)P(C|K) \\ &= (1.5) \cdot \frac{1}{3} + (3.5) \cdot \frac{1}{3} + (5.5) \cdot \frac{1}{3} = .5 + \frac{3.5 + 5.5}{3} = .5 + \frac{9}{3} = 3.5, \end{aligned}$$

again.

Notice that beyond the rules themselves, the key to being able to arrive at the expected value is to be able to find a partition of unity where all statement in the partition are equally likely, so that the Principle of Indifference can be applied to give us a Model of Equally Likely Outcomes, and then to be able to calculate the expected value given each of the conditions of the partition. This is the technique that is used in all theoretical probability and expectation calculations, of which we will see many examples in the future.

Of course, in terms of guessing, to guess the number of spots on the top face of the dice is 3.5 is to know you are wrong right off the bat. To see why that would be a good guess, we need to consider the error, R, given as

$$R = X - 3.5.$$

Thinking of the robot, it can only calculate expected values, so it might try to guess his error as E. But, by the Linearity Rule,

$$E(R|K) = E(X - 3.5|K) = E(X|K) - E(3.5|K) = 3.5 - 3.5 = 0.$$

However, the robot can also consider the Squared Error $R^2 = (X - 3.5)^2 \geq 0$. Thus, by the Order Rule,

$$E(R^2|K) \geq 0.$$

But, notice that as all values of R^2 are positive, in fact, we have in this case that

$$R^2 \geq .25.$$

Therefore, by the Order Rule, we know that

$$E(R^2|K) \geq .25.$$

We see from this that the robot will have to consider his squared error to see its error. Could this consideration lead it to change its guess??

5. FRIDAY 24 JANUARY 2014

The main result of our Expectation Rules is that if our information K tells us that the unknown X has possible values only in a finite set $F \subset \mathbb{R}$, and if our information K tells us the probability that X has value r for each of the values in F, then we take each value in F and multiply it by its probability and sum up these products to calculate $E(X|K)$. Thus,

$$E(X|K) = \sum_{r \text{ in } F} r \cdot P(X = r|K).$$

Moreover, in case K gives no preference to any of the values in F, then all the probabilities must be the same, so if F has n numbers, then

$$P(X = r|K) = \frac{1}{n}, \text{ for all } r \text{ in } F,$$

and therefore in this case, $E(X|K)$ is simply the **Average** of all the possible values.

Last time we ended with the problem of our expectation rules dictating that we guess a value we know is wrong. Thinking in terms of a robot doing the guessing, if X is any unknown, and if we set

$$\mu_X = E(X|K),$$

then the error, which is called the **Deviation of X from its mean**, is R given by

$$R = X - \mu_X.$$

The robot would begin by trying to guess the error R, that is it would calculate $E(R|K)$. But,

$$E(R|K) = E(X|K) - E(\mu_X|K) = \mu_X - \mu_X = 0,$$

so the robot's first attempt at guessing its error is to think it must be zero. However, consider what happens when the robot tries to guess the **Squared Error**, $E(R^2|K)$. Then,

$$R^2 \geq 0, \text{ so } E(R^2|K) \geq 0,$$

by the Order Rule. In fact, if X has only a finite number of possible values, then the robot can easily see the minimum value possible for R^2, and if μ_X is not one of the possible values, then this minimum possible value is itself a positive number, say e^2. Thus,

$$R^2 \geq e^2, \text{ and therefore } E(R^2|K) \geq e^2,$$

again, by the Order Rule. Therefore, the robot will calculate a positive value for his **Expected Squared Error** which is also called the **Variance** of X. We see here that multiplication is entering the problem of analyzing the error. In general, for any unknowns X, Y, we define the **Covariance** of X and Y, denoted $Cov(X, Y \;|K)$, as

$$Cov(X, Y \;|K) = E([X - \mu_X][X - \mu_Y] \;|K).$$

The variance of X, denoted $Var(X \;|K)$, is then

$$Var(X \;|K) = Cov(X, X \;|K) = E([X - \mu_X]^2) \geq 0.$$

In many situations, the information K being the background information, is understood, so we leave it out of the notation, thus if B is any statement other than the background K, then, assuming K to be understood,

$$E(X) = E(X|K), \; E(X|B) = E(X|K \;\&\; B), Var(X) = \; Var(X \;|K), \text{ and, } Cov(X, Y) = Cov(X, Y \;|K).$$

Using a little algebra, we have

$$[X - \mu_X][Y - \mu_Y] = XY - X \cdot \mu_Y - \mu_X \cdot Y + \mu_X \cdot \mu_Y.$$

But, μ_X and μ_Y are actual real numbers in $\mathbb{R}$, so

$$E(\mu_X \cdot Y) = \mu_X E(Y) = \mu_X \cdot \mu_Y,$$

and

$$E(X \cdot \mu_Y) = E(\mu_Y \cdot X) = \mu_Y E(X) = \mu_Y \cdot \mu_X = \mu_X \cdot \mu_Y,$$

and

$$E(\mu_X \cdot \mu_Y) = \mu_X \cdot \mu_Y.$$

Applying these simplifications, we have

$$Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y - \mu_X \cdot \mu_Y + \mu_X \cdot \mu_Y = E(XY) - \mu_X \cdot \mu_Y,$$

so after canceling we have the simplification

$$Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y = E(XY) - E(X)E(Y).$$

In particular, for variance, we have

$$Var(X) = Cov(X,X) = E(X^2) - \mu_X^2 = E(X^2) - [E(X)]^2.$$

Alternately, we can write these equations as

$$E(X,Y) = \mu_X \cdot \mu_Y + Cov(X,Y) = E(X)E(Y) + Cov(X,Y),$$

and

$$E(X^2) = \mu_X^2 + Var(X) = [E(X)]^2 + Var(X).$$

Notice that these equations are telling us how far the multiplication rule for general products of unknowns actually fails.

Now, getting back to the robot analyzing its squared error, we might think it could reduce its squared error by guessing some value g for X instead of μ_X. If we try this, then our expected squared error would be

$$R_g = X - g = X - \mu_X + \mu_X - g.$$

To have fewer symbols to deal with, let us put $\mu_X - g = h$. Then,

$$R_g = R + h,$$

so,

$$R_g^2 = R^2 + h^2 + 2hR.$$

But, remember that

$$E(R) = 0,$$

and therefore

$$E(hR) = hE(R) = 0,$$

which means that

$$E(R_g^2) = E(R^2) + h^2 \geq E(R^2) = Var(X).$$

In other words, any change in the guess causes an increase in the guess for the squared error. The robot can only guess anything, as far as values of unknowns are concerned, so it tries to guess so as to minimize what it guesses its squared error to be. If we think of the squared error as being proportional to a pain for being wrong, then the robot is guessing so as to minimize what he guesses will be the pain for being wrong.

For any unknown X, we define the **Standard Deviation** of X, denoted σ_X, as

$$\sigma_X = \sqrt{Var(X)}, \text{ so } Var(X) = \sigma_X^2.$$

Notice that if X is an unknown with $\sigma_X = 0$, and if X has only a finite number of possible values, and if $\sigma_X = 0$, then as R^2 has only a finite number of non-negative values, the only way it can have an expected value of zero is for it to actually be zero which would mean that $X = \mu_X$ is merely a number, that is the information K would now tell us the actual value of X, not just a guess. In particular, for the dice in the box, there are several possible values and the information K definitely does not tell us the number of spots on the top face, so σ_X cannot be zero and therefore,

$$E(X^2) \neq [E(X)]^2.$$

As another example, suppose that X has possible values in the set $F = \{-1, 0, 1, 3, 4, 5\}$. For simplicity, let us simply write $P(r) = P(X = r|K)$, here. Further, suppose that K tells us that

$$P(-1) = .2,\ P(0) = .1,\ P(1) = .2,\ P(3) = .1,\ P(4) = .3,\ P(5) = .1.$$

Then,

$$\begin{aligned} E(X) &= (-1)(.2) + (0)(.1) + (1)(.2) + (3)(.1) + (4)(.3) + (5)(.1) \\ &= (-.2) + (0) + (.2) + (.3) + ((1.2) + (.5) = 2, \end{aligned}$$

so

$$E(X) = 2.$$

Notice $E(X)$ is not a possible value of X.

To compute the variance, we begin by computing $E(X^2)$. To do this, simply square the values of X using the same probabilities, so

$$E(X^2) = (1^2)(.2) + (0^2)(.1) + (1^2)(.2) + (3^2)(.1) + (4^2)(.3) + (5^2)(.1) = .2 + 0 + .2 + .9 + 4.8 + 2.5 = 8.6$$

Thus,

$$E(X^2) = 8.6.$$

To complete the calculation of the variance of X, we use the formula we derived which says

$$Var(X) = E(X^2) - [E(X)]^2,$$

so here,

$$Var(X) = 8.6 - (2^2) = 8.6 - 4 = 4.6,$$

and therefore the standard deviation of X is

$$\sigma_X = \sqrt{4.6}, \text{ which is approximately equal to } 2.144761059,$$

or,

$$\sigma_X = 2.14, \text{ to three significant digits }.$$

In general, we can ask for some estimate of the probability that the error can have absolute value more than a positive number $\epsilon > 0$. That is, we ask for the probability that

$$|X - \mu_X| \geq \epsilon.$$

Let A be the statement of the above inequality. Notice that A is logically equivalent to the statement

$$\epsilon^2 \leq (X - \mu_X)^2,$$

since for positive numbers, squaring preserves order. Now let us compare the unknowns

$$\epsilon^2 I_A \text{ and } (X - \mu_X)^2.$$

If A is false, then the left unknown is zero and the right unknown being a square is non-negative so if A is false, the left unknown is no more than the squared deviation. On the other hand, if A is true, then the indicator of A has value 1, so the left unknown has value ϵ^2, but when A is true, $\epsilon^2 \leq (X - \mu_X)^2$, so in case A is true, the left unknown is no more then the squared deviation, that is, we have the general inequality

$$\epsilon^2 I_A \leq (X - \mu_X)^2$$

holds as an inequality between the two unknowns, so by the Order Rule,

$$E(\epsilon^2 I_A) \leq E((X - \mu_X)^2) = \sigma_X^2.$$

But,

$$E(\epsilon^2 I_A) = \epsilon^2 E(I_A) = \epsilon^2 P(A),$$

Therefore, we conclude that, in complete generality,

$$\epsilon^2 P(|X - \mu_X| \geq \epsilon) \leq \sigma_X^2, \text{ for any } \epsilon > 0 \text{ and for any unknown } X,$$

an inequality that is known as the **Tchebeychev Inequality**. If $E(X^2) = [E(X)]^2$, then $\sigma_X^2 = 0$, and as $\epsilon > 0$, the Tchebeychev Inequality would imply that $P(|X - \mu_X| \geq \epsilon) = 0$. That is,

$$\text{If } E(X^2) = [E(X)]^2, \text{ then } P(|X - \mu_X| < \epsilon) = 1, \text{ for every positive real number } \epsilon.$$

In particular, if our information K tells us that X has only a finite set of possible values, then we can choose ϵ to be a positive number which is smaller than the absolute value of all the positive absolute values of possible deviations from μ_X, and then we see that our information is then telling us that μ_X is the value of X for sure. Thus, if the general multiplication rule were to hold telling us that all expected values of products of unknowns are simply found by multiplying the expected values of the factors, then every standard deviation would be zero, and our information would be telling us the value of every unknown. Since we see plenty of examples where our information is definitely not telling us the value of some unknowns, this means such a general multiplication rule cannot possibly exist. There can be no general rule for calculating $E(XY)$ from the two numbers $E(X)$ and $E(Y)$.

Finally, we see that our general rules will reduce the calculation of all expected values to the problem of calculating probabilities. Of course, the most useful rule for calculating probabilities is the Principle of Indifference which says that if our information does not allow us to have any preference for any one outcome, all must have the same probability, that is, we have the model of equally likely outcomes. In gambling situations, gamblers bet on certain outcomes happening, and the game is called a **Fair Game** if all the possible outcomes are equally likely. For instance, if we draw card from a standard deck of 52 cards, then the chance the ace of spades is the top card on the deck is $1/52$, when our information does not tell us any card is more likely to be on top than any other. Thus, shuffling the deck generally insures this state of our information. The probability that the second card from the top is the ace of spaces is likewise $1/52$, as is the probability it is on the bottom of the deck. If I ask for the probability that the second card is the ace of spades given that the third card is the ace of hearts, the probability is reduced to $1/51$, since there are only 51 places in the deck which are unknown to us. If we ask for the probability that the second card is a heart given that the first is a heart and the third is a spade, then there are only 50 positions unknown to us in the deck, and there are only 12 hearts possible for the second position, so the probability is

$$P(\text{second is a heart} \mid \text{first is a heart and third is a spade}) = \frac{12}{50}.$$

Notice that if someone deals the cards from the top of the deck, the results are the same, as information wise it is the same. Thus

$$P(\text{second card dealt is a heart} \mid \text{first dealt is a heart and third dealt is a spade}) = \frac{12}{50}.$$

Likewise, the probability

$$P(\text{fifth card dealt is a heart }) = \frac{13}{52} = \frac{1}{4}.$$

Of course, we can see this also from the fact that as all suits have the same number of cards, no suit has a preferred status, so each has probability 1/4 of being in a specific location in the deck. This applies equally well to problems of drawing blocks from a box. For instance, if a box contains 3 red blocks and 2 blue blocks, and someone draws the blocks from the box one after another, then

$$\begin{aligned} P(\text{third block drawn is red} \mid \text{fifth block drawn is blue}) = \frac{3}{4} \\ = P(\text{fifth block drawn is red} \mid \text{third block drawn is blue}), \end{aligned}$$

since the situation is the same as far as the information is concerned as if we stack the blocks and draw blocks from the top of the stack repeatedly. After all, a deck of cards is really just a stack of blocks. If you cannot see the stack of blocks, you are in the same situation as dealing from a shuffled deck of cards.

6. MONDAY 27 JANUARY 2014

We have worked out our fundamental rules of expectation and probability and noticed that the calculation of expected value usually involves the calculation of probabilities. For instance, whenever an unknown has only finitely many possible values, if we know the probability of each possible value, then we multiply each value by its probability and sum all the resulting products to get the unknown's expected value. Today we will use our rules of expectation and probability to calculate some probabilities. In general, when dealing with an experiment with only a finite number of possible outcomes, if our information gives no preference to any specific outcome, then we know all are treated as equally likely. We then say we are dealing with the **Model of Equally Likely Outcomes**. However, in many situations, we can see clearly that not all outcomes are equally likely.

For instance, if we toss a pair of dice and look to see the total number of spots on top, then we know the outcome is one of the whole numbers bigger than 1 but not more than 12. There are eleven possible outcomes, but obviously not all are equally likely, since if we ask for the probability that the total is 2, then both dice must have landed with only one spot on top, whereas if we ask for the probability of getting a total of 7 spots on top, there are lots of ways that can happen, so obviously, it must have higher probability than a total of only 2 spots on top. Let us imagine that we have a single red dice and a single blue dice, and list all the possible results for each of these dice. Now we actually do not need a pair of dice here. If we only have a single dice, we can simply roll it once and then roll it again. Let X be the number up on the first roll and let Y be the number up on the second roll. We can specify a typical outcome with a pair of numbers, say (R, B), where R is the number of spots on top of the red dice and B is the number of spots on top of the blue dice, or equivalently, we can specify an outcome with a pair (X, Y), where X is the number up on the first toss and Y is the number up on the second toss. Obviously there are 36 such pairs of numbers which can be naturally arranged in a six by six array of pairs, where the first row is all pairs with the red dice having one spot on top, the second row is all the pairs with the red dice showing two spots on top, and so on. Thus there are six rows and each row has six pairs, so the total number of such pairs is $6 \cdot 6 = 36$. Now when we look to see which pairs have the total of five spots, we see it forms a diagonal in the array containing exactly 4 pairs, so the probability of rolling a total of five is $4/36$. In fact, we can see easily from the array that if k is a whole number, then

$$P(\text{total } = k) = \frac{k-1}{36},\ 1 \leq k \leq 7,$$

whereas

$$P(\text{total } = k) = \frac{13-k}{36},\ 7 \leq k \leq 12.$$

This pattern of probabilities is easy to remember visually, and all gamblers are very familiar with it:

$$P(\text{total } = 2) = \frac{1}{36} = P(\text{total } = 12),$$
$$P(\text{total } = 3) = \frac{2}{36} = P(\text{total } = 11),$$
$$P(\text{total } = 4) = \frac{3}{36} = P(\text{total } = 10),$$
$$P(\text{total } = 5) = \frac{4}{36} = P(\text{total } = 9),$$
$$P(\text{total } = 6) = \frac{5}{36} = P(\text{total } = 8),$$
$$P(\text{total } = 7) = \frac{6}{36} = P(\text{total } = 7),$$

so,

$$P(\text{total } = k) = P(\text{total } = l),\ \text{whenever } k + l = 14.$$

Notice that we can also view R and B as unknowns here, and the total is simply

$$T = R + B = X + Y,$$

so

$$E(T) = E(X) + E(Y) = E(R) + E(B) = 3.5 + 3.5 = 7,$$

but from our knowledge of the probabilities, we see that not only is 7 the expected total, it is also the most likely total. For any unknown, a most likely value is called a **Mode**. We also see from these probabilities that

$$P(T \leq 7) = \frac{21}{36} = P(T \geq 7) \text{ which is greater than or equal to } \frac{1}{2},$$

so 7 is called the median total. For any unknown X, if m is a number (not necessarily a possible value) with the property that

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2},$$

then m is called a **Median** for X. Notice for T, the total for the pair of dice, the (only) median is 7.

We can use these probabilities for the pair of dice to analyze a simple game of dice. The game played in the casino is similar with some complications that we will ignore for simplicity sake. The game goes like this. You roll the pair of dice, and if you get $T = 7$, then you win right away. If you do not roll a 7 on your first toss, then we let t denote the specific value of T on the first roll. You then keep rolling until you either get another $T = t$ or you get $T = 7$. If you get $T = 7$ before getting $T = t$, then you lose the game whereas if you get $T = t$ before you get $T = 7$, then you win. We would like to calculate the probability that you end up winning the game for each possible value of t. Clearly, this could take a while to determine if the player wins or loses, but we should keep in mind that **All Probability is Conditional**. At the casino, if you toss the dice and they bounce off the table and stop underneath the Blackjack table, we do not climb under to see what the total is. It is a do over, that toss would not count. Thus, all the results are conditioned on the toss satisfying certain criteria, and if not it is a do over. These criteria are contained in the statement of the background information that is understood. Well, in our little game, we can like wise think of the criteria as having been modified to require that the total be either 7 or t, and anything else is just a do over. That is, we see

$$P(\text{win}) = P(T = t | T = t \text{ or } T = 7).$$

From the multiplication rule, that is the rule of conditional probability, we have

$$P(T = t) = P(T = t \;\&\; [T = t \text{ or } T = 7]) = P(T = t \mid [T = t \text{ or } T = 7])P([T = t \text{ or } T = 7])$$

hence,

$$P(win) = \frac{P(T = t)}{P([T = t \text{ or } T = 7])}.$$

More generally, if A and B are statements about outcomes of a repeatable experiment, and if A and B cannot both happen, we say they are **Exclusive Events**, and in repeated trials, the probability that A happens before B is simply

$$P(A|A \text{ or B}) = \frac{P(A)}{P(A \text{ or } B)}.$$

From the pattern of probabilities, we see for instance that if $t = 5$, then the probability of going on to win the game is

$$P(\text{win} \mid \text{first roll 5}) = \frac{4/36}{[4/36] + [6/36]} = \frac{4}{10} = P(\text{win} \mid \text{first roll 9}),$$

where we note that all factors of 36 just cancel. Likewise,

$$P(\text{win} \mid \text{first roll } 4) = \frac{3}{9} = P(\text{win} \mid \text{first roll } 10),$$

and so on.

One thing we see is that calculating probabilities using the model of equally likely outcomes boils down to simple counting. Unfortunately, there are many counting problems that are very difficult. As an example, if we are to be dealt a five card hand from a standard deck of cards, and if we ask for the probability that all five cards are hearts, then we have a much more difficult counting problem than we had with the dice. Here, we introduce some useful notation. If S is a set of n things, then we denote by $C(n, k)$ the number of ways to choose a subset of exactly k members of S. Such a choice is called a **Combination**. Thus, the number of possible five card hands we can be dealt is $C(52, 5)$. But, the number of five card hands containing only hearts is only $C(13, 5)$, since there are only 13 hearts in the deck. Therefore,

$$P(\text{dealt all hearts }) = \frac{C(13, 5)}{C(52, 5)}.$$

If instead we say that we are going to deal the cards out face up in a sequence and that the order in which the cards appear is important, then we actually have to count in a way that includes this extra information. We now must count all possible ways to arrange five cards taken from the deck. For instance if we are going to bet after each card is dealt, then the order in which they appear is very important. Such an arrangement is called a **Permutation**. We denote by $P(n, k)$ the number of arrangements of k things taken from a set of n things. Thus, if we want the probability that the cards come out so that the first is a 2, the second is a three, the third is a four, the fourth is a five, and the last is a six, then we have to count all possible arrangements of five cards taken from the deck of 52 cards. This number is simply $P(52, 5)$. We can easily see that when we go to get the first card, there are 52 possibilities and once that is chosen, there are only 51 possible left for the second card and so on. This means

$$P(52, 5) = (52)(51)(50)(49)(48).$$

On the other hand, there are four deuces, four threes and so on, so the number of ways to get the result we are asking for is 4^5, and therefore

$$P(2, 3, 4, 5, 6 \text{ in order }) = \frac{4^5}{P(52, 5)}.$$

Now, you might have noticed that we did not calculate $C(52, 5)$. On the other hand, we could use our combination to form an arrangement. Notice that if you are given five cards, there are $P(5, 5)$ ways to arrange all five cards. So, to get an arrangement of five cards from the deck we could as a first step simply choose any five cards, which can be done in $C(52, 5)$ ways, and then as a second step simply arrange the five chosen cards. That is, it must be that

$$P(52, 5) = C(52, 5) \cdot P(5, 5).$$

Since

$$P(52, 5) = (52)(51)(50)(49)(48),$$

and

$$P(5, 5) = (5)(4)(3)(2)(1),$$

it follows that

$$C(52, 5) = \frac{(52)(51)(50)(49)(48)}{(5)(4)(3)(2)} = (13)(51)(10)(49)(8).$$

Using the equation

$$(n+1)(n-1) = n^2 - 1,$$

we have

$$(49)(51) = (50)^2 - 1 = 2500 - 1 = 2499.$$

This means that

$$C(52,5) = (13)(2499)(8)(10) = 2598960.$$

On the other hand,

$$P(5,5) = (5)P(4,4) = (5)(4)P(3,3) = (20)(3)P(2,2) = (20)(3)(2) = 120,$$

or

$$P(5,5) = 120,$$

so

$$P(52,5) = C(52,5)P(5,5) = (2598960)(120) = 311875200.$$

Notice the effect of including order is a vast increase in the number of possibilities.

We can effectively use this notation to calculate more complicated probabilities by simply thinking in terms of sequences of choices and counting using the $C(n,k)$'s. As an example, suppose we wanted to calculate

$$P(\text{three aces and two 8's} \mid \text{dealt 5 cards }).$$

we know its a fraction with denominator $C(52,5) = 2598960$, so we just have to count the number of five card hands which contain three aces and two 8's. To make such a hand, we have to get the four aces from the deck and choose 3 of them, and then get the four 8's from the deck and choose two of them, so this can be done in $C(4,3) \cdot C(4,2)$ ways. Therefore,

$$P(\text{three aces and two 8's} \mid \text{dealt 5 cards }) = \frac{C(4,3) \cdot C(4,2)}{C(52,5)}.$$

More generally, in Poker, a five card hand which has three of a kind and a pair is called a **Full House**. To make such a hand, as a first step we can decide which two denominations we will use. For instance, in the previous hand we had a full house said to be aces over 8's. If there had been three 8's and a pair of aces it would have been said to be 8's over aces. In any case, the denominations used for this full house are aces and 8's. There are always two denominations in a full house, but there are 13 possible denominations in the deck, so choosing two denominations can be done in $C(13,2)$ ways. We then have to decide which of the two chosen denominations will be used to make three of a kind. This means we must choose one of the two chosen denominations and this can be done in $C(2,1)$ ways. At this point we have decided which denominations are to be used for the three of a kind and which to use for the pair. Now we go into the deck get the four cards of the denomination to be used for the three of a kind and actually choose three of them, which can be done in $C(4,3)$ ways, and then we get the four cards of the denomination to be used for the pair and choose the pair which can be done in $C(4,2)$ ways. Looking back at our decision counting, we see

$$P(\text{full house} \mid \text{dealt five cards}) = \frac{C(13,2) \cdot C(2,1) \cdot C(4,3) \cdot C(4,2)}{C(52,5)}.$$

Looking at the numerator of the expression above, we can almost know what sequence of decisions was made by a person who would get this result. However, this is not the only way to accomplish the task of making a full house via a sequence of decisions. However you try to do it, for this counting method to be valid, your procedure must have the property that changing any decision in the sequence changes the final outcome. For instance, in the sequence we use above, if we change the pair of denominations, we get a different full house. If we change which of our pair of denominations is used to make the three of a kind, it changes the full house formed. If we change which three of the four cards of the denomination used for the three of a kind, it changes the hand, and likewise, finally, if we change our choice of the two cards chosen of the four of the denomination used to make the pair, then we get a different full house.

Let's look at another sequence of decisions which will lead to a full house. First choose any card, and then get two more of the same denomination. Then choose another card of a different denomination and choose another card of that denomination. This sequence of decisions can be done in

$$C(52,1) \cdot C(3,2) \cdot C(48,1) \cdot C(3,1) \text{ ways.}$$

Does changing any decision change the final outcome? Suppose step one you happen to get the ace of hearts as the result of your first choice. Then you choose the ace of spades and the ace of diamonds for the second step. Next you happen to choose the 8 of clubs for the third step, and then you get the 8 of hearts for the last card, resulting in aces over 8's. On the other hand, suppose your first choice was the ace of spades, instead of the ace of hearts, so we have changed our first decision. Next get the ace of hearts and the ace of diamonds for the second step. Thus, we have done the second step differently, but we have arrived at the same three aces. Then get the 8 of hearts for the next step and the 8 of clubs for the last card. We have changed our sequence of decisions, but we have produced the exact same full house. This means we would be over counting, as we would be counting different sequences of decisions as resulting in different full houses when in fact they are the same full house. If any one decision change fails to change the final outcome you will over count. You need to be careful when you think of a decision sequence to make sure that at each step, any change in decision will change the final outcome. Here is another sequence of decisions which will produce a full house. First choose a denomination, so there are obviously $C(13,1) = 13$ possibilities. Then choose three of the four cards having that denomination, which can be done in $C(4,3)$ ways. Then of the remaining twelve denominations, choose one, which can be done in $C(12,1) = 12$ ways, and then choose two of the four cards of that denomination, which can be done in $C(4,2)$ ways. This would lead to a sequence of decisions done in

$$(13) \cdot C(4,3) \cdot (12) \cdot C(4,2) \text{ ways.}$$

Notice that if you change any decision at any step here, it will change the resulting full house. Our original sequence was done in

$$C(13,2) \cdot C(2,1) \cdot C(4,3) \cdot C(4,2) \text{ ways.}$$

If these are the same, it means that we must have

$$(13) \cdot (12) = C(13,2) \cdot C(2,1).$$

But, in fact,

$$(13) \cdot (12) = P(13,2) = C(13,2) \cdot P(2,2),$$

and

$$P(2,2) = (2)(1) = 2 = C(2,1),$$

so in fact the two different procedures have the same number of ways of being done. It does not matter what sequence of decisions you use, if your sequence has the property that it accomplishes the task and any change will change the final outcome and if any desired outcome can be accomplished via your sequence, then it will give the correct number of ways.

7. FRIDAY 31 JANUARY 2014

We have discussed some formulas convenient for counting. In particular, a useful notation we will use is $C(n,k)$ as the symbol for the number of ways to **choose** k things from a set of n things. Notice here, there is not consideration as to any ordering of the chosen objects. We will use the notation $P(n,k)$ to denote the number of ways to **arrange** k things chosen from a set of n things, in the sense that each chosen object now has a specific destination, so there are k different destinations and which object ends up in each of the final locations is important. Thus here the order definitely matters. As an example, if a mailman has 12 letters and he chooses 5 to put in a mailbox, then it can be done in $C(12,5)$ ways, whereas if he has 5 different mailboxes and puts one letter in each mailbox, then that can be done in $P(12,5)$ ways, as which letter goes into which mailbox is now important in counting the number of ways to accomplish the task. We can notice that the task of arranging k things taken from a set of n things can be accomplished in two steps, where the first step is to simply choose k things from the set and then the second step is to arrange the k chosen things. The first step can be done in $C(n,k)$ ways, whereas the second step can now only be done in $P(k,k)$ ways. From this we see that

$$P(n,k) = C(n,k) \cdot P(k,k).$$

Last time we also observed that

$$P(k,k) = k! = 1 \cdot 2 \cdot 3 \cdots k,$$

and

$$P(n,k) = n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!},$$

so

$$C(n,k) = \frac{n!}{k!(n-k)!}.$$

For computing $C(n,k)$ when n and k are small, say less than ten, we can now observe they can be rapidly computed using **Pascal's Triangle**,

$$C(0,0)$$

$$C(1,0)\ C(1,1)$$

$$C(2,0)\ C(2,1)\ C(2,2)$$

$$C(3,0)\ C(3,1)\ C(3,2)\ C(3,3)$$

$$C(4,0)\ C(4,1)\ C(4,2)\ C(4,3)\ C(4,4)$$

and so on.

We easily see that this array is

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

and so on.

Notice that each number is the sum of the two numbers just above it on the left and right. The principle of Pascal's Triangle is

$$C(n+1, k+1) = C(n, k) + C(n, k+1), \ k \leq n.$$

To see that this is true, imagine a bucket containing n white blocks and one red block, for a total of $n+1$ blocks. Your job is to choose $k+1$ blocks from the bucket with $k \geq 0$. In any given choice, you either did or did not get the red block, so the total number of ways to do this is the number of ways to do it where you do get the red block plus the number of ways where you do not get the red block. Now to choose $k+1$ blocks where one of them is the red block, you have to first get the red block (only one way to do this), and then choose k of the white blocks which can be done in $C(n, k)$ ways. Thus there are $C(n, k)$ ways to choose $k+1$ blocks making sure one of them is the red block. On the other hand, to choose $k+1$ blocks so as to not get the red block, then you must choose all $k+1$ blocks from the available n white blocks, which can be done in $C(n, k+1)$ ways.

As an example, when we calculate probabilities for various 5 card hands from a standard deck, the denominator is always going to be $C(52, 5)$ which we calculated once and for all and we know it is

$$C(52, 5) = 2598960,$$

whereas most of the other choices we need to calculate will often be for $C(n, k)$ where $n = 4$, which we easily get from Pascal's Triangle.

Remember, that $C(n, k)$ counts choices where order does not matter and $P(n, k)$ counts arrangements with order mattering. Midway between these two would be **Partial Ordering**. Suppose that we have a set of n objects and we have r numbered locations (think of buckets or mailboxes) numbered successively with the numbers $1, 2, \cdots, r$. Suppose that we are given whole numbers $n_1, n_2, n_3, \cdots n_r$ with

$$n_1 + n_2 + n_3 + \cdots + n_r = n,$$

the task of putting n_1 things from the set into location 1, n_2 things from the set into location 2, $\cdots$,and finally, n_r things from the set into location r. Each way of doing this is called a **Partition** of the set into r named subsets of $(n_1, n_2, n_3, \cdots, n_r)$ objects. We denote the number of ways to do this by $C(n; n_1, n_2, n_3, \cdots, n_r)$.

Thus,

$$C(n; n_1, n_2, n_3, \cdots, n_r) = \text{ the number of partitions of a set of } n \text{ objects}$$
$$\text{into } r \text{ named subsets of } (n_1, n_2, n_3, \cdots, n_r) \text{ objects.}$$

As an example, suppose that we have a bucket containing 4 red blocks, 5 blue blocks, and 6 green blocks. We are given three buckets labelled with the numbers 1,2,3. We have to put 4 blocks in the bucket number 1, 5 blocks in the bucket number 2, and 6 blocks in the bucket number 3, and we are completely color blind. What is the probability that we get the 4 red blocks in the first bucket, the 5 blue blocks in the second bucket, and the 6 green blocks in the third bucket? Thinking in terms of the number of ways to accomplish the task of putting 4 blocks in bucket number 1, 5 blocks in bucket number 2, and 6 blocks in bucket number 3, we see this is $C(15; 4, 5, 6)$. This means the probability of getting this done correctly if we are color blind is $1/C(15; 4, 5, 6)$.

We need a formula for $C(n; n_1, n_2, n_3, \cdots, n_r)$. Now, we notice that the partition actually partially arranges the n objects. We can complete the arrangement by arranging the n_1 objects in the first location, then arranging the n_2 objects in the second location, and so on, finally arranging the n_r objects in the last location. This means that

$$n! = P(n, n) = C(n; n_1, n_2, n_3, \cdots, n_r) \cdot (n_1!)(n_2!)n_3!) \cdots (n_r!),$$

and therefore

$$C(n; n_1, n_2, n_3, \cdots, n_r) = \frac{n!}{(n_1!)(n_2!)n_3!) \cdots (n_r!)}.$$

For example, by a **word** in mathematics, we mean only a string of symbols. If we are to arrange the letters in the word FAST into a four letter word, then obviously there are 4! ways, but if we are to arrange the letters in the word LOOK, then there are not as many arrangements, since the O's are indistinguishable. To make the O's distinguishable, we can imagine that we tag the O's with numbers 1 and 2, that is one of the O's is tagged with a 1 and the other O is tagged with the number 2. With the tags, the O's become distinguishable, so there are 4! ways to arrange the letters with the tags. But, we can first arrange the untagged letters which can be done in say x ways, and then put the tags on the O's which can be done in two ways, so it must be that

$$2x = 4!, \text{ so } x = \frac{4!}{2}.$$

If we had to arrange the letters in LOOOK, then there are 3! ways to attach the tags, and the result would be

$$x = \frac{5!}{3!}.$$

To arrange the letters in the word MISSISSIPPI, we see then that number of words is $C(11; 1, 4, 4, 2)$, since we can also tag the M, and there is only one way to do that.

In reviewing probability, remember that the rules which always hold are

$$P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B).$$
$$P(A|B)P(B) = P(A \,\&\, B) = P(B \,\&\, A) = P(B|A)P(A).$$
$$P(\text{Sure}) = 1, \text{ and } P(\text{not } A) = 1 - P(A).$$

Notice that in particular, the conditional probability formulas allow for a kind of reversal:

$$\frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)},$$

so,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

As an example, suppose that ACME WIDGET CORP has three factories:1,2,3, that factory 1 produces 60 percent of the widgets, factory 2 produces 30 percent of the widgets, and factory 3 produces 10 percent of the widgets. Of course, some of the widgets produced will be defective, and we suppose that 2 percent of the factory 1 output is defective, 3 percent of the factory 2 output is defective, and 6 percent of the factory 3 output is defective. Darn that factory 3. A customer calls and complains he received a defective widget. Should we immediately blame it on factory 3? We should calculate the probability a widget comes from each factory given that it is defective. Let D be the statement a widget is defective and k be the statement that a widget comes from factory k. Our information is

$$P(1) = .6,\ P(2) = .3,\ P(3) = .1,$$

$$P(D|1) = .02,\ P(D|2) = .03,\ P(D|3) = .06.$$

We need to calculate the probabilities

$$P(1|D),\ P(2|D),\ P(3|D).$$

To do this, we need to calculate $P(D)$. But,

$$P(D) = P(D\ \&\ 1) + P(D\ \&\ 2) + P(D\ \&\ 3)$$

$$= P(D|1)P(1) + P(D|2)P(2) + P(D|3)P(3) = .012 + .009 + .006 = .027,$$

which means, finally,

$$P(D) = .027.$$

Thus,

$$P(1|D) = \frac{12}{27} = \frac{4}{9},\ P(2|D) = \frac{9}{27} = \frac{3}{9},\ \text{ and } P(3|D) = \frac{6}{27} = \frac{2}{9},$$

and this means that factory 3 is the least likely to be at fault for producing the defective widget causing the specific complaint, and it most likely came from factory 1.

APPENDIX: WORDS AS FUNCTIONS

In dealing with words generally, we can see that any word can be viewed as a function. The set of symbols allowed in the word is the **alphabet** used to make the word, and the number of letters in the word determines the domain, thought of as the set of positions available in which to put letters. Thus, a word is a function which assigns each position a letter from a specified alphabet. For instance, if we consider 5 letter words, then we can form the domain set $\mathbf{5} = \{1, 2, 3, 4, 5\}$, the set of available positions, and if the alphabet is $\mathbf{a} = \{K, M, O, U\}$, then each function $f : \mathbf{5} \longrightarrow \mathbf{a}$ determines a word according to the scheme

$$f : \mathbf{5} \longrightarrow \mathbf{a} \text{ determines the word } f(1)f(2)f(3)f(4)f(5).$$

For instance, if $f(1)$ =M, $f(2)$ =O, $f(3)$ =O, $f(4)$ =K, $f(5)$ =M, this determines the word MOOKM. However, the assignment of positions can be viewed physically as accomplished by having buckets labeled with the alphabet letters in set **a**, where assigning a position to a letter is accomplished by putting that position number into the bucket with that letter. Thus, a partition is a particular form of function where the number of positions to be put in each bucket is specified. Thus, the set of all functions from **5** to **a** is $\mathbf{a^5}$, so it follows that there are 4^5 functions, which means there are 4^5 five letter words which can be made with this alphabet, whereas $C(5; 3, 2, 0, 0)$ is the number of 5 letter words having three K's, two M's, zero O's and zero U's. Thus,

$$P(3 \text{ K's and } 2 \text{ M's} \mid \text{ five letter word from alphabet K,M,O,U}) = \frac{C(5; 3, 2, 0, 0)}{4^5} = \frac{5!}{(4^5)(3!)(2!)},$$

keeping in mind the convention that $0! = 1$. In general, if **a** is any alphabet and $\mathbf{n} = \{1, 2, 3, \cdots, n\}$, then the set of $n-$letter words which can be made is $\mathbf{a^n}$ and if the alphabet **a** has exactly r symbols, then there are r^n possible $n-$letter words which can be made with this alphabet.

We can also note here, that if we include the blank space and punctuation marks in our alphabet and all lower case and capital letters, we get a set of about forty symbols and a 1000 page novel with 100 lines per page and 100 characters (including blank spaces and punctuation) per line would simply be a word of length 10^{11}, a word of length one hundred billion. Then, the number of such words would be forty raised to the power one hundred billion, which is a number which is astronomical even in comparison with the size of the universe. The chance that such a word would actually make sense as a one thousand page novel is thus astronomically small.

We noted in our discussion above that by convention $0! = 1$. To see why this makes sense, recall that

$$C(n, k) = \frac{n!}{k!(n-k)!},$$

and in our discussion of Pascal's Triangle, we have $C(0, 0) = 1$. If this is to be reconciled with the previous formula, then

$$1 = C(0, 0) = \frac{0!}{0!0!},$$

so

$$(0!)^2 = 0!.$$

This means that 0! is one or zero. However, it cannot be zero, as our expression for $C(0, 0)$ in terms of factorials would be zero divided by zero which cannot possibly make sense, so the only possibility which does make sense is $0! = 1$.

DEPARTMENT OF MATHEMATICS, TULANE UNIVERSTIY, NEW ORLEANS, LA 70118
E-mail address: `mdupre@tulane.edu`