THE EXPECTATION PRIMER
EXPECTATION COVARIANCE PROBABILITY V3.2

MAURICE J. DUPRÉ

1. INTRODUCTION

These notes are designed to complement an elementary course in probability and/or statistics. The aim is to develop all the mathematical tools required using only elementary algebra starting from five basic assumptions about expectation which are easily motivated. No attempt is made to provide sufficient examples or exercises. Later sections present more advanced material which requires more detail, but nothing more than elementary algebra is ever required of the reader. A more sophisticated reader can easily supply the more advanced interpretations required. For more background on the foundation of the five basic assumptions including the multiplication rule for conditional expectation and conditional probability, the interested reader should consult references [1], [2], and [3] in the bibliography. In particular, here we have not restricted attention to the usual notion of random variable, but rather, as in [1] and [2], the foundations are developed for a general notion of unknown which in particular leads to a more generally applicable notion for expectation and the more general Laplacian or Bayesian notion of probability. This more general foundational development given in [2] and with a more elementary treatment in [1] includes the proof of the five basic properties of expectation including the multiplication rule for conditional probability based on not much more than the requirement that logical consistency should be maintained. In the section below on conditional expectation, the multiplication rule for conditional expectation is made plausible for unknowns and random variables, purely under the assumption that conditional expectation should relate to "unconditional" expectation. That is to say expectation only under the background information, which for random variables works through a simple restriction and extension process and which leads to the multiplication rule following from a simple renormalization procedure.

2. UNKNOWN NUMBERS

Probability and statistics deals with unknown numbers and their relationships to one another. Generally, we can think of an UNKNOWN NUMBER as a description of a numerical value insufficient possibly to determine its exact value. If a dice is in a box so we cannot see which face is on top, then the number on top is an an example of an unknown number. One of the main aims in dealing with unknown numbers here is to know the likelyhood of various possible values. For instance, for the dice in the box, as described, there is no information allowing us to be able to guess any value as being more likely than any other, so as specified, all of the six possible values must be considered equally likely. For unknowns in general, it turns out that this problem of determining how likely various values are to be observed can mathematically be reduced to the problem of determining an optimal guess for each of the

Date: April 11, 2007.

unknown numbers in a related family of unknowns and in many situations, the optimal guess and a measure of the guess for the squared error of the guess suffice to give a practical way to determine how likely various values are to be observed. We call this optimal guess the MEAN or EXPECTED VALUE of the unknown. In general, for any given situation, there are many unknown numbers. Clearly given two of them, say X and Y, we can add and multiply them by doing so to their values, forming $X + Y$ and XY. As an example, for the dice in the box, if X denotes the number that comes up on the top face and Y denotes the number on the bottom face, then $X + Y$ and XY make sense. If you are familiar with dice, you will recognize that for the dice example here, in fact always $X + Y = 7$, that is the numbers on opposite faces of a standard dice always add up to 7. In particular, we see ordinary numbers or what we call constants, are unknowns of the simplest kind, since we actually know their definite value. Of course, the same considerations apply to unknowns generally, that is in any situation or set up, there are usually many unknown numbers that can be specified, and such unknown numbers can be added and multiplied. In some situations, we have a physical set up in which there is a clearly defined set of possibilities, such as for the dice in the box. In such a case, it is customary to refer the the unknown as a RANDOM VARIABLE. To handle this mathematically, we describe the set of all possible outcomes with the precision to capture the essential information about the outcomes. The resulting set is called the SAMPLE SPACE and is customarily given the symbol S. Then a random variable assigns a number to each outcome in a definite way. For instance, if S is a set of people, we can consider the experiment of choosing a person at random from S. If the random variable X is weight in pounds, then X assigns each person in S his weight in pounds. If Joe belongs to S, then $X(Joe)$ is used to denote Joe's weight in pounds. Notice that there is nothing random about Joe's weight. The randomness in a random variable comes from the randomness of the outcome as to whose weight we are to observe. Mathematically, this means that a random variable is in particular a FUNCTION whose domain is the sample space S and whose values are in the set of all real numbers. If all we know about a is that a belongs to S, then $X(a)$ is an unknown number. Thus if S is a set of people and X is weight as before, then $X(a)$ is the weight of person a so is an unknown number, since we do not have any information about who a is other than a belongs to S. Thus, in this situation, the a is really playing no role other than as a place holder for something in S and may as well be dropped from the notation, and we will say X itself is the unknown here. If S is a finite set, then every real valued function whose domain is S constitutes a random variable. If S is an infinite set, then such is not the case-there may be real-valued functions on S which are not random variables. However, sums and products of random variables are again random variables. If $A \subset S$ and X is a random variable on S, then it is useful to denote by $X(A)$ the set of values that X assumes for outcomes in A. In particular, $X(S)$ is the set of all values which X may take on. We denote by $\mathbb{R}$ the set of all real numbers, so $X(A) \subset \mathbb{R}$. Obviously, if S is finite, then so is $X(S)$. In general, for an unknown X, there may not be any reasonable formulation of a sample space, but we can still speak of the possible values which X may have, which of course can depend on our information. That is, we know that all possible values of X belong to the set of all real numbers, or, in other words, the set of possible values of X is a subset of $\mathbb{R}$. Suppose that V denotes the set of all possible values of X given our information about X. We then certainly have $V \subset \mathbb{R}$. We say X is SIMPLE, if V is finite. Clearly when we add or multiply simple unknowns, the results are simple unknowns. In particular, if we have a sample space S which is finite, then all random variables on S are simple. Consider the experiment where we repeatedly roll a dice over and over until we roll a 5 and X is the number of rolls it takes to roll a 5. The set $V = X(S)$ of possible values for X in this case is the set of all positive whole numbers, which is infinite. However, we say that this set forms a countable infinity. Whenever the members of a set can be listed in a possibly infinite list, we say the set is COUNTABLE. If V is a countable set, then we say that X is DISCRETE. In applications, values of unknowns and random variables are usually observed either by counting or by measuring. In the case where we count, the unknown will then be discrete. Even though the set of values in this case may be infinite (as in the example of rolling the dice until we get a 5), the infinity is countable. Here, it can be shown that again, if we add or multiply discrete unknowns, the results will be discrete unknowns. In the case where a measurement process is used, such as a scale to measure weight or a ruler to measure length, the possible values can be considered to cover a "continuous" range of possibilities, so we say the unknown is CONTINUOUS. Unfortunately, a continuous infinity of possible values is not countable. Now in a measurement process we always have to deal with the level of accuracy of measurement. For instance, if we are weighing people, we need to decide in advance the number of decimal places used in reporting the weights and in measuring. To deal with this, if X is an unknown, and n is a positive integer, we can denote by $R_n(X)$ the result of ROUNDING OFF X to n decimal place accuracy. Naturally, we must assume that if X is a random variable, then so is $R_n(X)$. Besides using the usual rounding process, we can instead just round down (changing all digits past the n decimal place to 0) to get the LOWER ROUND which we denote by $L_n(X)$, or we could round up getting the UPPER ROUND which we denote by $U_n(X)$, and which just raises the digit in the n decimal place (if that digit is less than 9, or in case the n decimal place is 9, we change it to 0 and consider the previous or $n-1$ decimal position). It is useful to be able to use any one of these round off procedures to any number of decimal places and get a random variable as a result if we start with a random variable. That is, if X is a random variable, then so are $L_n(X)$,$R_n(X)$, and $U_n(X)$, for each nonnegative integer n. Actually, it suffices to know that we can take the integer part of a random variable and the result is a random variable. For, if $int(X)$ denotes the largest integer less than or equal to X, then $L_n(X) = (1/10^n)(int(10^nX))$, whereas $U_n(X) = (1/10^n)(int(1 + 10^nX))$. Also, $R_n(X) = (1/10^n)(int(.5 + 10^nX))$, so as we can always multiply random variables by ordinary numbers, if the integer part of a random variable is again a random variable, then these equations show that the various round offs will again be random variables. Of course, since the integer part of an unknown is a discrete unknown, this means that the various round offs of an unknown are discrete. It is useful to notice that for any unknown X, we have both

$$L_n(X) \leq X \leq U_n(X) \tag{2.1}$$

and

$$L_n(X) \leq R_n(X) \leq U_n(X). \tag{2.2}$$

Consequently, using absolute value

$$|x| = \sqrt{x^2}, \tag{2.3}$$

and remembering that $|x - y|$ =distance from x to y on the number line, as both X and $R_n(X)$ are between $L_n(X)$ and $U_n(X)$, it must be the case that

$$|X - R_n(X)| \leq U_n(X) - L_n(X) \leq \frac{1}{10^n}. \tag{2.4}$$

This will mean that in case of a continuous unknown, we can approximate to whatever degree of accuracy we might require with unknowns which are discrete (notice again that $L_n(X)$,$R_n(X)$, and $U_n(X)$ only assume at most a countable infinity of values). We say that the unknown X is BOUNDED if there is a non negative number b so that $|X| \leq b$. We can note here that if X is a bounded unknown, then all the various round offs of X are simple unknowns.

In general, if X is an unknown, then its description is a statement which we understand with our background information and that we will assume can be put in the form of a statement B. In general, if A is any statement, our background information B may not tell us whether A is true or false. In this situation, we can form an unknown I_A called the INDICATOR of A, defined by saying I_A has value 1 if A is true whereas I_A has value 0 if A is false. Thus the value of I_A exactly tells us whether A is true or false-knowing the value of I_A is therefore equivalent to knowing whether A is true or false. For instance, notice that XI_A is a new unknown and has the value that X itself has if A is true, whereas XI_A is simply 0 if A is false. We shall see this new unknown XI_A is the key to how we learn about X as new information about X is discovered. It also turns out to be the key to how to best guess a value for X, as we shall see next.

3. GUESSING AND EXPECTATION

As mentioned in the previous section, the first thing of interest to be determined about an unknown is an optimal guess, based on the information at hand. The sense in which our guessing is optimal will become clear later as it turns out to be a result of the theory we are developing. We will think of the information at hand as consisting of a statement B of what we can think of as background information which we use to make our guess. If X is an unknown we use $E(X|B)$ to denote our optimal guess which we will call the MEAN or EXPECTED VALUE of X GIVEN B. We also often find it convenient to use the Greek letter μ, pronounced "mu", to denote the mean, so μ_X denotes $E(X|B)$, or we simply use μ or $E(X)$ for $E(X|B)$ if there is no confusion as to what X is or what B is as the case may be. The expectation E has some simple fairly obvious properties (here X and Y are any unknowns, c is any constant, A and B are any statements of information):

$$E(X+Y|B) = E(X|B) + E(Y|B), \tag{3.1}$$

$$E(cX|B) = cE(X|B), \tag{3.2}$$

$$E(X|B) \geq 0, \quad \text{if} \quad X \geq 0, \tag{3.3}$$

$$E(1|B) = 1, \tag{3.4}$$

and finally,

$$E(XI_A|B) = E(X|A\&B)E(I_A|B). \tag{3.5}$$

These five properties are the BASIC PROPERTIES OF EXPECTATION which we shall assume. As immediate consequences of these basic properties we have, with fixed background information B, for any constants a and b that:

$$E(aX \pm bY) = aE(X) \pm bE(Y) \tag{3.6}$$

$$E(X) \geq E(Y) \quad \text{if} \quad X \geq Y \tag{3.7}$$

$$E(c) = c. \tag{3.8}$$

For instance, [3.6] is the result of combining [3.1] and [3.2], whereas [3.7] follows from[3.3] and [3.6] and the fact that $X \geq Y$ is the same as $X - Y \geq 0$. On the other hand, [3.8] is an immediate consequence of [3.2] and[3.4].

More generally, if $X, X_1, X_2, X_3, ..., X_n$ are all unknowns and $c_1, c_2, c_3, ..., c_n$ are any numbers and if

$$X = c_1X_1 + c_2X_2 + c_3X_3 + ... + c_nX_n, \tag{3.9}$$

then repeated application of [3.6] gives

$$E(X) = c_1E(X_1) + c_2E(X_2) + c_3E(X_3) + ... + c_nE(X_n). \tag{3.10}$$

In mathematics, it is customary to call the expression such as for X on the right hand side of [3.9] or for $E(X)$ on the right hand side of [3.10], a LINEAR COMBINATION. Thus [3.10] says to compute the expected value of a linear combination just use the same corresponding linear combination of expected values or means:

$$E(c_1X_1 + c_2X_2 + c_3X_3 + ... + c_nX_n) = c_1E(X_1) + c_2E(X_2) + c_3E(X_3) + ... + c_nE(X_n). \tag{3.11}$$

For instance, if we know that $E(X) = 5, E(Y) = 6$, and $E(W) = 7$, then

$$E(4X - 3Y + W) = 4(5) - 3(6) + 7 = 9.$$

The first basic property of expectation, [3.1], is known as the ADDITIVITY PROPERTY. The second basic property, [3.2], is known as the HOMOGENEITY PROPERTY. The third basic property of expectation, [3.3], is known as the POSITIVITY PROPERTY, and the fourth basic property of expectation, [3.4], is known as the NORMALIZATION PROPERTY. Finally, the last property of expectation is known as the MULTIPLICATION RULE. The first two properties together are called LINEARITY. That is to say that if a proposed expectation model (meaning a guessing method) is known to satisfy the first two basic properties of being additive and homogeneous, then it is called LINEAR. If in addition the positivity property is satisfied, then it is POSITIVE LINEAR. If the normalization property and linearity properties are satisfied, then we have a normalized positive linear model. Thus, keeping the background information fixed we can say the expectation given a fixed B is a normalized positive linear model. The purpose of the multiplication rule is then to give the expectation the ability to learn to make "better" guesses when new or "better" information is available. In some sense then, the five rules together are a model for a process for learning about the world.

MOTIVATION OF BASIC PROPERTIES. As for motivation of the basic properties of expectation, it often becomes clearer when we think in terms of money. Suppose that X is the value in dollars of a retail business which we must buy and which must be resold, hopefully at a profit. Our information B may not tell us the exact value of the business, but we may be able to use it to arrive at an optimal guess which is then $E(X|B)$. Notice that we must offer enough to be able to strike a deal, but not too much since we want to make a profit. Thus, there is a penalty for being wrong and we see that the bigger our error, the bigger the penalty. Suppose that Y is the value of another retail business which we are thinking of buying for the same purpose. Suppose that we must sell the business we buy within two weeks and possibly take a loss. We need to be careful. If we think that the first business is worth 2 million dollars and the second business is worth 3 million dollars, then it is certainly reasonable that we would think both together are worth 5 million dollars. This is exactly what additivity, [3.1], says. If we wish to consider the value of the first retail business in French Francs and the exchange rate is 4 Francs to the dollar, then clearly we must think that the value in Francs is 8 million Francs. We do not have to completely rethink our guess in terms of Francs, we just convert our guess in dollars directly to Francs. This is what [3.2] says. If instead, X is the net profit in dollars from selling the first retail business and if our advance information tells us $X \geq 0$, then certainly, $E(X|B) \geq 0$, and this is clearly the content of positivity, [3.3]. If in fact B tells us that $X = 1$, then certainly $E(X|B) = 1$, and in particular, $E(1|B) = 1$, which is the normalization property, [3.4].

For a random variable view, using a different example, we can instead think of X as the daily net profit of a retail store and Y as the daily net profit of a nearby coffee shop, then a businessman who happens to own both of these knows that $X+Y$ is the total daily net profit from the two businesses. One thing this business man would be interested in is the long run average daily net profit. If you ask him to guess the daily net profit for the retail store on Tuesday after next, and if he has no reason to think of that day as anything special, he will probably guess the long run average. In this case, we are thinking of the sample space as the set of all possible days for which both of these businesses operate. Clearly, if the business man thinks the retail store (in the long run) averages a net profit of 3000 dollars a day and the coffee shop averages a daily net profit of 7000 dollars a day, then he should think the two businesses together on average are netting $10,000$ dollars. He does not have to go back into the records and add up the daily net profits for the two businesses (over a long run) and average the resulting totals. This is exactly what additivity, [3.1], says. Moreover, if you ask him to guess the profit from the retail store for some particular day which he has no reason to think of as being anything other than ordinary, he will likely think it reasonable that 3000 dollars is a good guess. If a German business man inquiring about buying his retail store wants to know the average daily net profit in German marks, and if a dollar is worth 3 German marks (we wish), then the store owner knows immediately that the average daily net profit in German marks is simply 3 times 3000, or 9000 German marks. Again, there is no need to go back into the records and convert each days net profit into German marks and recompute the long run average. Alternately, we are saying here that the long run average can be calculated with any new units by converting the long run average value directly to the new units. This is what [3.2] says. If the coffee shop never suffers a net loss, then obviously its long run average daily net profit cannot be negative,and this is clearly the content of positivity, [3.3]. In fact, if it nets 1 dollar every day, then the long run average daily net profit must be 1 dollar, which is the normalization property, [3.4]. This indicates that one way to arrive at a guessing procedure for expectation is to try to guess the long run average, in the case of random variables. In either case, we therefore see that the first four properties of the expectation are virtually undeniable and so we have simply taken them as axioms, like in Euclidean geometry.

The multiplication property is more problematic. We will see later that for random variables we can view the multiplication rule as simply a way to restrict our attention to the case where the new information is true so as to simultaneously preserve the normalization property. However, the situation for the more general unknowns is really based on the idea of preserving logical consistency of our guessing procedure and for this discussion we refer the interested reader to [2] and [1].

Let us consider what rounding off does to the process of observing unknowns and expectations in more detail. In fact, we can use the round off procedure to approximate expectations to any degree of accuracy we like. Indeed, by [2.1], [2.2], and [3.7], we see that for any n we know that both $E(X)$ and $E(R_n(X))$ are between $E(L_n(X))$ and $E(U_n(X))$, and that

$$E(U_n(X)) - E(L_n(X)) = E(U_n(X) - L_n(X)) \leq E(10^{-n}) = 10^{-n}$$

so

$$|E(X) - E(R_n(X))| \leq \frac{1}{10^n}. \tag{3.12}$$

What this means is that no matter how large n is chosen, if we measure to n decimal place accuracy then expectations come out to n decimal place accuracy. For instance, if we are using information about weights in pounds measured to three decimal place accuracy in order to arrive at an optimal guess for a person's weight in pounds, then the result will in fact be the optimal guess to three decimal place accuracy.

Finally, we must add a note of caution. The background information is actually very important. Given an experiment, our information about the specific type of experiment (such as tossing a dice) determines the sample space and the set of all unknowns or random variables on the sample space, but our information as to the specific physical characteristics of the experiment (such as how the dice is loaded) determine the expectations of the random variables. For instance in the dice example, changing the loading of the dice will generally change the expectation of some or all of the random variables (except constants of course). Thus when we deal with a specific loading of the dice, we are dealing with a specific expectation model. In reality, what we are dealing with is different more specific background information about the dice. For if we are considering the example of the dice in a box, clearly it does not matter how the dice is loaded, it is just sitting there in the box. Alternately, if we have some information as a result of a very brief view inside the box, that changes the state of our information. Whatever the model, however, the expectation must obey the preceding rules and all their consequences. In applications of expectation to real physical problems, it is the job of the statistician, based on his information, to find an expectation model that reasonably approximates the true expectation model which may be difficult to know, in practice. For instance, by rolling a dice many times, the statistician begins to know approximately the loading of the dice, but (s)he can never know it with absolute certainty by just looking at the results of rolling the dice a finite number of times. For instance, suppose that I have a dice which I have loaded in such a way as to make the even numbers more likely than the odd numbers in a specific way which I know and you do not. If I give you the dice, if I do not tell you the dice are loaded and ask your guess as to what number will come up when the dice are tossed, then you should base your answer on the assumption that all outcomes are equally likely since you have no information which contradicts this. But after you toss the dice a finite number of times, you will begin to learn something about the way the dice are loaded, so your expectation model will be a better approximation of the true model which I know but you do not. No matter how many times you toss the dice, you can never really be certain as to how I have loaded the dice, so you will never be able to find the "true" expectation model, but rather you will learn an expectation model based on the information you receive from watching the dice. In general, the statistician must find a method of computing expected values which is approximately "correct" and in general, there are many possibilities, even when we restrict to expectation models satisfying the first four basic properties of expectation. When attempting to create a method of calculating expectations, given fixed background information, B, it is best to look for a way of assigning each unknown X a number $F_B(X)$ so that the first basic property of expectation-the additivity is satisfied: $F_B(X+Y) = F_B(X) + F_B(Y)$, because it is usually the most difficult to get and causes the most severe restrictions on what is possible to use for an expectation model. We would then check to see if the homogeneity property, that is the second basic property holds, which is usually very easy to determine, and then check the third basic property, the positivity property which is also usually easy to determine. Finally, we would look to see if the fourth basic property known as the normalization property holds, which means calculating $F_B(1)$ to see if we get 1. If the first three basic properties are satisfied but the normalization property is not, we do not really have a problem. If $F_B(1) \neq 0$, then as $1 \geq 0$ the positivity property guarantees that $F_B(1)$ is a positive number. We then define a new expectation model which we can denote E_{F_B} by requiring for any unknown X, that

$$E_{F_B}(X|B) = \frac{F_B(X)}{F_B(1)}. \tag{3.13}$$

It is now easy to calculate directly that if F_B satisfies the first three basic properties, then E_{F_B} satisfies the first four basic properties so is an expectation model for the given background information, B. That is if circumstances were pointing toward the model F_B, then we would choose E_{F_B} as our expectation model given B, so we would use $E(X|B) = E_{F_B}(X)$ for each unknown X. This process of forming E_{F_B} from F_B is called NORMALIZATION. Thus, the fourth basic property amounts to the assumption that normalization has been carried out. If it is the case that $F_B(1) = 0$, then it can be shown (see the appendix) that $F_B(X) = 0$ for every random variable X so such an attempted model would be useless from the start. It may seem from this discussion that the problem of determining the proper expectation model to use in a given situation is theoretically unsolvable. What is really going on however, as we shall see in more detail later, is that the expectation model is basically determined by our state of information about the physical setup of the experiment. As a statistician studies the actual physical experimental apparatus, he is continually refining his information about the apparatus which then causes a change in the expectation model used, through the multiplication rule, that is the last property. This will become clearer in a later section where we discuss conditional expectation and probability. In particular, in Jaynes, [3], general methods are discussed for determining the best expectation model determined by the available information.

4. PROBABILITY

So far, we have not mentioned probability at all. In fact, probability is really sort of a special case of expectation. That is, the expectation E applies to any unknown, but probability is the result of applying expectation to an indicator of a statement. Thus, if A is a statement and B is the statement of our background information, then we define the PROBABILITY of A GIVEN B, denoted $P(A|B)$, by the equation

$$P(A|B) = E(I_A|B), \text{ for any statements, } A, B. \tag{4.1}$$

Before we go any further we should point out that this means that the multiplication rule for expectation [3.5] says that

$$E(XI_A|B) = E(X|A\&B)P(A|B), \text{ for any unknown } X \text{ and any statements } A \text{ and } B. \tag{4.2}$$

Incidentally, when we have a sample space, statements about the outcome of an experiment are called EVENTS if their indicators are random variables. Notice an unknown X is an indicator (of some statement) if and only if it equals its own square, or in symbols,

$$X^2 = X. \tag{4.3}$$

This rather innocent equation means immediately that the only possible values X can have are 0 and 1, so for sure we have

$$0 \leq X \leq 1, \tag{4.4}$$

and therefore, we can conclude from [3.7] and [3.8] that

$$0 \leq E(X|B) \leq 1. \tag{4.5}$$

Therefore we have for probability

$$0 \leq P(A|B) \leq 1, \qquad \text{ALWAYS.} \tag{4.6}$$

Notice also, that if X is an indicator and if A is a statement, then to say $X = I_A$ is the same as saying the statement A is logically equivalent to the statement $X = 1$. In particular, the constants 0 and 1 are indicators. We will use the symbol $\emptyset$ to denote a statement which is false for sure, so that $I_\emptyset = 0$. We use the symbol S to denote a statement which is true for sure so that $I_S = 1$. It is customary to call S the SURE STATEMENT since it is true for sure, and we will call $\emptyset$ the SURELY FALSE STATEMENT since it is false for certain. Thus

$$P(S|B) = 1 \quad \text{and} \quad P(\emptyset|B) = 0, \quad \text{for any} \quad B. \tag{4.7}$$

When we have a sample space, any statement about the outcome of the experiment can be used to define a function on the sample space whose only values are 0 and 1. We simply declare that the truth value of the statement gives the value of the function via

$$TRUE = 1 \qquad and \qquad FALSE = 0. \tag{4.8}$$

This is clearly the indicator of the statement viewed as a function on the sample space. For instance, if we are dealing with a dice in a box which we cannot see, and if A is the statement that the number on top is even, then that statement has unknown truth value. But if the number up is 2, then I_A has the value 1, whereas if the number up is 3, then I_A

has the value 0. The truth value of A depends or is a function of the number on top, so likewise, I_A is a random variable on the set of outcomes. In general when we have a sample space, if the indicator of a statement is actually a random variable, then the statement is an EVENT. Thus, if the sample space is finite, then any statement about the outcome of the experiment is an event. In case we have a sample space, a statement determines a set of outcomes, namely, the set of all outcomes for which the statement is true, or equivalently the set of all outcomes for which the indicator value is 1. For instance, if we are dealing with a dice in a box, the sample space is considered to be the set $S = \{1, 2, 3, 4, 5, 6\}$ and we see that if we consider the set of outcomes for which A is true (here again we take the example where A says the number on top is even), we find it is $\{2, 4, 6\}$, as this is exactly the subset of outcomes for which A is true, or equivalently, the set of outcomes for which I_A takes the value 1. We will use capital letters near the beginning of the alphabet for events just as for statements in general. Thus, if A is an event, then we can think of A as simply consisting of the set of all the outcomes for which statement A is true or for which the value of I_A is 1, and thus also think of A in terms of its indicator I_A. All three ways are useful, and being able to transfer your thinking from one to the other can be very useful. For instance, as the sure statement S is true for sure so $I_S = 1$, it follows that whenever we have a sample space, then as a set it must be the case that S is the set of all possible outcomes, which is consistent with our previous discussion using S to denote the sample space.

Technically speaking, if S is uncountably infinite, as it is the case that possibly not every function with domain S defines a random variable, in fact, possibly not every indicator defines a random variable. This means that if the sample space is uncountably infinite, then possibly not every subset of S is an event, and consequently, possibly not every statement about the outcome is an event. We will not need to deal with these technicalities until later, however.

Now, let's see how indicators can combine to form new indicators and how the combinations relate to logic and set theory. If X and Y are indicators, then clearly so is XY, their product. As statements, suppose $X = I_A, \quad Y = I_B$, and $XY = I_C$. As XY is the indicator of the statement C, then C must be logically equivalent to $A\&B$, since the only way that XY can equal 1 (that is for C be true) is for both X and Y to be equal to 1 (that is for both A and B to be true), no matter what other background information we have. Thus we can regard XY as the indicator of the statement $A\&B$. If we have a sample space, as a set of outcomes, $(A\&B)$ is the set of outcomes common to both A and B, that is to say their common overlap which in set theory is known as their INTERSECTION, denoted $A \cap B$, so finally we have

$$I_A I_B = I_{(A\&B)} = I_{A\cap B}, \quad \text{and} \quad (A\&B) = A \cap B \tag{4.9}$$

giving the three ways to interpret the overlap of two events, the first as a random variable, the second as a statement, and the third as a set of outcomes. We can now use [4.9] and the multiplication rule [4.2] with $X = I_C$ to get a special form of multiplication rule known as the LAW OF CONDITIONAL PROBABILITY:

$$P(C\&A|B) = P(C|A\&B)P(A|B), \text{ for any statements } A, B, C. \tag{4.10}$$

Unfortunately, if we add two indicators we do not necessarily get an indicator. For notice that when we form the indicator $X + Y$, if both A and B are true, then the value of $X + Y$ will be 2, not allowed for an indicator. We therefore need to subtract the value 1 exactly when both are true. This means that $X+Y-XY$ is an indicator. It is now an easy exercise to interpret this indicator as the indicator of the statement (A or B) and in case there is a sample space, as $A \cup B$ as a set of outcomes, that is the UNION of A and B as a set of outcomes. To summarize,

$$(A \text{ or } B) = A \cup B \quad \text{and} \quad I_{A \text{ or } B} = I_A + I_B - I_A I_B. \tag{4.11}$$

Since the probability of a statement is just its expectation as an indicator, from [4.11] and the first consequences [3.6] and [3.8] of the basic laws of expectation we have

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \ \& \ B) \tag{4.12}$$

We say that statements A and B are MUTUALLY EXCLUSIVE when $A\&B = \emptyset$, that is when $A\&B$ is certainly false. In case where we have a sample space, this means as sets, A and B are disjoint, or do not overlap. For this special case we have, since $P(\emptyset) = E(0) = 0$, that [4.12] reduces simply to

$$P(A \text{ or } B) = P(A) + P(B), \quad \text{when } A \text{ and } B \text{ are mutually exclusive.} \tag{4.13}$$

The basic laws of probability are [4.6],[4.13], and the first half of [4.7], since using algebra these can be used to construct the expectation. That is, it is mathematically the case that once the probability of all events are determined, then the expectation of all random variables can be mathematically determined. This however requires a lot more mathematical effort than getting the laws of probability from the rules of expectation which are themselves ([3.1],[3.2],[3.3],[3.4]) very simple and easy to understand.

Besides using "and" and "or" to combine statements, we can also use "but not". So, if A and B are statements or events, then so is (A but not B). If we have a sample space, as a set this is easily seen to be what in set theory is called the RELATIVE COMPLEMENT of B in A which is usually denoted by $A \setminus B$, that is the set of outcomes in A which do not belong to B. In the indicator picture, it's indicator is clearly just $I_A - I_A I_B$. We therefore have

$$(A \,\text{but not}\, B) = A \setminus B = A \setminus (A \cap B), \text{ and } I_{A \,\text{but not}\, B} = I_A - I_A I_B, \tag{4.14}$$

so, by [3.6],

$$P(A \,\textit{but not}\, B) = P(A \setminus B) = E(I_A - I_A I_B) = P(A) - P(A \cap B). \tag{4.15}$$

As a special case we have a logical operation on statements that we have not mentioned so far which is the simple act of negation. If A is a statement, then we can form its negation (not A), which is true exactly if A is false. If we have a sample space, as a subset of S, then (not A), consists of all outcomes for which A is false, namely the complement of A in S denoted $S \setminus A$. On the other hand, in terms of indicators, it is easily seen that simply

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

Thus we have

$$I_{\text{not} A} = 1 - I_A, \ (\text{not}\, A) = S \setminus A, \tag{4.16}$$

and consequently by [3.6] and [3.8],

$$P(\text{not} A) = 1 - P(A). \tag{4.17}$$

Notice that if A is any statement, and if $C = (\text{not}\, A)$, then $(A \,\text{or}\, C) = S$, whereas, A and C are mutually exclusive. More generally we say that statements $A_1, A_2, A_3, ..., A_n$ partition

S if they are all mutually exclusive, meaning any two are, and if $(A_1 \text{ or } A_2 \text{ or } A_3 \text{...or } A_n) = S$, which is to say at least one of these must be true. Thus exactly one is true and all the others are false. Then clearly we must have for their indicators

$$I_{A_1} + I_{A_2} + I_{A_3} + ... + I_{A_n} = I_S = 1, \tag{4.18}$$

so for probability we then have

$$P(A_1) + P(A_2) + P(A_3) + ... + P(A_n) = 1. \tag{4.19}$$

Better still, we see that as for any unknown X we have $X = 1X$, it follows that on multiplying both sides of [4.18] by X we have

$$X = XI_{A_1} + XI_{A_2} + XI_{A_3} + ... + XI_{A_n}, \tag{4.20}$$

and therefore, with any background information B,

$$E(X|B) = E(XI_{A_1}|B) + E(XI_{A_2}|B) + E(XI_{A_3}|B) + ... + E(XI_{A_n}|B). \tag{4.21}$$

Next, let us apply the multiplication rule [4.2] to each term on the right hand side of [4.21]. Then for each k with $1 \leq k \leq n$ the multiplication rule gives

$$E(XI_{A_k}) = E(X|A_k \& B)P(A_k|B),$$

and substituting this result for the terms on the right hand side of [4.21] for k taking values $1 \leq k \leq n$, gives

$$E(X|B) = E(X|A_1 \& B)P(A_1|B) + E(X|A_2 \& B)P(A_2|B) + ... + E(X|A_n \& B)P(A_n|B). \tag{4.22}$$

In particular, by taking $X = I_C$ to be the indicator of a statement C, in view of [4.1] we get

$$P(C|B) = P(C \& A_1|B) + P(C \& A_2|B) + P(C \& A_3|B) + ... + P(C \& A_n|B), \tag{4.23}$$

and consequently, in view of [4.9],

$$P(C|B) = P(C|A_1 \& B)P(A_1|B) + P(C|A_2|B)P(A_2|B) + ... + P(C|A_n \& B)P(A_n|B), \tag{4.24}$$

a very useful way in many situations of breaking up a probability computation into several mutually exclusive cases. As another useful application, if X is a simple unknown whose values are the numbers $x_1, x_2, x_3, ..., x_n$, we can let A_k be the statement that $X = x_k$ for each k. Then $XI_{A_k} = x_k I_{A_k}$ for each k, hence the $k-$th term on the right hand side of [4.20] becomes $x_k I_{A_k}$, so [4.20] becomes

$$X = x_1 I_{A_1} + x_2 I_{A_2} + x_3 I_{A_3} + ... + x_n I_{A_n}, \tag{4.25}$$

so by [3.6]

$$E(X|B) = x_1 P(A_1|B) + x_2 P(A_2|B) + x_3 P(A_3|B) + ... + x_n P(A_n|B). \tag{4.26}$$

Here we see that [4.26] tells us how to compute the expectation for simple random variables once we know the probability of each of its values. That is to say, [4.26] reduces the computation of expectation to the computation of probability. Notice that if you look in any elementary book on probability or statistics, you will find [4.26] is the usual way of defining expectation in terms of probability for simple unknowns. Also, [4.26] together with [3.12] tells us that the expectations of all unknowns in a given situation are actually determined by the probabilities of all relevant statements, which, in the case of an experiment with only finitely many outcomes, is in turn determined by the probabilities of the individual outcomes. We can also see that [4.26] is a special case of [4.22], because obviously for each k with $1 \leq k \leq n$ we have $E(X|A_k \& B) = x_k$, since if A_k is true, then we know $X = x_k$ and therefore x_k must be our guess for the value of X, which is $E(X|A_k \& B)$.

For calculations using the TI-83, we simply put the possible values in a list in the statistical editor and in the list right beside we put the corresponding list of probabilities in the same order so each value is beside its probability. We then use the 1-var stat operation at the top of the statistical calculation menu followed by the value list and probability list separated by a comma-always put the probability list last. When we hit the enter button, the TI-83 then gives the mean μ_X as $\bar{x}$, and the standard deviation σ_X as σ_x.

As we remarked above, in case the sample space is finite, as soon as we determine the probability of each individual outcome of the experiment we can determine the expected value of all the random variables. For instance if the experiment is to toss a dice, then we can take $S = \{1, 2, 3, 4, 5, 6\}$ and as soon as the probabilities

$$P(\{1\}), P(\{2\}), P(\{3\}), P(\{4\}), P(\{5\}), P(\{6\})$$

are determined, then all expected values of all the random variables are determined. But, what these six probabilities are depends on the specific loading of the dice. We say the dice is FAIR if all these six probabilities are equal, which is to say all outcomes are equally likely. More generally, if S is any finite sample space, we call the model of equally likely outcomes the assumption that all outcomes are equally likely. For any model, the rules of probability now dictate that the probability of each outcome is a non negative real number less than or equal to 1 and that the sum of these probabilities exactly add up to 1. Consequently, if S has exactly n outcomes, then in the model of equally likely outcomes each individual outcome must have probability $1/n$. In gambling situations, devices which operate according to the model of equally likely outcomes are usually described as fair. That is in gambling situations, gamblers are presented with some sort of physical experiment in which it appears that all outcomes should be equally likely, and consequently use that assumption to judge their bets. Consequently, when it appears all outcomes are equally likely, then we would say the device is fair if in fact all the outcomes are equally likely. Thus we speak of a fair dice or a fair roulette wheel or a fair drawing and so forth. There is however, a quite different meaning of the word fair in gambling which is used to determine payoffs in gambling, and in this other sense, casino gambling games are never fair. If person K is betting against person M in a gambling situation, if W is the random variable which gives the net amount K wins, then we say the game is fair provided that $E(W) = 0$. This means that neither party has an outright advantage and who wins is up to chance. Suppose that K is willing to bet b dollars that A will happen against a bet of c dollars by M that A will not happen. Now, $W = cI_A - bI_{(\text{not}A)}$, so if $P(A) = p$ and $P(notA) = q$, then $p + q = 1$ and $E(W) = cp - bq = cp - b(1 - p)$. For the game to be fair, then it must be the case that $0 = E(W) = cp - bq$, or $cp = bq$ which means $(b/c) = (p/q)$. Such ratios are what are usually referred to as betting odds. Thus if both K and M are willing to bet like this, if the bet is fair, then we would say that from their point of view the odds of A happening are b to c. For instance at the race track when the odds are ten to one on Gluefoot to win the race, that means that if you bet one dollar on

Gluefoot to win, then you win ten dollars in case Gluefoot wins the race (here K is the race track and M is the racetrack gambler). It is easy to see that specifying the betting odds on an event is equivalent to specifying the probability of the event that the gamblers must be tacitly agreeing on provided they both believe the bet to be fair. Actually if typical gamblers agree to bet, they each believe the probability of the event differs from this value, but each believes the difference is in his favor-they believe the bet is not fair. But at the racetrack, the odds are simply set by the overall bets placed by the gamblers themselves-a system known as paramutual betting. If $r = b/c = p/q$ is given, then as $q = 1 - p$ we have $(1 - p)r = p$, so $r - pr = p$, and thus $r = p + pr = p(1 + r)$. Therefore, $p = r/(1 + r) = b/(b + c)$ and $q = 1 - p = c/(b + c)$. That is if K and M both agree that the bet is fair, then they are agreeing that

$$P(A) = \frac{b}{b+c},\ P(notA) = \frac{c}{b+c}. \tag{4.27}$$

For instance for Gluefoot, if the odds are 10 to 1 that he wins the race then the probability that Gluefoot wins the race is 1/11 as seen by gamblers who think this a fair bet. In paramutual betting it means that for every dollar bet on Gluefoot to win the race there were 10 dollars bet on the other horses to win the race. Consequently, at the racetrack, the odds are changing as the bets are placed. This means that it is advantageous to wait until the last minute to place a bet so as to best know the betting odds, which of course all the gamblers know. The result is that everyone tries to place bets at the last minute and the odds can change very wildly in the last seconds before the horse race begins. It is interesting to note here that it is known that in horse racing the favorite horse wins roughly one third of the time-there is some accuracy in the overall evaluation of the probabilities given by the paramutual betting of a lot of gamblers.

Finally, we are now in a position to compute the optimal guess for the number up on the dice in the box. Since we cannot see inside the box, our information gives no preference as to any specific value, any of the numbers in $S = \{1, 2, 3, 4, 5, 6\}$ could be on top and all are equally likely according to our background information B. Thus, it must be the case, that if X is the number up, then $P(X = k) = 1/6$, for each $k \in S$. Then by [4.26], we find here

$$E(X|B) = (1)(1/6)+(2)(1/6)+(3)(1/6)+(4)(1/6)+(5)(1/6)+(6)(1/6) = 21/6 = 7/2 = 3.5, \tag{4.28}$$

so our optimal guess for the number on top is 3.5, according to our theory. One interesting thing we notice here is that the number up on the dice cannot be 3.5, but our theory is telling us that in some sense it is the optimal guess. One of our next jobs is to see why this is the case.

5. DEVIATION AND COVARIANCE

Even though we have seen that the computation of expectation can be reduced to the computation of probability in many cases, the expectation itself still has more to say. For most of this section, the background information is fixed, so we will drop from the notation unless specifically needed. So, in particular, if X is an unknown or random variable, besides simply finding its mean, the optimal guess, we would also like to know how X differs from its mean. In the case of a random variable, we can think of the mean as the long run average value over many repeated observations, and here we would like to know how these various observed values tend to differ from the mean. If we are thinking of the mean as an optimal guess, then we would like to know how far off our guess is, and of course, here we can only guess how far off we are, as well. For instance, if we toss a fair dice, the mean or expected number up is 3.5, as seen in [4.28], but when we actually toss it many times, we get many different results. The difference $D_X = X - \mu_X$ is called the DEVIATION of X from its mean. Notice it is a new unknown or random variable. Since the number $E(X) = \mu_X$ is itself a constant, it follows from [3.6] and [3.8] that $E(D_X) = E(X - \mu_X) = \mu_X - \mu_X = 0$, which in words says that the mean deviation is always zero. Suppose that both X and Y are random variables. Then $D_{X+Y} = X + Y - E(X+Y) = X + Y - (E(X) + E(Y)) = (X - \mu_X) + (Y - \mu_Y) = D_X + D_Y$. Consequently,

$$D_{X+Y} = D_X + D_Y, \tag{5.1}$$

which in words says that the deviation of a sum is just the sum of the deviations as unknowns or random variables. On the other hand, if c is a constant, then $D_{cX} = cX - E(cX) = cX - cE(X) = cX - c\mu_X = c(X - \mu_X) = cD_X$. Therefore,

$$D_{cX} = cD_X, \tag{5.2}$$

which says that multiplying the unknown by a constant multiplies the deviation unknown by that same constant. Notice the similarity with the first two rules for expectation. If you are having trouble keeping these equations straight in your mind, suppose that we take the example of an unknown X with mean equal to say 7. Then $D_X = X - 7$, so $E(D_X) = E(X - 7) = E(X) - 7 = 7 - 7 = 0$. This is just telling us that if we use our optimal guess, then our guess for the deviation from that optimal guess is zero. This certainly seems to make sense-at least it seems logically consistent, but it does not help us see how our optimal guess can be different from the actual value. What is happening in the long run average case is that some values are below the mean and some are above the mean, so some of the deviations are positive and some are negative and in the long run average of the deviations, they just cancel each other all out-the positives cancel the negatives.

Since the only tool we have to assess the overall deviation of X from its mean is E, the expectation, and since $E(D_X) = 0$, we go further and consider D_X^2. Since $D_X^2 \geq 0$, we know $E(D_X^2) \geq 0$, so we can try looking at the squared deviation. Let us begin with a very simple example. Suppose that we have a coin on the table which can either be heads side up or tails side up, denoted H and T, respectively. Suppose that we have no way of knowing which of the two sides is actually up. Also, suppose that the unknown X has value 5 if tails is up and has the value 7 if heads is up. Then, we can see that we have no way to know one side being more likely to be up over the other. Thus, $P(H) = 1/2 = P(T)$, and therefore by [4.26], $E(X) = 6$. Now, $D_X = 1$ if $X = 5$, whereas $D_X = -1$ if $X = 7$, and therefore, $D_X^2 = 1$. This means $E(D_X^2) = E(1) = 1 > 0$. Notice that now we are getting some indication of the variable nature as to the value of X. We cannot get this by considering only D_X, rather we must look at D_X^2. However, we want to be able to analyze a sum in terms of the parts making it up, so for the simplest sum of two terms, say when dealing with $X + Y$, since $D_{X+Y}^2 = (D_X + D_Y)^2$ and since

$$(D_X + D_Y)^2 = D_X^2 + D_Y^2 + 2D_X D_Y,$$

we cannot avoid dealing with the product of deviations $D_X D_Y$. To see how to better deal with assessing overall deviation from the mean, it will save time to see how we deal with two random variables at the same time, right to start.

We are now going to consider how to deal with two unknowns or random variables. If we have two unknowns or random variables they may be related in some rough way. Think of height and weight for people or length and weight for fish, for instance. Notice the product of deviations will tend to be positive if the two variables tend to increase together. For instance, longer fish tend to weigh more, whereas shorter fish tend to weigh less. Put another way, a fish which is longer than average (positive deviation) will tend to be heavier than average (again positive deviation), whereas a fish which is shorter than average (negative deviation) will tend to weigh less than average (again negative deviation). We see that in either of these two cases, the product of the deviations is nonnegative. Keep in mind that these are just tendencies here-the product of deviations would reasonably seem to usually be positive, and therefore the expected value for the product of these deviations should be positive. On the other hand, the product of deviations will tend to be negative if the one variable tends to go down when the other goes up in value. For instance with traffic in the street, when the number of cars per block gets larger the speed of the cars tends to go down. In attempting to get an overall measure of this we take the expected product of the deviations. Thus if X and Y are any unknowns or random variables, we form their COVARIANCE, denoted $Cov(X,Y)$, as

$$Cov(X,Y) = E(D_X D_Y) = E((X-\mu_X)(Y-\mu_Y)). \tag{5.3}$$

In particular, X certainly relates to itself perfectly, so we define the VARIANCE of X, denoted $Var(X)$, as its covariance with itself. Thus using [3.3],

$$Var(X) = Cov(X,X) = E(D_X D_X) = E(D_X^2) = E((X-\mu_X)^2) \geq 0. \tag{5.4}$$

We can use simple algebra to see that

$$(X-\mu_X)(Y-\mu_Y) = XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y, \tag{5.5}$$

so using properties [3.6] and [3.8] on [5.5] we get after some cancellations

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y, \tag{5.6}$$

an expression which is often more efficient and useful in calculations. In particular, when we take $X = Y$ in [5.6] we find a useful formula for variance of a random variable:

$$Var(X) = E(X^2) - \mu_X^2, \tag{5.7}$$

so variance is the "mean of the square minus square of the mean", a phrase which almost sounds like it's from Gilbert and Sullivan. Using [3.6], [5.1], [5.2] and [3.8] and simple algebra, we find that covariance has the following useful simple properties:

$$Cov(W \pm X, Y) = Cov(W,Y) \pm Cov(X,Y) \tag{5.8}$$

$$Cov(X,Y) = Cov(Y,X) \tag{5.9}$$

and more generally than [5.8] we have for any constants a and b:

$$Cov(aW \pm bX, Y) = aCov(W,Y) \pm bCov(X,Y). \tag{5.10}$$

We can realize that [5.8] and [5.9] are sort of like the distributive and commutative laws (respectively) for multiplication. This means that if we think of covariance as a type of multiplication, then variance is squaring with this multiplication. The same use of the commutative and distributive laws which in high school algebra give

$$(a \pm b)^2 = a^2 + b^2 \pm 2ab, \tag{5.11}$$

when applied to covariance and variance tell us that

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y). \tag{5.12}$$

This means that unlike the mean, when dealing with variance, to find the variance of a sum in terms of the variances of the summands, we cannot escape having to deal with their covariance in some way. Sometimes in practice, the covariance is not easy to find, so various assumptions about it can be made. The simplest assumption here is that the covariance is exactly zero. Of course this is not always true, but in many useful situations it is true. We say X and Y are UNCORRELATED when $Cov(X, Y) = 0$. Stronger than being uncorrelated is being INDEPENDENT, a very useful condition to be discussed later. That is, independent unknowns or random variables must be uncorrelated. This means a useful special case of [17] is

$$Var(X + Y) = Var(X) + Var(Y), \quad \text{if} Cov(X, Y) = 0, \tag{5.13}$$

that is when X and Y are uncorrelated and hence also if they are independent. In any case, notice if we are given standard deviations for both X and Y, then we must convert to variance by squaring before combining using [5.12] or [5.13]. Another useful fact about covariance is that adding a constant to either of the random variables makes no change in their covariance. This is because $E(c) = c$ by [3.8], so if c is a constant then $D_c = 0$, that is the deviation of c from its mean is 0. Therefore by [5.3]

$$Cov(c, X) = 0 = Cov(X, c), \tag{5.14}$$

so using [5.8] we get

$$Cov(X + c, Y) = Cov(X, Y) = Cov(X, Y + c), \tag{5.15}$$

and for variance

$$Var(X + c) = Var(X), \quad Var(c) = 0. \tag{5.16}$$

The variance of X is a measure of how likely we expect X to differ from $E(X)$, but it involves squaring. To compensate for this in the end, we define the STANDARD DEVIATION of X to be the square root of its variance. We denote the standard deviation of X by $SD(X)$ or more commonly by σ_X. As with the symbol for the mean, if their is no confusion as to the particular random variable X we are discussing, its standard deviation will simply be denoted by σ. By [5.16] we can conclude that for standard deviation we have for any constant c,

$$SD(X + c) = SD(X) \quad \text{or} \quad \sigma_{X+c} = \sigma_X \tag{5.17}$$

This means that X and $X + c$ both have the same standard deviation, or that standard deviation is unchanged by adding a constant value to all scores in a population. On the other hand if we multiply by a constant the situation is very different. From [5.10] and [5.9] applied to variance we see that

$$Var(cX) = c^2 Var(X), \tag{5.18}$$

so taking square roots of both sides of [5.18] gives

$$SD(cX) = |c|SD(X) \qquad \text{or} \qquad \sigma_{cX} = |c|\sigma_X. \tag{5.19}$$

Thus if $Y = mX + c$, then $\sigma_Y = |m|\sigma_X$. On the other hand for the expectation we would have $E(Y) = mE(X) + c$. So we notice that the added constant does not change the standard deviation but the multiplying constant does, whereas both constants change the expected value or mean. In fact the same is true of the population median. If X is any random variable, then roughly speaking, the median of X is the number for which an observed value is just as likely to be below it as above it. More precisely, we say x is the MEDIAN of X provided that the chance that an observed value of X is less than or equal to x is at least one half and the chance that an observed value of X is greater than or equal to X is also at least one half. If x is the median of X and y that for Y, then at least half the X population is at or below x and at least half is at or above x so if $Y = mX + c$, then at least half the Y population is at or below $mx + c$ and at least half of the Y population is at or above $mx + c$, which means $y = mx + c$. That is we have for any constants b and c,

$$median(bX + c) = b(median(X)) + c. \tag{5.20}$$

We say X and Y are independent loosely speaking when no prior information as to what the value of X will be is of any use in predicting what the value of Y will be. As an example, if we roll a pair of dice, say one is blue and the other is red, knowing that the red dice will come up 4 still gives no clue as to what the blue dice will come up. This is because we know intuitively that the two dice are "rolling independently" of each other. If the two dice are attached to each other by a rigid connecting wire, then clearly they will no longer roll independently of one another and we could soon figure out how to predict what is up on the blue dice from knowing what came up on the red dice. Suppose instead, the two dice are connected by a wobbly spring. They will still not be rolling independently of one another, but depending on the properties of the spring, the values on the red dice will still be somewhat useful in trying to predict what comes up on the blue dice, and in this case the covariance would not be zero. At this point, you should be thinking that even though you do not know how to precisely define independence in the mathematical sense, that at least in many applications you will be able to "know it when you see it".

Returning to covariance in general terms, we began in an attempt to measure the extent to which X and Y relate in some sense. However, from the defining formula [5.3] we can see that the covariance of X and Y can be artificially large if either X or Y has lots of large deviations from its own mean. We compensate for this by dividing by standard deviation. Recalling that the standard deviation is defined as the square root of the variance, we have formulas

$$SD(X) = \sigma_X \tag{5.21}$$

$$Var(X) = \sigma_X^2 \qquad or \qquad \sigma_X = \sqrt{Var(X)} \tag{5.22}$$

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{5.23}$$

and here in [5.23] we have the definition of the CORRELATION COEFFICIENT which is denoted with the lower case Greek letter rho, that is ρ. We can notice that if we are given the standard deviations and the correlation, we can recover the covariance. Because on multiplying the denominators out of the right side of [5.23] we find

$$Cov(X, Y) = \rho\sigma_X\sigma_Y. \tag{5.24}$$

This means any time you are given the variance of X, the variance of Y and the correlation coefficient, ρ, of X and Y, we can use [5.24] to compute the covariance of X with Y and then apply [5.12] to find the variance of $X \pm Y$. Using a more involved argument (see the section on regression,below) with the preceding properties, it can be shown that always

$$-1 \leq \rho \leq 1. \tag{5.25}$$

Since the square of any number is never negative, this is equivalent to

$$0 \leq \rho^2 \leq 1 \tag{5.26}$$

and it turns out that ρ^2 can in fact be thought of as the fraction of variation in say Y that can be accounted for when we use optimal linear regression methods to best predict Y from X.

6. DIGRESSION ON REGRESSION

Let us now return to the problem of how two unknowns or random variables may relate to each other. In more detail, what we would like is a simple equation, called a (LINEAR) REGRESSION EQUATION, of the form

$$y = mx + c \tag{6.1}$$

so that given the observed value x of X, the equation [6.1] can be used to get the best guess for the value y which Y will have. You may recall that the graph of an equation like [6.1] must be a straight line. For instance, if X is length and Y is weight in a population of fish, then given the length, x, of a fish, with the correct values of m and c we use [6.1] to make a best guess for the weight, y, of the fish. Of course, the immediate problem would be how to best choose the numbers m and c once and for all so as to be able to predict a fish's weight from its length. In other words, what is the best straight line. If we think of plotting all the pairs (x, y) for each and every thing in the population (every fish for example), then we would have a massive number of points in the plane and we are trying to find the straight line which as well as possible simultaneously passes nearest to all the points. Clearly, in general there is no straight line passing through all the points (the points would appear as a cloud of points in the plane). The criterion for passing close to a point is to consider $Y - (mX + c)$, the difference between what the equation [6.1] predicts and what the actual value is. For ease of notation, let us set

$$W = mX + c.$$

Thus the difference we are considering is just $Y - W$. But we need to consider this difference for everything in the population and we need to evaluate overall how well this equation [6.1] works. This difference is called the RESIDUAL and we can denote this by R. Thus we have

$$R = Y - (mX + c) = Y - W \tag{6.2}$$

and our objective is to overall minimize the residuals. However, we cannot allow positive residuals to cancel negative residuals for then we might overall have 0 for the total of the residuals when in fact all of them are enormous in absolute value. To keep that from happening we consider the squared residuals, and as you might guess, the overall is computed with the expectation, E. We therefore seek to choose m and c so as to minimize $E(R^2)$, that is we want to minimize the MEAN SQUARED RESIDUAL. From [5.7] we see that

$$E(R^2) = (E(R))^2 + Var(R). \tag{6.3}$$

Notice one thing we can try is choosing m and c so that $E(R) = 0$ so that the first term in [6.3] is 0 which is the smallest that it or any square can be. This means that by [3.6] and [6.2] we should choose m and c so that

$$E(Y) = E(W) = mE(X) + c \qquad \text{or,} \qquad c = E(Y) - mE(X). \tag{6.4}$$

Therefore, if the means of X and Y are known and if the best m can be found, then the best value for c will follow from [6.4]. In any case, since $Var(R) \geq 0$ from [5.4], we can apply our formulas [5.10],[5.9], [5.12], and [5.18] to find

$$Var(R) = Var(Y) + Var(W) - 2Cov(W,Y) = Var(Y) + m^2Var(X) - 2m\rho\sigma_X\sigma_Y \tag{6.5}$$

or alternately using the Greek symbols, with minor algebraic revision,

$$\sigma_R^2 = \sigma_Y^2 + (m\sigma_X)^2 - 2(m\sigma_X)(\rho\sigma_Y), \tag{6.6}$$

and we notice that this only involves the correlation coefficient, ρ, the variances (see [5.22]) and m. On the other hand, using $1 = 1 - \rho^2 + \rho^2$ and multiplying by σ_Y^2, we can substitute the result in the first term on the right of [6.6] to get

$$\sigma_R^2 = (1 - \rho^2)\sigma_Y^2 + (\rho\sigma_Y)^2 + (m\sigma_X)^2 - 2(m\sigma_X)(\rho\sigma_Y). \tag{6.7}$$

Now we can notice that the last three terms of [6.7] can be combined using [5.11] to give simply $(\rho\sigma_Y - m\sigma_X)^2$ which when substituted for the last three terms in [6.7] give us

$$\sigma_R^2 = (1 - \rho^2)\sigma_Y^2 + (\rho\sigma_Y - m\sigma_X)^2. \tag{6.8}$$

When we combine this with [6.3], we get finally

$$E(R^2) = (E(R))^2 + (1 - \rho^2)\sigma_Y^2 + (\rho\sigma_Y - m\sigma_X)^2. \tag{6.9}$$

Since squares can never be negative, the smallest that $E(R^2)$ can be is when m and c are chosen so that the first and last terms on the right are exactly 0. Therefore, if we set $\rho\sigma_Y - m\sigma_X = 0$, we get the optimal regression slope

$$m = \rho\sigma_Y/\sigma_X, \tag{6.10}$$

whereas setting $E(R) = 0$ gives the intercept c for the regression line as given in [6.4]. This means that m is indeed determined only from the correlation and variances whereas c is determined only from the means of X and Y and the value of m. When we use these optimal values for m and c in [6.9], the first and last terms on the right side drop out and the equation for the mean square residual in regression becomes (in view of [6.3]), simply

$$Var(R) = E(R^2) = (1 - \rho^2)\sigma_Y^2. \tag{6.11}$$

Notice that since the left hand side of [6.11] can never be negative and as σ_Y^2 cannot be negative, it follows that $(1 - \rho^2)$ can never be negative, and this is just [5.26]. We also note here that [5.26] implies [5.25]. The inequality [5.25] is what is usually called the CAUCHY-SCHWARZ INEQUALITY in mathematics, which is also here obviously equivalent to

$$|Cov(X,Y)| \leq \sigma_X\sigma_Y. \tag{6.12}$$

We can also notice that [6.11] leads us naturally to interpret $1-\rho^2$ as the fraction of variance in Y unaccounted for by the regression (since that is $Var(R)$) and therefore, ρ^2 must be the fraction of variance in Y that is accounted for using X through the regression equation. In more detail, notice that using our equations for the optimal values of m and c, we have, by [6.11],

$$\sigma_Y^2 = (1-\rho^2)\sigma_Y^2 + \rho^2\sigma_Y^2 = E(R^2) + m^2\sigma_X^2 = \sigma_R^2 + \sigma_W^2. \tag{6.13}$$

Notice that in [6.13], with each expression the first terms are the same and the last terms are the same. Since σ_R^2 represents variation in Y which is not accounted for by the regression equation and obviously σ_W^2 represents variation in Y which is completely accounted for by the regression equation, we see that $\rho^2\sigma_Y^2 = \sigma_W^2$ represents the part of the variation in Y accounted for by the variation in X through regression and therefore,

$$\rho^2 = \frac{\sigma_W^2}{\sigma_Y^2} \tag{6.14}$$

represents the fraction of variation in Y accounted for from variation in X through the regression equation. Notice that [6.14] also follows immediately from [6.10], as by [6.10], we have

$$|\rho| = \frac{|m|\sigma_X}{\sigma_Y} = \frac{\sigma_{mX+c}}{\sigma_Y} = \frac{\sigma_W}{\sigma_Y}. \tag{6.15}$$

Another useful observation is that since $Y = R + W$, if we combine [5.12] with [6.13], then we have

$$Var(R) + Var(W) + 2Cov(R, W) = Var(Y) = Var(R) + Var(W). \tag{6.16}$$

Notice that an immediate consequence of [6.16] is that

$$0 = Cov(R, W) = mCov(R, X), \tag{6.17}$$

and therefore by [6.17], if m is not equal to 0, then

$$Cov(R, X) = 0. \tag{6.18}$$

That is with the optimal regression line, by [6.18], the explanatory variable X must be uncorrelated with R, the residual variable. It is also interesting to notice that if we choose m so that [6.18] holds (say with a lucky guess), then that also gives us the correct value of m in [6.10]. To see this, notice that [6.18] implies [6.17], so as $Y = R + W$ with $W = mX + c$, we get

$$\rho\sigma_X\sigma_Y = Cov(X, Y) = Cov(X, R) + Cov(X, W) = m\sigma_X^2. \tag{6.19}$$

Thus, solving [6.19] for m gives [6.10].

7. CALCULATIONS WITH THE TI-83

Using the TI-82/3 we can calculate the correlation coefficient in the case of a finite population by putting the population data for the two variables in lists and using the lin reg(ax+b) command in the statistical calculation menu. With the diagnostics on, the readout reports the value of m as a, the value of c as b, and the value of ρ as r. If we only have sample data, the same calculations give approximate values for m and for c, and then r is called the SAMPLE CORRELATION COEFFICIENT. With the data plotted in a scatterplot, we can roughly draw the regression line by eye, and usually the result is accurate enough for simple applications.

8. USEFUL PHYSICAL INTERPRETATION

The optimal line can be imagined as the result of attaching springs to the line from each one of the points in the scatterplot and allowing the springs only to pull vertically so as to move the line to an equilibrium position. Now from elementary physics we know from Hook's Law that the energy in a stretched spring is proportional to the square of how far it is stretched from its equilibrium position. That is, the energy in the system is proportional to the total of the squared residuals which in turn is proportional to the mean squared residual when the sample size, n, is fixed, so the system would come to rest in the position of minimum energy, which is the one minimizing the mean of the squared residuals. For this reason, the regression line given by [6.4] and [6.10] is often referred to as the least squares regression line or the least squares best fit. Even though the derivation of the regresson equation here takes a substantial amount of elementary algebra, it is a lot simpler than the derivation given in many standard statistics texts using multivariable calculus. Moreover, our result is perfectly general applying to both finite and infinite populations with any distributions. The physical interpretation is useful for thinking about how changes in the data will affect the resulting regression line. For instance, from [6.4] and [6.10] we see that (μ_X, μ_Y) is always a point on the true regression line (if we put $x = \mu_X$ in [6.1], then we get $y = \mu_Y$ using [6.4] and [6.10]), so any point in the scatterplot which is far from (μ_X, μ_Y) has a lot of leverage to turn the regression line and therefore any outliers far from this point in the scatterplot which are erroneous data will cause the regression line calculated from sample data to diverge substantially from the true regresson line, whereas erroneous data near (μ_X, μ_Y) will not have much of an effect as there is not much leverage.

9. DISTRIBUTIONS AND TCHEBEYCHEV'S INEQUALITY

Suppose now that X is a random variable. Let $\mathbb{R}$ denote the set of all real numbers, which we can picture as an infinite line. By the DISTRIBUTION of X we mean the information as to how likely various values of X are to be observed. We can capture most of this information in a single real-valued function on $\mathbb{R}$ itself, called the CUMULATIVE DISTRIBUTION FUNCTION(or CDF, for short) of X and denoted by F_X, and whose value at any $x \in \mathbb{R}$ is given by

$$F_X(x) = P(X \leq x). \tag{9.1}$$

Notice that if $a, b \in \mathbb{R}$, then

$$P(a < X \leq b) = F_X(b) - F_X(a). \tag{9.2}$$

In case S is finite, then X can only assume a finite number of values, so if b is a possible value of X, and $a < b$ is chosen so it is not a possible value of X and so that no possible value of X is between a and b, then $P(X = b) = P(a < X \leq b) = F_X(b) - F_X(a)$, so the cdf of X tells us the probabilities of the various values of X. Conversely, in this case where X has only finitely many values, if we know the probability of each possible value, then we know $P(a < X \leq b)$ for any numbers $a, b \in \mathbb{R}$ and therefore we know the cdf for X. If S is infinite, then since $X \leq x$ must be an event for [9.1] to make sense, we must assume that $X \leq x$ is an event, for any real number x. In fact, the technical definition of a random variable insures that this is always the case, so we have nothing to worry about. That is, if X is a random variable on the sample space S, then by definition, $X \leq x$ is an event in S no matter how x is chosen.

Now, Tchebeychev's inequality is a simple inequality that for any random variable X limits how likely an observed value can be far from the mean relative to the standard deviation. Let $E(X) = \mu$ and $Var(X) = \sigma^2$. Remember that for any two numbers $a, b \in \mathbb{R}$, the absolute value of their difference, $|a - b|$ gives their separation distance on the number line, and

$$|a| = \sqrt{a^2}, \; a^2 = |a|^2. \tag{9.3}$$

Notice taking absolute value of a number just drops its algebraic sign. Suppose now that k is any non negative real number and we consider whether

$$k\sigma \leq |X - \mu|. \tag{9.4}$$

If this turns out to be true, then the observed value is at least k standard deviations from the true mean. We want to be able to say something about how likely this is. Now [9.4] defines a statement about the outcome so is an event which we simply denote by A. The question here is how big can $P(A)$ actually be. Clearly, if k is large, then we would like to be able to know that $P(A)$ is small. Therefore, even if we do not know $P(A)$ exactly, we would, at least, like to find a number which we are sure is bigger, and which is actually useful. This is what Tchbeychev's inequality does. Notice that squaring both sides of the inequality [9.4] gives the equivalent inequality

$$k^2\sigma^2 \leq (X - \mu)^2 \tag{9.5}$$

defining event A. Now, let us consider the indicator I_A of A, which is an unknown. Thus the only possible values of I_A are $0, 1$, and I_A takes the value 1 exactly when the inequality [9.5] is actually true. Now $k^2\sigma^2$ is actually a constant, so $(k^2\sigma^2)I_A$ is also an unknown and we can compare it to the right side of [9.5]. That is, consider the inequality between unknowns

$$(k^2\sigma^2)I_A \leq (X - \mu)^2. \tag{9.6}$$

Notice that if A does not happen, then the left side is 0, and as the right side is squared it must be ≥ 0, so in this case the inequality is true. On the other hand, if A does happen, then $I_A = 1$ and the statement immediately becomes [9.5], which by definition is true when A happens. This means that [9.6] is always true no matter the outcome, so it expresses an inequality between random variables and we can apply [3.7], the order preserving property of expectation to conclude that

$$E((k^2\sigma^2)A) \leq E((X - \mu)^2). \tag{9.7}$$

But, $E((X - \mu)^2) = \sigma^2$. Moreover, as A is a statement, we know $E(I_A) = P(A)$, and as $k^2\sigma^2$ is just a constant, we can use [3.2] to get

$$(k^2\sigma^2)P(A) = E((k^2\sigma^2)I_A) \leq E((X - \mu)^2) = \sigma^2. \tag{9.8}$$

Thus

$$(k^2\sigma^2)P(A) \leq \sigma^2. \tag{9.9}$$

Assuming that the standard deviation of X is not 0, we can cancel the σ^2 from both sides of [9.9] and get

$$(k^2P(A) \leq 1, \tag{9.10}$$

which means that if $k > 0$, then

$$P(A) \leq \frac{1}{k^2}. \tag{9.11}$$

That is, in view of the definition of A,

$$P(k\sigma \leq |X - \mu|) \leq \frac{1}{k^2}. \tag{9.12}$$

It is [9.12] that is known as Tchebeychev's inequality. Notice that we are really giving the chance that X is a certain number of standard deviations from its mean with Tchebeychev's inequality. That is, the inequality is better expressed when we change units to standard deviation units. In general, if X is any unknown, then we define its STANDARDIZATION as Z_X given by

$$Z_X = \frac{D_X}{\sigma_X} = \frac{X - \mu_X}{\sigma_X}. \tag{9.13}$$

Then [9.12] becomes simply

$$P(|Z_X| \geq k) \leq \frac{1}{k^2}. \tag{9.14}$$

10. CONDITIONAL EXPECTATION AND PROBABILITY

We have assumed that the multiplication property [3.5] holds for the expectation, which of course leads immediately to the rule [4.2]. This rule shows us that when new information A becomes available in addition to the basic background information B that we start with, then we can relate the positive linear normalized expectation model with fixed background information $A\&B$ to the positive linear normalized model with fixed background B. For notice that the multiplication rules [3.5] and [4.2]

$$E(XI_A|B) = E(X|A\&B)P(A|B) = E(X|A\&B)E(I_A|B) \tag{10.1}$$

can also be written

$$E(X|A\&B) = \frac{E(XI_A|B)}{P(A|B)} = \frac{E(XI_A|B)}{E(I_A|B)}. \tag{10.2}$$

The useful thing to notice here is that the right hand fractions are computed entirely with only B as background information, whereas on the left side we are computing with the new information in addition to the background information, $A\&B$. The equation [10.2] should be thought of as how the expectation models the learning process. It shows how all expected values or optimal guesses become modified when we are presented with new information which may itself be uncertain.

In order to examine the role played by the information used to guess values or arrive at expected values, we need to be more precise about how we handle the unknowns and statements we deal with. Suppose that we think of the unknowns we are interested in as forming a set $\mathcal{A}$, and the statements we are interested in form a set $\mathcal{S}$. Of course, we will axiomatically assume that

$$X + Y \in \mathcal{A} \,\text{and}\, XY \in \mathcal{A} \,\text{for all}\, X, Y \in \mathcal{A}, \tag{10.3}$$

and that

$$\mathbb{R} \subset \mathcal{A} \,\text{and}\, I_A \in \mathcal{A} \,\text{for each statement}\, A \in \mathcal{S}. \tag{10.4}$$

Now, when the background is fixed as $C \in \mathcal{S}$, then we can set $E_C(X) = E(X|C)$, for all $X \in \mathcal{A}$. Thus, we can think of E_C as the expectation model determined by the information $C \in \mathcal{S}$. Notice then that E_C satisfies the first four properties of expectation-it is a positive linear normalized function on $\mathcal{A}$. If we have worked out how to compute $E_C(X)$ for each $X \in \mathcal{A}$, and if we have another statement $K \in \mathcal{S}$, then putting $A = K\&C$, we want to know how to compute $E_A(X)$ for each $X \in \mathcal{A}$. Of course this is just what the multiplication rule [10.2] tells us how to do, but our aim here is to see why this makes sense. Remember, an unknown is a description of a number, such as the number up on a dice in a box. The background information could be that we cannot see inside the box and otherwise have no information making any of the six possible values more likely than any other. Let X be the number up and B be the background information. Let K be the statement that a reliable friend has looked in the box and says the number up is even. We can think of $X|B$ as the unknown together with the background information as its description, which of course is a new unknown. Likewise, $X|A$ is a new unknown, where $A = K\&B$. In effect, if $\mathcal{A}|B = \{X|B : X \in \mathcal{A}\}$, then E_B in a sense is really giving guessed values for unknowns in $\mathcal{A}|B$. Likewise, E_A is really giving guessed values for unknowns in $\mathcal{A}|A$. So the real question here is once we "know" how to compute with E_B for unknowns in $\mathcal{A}|B$, why and how should this tell us how to compute E_A for unknowns in $\mathcal{A}|A$. One thing we can notice is that $XI_K|B$ belongs to $\mathcal{A}|B$ and with the background information B it has the same value as $X|B$ provided that K is true, whereas it is simply 0 if K is false. Thus, it seems that $E_B(XI_K)$ must be based on both the guess for $X|B$ as well as the guess for "how true" K is. Also, it seems to partially arrive at a guess for X based on both K and B, as the only value allowed for the case where K is false is 0. So, one possible choice for a method of guessing values for unknowns in $\mathcal{A}|A$ would be to use the value $E_B(XI_K)$. Thus, we are led to consider the function F_A defined for unknowns in $\mathcal{A}$ by the rule $F_A(X) = E_B(XI_K) = E(XI_K|B)$. Now clearly, F_A is a positive linear function on $\mathcal{A}$, but it may not be normalized. In fact, $F_A(1) = E(1I_K|B) = E(I_K|B) = P(K|B)$. Now by [3.13] we know that $F_A/P(K|B)$ is a normalized positive linear function on $\mathcal{A}$, so it is reasonable to use it to compute guesses for unknowns in $\mathcal{A}$ when we assume that K is true. Therefore, since it is built out of E_B, we should guess that $E_A = F_A/P(K|B)$. When we compute the value $E_A(X)$, we now find

$$E(X|K\&B) = E(X|A) = E_A(X) = \frac{F_A(X)}{P(K|B)} = \frac{E(XI_K|B)}{P(K|B)}, \tag{10.5}$$

so finally,

$$E(X|K\&B) = \frac{E(XI_K|B)}{P(K|B)}. \tag{10.6}$$

which is the multiplication rule in the form [10.2]. For the example of the dice in the box, with X the number up, with B the background giving no preference to any of the six possible values for X and K the additional information that the value is even, we see that there are only three possible values: 2,4,6, and all are equally likely as we still have no preference for any one value over another beyond the information that the value is even. Consequently, by [4.26], we must have $E(X|K\&B) = 4$. On the other hand, XI_K has the possible values:

0,2,4,6 and these given B are not equally likely, as all six possible values of X are equally likely given B, and the value 0 happens for XI_K if the value of X is any odd value. Therefore, $P(XI_K = 0|B) = 1/2$ whereas the probability that XI_K has any given positive value is only $1/6$. It follows that $E(XI_K|B) = (2+4+6)/6 = 2$. On the other hand, since the information B tells us all six values are equally likely and as half of them are even, it must be the case that $P(K|B) = 1/2$. We therefore have in this example

$$E(X|K\&B) = 4 = \frac{2}{1/2} = \frac{E(XI_K|B)}{P(K|B)},$$

giving an explicit example of the multiplication rule.

Let us now examine this construction of the multiplication rule through normalization more carefully for the case of random variables when we have a sample space. For any sample space S, any event $A \subseteq S$, and any random variable X on S, we can restrict our attention to outcomes in A for observations of X. In effect, A becomes the sample space, and X restricted to this subset is denoted by $X|A$. If X and Y are both random variables on S, then clearly we have

$$(X+Y)|A = (X|A) + (Y|A) \tag{10.7}$$

$$(XY)|A = (X|A)(Y|A). \tag{10.8}$$

Notice also, that for the indicator I_A as a random variable on S we have

$$I_A|A = 1 \tag{10.9}$$

$$I_{\text{not}A}|A = 0. \tag{10.10}$$

In particular, this means that for any random variable X on S, we have

$$(XI_A)|A = X|A \tag{10.11}$$

$$(XI_{(S\setminus A)})|A = 0. \tag{10.12}$$

To get the expected values of random variables on the sample space A, we need some expectation model satisfying the first four properties of expectation-it must be a positive linear normalized function on the random variables on A. If we have such an expectation model, say E_A, then we could try to define the conditional expectation of X given A by setting it equal to $E_A(X|A)$. The problem here is that there are possibly many such models, and the model we seek should somehow depend on the original expectation for random variables on S, which we start with due to our background information B. We shall denote this by E_S. Thus we want to manufacture a particular E_A using E_S.

Remember, conditional expectation given A must be some kind of ordinary expectation model-that is a positive linear normalized function for random variables on S restricted to the sample space A. In particular, conditional expectation must satisfy the first four properties of ordinary expectation that we have developed so far-it is a positive linear normalized model. Of course this conditional expectation is really a whole new expectation model. Here we recall that in general, there are many expectations with various backgrounds held fixed for a given experiment, since we only require that the expectation assign each random variable a number called its mean in such a way as to satisfy the first four basic properties:[3.1], [3.2], [3.3], and [3.4]. For instance, for the dice example, the mean number up when rolled will clearly depend on how the dice is loaded. Changing the loading of the dice may change some or all of the expected values of the random variables on this experiment. So, what we really need is a way to relate conditional expectation to ordinary expectation. To begin, we notice we can view our notation $E(X|A)$ as an expectation model for the random variables on the sample space A. Since we are assuming an expectation model E_S in operation for the random variables on S when we write expressions like $E(X)$, what we are really interested in is a natural way to produce the expectation model for random variables on A from the expectation model we already have for random variables on S. For instance, if G is just any expectation model for random variables on A so that $G(W)$ is the expected value of the random variable W on A satisfying the basic properties of expectation, then there need be no connection at all to the expectation model E operating on random variables on S. To produce an expectation model for random variables on A by using the expectation model for random variables on S we need to look for a natural way to extend each random variable on A to be a random variable on all of S. For instance, suppose that S is a population consisting of people and rocks. Suppose that A is the set of people and the random variable W on A is blood pressure. If we try to extend this random variable to all of S, then we are in a sense trying to make sense of the blood pressure of the rocks as well as of the people. You can easily see that in general, there may be no really natural extension to all of S for a random variable defined on A. However, mathematically we can always extend the random variable to all of S by simply declaring the variable to have value 0 on the complement of A in S. In our example, this amounts to agreeing that all rocks have blood pressure zero, which in some sense is perfectly reasonable. For another example, suppose that S is the population of adults in the United States, that A is the subset consisting of U.S. citizens and resident aliens with legal working papers. Let W assign each member of A the last four digits of his Social Security Number considered as a four digit whole number. If a person is randomly chosen from S he may not even have a Social Security Number, if he is not in A. Moreover, here the mathematical extension seems completely arbitrary, that is if we agree to define the value of W to be zero for anyone in $S \setminus A$, then there is no apparent reason why this should make sense in any ordinary way. We have here a purely mathematical fabrication. If W is any random variable on A, let us denote by $h_A^S(W) = h(W)$ the extension of W to all of S gotten by simply declaring that the value should be zero for outcomes in $S \setminus A$ but unchanged for outcomes in A. We can then easily see that for any random variables U and W on A,

$$h(U+W) = h(U) + h(W), \tag{10.13}$$

$$h(UW) = h(U)h(W), \tag{10.14}$$

and

$$h(U) \geq 0, \quad \text{if} \quad U \geq 0. \tag{10.15}$$

Now, because of [10.13], [10.14], and [10.15], it is easily seen that if we try to define an expectation model F operating for random variables on A by setting $F(W) = E(h(W))$, then F satisfies all the first four of the basic properties of an expectation model except that we may not have $F(1) = 1$. Because when we write $F(1)$ we are refering to the constant random variable defined on A with value always 1, but when we form $h(1)$ we get the random variable on S that has value 1 on all of A and value 0 on $S \setminus A$. But this random variable is none other than I_A, the indicator of A, considered as a random variable on S. That is, we have $h(1) = I_A$. In fact it is easy to see that for any random variable X on S we have

$$h(X|A) = XI_A. \tag{10.16}$$

This means that $F(X|A) = E(h(X|A)) = E(XI_A)$ and in particular, $F(1) = E(I_A) = P(A)$. If $P(A)$ is not zero, then as usual we can easily fix this problem by normalizing (see [3.13]), that is by dividing all the values of F by $P(A)$. This means that the natural expectation model E_A for random variables on A produced from the expectation model E for random variables on S is just gotten by requiring that $E_A(W) = E(h(W))/P(A)$ for any random variable W defined on A. This means that $E(X|A)$ should be defined as $E_A(X|A) = E(h(X|A))/P(A) = E(XI_A)/P(A)$. Since the A appears twice on the left side of this last equation, the subscript becomes unnecessary, and we can always just write $E(X|A)$ for this expectation. It is finally then the natural construction for the CONDITIONAL EXPECTATION OF X GIVEN A, to repeat, the multiplication rule

$$E(X|A) = \frac{E(XI_A)}{P(A)}. \tag{10.17}$$

Now the drawback of this construction is that it has been obtained purely by abstract mathematical criterion, so we need to see if it really gives us what we are intuitively thinking of. Before doing this, we can first observe that, as long as $P(A) \neq 0$, our definition [10.17] is equivalent to requiring that

$$E(XI_A) = E(X|A)P(A). \tag{10.18}$$

Combining [4.20] with [10.18], we recall that if $A_1, A_2, A_3, ..., A_n$ is a partition of S, then we get the GENERAL PARTITION PROPERTY of expectation

$$E(X) = E(X|A_1)P(A_1) + E(X|A_2)P(A_2) + E(X|A_3)P(A_3) + ... + E(X|A_n)P(A_n) \tag{10.19}$$

holds, allowing us to compute the overall expectation of X from its various conditional expectations in each of the partitioning circumstances. As a special case of the general partition property we have the PARTITION PROPERTY

$$E(X) = E(X|A)P(A) + E(X|(S \setminus A))P(S \setminus A), \tag{10.20}$$

which holds for any event A in S. Keeping this partition property in mind, let us think of the business example where X is the daily net profit of a coffee shop. Suppose A is the event of a rainy day, then $E(X|A)$ is the long run average net profit for rainy days. With $B = S \setminus A$, then $E(X|B)$ is the long run average net profit for days that are not rainy. Suppose for the moment that $E(X|A) = 500$ dollars and $E(X|B) = 1000$ dollars. Thus the coffee shop averages 500 dollars a day for rainy days and averages 1000 dollars a day for the days that are not rainy. Now it should be intuitively clear that if overall half the days are rainy, that is $P(A) = 1/2$, then the overall daily net profit is just the average of 500 and 1000, that is $E(X) = 750$ dollars. If only thirty percent of the days are rainy, overall, that is if $P(A) = 0.3$, then $P(B) = 0.7$, then clearly we need to form the weighted average, $E(X) = (500)(0.3) + (1000)(0.7)$. This leads us naturally to the partition property [10.20], as a general relationship, which should always hold relating conditional expectation and the original overall expectation. In particular, as from [10.11] we have $(XI_A)|A = X|A$ and $(XI_A)|(S \setminus A) = 0$, we must have [10.18] as a special case of the partition property, since the second term of [10.20] drops out when X is replaced by XI_A. And as [10.18] is already seen equivalent to our definition of conditional expectation [10.17], we can safely conclude we have made the right choice with our mathematical definition for conditional expectation. That is, the partition property is the intuitive property that tells us the definition of conditional expectation, but already it is a consequence of the definition of conditional expectation, which is to say that assuming the partition property is equivalent to assuming the mathematical definition of conditional expectation.

As in the case of ordinary probability, to define conditional probability, we just restrict the conditional expectation to apply to events. Thus if B is any event in S, then, remembering that $AB = A \cap B$ for events, we define the CONDITIONAL PROBABILITY OF B GIVEN A by the formula

$$P(B|A) = E(I_B|A) = \frac{P(B \cap A)}{P(A)}. \tag{10.21}$$

Consequently, by reversing the roles of A and B in [10.21] we see that

$$P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A). \tag{10.22}$$

In effect [10.22] allows us to find $P(A|B)$ from $P(B|A)$ when $P(A)$ and $P(B)$ are also known. In addition, if X in [10.19] is replaced by an event B, then each conditional expectation in [10.19] becomes a conditional probability, and [10.19] takes the form

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + ... + P(B|A_n)P(A_n), \tag{10.23}$$

an equation known as Baye's Theorem.

As another application of [10.19], if X is a simple discrete random variable, that is if X has only a finite set of possible values, say $x_1, x_2, x_3, ..., x_n$, and if for each k we take $A_k = (X = x_k)$, meaning A_k is the event that X takes the value x_k, for $1 \leq k \leq n$, then as now $E(X|A_k) = x_k$ for each k, by[10.19] we find

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + x_3 P(X = x_3) + ... + x_n P(X = x_n), \tag{10.24}$$

which is the same as [4.26], which is to say, [4.26] is a special case of [10.19].

Conditional probability can usefully be applied in situations where an experimental procedure can be broken down into a sequence of steps where the probabilities of later steps depend on what happens in those leading up to it. For instance if a box has 10 blocks inside, 4 are red and 6 are green, and if the experiment is to draw two blocks from the box at random, then we can imagine this being accomplished by first randomly drawing a single block and then without replacing that block we go on to draw a second block at random from the box. To calculate the probability that both blocks drawn are red, we can easily calculate the probability that the first block drawn is red (clearly = 4/10), and then assuming a red block removed in the first draw, the probability that the second block drawn is red given that the first block drawn is red must be simply 3/9. The product of these two then gives the probability that both are red, namely 12/90. What is often not obvious on first encountering conditional probability is that the time ordering of the sequence of outcomes is really not material. That is, if we ask for the conditional probability that the first block drawn is red given that the second block drawn will be red, we get the same result, again 3/9. This is because probability is really just about the state of our information and not about time ordering. For instance, if 4 blocks are drawn from the box and we ask for the conditional probability that the third block drawn is red given that the first block is green and the last block drawn is red, then that must be simply 3/8. This is because the information given is that two of the blocks are ruled out of being the third block drawn and we know the colors of those blocks, so when we draw the third block we have 8 possibilities of which only 3 are red. These kind of calculations can be done using Baye's theorem with a lot more computation, but the result is the same either way. One way to see easily why the time ordering in drawing blocks from a box is immaterial is to think of the process being performed by someone else other than ourselves who in fact sequentially takes all the blocks from the box and then gives us information concerning some of the results. If he tells us the colors of the second and fifth blocks, and we are asked for the probability the third is red given the information about the colors of the second and fifth blocks it should be clear what the result is. Alternately, think of drawing cards sequentially from a shuffled deck of cards. Assuming the cards are in random order to begin, the drawing can be performed by removing the card on top, then the next card (the new top card) and so on. If we are given that the card underneath the top card is a heart and the fifth card from the top is a spade, then for a standard deck of cards, the probability that the third card from the top is a heart given the information stated would clearly be 12/50. Likewise, the probability that the seventh card is a spade given that the second, third and tenth cards down in the pile are spades and the fifth card down is a heart would be simply 10/48. But, clearly this must be the same as the probability that the seventh card drawn is a spade given that the second, third and tenth cards drawn are spades and the fifth card drawn is a heart, because with the cards in the order shuffled with the given information, the card position in the deck exactly corresponds to the order of results for drawing cards one after another. Likewise, for blocks in a box, we can think of the random sequential drawing of the blocks from the box as being done by randomly stacking the blocks and then taking blocks sequentially from the top of the stack.

Another useful thing to realize is that all probability is really conditional. If the experiment is rolling a pair of identical red dice to see what comes up on the top faces, and on a given roll the result is that the dice fall into a bucket of identical dice so we can no longer tell which dice we actually rolled, then the experiment must be repeated. That is, it is a "do over". Let T be the total up on the pair of dice when rolled. If we ask what is the chance that in repeated rolls of the pair of fair dice we roll a 7 before we roll an 8, then we can regard the sample space as having been restricted to just those 11 outcomes which give a 7 or an 8. Since not getting either of these two results is a "do over", we are really just dealing with the conditional probability that we roll a 7 given that we roll 7 or 8. More generally, if A and B are mutually exclusive events and we perform our experiment over and over (a sequence of independent trials) to see which event happens first, then the chance that A happens before B is just $P(A|A \cup B)$.

Finally, in many situations, part of our information is in the form of conditional probabilities, and we can use [10.23] together with [10.22] effectively to "reverse" the order of conditioning. As an example, if a student takes a multiple choice test and each question has five choices for an answer, we can assume that the student guesses those which he does not know and therefore has a twenty percent chance of correctly marking a question given that he does not know the answer. If K is the event he knows the correct answer and C is the event he marks the correct answer on the answer sheet, then given $P(K)$ and $P(C|K)$, we can calculate $P(notK) = 1 - P(K)$ and as $P(C|notK) = .2$, we can use Baye's theorem, [10.23], to calculate $P(C)$, and then use [10.22] to calculate $P(K|C)$. That is, if we know his error rate for marking questions for which he knows the answers, if we know the percentage of questions for which he knows the answers, then for a question we see correctly marked we can calculate the chance he actually knew the answer to that question.

11. COUNTING

We have already observed that the expectation model for an experiment is completely determined by the probability model, and in case of a finite sample space, the probability model is determined as soon as the probability of each individual outcome is determined. Of course, the only general theoretical restriction on the probabilities of the outcomes is that they must be nonnegative numbers summing up to one. The simplest case then is when all outcomes are equally likely. In that case, recall, we say we have the model of equally likely outcomes. Let us use the notation $card(S)$ to denote the cardinality of the set S, which merely means the number of things in the set S for the case where S is finite. Then, in case S is finite, the rules of probability and the model of equally likely outcomes force the equation

$$P(A) = \frac{card(A)}{card(S)}, \tag{11.1}$$

for any event $A \subseteq S$. Thus, for a finite sample space, the model of equally likely outcomes reduces the problem of computing probabilities to the problem of counting. As easy as that sounds, in fact there are unsolved problems in counting. In order to deal with problems in counting, it is useful to keep in mind some very basic rules which are fairly obvious. The first rule is that if A and B are disjoint, then

$$card(A \cup B) = card(A) + card(B). \tag{11.2}$$

Notice that [11.2] is actually a consequence of one of the basic rules of probability, namely that the probability of the union of disjoint events is the sum of their individual probabilities-just divide both sides of the equation by $card(S)$ assuming that A and B are both subsets of some finite sample space S, and apply [11.1]. In what follows we will refer to [11.2] as the ADDITION RULE. One of the things we often need to count is the number of ways to choose r things from a set of n things. It is not immediately obvious how to count this. In mathematics, when we do not know what a number's actual value is, we give it a symbol and work with that symbol as if we actually knew what its value is, and hope that we can develop relationships with things we know in order to find the value. In that spirit, we use the symbol $C(n, r)$ for the number of ways to choose r things from a set of n things. Notice there is only one way to choose nothing from a set containing n things, so $C(n, 0) = 1$, even if $n = 0$. On the other hand, there are clearly n ways to choose one thing from a set of n things, so $C(n, 1) = n$. Another observation we can make is that if $0 \leq r \leq n$, then the number of ways to choose r things from a set of n things is the same as the number of ways to choose $n - r$ things from the set of n things, since each such choice of the r things is completely determined by the things not chosen-that is we could equally well decide which things are to be "left behind". This means

$$C(n, n - r) = C(n, r), 0 \leq r \leq n. \tag{11.3}$$

In particular, we see that by [11.3] we have $C(n, n) = C(n, 0) = 1$ and $C(n, n - 1) = C(n, 1) = n$ for every n. Suppose that we know all values of $C(n, r)$ for $0 \leq r \leq n$. We can then use that information to determine all values of $C(n + 1, r)$ for $0 \leq r \leq n + 1$, by applying the addition rule for disjoint sets. To do this, let us imagine that we begin with a set B having exactly $n+1$ things in it, think of a box containing $n+1$ blocks. Suppose that n of the blocks are blue and one is red. Suppose that $0 \leq r \leq n$. Then $r+1 \leq n+1$. Now let S be the sample space consisting of all subsets of B which contain exactly $r+1$ things. Thus $card(S) = C(n+1, r+1)$. Now consider the event A which as a statement is "the red block is in the subset of blocks chosen". Notice that for any way we choose $r+1$ of the blocks, we either get the red block or we do not get the red block. This means $A \cup notA = S$ and therefore, $C(n+1, r+1) = card(A) + card(notA)$. But notice, that the number of ways to choose $r+1$ blocks so as to get the red block is the same as the number of ways to choose r of the blue blocks and there are exactly n blue blocks. Thus, $card(A) = C(n, r)$. On the other hand, the number of ways to choose $r+1$ blocks so as not to get the red block is the number of ways to choose $r+1$ blue blocks from the n blue blocks. This means that $card(notA) = C(n, r+1)$. Combining these three equations we see

$$C(n+1, r+1) = C(n, r) + C(n, r+1), 0 \leq r \leq n. \tag{11.4}$$

We will refer to [11.4] as the FORMULA FOR PASCAL'S TRIANGLE, since the triangular arrangement of all the numbers $C(n, r)$ generated by [11.4] is called Pascal's triangle and is very convenient for computing these numbers when n and r are small, say not more than ten. To generate Pascal's triangle, we start at the top of the page and in the center we write $C(0, 0)$ which is of course just 1. For simplicity, let us refer to this top line as line zero. Then underneath on the next line, called line one, we start diagonally to the left of center and put $C(1, 0)$ followed by $C(1, 1)$ so that this last is diagonally to the right of center. Of course, both of these numbers are equal to 1. On the next line, that is line two, we begin diagonally to the left of the position directly under $C(1, 0)$ and put $C(2, 0)$ which is of course equal to 1, and then in the center position we put $C(2, 1)$ and follow that with $C(2, 2)$, which of course is just equal to 1. Now notice that by the formula for Pascal's triangle $C(2, 1) = C(1, 0) + C(1, 1)$ by taking $r = 0$ and $n = 1$ in the formula. That is, to get the value of the number at the center position on line two we just look at the two numbers above, diagonally to the left and to the right, and add those two numbers together. Thus $C(2, 1) = 2$ which is of course obvious. Continuing this pattern on down we get a triangular array of numbers known as Pascal's triangle. To compute $C(n, r)$ using this array, always begin counting with zero. So counting down lines to line n, we count from the left starting with counting the 1 in the starting position as position zero on line n and then count over to position r to get the value of $C(n, r)$ by simply adding the two numbers above and diagonally to the left and right of that position. Of course it would be useful to have a formula for $C(n, r)$ to calculate directly. To get this formula, we need to consider more rules for counting the number of ways to perform complicated processes.

In general, when dealing with a complicated process or procedure, it is useful to break it down into a sequence of simple steps. Suppose that S_1 is the set of all possible outcomes for the first step of a process which takes m steps. Now for each outcome x_1 in S_1 let $S_2(x_1)$ be the set of possible outcomes for the second step as a result of having x_1 as outcome for the first step. Next, for each outcome x_2 in $S_2(x_1)$, let $S_3(x_1, x_2)$ denote the set of possible outcomes for the third step when the result of the first two steps is x_1 followed by x_2. In general then, we could write $S_r(x_1, x_2, ..., x_{r-1})$ for the set of outcomes for step r when the result of the first $r-1$ steps is the sequence of outcomes $(x_1, x_2, ..., x_{r-1})$ for $1 \leq r \leq m$ and where for $r = 1$ we interpret $S_r(x_1, ..., x_{r-1}) = S_1$ which is the set of outcomes for the first step of the process. Now how can we count say the number of ways to perform the first two steps of the process? Notice that for each outcome x_1 in S_1 we have $card(S_2(x_1)$ ways to perform the second step, so in all, if $S_1 = \{a_1, a_2, ..., a_r\}$ we have

$$card(S_2(a_1)) + card(S_2(a_2)) + ... + card(S_2(a_r)) \tag{11.5}$$

as the number of ways to perform the first two steps. If we wanted to count the number of ways to perform the first three steps, we would have to consider the possible results for the first two steps and for each such possibility, we would add all the numbers of the form $card(S_3(x_1, x_2))$ for (x_1, x_2) a possible sequence of outcomes for the first two steps. In general, to get the number of ways to perform the first r steps we would have to add all numbers of the form $card(S_r(x_1, x_2, ..., x_{r-1}))$ for all possible sequences of outcomes $(x_1, x_2, ..., x_{r-1})$ for the previous steps. This looks terribly complicated and in general it is. If there are not too many possibilities at each number of steps, the counting can be done with the help of what is called a TREE DIAGRAM. We put a dot at the center of the top line to be the start. On the next line underneath, we list the outcomes for the first step leaving plenty of space between each outcome. Then on the next line, under each outcome above we list the outcomes of the second step when the the first step outcome is the outcome above. To keep things straight, it helps to draw a line from the outcome on the previous line to each of the outcomes at the next step of the process which could result from that previous sequence of outcomes. Doing this for each of the steps results in a diagram looking like an upsidedown tree. At each outcome at the step r you have branches leading to all the possible outcomes for the next step. Then, the number of outcomes all the way across the line for step r gives the number of ways to perform the first r steps of the process. In many cases of interest, the number $card(S_r(x_1, x_2, ..., x_{r-1}))$ does not depend on the history $(x_1, x_2, ..., x_{r-1})$ of what outcomes came before even though the actual set $S_r(x_1, x_2, ..., x_{r-1})$ may depend on the history of what came before on previous steps. For instance if we are drawing blocks from a box without replacement, one after another, then if there are 8 blocks to start and we have taken out 3 already, then when we go to draw the fourth block we know there are 5 possible outcomes no matter which three blocks were drawn on the first three steps of the process. Of course which of the 8 blocks are possible on the fourth draw does depend on which three came out on the first three steps. That is, the number of possibilities does not depend on the history at each stage of the process even though the actual set of possibilities at each stage does depend on what happened before. In this case, where the number of possibilities at each step is independent of the results of previous steps, then letting n denote the number of ways to perform the whole $m-$step process and for $1 \leq k \leq m$, letting n_k denote the number of possibilities for performing step k, given outcomes for the previous steps, we then find that the number of ways to perform the whole $m-$step process is simply

$$n = n_1 n_2 n_3 ... n_m \tag{11.6}$$

so we just multiply all the numbers of ways to do each stage together to get the number of ways to perform the overall process. We therefore call this the MULTIPLICATION PRINCIPLE for counting.

As an application of the multiplication principle we consider the problem of counting the number of arrangements of r things chosen from a set of n things. We do not have immediately what this number is, so we denote it by $P(n, r)$. For instance, $P(5, 3)$ could be the number of three letter lists made by listing letters chosen from the set containing the letters A, B, C, D, F, where each letter is only used once. Or, we could have a box of blocks which are all numbered differently, and we consider the number of ways to arrange three of them in a row starting with five in the box. Notice there are two procedures for doing this. One way is to first choose three blocks from the box and then as a second step arrange the three chosen blocks. The second way is to first choose one of the blocks for the first position in the row, then choose a second block for the next position in the row, and finally a third block for the last position. Using the last procedure, we see immediately from the multiplication principle that $P(5,3) = (5)(4)(3) = 60$ whereas using the first procedure we see that $P(5,3) = C(5,3)P(3,3)$. More generally, this same line of reasoning, thinking of a box containing n numbered blocks, tells us that

$$P(n,r) = C(n,r)P(r,r) \tag{11.7}$$

because we can accomplish the job of arranging r things chosen from the set of n things by first choosing the r things and as a second step arranging all r of the things chosen on the first step. Notice that our stepwise procedure of forming the arrangement by sequentially choosing blocks one at a time and arranging them in a row gives

$$P(r,r) = (r)(r-1)(r-2)(r-3)...(3)(2)(1) = (1)(2)(3)...(r).$$

It is useful to have a short hand notation for this. We write $r!$ for the product of the first r consecutive positive integers, so

$$r! = P(r,r) = (r)(r-1)(r-2)(r-3)...(3)(2)(1) = (1)(2)(3)...(r). \tag{11.8}$$

(By convention and consistent with the preceding, we define 0!=1.) On the other hand, another procedure we can use to arrange all n of the blocks in the box is to stop momentarily after arranging r of the blocks for a short rest, and then resume by arranging the remaining $n-r$ blocks in the box. Now the multiplication principle ([11.6]) tells us that

$$P(n,n) = P(n,r)P(n-r,n-r). \tag{11.9}$$

Consequently, by combining [11.8] and [11.9] we find

$$P(n,r) = \frac{P(n,n)}{P(n-r,n-r)} = \frac{n!}{(n-r)!}. \tag{11.10}$$

But, now by [11.7] and [11.10] we have

$$C(n,r) = \frac{P(n,r)}{P(r,r)} = \frac{n!}{r!(n-r)!}. \tag{11.11}$$

Thus [11.11] gives a handy formula for computing the number of ways to choose r things from a set of n things.

Suppose we are interested in computing the probabilities for various hands in the game of poker. A standard deck of cards consists of 52 cards of which there are four suits of thirteen cards each: hearts, diamonds, clubs, and spades. Likewise there are nine numbered cards in each suit with the numbers 2 through 10, and in addition each suit has an ace and the three face cards: jack, queen, and king. The ace can be counted as one or as higher than a king. We will call these the thirteen denominations, so there are four cards of each denomination, one of each suit. A hand in poker consists of exactly five cards. In games where you are dealt more than five cards you are allowed to choose the best five to make your hand. For instance a STRAIGHT in poker is a hand containing a sequence of 5 consecutive denominations starting at any denomination at or below ten-the ace can be used as the highest card for a straight beginning with ten or can be used as the starting denomination for a straight ending in denomination 5. In poker language if the highest card in a straight is the denomination jack, then it would be called a jack-high straight. Thus the straight starting with ace as one is a five high straight, whereas a straight starting with the ten is an ace-high straight. If all the cards in the five card hand are of the same suit, then the hand is called a FLUSH. If all the cards in a straight are of the same suit, the hand is called a STRAIGHT FLUSH. An ace-high straight flush is called a ROYAL FLUSH. Four cards of the same denomination are called four of a kind, three cards of the same denomination are called three of a kind, whereas two cards of the same denomination is called a pair. If a five card hand contains four of a kind the hand is said to be FOUR OF A KIND. If a five card hand contains three of a kind and a pair with two different denominations it is called a FULL HOUSE. If a five card hand contains three of a kind but does not qualify to be a full house it is called THREE OF A KIND. If a five card hand contains two pair but does not qualify to be either four of a kind or a full house, then it is called TWO PAIR. If a five card hand has a single pair but does not qualify to be either two pair or three of a kind, then it is called a PAIR. In poker, to compare hands, a straight flush beats four of a kind which beats a full house which beats a flush which beats a straight which beats three of a kind which beats two pair which beats a single pair. In case of a tie in the sense of two hands of the same type one first compares the highest card in the hands of the cards that are important to forming the type of hand, so for instance three tens beats three eights even if the hand with three eights also contains an ace and a king. However, in case of a full house, one compares the denomination of the three of a kind first and if there is still a tie, then one compares the denominations of the pairs. Thus a full house consisting of three nines and a pair of fours would beat a full house consisting of three sixes and a pair of aces. If two hands tie with all the cards determining the hand type then we compare the remaining highest cards until the tie is broken. Thus, a hand containing a pair of tens, a jack, a queen, and a king beats a hand containing a pair of tens, a nine, a queen, and a king. Now, we can use the counting rules and formulas to compute the probabilities of the various hands in poker. For instance, there is only one royal flush in each suit, so there are four royal flushes. On the other hand, there are $C(52, 5)$ ways to choose five cards from the standard 52-card deck, and $C(52, 5) = 2,598,960$ or about 2.6 million five card poker hands. Thus the chance of drawing five cards from the deck and getting a royal flush is 1/649740 or about 0.0000015390772. Clearly there are ten times as many (or 40) straight flushes as royal flushes so the chance of a straight flush (including a royal flush) is 10/649740=1/64974 or about 0.000015390772. Two get four of a kind, there are thirteen denominations and such a hand will contain two. We can first choose the denomination to be the four of a kind and then as a second step choose the remaining card from the remaining 48 cards in the deck. This means there are (13)(48)=624. Alternately, we can choose the two denominations which will appear which can be done in $C(13, 2) = 78$ ways and then choose one of those two to be the four of a kind which can be done in $C(2, 1) = 2$ ways, and then choose one of the four cards of the remaining denomination which can be done in $C(4, 1) = 4$ ways, so here we find the number of four of a kind poker hands is (78)(2)(4)=624, again. Two count the number of full houses, this last method is easier. A full house contains two different denominations with one being the three of a kind. We can form such a hand by first choosing the two denominations which can be done again in 78 ways, then choose the denomination to be the three of a kind which can be done in 2 ways, then choose three cards of the chosen denomination which can be done in $C(4, 3) = C(4, 1) = 4$ ways, then for the remaining denomination choose a pair which can be done in $C(4, 2) = 6$ ways. Thus there are (78)(2)(4)(6)=3744 ways to form a full house.

Sometimes we are interested in computing the number of arrangements of a list of symbols of which there are repeats. For instance, suppose that we wish to compute the number of ways to arrange the letters in the word "MISSISSIPPI". There are 11 letters in the word, but some of the rearrangements of these letters are indistinguishable because of the repeated letters. In order to work out the number of distinguishable arrangements of a sequence of letters or objects of any kind in case there are repeats in the sense that some are considered to be indistinguishable, give each object a symbol so that the indistinguishable objects get the same symbol. For instance, suppose that this results in $AAABBBBCCDDDDD$. To see how many arrangements there are for these symbols, we begin by denoting this number by x. Then as a next step, we tag the indistinguishable symbols with subscripts making them distinguishable, which in the example results in $A_1A_2A_3B_1B_2B_3B_4C_1C_2D_1D_2D_3D_4D_5$. Now these symbols are all distinguishable and therefore the number of arrangements of these distinguishable symbols is $n!$, where n is the number of symbols we began with, and which in our example is just $n = 14$. Our next step is to realize that the job of rearranging the tagged distinguishable symbols could be accomplished in two steps. In the first step we arrange the original symbols, and then in the second step we attach the subscript tags to the symbols which are alike to make all of them different again. Suppose that there are m different distinguishable symbols of which there are say n_1 of the first kind, n_2 of a second kind, and so on so finally there are n_m of the last kind. For instance, in our example there are 4 distinguishable symbols so $m = 4$ and $n_1 = 3, n_2 = 4, n_3 = 2, n_4 = 5$. In the example, if we are presented with an arrangement of the untagged symbols, then we have 3! ways to attach the subscripts to the $A's$, we have 4! ways to attach the subscripts to the $B's$, we have 2! ways to attach the subscripts to the $C's$, and finally 5! ways to attach the subscripts to the $D's$. That is in general, there would on seeing an arrangement of the unsubscripted symbols be $n_1!n_2!...n_m!$ ways to attach subscripts to the symbols so as to result in an arrangement of the subscripted symbols. Therefore by the multiplication principle, [11.6], there must be $xn_1!n_2!...n_m!$ ways to arrange the tagged symbols, so this must equal $n!$, that is we know now $n! = xn_1!n_2!...n_m!$ must be true. Solving this last equation for x we have

$$x = \frac{n!}{n_1!n_2!...n_m!}$$

for the number of ways to arrange the original symbols given to us. For instance, in our example we have 14!/3!4!2!5! ways to arrange the symbols $AAABBBBCCDDDDD$. In general it is convenient to denote this number by

$$x = C(n; n_1, n_2, ..., n_m)$$

where $n = n_1 + n_2 + ... + n_m$. We therefore have that

$$C(n; n_1, n_2, ..., n_m) = \frac{n!}{n_1!n_2!...n_m!} \tag{11.12}$$

gives the number of distinguishable arrangements of n things of which there are m types where things of the same type are considered indistinguishable, and with n_1 of a first type, n_2 of a second type, and so on with finally n_m of the last type. Because of the convention that 0!=1, the preceding remains true even if some of the $n_k's$ are zero. For instance this means that for the distinguishable arrangements of the letters in "MISSISSIPPI" there are 11!/(1!4!4!2!), since there is one "M", 4 "I's", 4"S's", and 2"P's". As an application of this formula, let us consider an algebra problem. We suppose that we have m unknowns $x_1, x_2, x_3, ..., x_m$ and suppose that we want to expand $(x_1 + x_2 + x_3 + ... + x_m)^n$. For a moment let us not allow the commutative law of multiplication, just the associative law for multiplication and the distributive law. Notice that in case $m = 3$ we get

$$(x_1 + x_2 + x_3)^n = (x_1 + x_2 + x_3)(x_1 + x_2 + x_3)^{n-1}$$

$$= x_1(x_1 + x_2 + x_3)^{n-1} + x_2(x_1 + x_2 + x_3)^{n-1} + x_3(x_1 + x_2 + x_3)^{n-1}.$$

In general, we see that if such application of the distributive law is continued until the expansion is complete, we will have a sum of terms of which each is a "word" of length n formed with the variables $x_1, x_2, x_3, ..., x_m$ as symbols. In such a word, let n_1 be the number of occurrences of x_1, let n_2 be the number of occurrences of x_2, let n_3 be the number of occurrences of x_3, and so on, finally, let n_m be the number of occurrences of x_m, so then we must have $n = n_1 + n_2 + ... + n_m$. There are then $C(n; n_1, n_2, ..., n_m)$ of these words, which will all be equal to each other if we suppose the commutative law to apply, that is all such words equal $x_1^{n_1} x_2^{n_2} x_3^{n_3} ... x_m^{n_m}$. This even applies in case some of the $n_k's$ are zero because $x_k^0 = 1$ and 0!=1. This means that

$$(11.13) \qquad (x_1 + x_2 + x_3 + ... + x_m)^n = \sum_{\binom{n_1, n_2, ..., n_m \geq 0}{n = n_1 + n_2 + ... + n_m}} C(n; n_1, n_2, ..., n_m) x_1^{n_1} x_2^{n_2} x_3^{n_3} ... x_m^{n_m}$$

gives the equation for expanding the power of the sum as a sum of products, and it is called the MULTINOMIAL EXPANSION. In case that $m = 2$ we can see immediately that $C(n; r, n - r) = C(n, r)$ and the multinomial expansion reduces to the usual BINOMIAL EXPANSION

$$(11.14) \qquad (x + y)^n = \sum_{k=0}^{n} C(n, k) x^k y^{n-k}.$$

Because of [11.13], it is customary to refer to $C(n; n_1, n_2, ..., n_m)$ as a MULTINOMIAL COEFFICIENT, and likewise from [11.14], it is the case that $C(n, r)$ is called a BINOMIAL COEFFICIENT. For instance in a sequence of n independent trials to see how many times event A happens, the outcome could be recorded as a word in the letters "s,f", where we put an "s" for a trial on which A happens and "f" otherwise. Thus the sequence or word "ssfsf" denotes the outcome of five trials where on the first two trials A happened, the third trial resulted in A failing to happen, on the fourth trial A did happen, and on the fifth trial A failed to happen. Notice that if $P(A) = p$ and $P(S \setminus A) = q$, then the probability of the sequence "ssfsf" is $ppqpq = p^3q^2$. In fact the probablity of each sequence where A happened exactly 3 times out of the five independent trials would be the same, and there are $C(5, 3)$ of these sequences. Thus, the probability that A happens 3 times in 5 independent trials is $C(5, 3)p^3q^2$. More generally, if we make a sequence of n independent trials, then the probability that A happens exactly k times is the number of words of length n having "s" appear k times and "f" appear $n - k$ times multiplied by the probability of any one of these sequences, and each such sequence has probability $p^k q^{n-k}$. Thus, we can say that if T_n denotes the total number of times that A happens in n independent trials, then

$$(11.15) \qquad P(T_n = k) = C(n, k) p^k q^{n-k}.$$

We call [11.15] the formula for the BINOMIAL DISTRIBUTION, and we say that in this case, T_n has the binomial distribution.

Notice that the only possibilities for values of T_n are the whole numbers $0, 1, 2, 3, ..., n$, and [11.15] gives the probability of each. If we are not making independent trials but are choosing from a finite sample space of size N without replacement where all outcomes are equally likely, and if A contains exactly R outcomes, then $p = R/N$ and $q = (N - R)/N$. Now if T_n denotes the number of times A happens, that is the number of choices from S which resulted in an outcome from A, then since we are choosing without replacement,

$$P(T_n = k) = \frac{C(R,k)C(N-R,n-k)}{C(N,n)}, \tag{11.16}$$

which is the formula for the HYPERGEOMETRIC DISTRIBUTION, and we say that T_n has the hypergeometric distribution. If we choose with replacement, then the choices become independent of each other, and the distribution must be binomial, whereas when we choose without replacement from a finite population, the distribution is hypergeometric. This distinction between independence and dependence will be made mathematically precise below in what follows.

12. NOTATION

Henceforth it will be convenient to identify statements and events with their indicator unknowns and random variables, so if a statement or event is in a place that requires a random variable, it should be interpreted as the indicator. Thus, in particular, if A and B are statements, then $E(A|B) = E(I_A|B) = P(A|B)$, $Cov(A,B) = Cov(I_A, I_B)$, and $Var(A) = Var(I_A)$.

13. INDEPENDENCE AND UNCORRELATION

Suppose that X and Y are both unknowns and recall from [5.6] that $Cov(X,Y) = E(XY) - E(X)E(Y)$. Also recall that X and Y are uncorrelated if and only if $Cov(X,Y) = 0$. We can therefore see that X and Y are uncorrelated if and only if $E(XY) = E(X)E(Y)$. Notice that for statements A and B this is the same as $P(A \cap B) = P(A)P(B)$. Since $Cov(X,1) = 0$ and $P(notA) = 1 - P(A)$, it is easy to see from [5.8] that if A and B are uncorrelated, then $notA$ and B are uncorrelated, likewise A and $notB$ are uncorrelated, and so $notA$ and $notB$ are uncorrelated. Thus, these are all equivalent forms of saying A and B are uncorrelated. Moreover, from [10.21] we see that $P(A \cap B) = P(A)P(B)$ says the same thing as $P(A|B) = P(A)$ and therefore also says the same thing as $P(B|A) = P(B)$. To summarize, we can say that for statements A and B to be uncorrelated is equivalent to any one of the following equations

$$Cov(A,B) = 0, \tag{13.1}$$

$$Cov(notA,B) = 0, \tag{13.2}$$

$$Cov(A,notB) = 0, \tag{13.3}$$

$$Cov(notA,notB) = 0, \tag{13.4}$$

$$P(A \cap B) = P(A)P(B), \tag{13.5}$$

$$P(A|B) = P(A), \tag{13.6}$$

$$P(B|A) = P(B), \tag{13.7}$$

which means if ANY ONE of these are TRUE, THEN they are ALL TRUE.

Let us now consider in more detail what independence means so we can formulate it in a precise mathematical way. To say unknowns X and Y are independent, as indicated previously, means that no information about the value of X will be is of any use in guessing something about the value of Y. More precisely, if A is a statement about the value of X and B is a statement about the value of Y, then for X and Y to be independent it must be the case that knowing whether A is true is of no use in determining whether B is true, no matter how A and B are chosen as statements about X and Y, respectively. But this just says that for any statement A about X and any statement B about Y it is the case that $P(B|A) = P(B)$ which is equivalent to saying that A and B are uncorrelated. If X and Y are just statements themselves, then as the only thing to say about the value of an indicators is whether or not it has the value 1, from [13.1] we see that for statements, independence and uncorrelation are exactly the same thing. In the case of discrete unknowns, we must require that for any real numbers c, d we have the events $(X = c)$ and $(Y = d)$ are themselves uncorrelated. For more general random variables, we have to consider what the possibilities are for what can be said about the value of an unknown. This becomes somewhat complicated when dealing with possibly continuous unknowns, but as we can use decimal approximation, to an arbitrary degree of accuracy, we can finally realize that in general for X and Y to be independent, we need the events $(a \leq X \leq b)$ and $(c \leq Y \leq d)$ to be uncorrelated for any choice of real numbers a, b, c, d. We can now show that independent simple unknowns are uncorrelated. For by [4.25] applied to both X and Y, we see that $Cov(X, Y)$ will be a sum of terms each containing a factor of the form $Cov((X = c), (Y = d))$ for c, d possible values of X and Y respectively; but by the assumption of independence of X and Y, all such covariances are zero. The same is true if X and Y are discrete, with an argument that is a little more careful because of the possible infinite number of terms. In the general case, we know that we can approximate with discrete unknowns by rounding off to any desired degree of accuracy, so the general case follows from the discrete case. In essence, the fact that independent unknowns are uncorrelated follows from the fact that independent simple unknowns are uncorrelated. Let us look at this argument in a little more detail for simple unknowns. Suppose that X and Y are simple unknowns. Then we can find a partition of the sure statement into mutually exclusive statements $A_1, A_2, A_3, ..., A_m$ and distinct constants $a_1, a_2, a_3, ..., a_m$ so that

$$X = a_1A_1 + a_2A_2 + a_3A_3 + ... + a_mA_m. \tag{13.8}$$

Likewise, we can do the same for Y, that is we can find a partition $B_1, B_2, B_3, ..., B_n$ and distinct constants $b_1, b_2, b_3, ..., b_n$ so that

$$Y = b_1B_1 + b_2B_2 + b_3B_3 + ... + b_nB_n. \tag{13.9}$$

Now, $X(S) = \{a_1, a_2, a_3, ..., a_m\}$ is the set of possible values of X and for each k, we have as events $(X = a_k) = A_k$, so $P(X = a_k) = P(A_k)$. Likewise, for Y we have as events $(Y = b_l) = B_l$, so $P(Y = b_l) = P(B_l)$, for each l. Thus the condition that for any real numbers c, d the events $(X = c)$ and $(Y = d)$ are uncorrelated (which is the condition for independence of X and Y) become simply the condition that for any k, l we have A_k and B_l are uncorrelated with each other, that is $Cov(A_k, B_l) = 0$. But then $Cov(X, Y)$ is a sum of terms of the form $Cov(a_kA_k, b_lB_l) = a_kb_lCov(A_k, B_l)$, so the condition that X and Y be independent now makes all these terms equal to zero. That is to say, if X and Y are independent, then they are uncorrelated. Notice the converse is not true, because for X and Y to be uncorrelated it is only necessary for the sum of these terms to be zero, which is not as stringent as requiring each term in the sum to be zero.

Now, let $\mathbb{R}$ denote the set of all real numbers. It can be thought of as a sample space for the experiment of choosing a real number. For this special sample space, it is customary to use lower case letters near the middle of the alphabet to denote random variables. Suppose that X is a random variable on S and f is a random variable on $\mathbb{R}$. We can then form a new random variable denoted $f(X)$ by declaring that its value is $f(x)$ when the value of X is x. That is, if s in S is the outcome of the experiment, then $X(s)$ is the value of X and therefore $f(X(s))$ is the value of $f(X)$ when the outcome is s. Notice that for the simple random variable X given by [13.10] we have

$$f(X) = f(a_1)A_1 + f(a_2)A_2 + f(a_3)A_3 + ... + f(a_m)A_m. \tag{13.10}$$

This means that for a simple random variable X, once we have the representation of X as in [13.10], then when we form $f(X)$, we can use the same partition of the sample space. Of course, the same applies to Y in case Y is another simple random variable, and consequently, since independence of the simple variables here just boils down to $Cov(A_k, B_l) = 0$ for all k, l, it follows that for simple random variables X and Y, if they are independent, then so are $f(X)$ and $g(Y)$ for any random variables f and g on $\mathbb{R}$. Conversely, if X and Y are simple and if for any random variables f, g on $\mathbb{R}$ it is the case that $f(X)$ and $g(Y)$ are uncorrelated, then we will demonstrate that X and Y must be independent. Begin by choosing for each k a random variable f_k which has the property that $f_k(a_l) = 0$ if l is not equal to k but $f_k(a_k) = 1$. The simplest way to do this is to take f_k to be the random variable which as an event pictured in set form is just $\{a_k\}$. Likewise, for each l choose a random variable g_l with $g_l(b_k) = 0$ if k is not equal to l but $g_l(b_l) = 1$. Then $f_k(X) = A_k$ and $g_l(Y) = B_l$ and therefore A_k and B_l are uncorrelated for each k, l. But this means that X and Y are independent. By using the decimal approximation carefully, we can say that for any random variables X and Y, the condition of being independent is the same as the condition that for any random variables f, g on the sample space $\mathbb{R}$ it is the case that $f(X)$ and $g(Y)$ are uncorrelated and this in turn is equivalent to the condition that $f(X)$ and $g(Y)$ be independent for any random variables f, g on $\mathbb{R}$. Consider now that when we form $f(X)$ we are encoding the values of X in some way, which can possibly destroy some of the information in X and emphasize other aspects of the information in X. In this view, the condition of independence of X and Y is that no matter how this is done, the results remain uncorrelated.

14. SAMPLING

Expected values and probabilities are approximated in practical situations by sampling. When we sample the random variable X we are making repeated observations of X and before the observations are made, we can think of the various observations as forming themselves new random variables. Thus we let X_1 be the random variable whose value on a given sample is the first observed value, X_2 be the random variable whose value on a given sample is the second observed value, X_3 be the random variable whose value on a given sample is the third observed value, and so on, so X_k is the random variable whose value on a given sample is the k^{th} observed value, $1 \leq k \leq n$. We notice right away that each X_k has the same distribution as X and therefore, for each k

$$E(X_k) = \mu_X \tag{14.1}$$

and

$$SD(X_k) = \sigma_X. \tag{14.2}$$

The reason for sampling is to estimate the population parameters we are interested in such as the population mean and the population variance. To do this we look for combinations of the sample data which have expected value equal to the population parameter we are interested in. First consider the population mean. A reasonable guess is that the mean of a sample should give an estimate. If we take the mean of a sample of size n, then in terms of the random variables

$$X_1, X_2, X_3, ..., X_n,$$

we see that we are totaling the variables and dividing by n. We can therefore define T_n as the total of these variables, so

$$T_n = X_1 + X_2 + X_3 + ... + X_n, \tag{14.3}$$

and as a random variable, the sample mean is $\bar{X}_n$, where

$$\bar{X}_n = \frac{1}{n} T_n. \tag{14.4}$$

But now, if $\mu = E(X)$, then repeated application of the basic properties of expectation tells us that

$$E(T_n) = n\mu, \tag{14.5}$$

and therefore

$$E(\bar{X}_n) = \mu. \tag{14.6}$$

That is, the expected value of the sample mean is the true population mean, just what we were looking for. Moreover, [14.5] just says that the expected value of a total of observations of X is simply the true mean of X multiplied by the number of observations made. The situation for variance of X in relation to sampling is more complicated. First, we see that unless we have an assumption about how the different observations of X are correlated, we cannot even calculate the variance of T_n. The simplest assumption to make is that all observations are independent of each other. In this case we say that we are doing INDEPENDENT RANDOM SAMPLING(IRS). If $Var(X) = \sigma^2$, then for each k we have $Var(X_k) = \sigma^2$, so by [5.13] we have

$$Var(T_n) = n\sigma^2, \tag{14.7}$$

and therefore

$$\sigma_{T_n} = \sqrt{n}\sigma. \tag{14.8}$$

Consequently, since $\bar{X}_n = (1/n)T_n$, we have by [5.19],

$$\sigma_{\bar{X}_n} = \frac{1}{\sqrt{n}} \sigma_X. \tag{14.9}$$

Squaring both sides of this last equation then tells us that

$$Var(\bar{X}_n) = \frac{Var(X)}{n} = \frac{\sigma_X^2}{n}. \tag{14.10}$$

By [14.6], the expected value of the sample mean is the true population mean. We want to find what combination of the observations will have the population variance as its expected value. Before giving the detailed calculation, let us use [5.7]. Applied to the variable X, recall it says that

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

whereas as a special case applied to a population of size n consisting of the independent observations $X_1, X_2, ..., X_n$ it says that

$$\overline{((X - \bar{X})^2)} = \bar{X^2} - (\bar{X})^2. \tag{14.11}$$

If we apply [14.6] to [14.11] we find that

$$E(\overline{((X - \bar{X})^2)}) = E(X^2) - E((\bar{X})^2),$$

and applying [5.7] to each term on the right together with [14.10] gives

$$E(\overline{((X - \bar{X})^2)}) = (\mu^2 + \sigma^2) - (\mu^2 + \frac{1}{n}\sigma^2).$$

Now, cancelling and simplifying gives the simple result

$$E(\overline{((X - \bar{X})^2)}) = \frac{(n-1)}{n}\sigma^2. \tag{14.12}$$

We say a variable whose expected value is the number v is an unbiased estimator of v. Thus $\bar{X}$ is an unbiased estimator of $E(X)$, the true mean of X. From [14.12] we see on multiplying both sides by the reciprocal of the coefficient of σ^2 that

$$E(\frac{n}{n-1}\overline{((X - \bar{X})^2)}) = \sigma^2 \tag{14.13}$$

giving us an unbiased estimator of the population variance.

We now want to look at a more detailed way of calculating this fact which will lead to a formula which is valid without necessarily using independent random sampling (IRS). First let us ask for the covariance of the sample observations with the sample mean, in IRS. That is, if $0 \leq k \leq n$, then what is $Cov(X_k, \bar{X}_n)$? To do this, we begin by calculating $Cov(X_k, T_n)$. By [5.10] and [14.3], we see that $Cov(X_k, T_n) = Cov(X_k, X_k) = Var(X) = \sigma_X^2$, and therefore

$$Cov(X_k, T_n) = \sigma_X^2, \ 1 \leq k \leq n \tag{14.14}$$

and

$$Cov(X_k, \bar{X}_n) = \frac{Var(X)}{n} = \frac{\sigma_X^2}{n}, \ 1 \leq k \leq n. \tag{14.15}$$

Since $E(X_k) = \mu = E(\bar{X}_n)$, it follows that $E(X_k - \bar{X}_n) = 0$, and therefore that

$$E((X_k - \bar{X}_n)^2) = Var(X_k - \bar{X}_n),$$

so by [5.12] and [14.10], we find

$$E((X_k - \bar{X}_n)^2) = \sigma_X^2 + \frac{\sigma_X^2}{n} - 2\frac{\sigma_X^2}{n}, \tag{14.16}$$

and cancelling the positive terms with negative terms leaves

$$E((X_k - \bar{X}_n)^2) = \sigma_X^2 - \frac{\sigma_X^2}{n} = \frac{(n-1)\sigma_X^2}{n}, \tag{14.17}$$

so finally,

$$E((X_k - \bar{X}_n)^2) = \frac{(n-1)\sigma_X^2}{n}. \tag{14.18}$$

Now, we almost have the objective we want. We define that sample variance random variable, S_n^2, by the equation

$$S_n^2 = \frac{(X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + (X_3 - \bar{X}_n)^2 + ... + (X_n - \bar{X}_n)^2}{n-1}, \tag{14.19}$$

so that

$$(n-1)S_n^2 = (X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + (X_3 - \bar{X}_n)^2 + ... + (X_n - \bar{X}_n)^2. \tag{14.20}$$

Thus we see that in our previous notation,

$$\overline{((X - \bar{X})^2)} = \frac{(n-1)}{n} S_n^2. \tag{14.21}$$

Now, the right hand side of [14.20] has exactly n terms and by [14.18], each has the same expected value, so multiplying the right hand side of [14.18] by n gives us the expected value of $(n-1)S_n^2$, and we see that $E((n-1)S_n^2) = (n-1)\sigma_X^2$. Since the $n-1$ is a factor common to both sides here, it too cancels, with the result that

$$E(S_n^2) = \sigma_X^2, \tag{14.22}$$

which is the same as [14.13] in view of [14.21].

With a more general sampling method, the various different observations may be correlated with each other to some extent, but this correlation effect should be independent of the particular observation, that is $Cov(X_k, T_n)$ should not depend on the particular choice of k. Let

$$c = \frac{Cov(X_k, T_n)}{\sigma_X^2}. \tag{14.23}$$

which we will refer to as the SAMPLING CORRELATION CONSTANT. Then

$$Cov(X_k, T_n) = c\sigma_X^2, \ 1 \leq k \leq n, \tag{14.24}$$

and as $\bar{X}_n = (1/n)T_n$,

$$Cov(X_k, \bar{X}_n) = \frac{c\sigma_X^2}{n}, \ 1 \leq k \leq n. \tag{14.25}$$

Since $T_n = X_1 + X_2 + X_3 + ... + X_n$, it follows that

$$Var(T_n) = Cov(T_n, T_n) = nc\sigma_X^2, \tag{14.26}$$

and therefore

$$Var(\bar{X}_n) = \frac{c}{n}\sigma_X^2. \tag{14.27}$$

Comparing [14.14] and [14.24], we see that for the case of IRS, that is in independent random sampling, we have $c_{IRS} = 1$. Notice this means the sampling correlation constant is just a simple correction factor for the ordinary IRS equations, [14.7] and [14.10], for variance of the random variables T_n and $\bar{X}_n$. Now the main ingredient of the sample variance is the sum of the squared deviations from the sample mean which we can denote by U_n, so

$$U_n = (n-1)S_n^2 = (X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + (X_3 - \bar{X}_n)^2 + ... + (X_n - \bar{X}_n)^2. \tag{14.28}$$

To compute $E(U_n)$ we need to compute the expected value of each term on the right side of [14.28], and as before we realize this means computing the variance of $X_k - \bar{X}_n$. But again, since we know the variance of X_k is just σ_X^2 and [14.27] tells us the variance of $\bar{X}_n$, we can now apply [5.12] to find

$$E((X_k - \bar{X}_n)^2) = \sigma_X^2 + \frac{c\sigma_X^2}{n} - 2\frac{c\sigma_X^2}{n} = \frac{(n-c)\sigma_X^2}{n}. \tag{14.29}$$

Consequently, multiplying the right side of [14.29] by n gives the expected value of U_n. Thus,

$$E(U_n) = (n-c)\sigma_X^2, \tag{14.30}$$

which tells us that

$$\sigma_X^2 = E(\frac{U_n}{n-c}). \tag{14.31}$$

That is, $(n-c)^{-1}U_n$ has σ_X^2 as its expected value, so that is what we would compute to estimate the true population variance from the sample data. As pointed out earlier, in case of IRS, we have $c_{IRS} = 1$, so $(n-c)^{-1}U_n$ reduces to S_n^2.

The most common sampling method is SIMPLE RANDOM SAMPLING (SRS). The requirement for simple random sampling is that all samples of size n from the population are equally likely to be the chosen sample. In practice, we are usually observing measurements on objects randomly chosen from a population of objects, so for SRS for each observation after observing the randomly chosen object, we remove it from the population, so that we do not ever observe anything twice. That is we sample WITHOUT REPLACEMENT. If we replace each randomly chosen object back in the population after observing it, then on a later random selection, we may end up observing the same object. When we sample in this manner, we are sampling WITH REPLACEMENT. When sampling with replacement it is clear that different observations are independent of each other so we have IRS, whereas when we sample without replacement we get SRS, and it is easy to see that there must be correlation between the different observations if the population is finite. For instance, if our population is a box full of envelops each containing a campaign donation in the form of a check, if the experiment is to randomly choose an envelop from the box and see how much the check is worth (assuming it is not a bad check), then when we sample without replacement, no envelop is observed twice. Getting a big check on observation k makes it less likely to get a big check on other draws. That is, when sampling without replacement in a finite population, as the large values get "used up", they become less likely. This means that we expect the different observations to be negatively correlated. On the other hand, if $i \neq j$ and $1 \leq i, j \leq n$, then $Cov(X_i, X_j)$ should not depend on the particular choice of i, j, so let

$$b = Cov(X_i, X_j),\ i \neq j,\ 1 \leq i, j \leq n. \tag{14.32}$$

In order to find b, the trick is to consider a sample of size $n = N$, for then T_N being the population total is constant and therefore $Cov(X_k, T_N) = 0$. But $Cov(X_k, T_N)$ is a sum of N terms of which $N-1$ are covariances of different observations and one term is the covariance of X_k with itself (its variance). Therefore we get

$$0 = Cov(X_k, T_N) = \sigma_X^2 + (N-1)b, \tag{14.33}$$

so $b = -\sigma_X^2/(N-1)$, and therefore we find that for SRS

$$Cov(X_i, X_j) = -\frac{\sigma_X^2}{N-1}, \; i \neq j, \; 1 \leq i, j \leq n. \tag{14.34}$$

We can see right away from [14.34] that the different observations are indeed negatively correlated in SRS, and as N is in the denominator, as N tends to infinity this correlation becomes negligible. Thus for an infinite population, the different observations are uncorrelated so the calculations above for IRS apply. On the other hand, for a finite N we see that we do have correlation between different observations which must be taken into account. For the sample of size n, the calculation of $Cov(X_k, T_n)$ now has n terms of which one is equal to the variance and the remaining $n-1$ terms all equal b, so

$$Cov(X_k, T_n) = \sigma_X^2 + (n-1)b = \sigma_X^2 - (n-1)\frac{\sigma_X^2}{N-1} = \frac{N-n}{N-1}\sigma_X^2. \tag{14.35}$$

By [14.23], we see that the sampling correlation constant for SRS is

$$c_{SRS} = \frac{N-n}{N-1}. \tag{14.36}$$

Thus by [14.26] and [14.27] we have for SRS the variances of T_n and $\bar{X}_n$ are

$$Var(\bar{X}_n) = \frac{(N-n)}{(N-1)}\frac{\sigma_X^2}{n} \tag{14.37}$$

and

$$Var(T_n) = \frac{(N-n)}{N-1}n\sigma_X^2. \tag{14.38}$$

For the estimation of population variance, by [14.31] we now should use $(n - c_{SRS})^{-1}U_n$, and by [14.36],

$$\frac{1}{n - c_{SRS}} = \frac{N-1}{nN - n - N + n} = \frac{N-1}{N(n-1)}, \tag{14.39}$$

and therefore for SRS we have the estimation of population variance is given by

$$\sigma_X^2 = E(\frac{N-1}{N(n-1)}U_n) = \frac{N-1}{N}E(S_n^2). \tag{14.40}$$

In practice, we would usually be using simple random sampling (SRS) to estimate population variance when the population is large, so the factor $(N-1)/N$ can be assumed to be nearly 1 and we can use S_n^2 to estimate population variance in SRS as well as in independent random sampling (IRS) when dealing with large populations.

Let us apply Tchebeychev's Inequality, [9.12], to these results for the case of independent random sampling (IRS). We are going to replace X in [9.12] by $\bar{X}_n$. Then by [14.9], we must replace σ_X in [9.12] by $\sigma_X/\sqrt{n}$. So that [9.12] becomes

$$P(\frac{k\sigma_X}{\sqrt{n}} \leq |\bar{X}_n - \mu_X|) \leq \frac{1}{k^2}. \tag{14.41}$$

This is the same as saying that

$$P(|\bar{X}_n - \mu_X| < \frac{k\sigma_X}{\sqrt{n}}) \geq 1 - \frac{1}{k^2}. \tag{14.42}$$

This means that by choosing k large enough, we can obtain any desired level of certainty (probabilistically speaking) that

$$|\bar{X}_n - \mu_X| < \frac{k\sigma_X}{\sqrt{n}} \tag{14.43}$$

is actually true. That is, we say that for large enough k we are ALMOST SURE that [14.43] is actually true. If we have an upper bound on how big σ_X can be, then since $\sqrt{n}$ is in the denominator on the right hand side of [14.43], we can by choosing n large enough insure that the right hand side is as small as we please. That is, by first choosing k large enough and then n large enough, we can get the sample mean to come as close as we like to the true mean with almost certainty. This means that in particular, thinking of $E(X)$ as the overall long run average over infinitely many observations of X is forced on us by [9.12] and our basic assumptions about the properties expectation must satisfy. In case that $X = A$ is an event, it is easy to see that $\bar{X}_n = \bar{A}_n$ is the relative frequency of occurrence of A in n independent trials of the experiment. This means we are virtually forced to accept the interpretation of the probability of an event as the relative frequency of occurrence of A over infinitely many independent trials of the experiment, merely as a consequence of the four basic properties we assumed for the expectation.

Finally, let us apply our sampling results to compute the mean and variance of the binomial and hypergeometric distributions. Thus we take our random variable X that we are sampling to be an event A. Then with $P(A) = p$ and $P(S \setminus A) = q$ we have that T_n can be viewed as the total number of times that A happens in the sample of size n. In case we use IRS, then T_n has the binomial distribution whereas if the population is finite and we use SRS, then T_n has the hypergeometric distribution. Since $E(A) = P(A) = p$, we see that by [14.5],

$$E(T_n) = np \tag{14.44}$$

Since $\sigma_A^2 = p - p^2 = pq$, we see by [14.7] that for the case of IRS,

$$Var(T_n)_{IRS} = npq. \tag{14.45}$$

Here we see that the computation of mean and variance of the binomial distribution given by the distribution formula, [11.15], is more difficult than by applying a little theory. If we count the number of times A happens using simple random sampling(SRS) in a finite population of size N instead of independent random sampling, then the distribution is hypergeometric instead of binomial, and our theory tells us that we still have the same mean, np, but now the variance is multiplied by the SRS correction factor $(N - n)/(N - 1)$ by [14.36]. On the other hand, if A has R outcomes and the whole sample space S has N outcomes, all equally likely, then $p = R/N$ and using SRS we see that it has mean

$$E(T_n) = np = nR/N \tag{14.46}$$

and from [14.36], it has variance

$$Var(T_n = k)_{SRS} = \frac{(N - n)npq}{(N - 1)} = \frac{(N - n)R(N - R)}{N^2(N - 1)}. \tag{14.47}$$

Again, the theory tells us the mean and variance (here for the hypergeometric distribution) more easily than computing with the distribution given by the formula [11.16].

15. SAMPLING DISTRIBUTIONS

By the SAMPLING DISTRIBUTION, we mean the distribution of $\bar{X}_n$ for samples of X of size n. In order to obtain the actual distributions of the sample mean random variable in the case of a normal random variable, we have to rewrite the sample variance as a sum of squares of uncorrelated standard variables, up to the factor σ^2. It is instructive to see how this is done for any sequence of random variables, the algebra is the same, and it shows how the general normal random variable has a determined sampling distribution. Let us begin with a positive real number σ and any sequence of random variables

$$X_1, X_2, X_3, ..., X_n, ...$$

For any natural number n, let us define

$$T_n = X_1 + X_2 + X_3 + ... + X_n, \tag{15.1}$$

$$\bar{X}_n = \frac{1}{n} T_n, \tag{15.2}$$

$$Y_n = X_{n+1} - \bar{X}_n, \tag{15.3}$$

$$Z_n = \frac{Y_n}{\sigma}\sqrt{\frac{n}{n+1}}. \tag{15.4}$$

We then have

$$\sigma^2 Z_n^2 = \frac{n}{n+1} Y_n^2, \tag{15.5}$$

$$T_n = n\bar{X}_n, \tag{15.6}$$

and

$$T_{n+1} - T_n = X_{n+1}. \tag{15.7}$$

Then using these equations,

$$\bar{X}_{n+1} - \bar{X}_n = \frac{1}{n+1} T_{n+1} - \bar{X}_n = \frac{T_{n+1} - (n+1)\bar{X}_n}{n+1} = \frac{T_{n+1} - T_n - \bar{X}_n}{n+1}, \tag{15.8}$$

$$\frac{T_{n+1} - T_n - \bar{X}_n}{n+1} = \frac{X_{n+1} - \bar{X}_n}{n+1} = \frac{Y_n}{n+1}, \tag{15.9}$$

and hence

$$\bar{X}_{n+1} - \bar{X}_n = \frac{Y_n}{n+1}. \tag{15.10}$$

On the other hand,

$$X_{n+1} - \bar{X}_{n+1} = X_{n+1} - \frac{T_{n+1}}{n+1} = \frac{nX_{n+1} + X_{n+1} - T_{n+1}}{n+1} = \frac{nX_{n+1} - T_n}{n+1}, \tag{15.11}$$

$$\frac{nX_{n+1} - T_n}{n+1} = \frac{nX_{n+1} - n\bar{X}_n}{n+1} = n\frac{X_{n+1} - \bar{X}_n}{n+1} = \frac{n}{n+1} Y_n \tag{15.12}$$

and therefore,

$$X_{n+1} - \bar{X}_{n+1} = \frac{n}{n+1} Y_n. \tag{15.13}$$

Now, since we are after sample variance, define

$$U_n = \sum_{k=1}^{n} (X_k - \bar{X}_n)^2. \tag{15.14}$$

Our aim is to show that

$$U_n = \sigma^2 \sum_{k=1}^{n-1} Z_k^2. \tag{15.15}$$

To accomplish this, we will observe that for any n,

$$U_{n+1} - U_n = \sigma^2 Z_n^2, \tag{15.16}$$

as the result [15.15] then follows from [15.16] by induction. Now, applying the factorization $a^2 - b^2 = (a+b)(a-b)$, we see

$$\begin{aligned} U_{n+1} - U_n &= \sum_{k=1}^{n} [(X_k - \bar{X}_{n+1})^2 - (X_k - \bar{X}_n)^2] + (X_{n+1} - \bar{X}_{n+1})^2 \\ &= \sum_{k=1}^{n} (X_k - \bar{X}_{n+1} + X_k - \bar{X}_n)(\bar{X}_n - \bar{X}_{n+1}) + (\frac{n}{n+1} Y_n)^2 \\ &= \sum_{k=1}^{n} (\bar{X}_{n+1} + \bar{X}_n - 2X_k)(\bar{X}_{n+1} - \bar{X}_n) + (\frac{n}{n+1} Y_n)^2 \\ &= \frac{Y_n}{n+1} \sum_{k=1}^{n} (\bar{X}_{n+1} + \bar{X}_n - 2X_k) + (\frac{n}{n+1} Y_n)^2 \\ &= \frac{Y_n}{n+1} [n(\bar{X}_{n+1} + \bar{X}_n) - \sum_{k=1}^{n} 2X_k] + (\frac{n}{n+1} Y_n)^2 \\ &= \frac{Y_n}{n+1} [n(\bar{X}_{n+1} + \bar{X}_n) - 2T_n] + (\frac{n}{n+1} Y_n)^2 \\ &= \frac{Y_n}{n+1} [n(\bar{X}_{n+1} + \bar{X}_n) - 2n\bar{X}_n] + (\frac{n}{n+1} Y_n)^2 \\ &= \frac{Y_n}{n+1} [n(\bar{X}_{n+1} - \bar{X}_n)] + (\frac{n}{n+1} Y_n)^2 \\ &= (\frac{Y_n}{n+1}) n (\frac{Y_n}{n+1}) + (\frac{n}{n+1} Y_n)^2 = \frac{n^2 + n}{(n+1)^2} Y_n^2 \\ &= \frac{n(n+1)}{(n+1)^2} Y_n^2 = \frac{n}{n+1} Y_n^2 = \sigma^2 Z_n^2. \end{aligned} \tag{15.17}$$

Thus finally we have [15.16] and therefore also [15.15].

Suppose that X is a random variable and that the preceding sequence of random variables are successive independent observations of X, and assume that $\mu = E(X)$ and $\sigma = \sigma_X$. Then

$$Cov(X_k, T_n) = \sigma^2, \ k \leq n$$
$$Cov(X_k, T_n) = 0, \ k > n.$$

Thus

$$Cov(X_k, \bar{X}_n) = \frac{\sigma^2}{n},\ k \leq n, \tag{15.18}$$

but is zero for $k > n$. Thus if $m \leq n$, then $Cov(T_m, \bar{X}_n) = (m/n)\sigma^2$, and therefore $Cov(\bar{X}_m, \bar{X}_n) = (\sigma^2)/n$ when $m \leq n$. Denoting the maximum of m and n by $m \vee n$, we therefore have

$$Cov(\bar{X}_m, \bar{X}_n) = \frac{\sigma^2}{m \vee n}. \tag{15.19}$$

Now, if $m < n$, then

$$\begin{aligned} Cov(Y_m, Y_n) &= Cov(X_{m+1} - \bar{X}_m, X_{n+1} - \bar{X}_n) \\ &= 0 + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} - 0 = 0. \end{aligned}$$

Thus the sequence

$$Y_1, Y_2, Y_3, ..., Y_n, ...$$

is a sequence of pairwise uncorrelated random variables. Notice also that as X_{n+1} and $\bar{X}_n$ are uncorrelated, it follows that

$$Var(Y_n) = Var(X_{n+1}) + Var(\bar{X}_n) = \sigma^2 + \frac{\sigma^2}{n} = \frac{n+1}{n}\sigma^2. \tag{15.20}$$

Moreover, as $E(X_{n+1}) = \mu = E(\bar{X}_n)$, it follows now that $E(Y_n) = 0$ for every n, and therefore

$$Z_n = \frac{Y_n}{\sigma}\sqrt{\frac{n}{n+1}}$$

is standard, so

$$Z_1, Z_2, Z_3, ..., Z_n, ...$$

is a sequence of pairwise uncorrelated standard random variables satisfying

$$(n-1)S_n^2 = \sigma^2 \sum_{k=1}^{n-1} Z_k^2. \tag{15.21}$$

We will not discuss multivariable joint distributions, but here we have the $Z-$sequence depends linearly on the $X-$sequence so that if X is normal, then so is Z_n for every n, and it is known that in this case pairwise uncorrelated implies pairwise independent. Now, the definition of the CHI-SQUARE DISTRIBUTION, denoted χ^2-distribution, for d degrees of freedom is

$$\chi_d^2 = \sum_{k=1}^{d} Z_k^2, \tag{15.22}$$

where $Z_1, Z_2, ..., Z_d$ are pairwise independent standard normal random variables. That is, if Z is a standard normal random variable, then χ_d^2 has the distribution of the sample total for IRS samples of Z^2 of size d. Thus if X is normal, then our preceding equations, [15.21] and [15.22], show

$$\frac{(n-1)S_n^2}{\sigma^2} = \chi_{n-1}^2, \tag{15.23}$$

that is to say $(n-1)S_n^2/\sigma^2$ has the χ^2-distribution for $n-1$ degrees of freedom, which is χ^2_{n-1}. Consequently,

$$\frac{S_n^2}{\sigma^2} = \frac{\chi^2_{n-1}}{n-1} \tag{15.24}$$

The STUDENT $t-$DISTRIBUTION for d degrees of freedom, denoted t_d, is by definition the distribution of

$$t_d = \frac{Z_0}{\sqrt{\frac{\chi_d^2}{d}}} \tag{15.25}$$

where $Z_0, Z_1, Z_2, ..., Z_d$ are independent standard normal random variables with $Z_1, Z_2, Z_3, ..., Z_d$ defining χ_d^2. Notice that the the term in the radical in the denominator here, $(\chi_d^2)/d$, is itself just the sample mean for a sample of Z^2 of size d. Now, if we define W_n by

$$W_n = \frac{\bar{X}_n - \mu_X}{\sqrt{\frac{S_n^2}{n}}}, \tag{15.26}$$

then to see that W_n has the $t-$distribution for $n-1$ degrees of freedom, we define the sequences (Y_m) and (Z_m) as above in terms of the sequence (X_m), and then we put $Y_0 = \bar{X}_n - \mu$. Then by [15.18] and [15.19], we see that

$$Cov(Y_0, X_k) = (\sigma^2/n) = Cov(Y_0, \bar{X}_{k-1})$$

and therefore

$$Cov(Y_0, Y_k) = 0, \;\; k = 1, 2, 3, ..., n-1.$$

It follows that as $Var(Y_0) = (\sigma^2/n)$, we have that the standardization of Y_0 is Z_0 where

$$Z_0 = \frac{Y_0}{\sqrt{\sigma^2/n}}.$$

Thus, by [15.24], we see

$$W_n = \frac{Z_0}{\sqrt{\frac{S_n^2}{\sigma^2}}} = \frac{Z_0}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = t_{n-1}. \tag{15.27}$$

This means that replacing population variance with sample variance in the formula for standardizing the sample mean random variable, in case X is normal, results in the $t-$distribution for $n-1$ degrees of freedom instead of the standard normal distribution.

16. TWO POPULATION SAMPLING DISTRIBUTIONS

When sampling two disjoint populations we are really dealing with two disjoint sample spaces S_1 and S_2, and two random variables X and Y say where X is a random variable on S_1 and Y is a random variable on S_2. The object here is to compare the means. To do this, we form a new sample space $S = S_1 \times S_2$ consisting of all pairs of outcomes (a, b) where a is an outcome in S_1 and b is an outcome is S_2. We then regard both X and Y as being random variables on S, by setting $X(a, b) = X(a)$ and $Y(a, b) = Y(b)$. It is easy to see that now X and Y become independent random variables on the same sample space. As a special case, we could be dealing with a pair of dice, where X is the number up on the first dice and Y is the number up on the second dice, and we would like to know if they are loaded the same way. When we compare by examining $X - Y$ it is clear then we have to be considering pairs of outcomes for this to make sense. Thus to estimate $\mu_X - \mu_Y$, we take a sample of observed values of X of size m and a sample for Y of size n, and estimate $\mu_X - \mu_Y$ with the difference in sample means $\bar{X}_m - \bar{Y}_n$. To know how good this is, then we need to deal with the distribution of $\bar{X}_m - \bar{Y}_n$. We will assume IRS(independent random sampling) here. Then as $Var(\bar{X}_m) = (1/m)\sigma_X^2$ and $Var(\bar{Y}_n) = (1/n)\sigma_Y^2$, and as $\bar{X}_m$ and $\bar{Y}_n$ are independent, we have

$$Var(\bar{X}_m - \bar{Y}_n) = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}, \tag{16.1}$$

and therefore,

$$SD(\bar{X}_m - \bar{Y}_n) = \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}. \tag{16.2}$$

If X and Y are normally distributed then so is $\bar{X}_m - \bar{Y}_n$ and therefore its standardization,

$$Z = \frac{(\bar{X}_m - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}, \tag{16.3}$$

is a standard normal random variable. We therefore can use the normal distribution to judge mean differences from sample data when the standard deviations σ_X and σ_Y are known, as for instance in hypothesis testing or forming confidence intervals. But this is realistic only in certain circumstances, such as when dealing with batches of machine products where the standard deviations are determined by the properties and built in tolerances of the cutting machines whereas the actual values are determined by the adjustments and settings of the cutting machines, which can be set independently. That is, in this case, long experience may lead us to know the standard deviations, or at least we know that changing the settings does not change the standard deviation for comparing outputs of a single cutting machine. In general, we have to estimate standard deviations using sample standard deviations, which as in the case of a single random variable leads to a $t-$distribution for some number of degrees of freedom. We are then dealing with the random variable W given by

$$W = \frac{(\bar{X}_m - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}, \tag{16.4}$$

where S_X^2 and S_Y^2 denote the sample variances for the $X-$sample and the $Y-$sample, respectively. We can then write

$$W = \frac{Z}{\sqrt{\frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)}{\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)}}}. \tag{16.5}$$

Now for dealing with degrees of freedom, let $df_X = m - 1$ and $df_Y = n - 1$ which are the degrees of freedom for the $X-$sample and the $Y-$sample, respectively. We then know that $((df_X)/\sigma_X^2)S_X^2$ has the $\chi^2_{df_X}-$distribution and $((df_Y)/\sigma_Y^2)S_Y^2$ has the $\chi^2_{df_Y}-$distribution. Since these sums of independent squares of standard normals are independent of each other, it follows that their sum is χ^2-distributed with degrees of freedom equal to the sum of the individual degrees of freedom, $df_{total} = (df_X) + (df_Y)$. That is,

$$\frac{(df_X)S_X^2}{\sigma_X^2} + \frac{(df_Y)S_Y^2}{\sigma_Y^2} = \chi^2_{df_{total}}. \tag{16.6}$$

Now the problem of getting a $t-$distribution to fit the distribution of W is in approximating the distribution of the expression inside the radical of [16.5] with a distribution of the form $(\chi_d^2)/d$ for some d.

There are then three cases to deal with. The first case, where the standard deviations are known is of course immediately handled by standardization, [16.3]. The second case is the case where the standard deviations σ_X and σ_Y are unknown but assumed to be equal. The third case is where we have no assumptions on the standard deviations.

For the second case, let us assume that $\sigma_X = \sigma = \sigma_Y$. Then the expression for Z simplifies to

$$Z = \frac{(\bar{X}_m - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{16.7}$$

and obviously, S_X^2 and S_Y^2 are both estimates of σ. Moreover, now [16.6] can be rewritten as

$$\frac{(df_X)S_X^2 + (df_Y)S_Y^2}{df_{total}} = \sigma^2 \frac{\chi^2_{df_{total}}}{df_{total}}, \tag{16.8}$$

and clearly the left hand side of [16.8] should give a more reliable estimate of σ^2 than either of the individual sample variances, since it averages the variances from both samples and weights according to degrees of freedom. We therefore use it to define the pooled variance S_p^2 given as

$$S_p^2 = \frac{(df_X)S_X^2 + (df_Y)S_Y^2}{df_{total}}, \tag{16.9}$$

so now we have by [16.8],

$$S_p^2 = \sigma^2 \frac{\chi^2_{df_{total}}}{df_{total}}. \tag{16.10}$$

Thus, replacing σ^2 in [16.7] by the estimate S_p^2 will give

$$\frac{(\bar{X}_m - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{S_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{Z}{\sqrt{\frac{\chi^2_{df_{total}}}{df_{total}}}} = t_{df_{total}}. \tag{16.11}$$

Thus, when assuming the two population standard deviations are equal, we should pool variances and add degrees of freedom for the appropriate $t-$distribution.

Consider now the third case, where we make no assumption about the standard deviations and must use the sample variances to estimate them. As noted above, from [16.5], in order to get a $t-$ distribution to best approximate the distribution of W, we must get the appropriate number of degrees of freedom, d, so that the expression in the radical in [16.5],

$$\frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)}{\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)}, \tag{16.12}$$

has a distribution which is approximately the distribution of $(\chi^2_d)/d$. Therefore, as we can set $S^2_X = (\sigma^2_X)(\chi^2_{df_X}/df_X)$ and $S^2_Y = (\sigma^2_Y)(\chi^2_{df_Y}/df_Y)$, making these substitutions in [16.12] gives

$$\frac{\left(\frac{S^2_X}{m} + \frac{S^2_Y}{n}\right)}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)} = \frac{\left(\sigma^2_X \frac{\left(\frac{\chi^2_{df_X}}{df_X}\right)}{m} + \sigma^2_Y \frac{\left(\frac{\chi^2_{df_Y}}{df_Y}\right)}{n}\right)}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)} \tag{16.13}$$

Now notice that the expected value of either side of [16.13] is 1 as is the expected value of χ^2_d/d for any number d of degrees of freedom. Hence we must go beyond matching expected values here to get the correct number of degrees of freedom. Thus, we will require that the number of degrees of freedom, d, be chosen so that the variance of the right side of [16.13] matches the variance of χ^2_d/d. Now, if Z is any standard normal random variable, then the distribution of $\overline{(Z^2)}_d$, the sample mean for an IRS of observations of Z^2 of size d is χ^2_d/d, hence we can conclude immediately that

$$Var\left(\frac{\chi^2_d}{d}\right) = \frac{Var(Z^2)}{d}. \tag{16.14}$$

Suppose that we denote $R_d = \chi^2_d/d$ and $Var(Z^2) = c$. Actually, $c = 2$, but this requires calculus and we do not need this fact here. Define a and b by

$$a = \frac{\frac{\sigma^2_X}{m}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} \ \mathit{and}\ b = \frac{\frac{\sigma^2_Y}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}, \tag{16.15}$$

so $a + b = 1$ and we have simply

$$Var(R_d) = \frac{c}{d},\ \mathit{and}\ aR_{dfX} + bR_{df_Y} = \frac{\left(\sigma^2_X \frac{\left(\frac{\chi^2_{df_X}}{df_X}\right)}{m} + \sigma^2_Y \frac{\left(\frac{\chi^2_{df_Y}}{df_Y}\right)}{n}\right)}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)}. \tag{16.16}$$

In terms of these new symbols, we are simply trying to choose d so that

$$Var(R_d) = Var(aR_{dfX} + bR_{df_Y}). \tag{16.17}$$

Here we can take R_{dfX} and R_{df_Y} to be independent of each other, since S^2_X and S^2_Y are independent of each other. Therefore, assuming [16.17] forces

$$\frac{c}{d} = Var(aR_{df_X} + bR_{df_Y}) = a^2 Var(R_{dfX}) + b^2 Var(R_{df_Y}) = a^2\frac{c}{df_X} + b^2\frac{c}{df_Y}, \tag{16.18}$$

so

$$\frac{c}{d} = a^2\frac{c}{df_X} + b^2\frac{c}{df_Y}. \tag{16.19}$$

Since $c > 0$, it cancels out in [16.19], giving finally

$$d = \frac{1}{a^2\frac{1}{df_X} + b^2\frac{1}{df_Y}} = \frac{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2}{\frac{1}{df_X}\left(\frac{\sigma^2_X}{m}\right)^2 + \frac{1}{df_Y}\left(\frac{\sigma^2_Y}{n}\right)^2}. \tag{16.20}$$

as the required number of degrees of freedom. Of course now the obvious problem is that in order to know the degrees of freedom, d, we have to know the population variances. A second problem is that the value of d need not be a whole number. Acceptable practice for getting around these problems is to use the sample variances in place of population variances in the preceeding formula, [16.20] for degrees of freedom and to use interpolation between whole number degrees of freedom when the formula [16.20] does not give a whole number. This means that in statistical practice, it is accepted that W of [16.4] has the $t-$distribution for d degrees of freedom where d is given as

$$d = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{1}{df_X}\left(\frac{S_X^2}{m}\right)^2 + \frac{1}{df_Y}\left(\frac{S_Y^2}{n}\right)^2}. \tag{16.21}$$

17. MORE ABOUT INDEPENDENCE AND UNCORRELATION

Let's recall what we have so far for the meaning of independence for the random variables X and Y. If X and Y are just events themselves, then as the only thing to say about the observation of an event is whether or not it happens, from [13.1] we see that for events independence and uncorrelation are exactly the same thing. In the case of discrete random variables, as already mentioned, we must require that for any real numbers c, d we have the events $(X = c)$ and $(Y = d)$ are themselves uncorrelated. For more general random variables, we have to consider what the possibilities are for what can be said about the outcome of observing a random variable. First of all, if A and B are statements about X, then so are the statements $(A\&B)$, $(A \, or \, B)$, and $(not \, A)$. Algebraically, this means that AB, $A + B - AB$, and $1 - A$ should be considered as statements about X if A and B are. From [5.8] we can see that if we can narrow down the statements to combinations of some basic simple statements where we combine using "and", "or", and "not", then to see that all such statements about X are uncorrelated with all such statements about Y, it will suffice to check it just for combinations using "and" and "not". This is because the union of sets can be written as a disjoint union via

$$A \cup B = (A \setminus B) \cup B = (A \cap (S \setminus B)) \cup B, \tag{17.1}$$

and disjoint union algebraically is just addition of events as random variables. Since the outcome of observing a random variable is always just a number, we just have to consider what can be said about a number. If x is a number, the simplest type of statement about x is to say something like $x \leq 4$ or $x \leq -5$, or more generally, $x \leq c$, where c is any given number. Such inequalities specify sets of numbers on the real line, $\mathbb{R}$. If we consider $\mathbb{R}$ as a sample space, then such inequalities are specifying events in that sample space. Combining these simple inequalities using "and" and "not", we see that for instance

$$(x \leq 7)\&(not(x \leq 3)) = (3 < x \leq 7) \tag{17.2}$$

so intervals of this form are simple events for the sample space $\mathbb{R}$. However, there is no way to say that $x = 4$, for instance, by combining such inequality statements using only "and", "or", and "not". It requires infinitely many such statements all of the form $c < x \leq 4$ to all be simultaneously true, for c any number less than 4. We can reduce this to a sequence of statements by requiring that $4 - (1/n) < x \leq 4$ for all positive integers n. That is to say: $4 - 1 < x \leq 4$ and $4 - (1/2) < x \leq 4$ and $4 - (1/3) < x \leq 4$, and so on. Thus here we have an infinite combination of "and" applications. Alternately, we have to notice that

$3 < x < 4$ is the same as saying that for some real number c satisfying $3 < c < 4$ we have $3 < x \leq c$, which can be reduced to a sequence of statements connected by "or" which can algebraically be expressed as an infinite sum of terms. Since the real line is an infinite sample space, we cannot avoid dealing with infinity, so we assume that these basic statements are events and that any union of a finite or infinite sequence of events is again an event. This means logically that we can also form the combination using "and" or "or" for any finite or infinite sequence of events, so $x = 4$ does indeed specify an event in $\mathbb{R}$. It turns out that as a consequence, some subsets of the real line cannot be considered to be events, but in essence we see that the subsets we easily specify in everyday circumstances are events. Now, the technical definition of a random variable, say X, on the sample space S for a general possibly infinite sample space, merely requires that for any event A on the real line, that is for any event $A \subseteq \mathbb{R}$, it must be true that the statement "X belongs to A" is an event of S. It suffices to check this for the basic interval events in $\mathbb{R}$, that is for X to be a random variable on S it is enough to know that for any real number c we have $(X \leq c)$ is an event of S. Because of these considerations, it is in fact always assumed in general probability theory that the probability of the union of an infinite sequence of disjoint events is equal to the sum of their probabilities. Moreover, when checking independence for random variables X and Y, it is enough to check that the events $(X\ in\ A)$ and $(Y\ in\ B)$ are uncorrelated, for any pair of finite intervals A and B contained in $\mathbb{R}$, and we can restrict consideration to intervals of the form $c < x \leq d$ where c and d are real numbers. This leads back to our earlier definition that for X and Y to be independent we require that

$$Cov((c_1 < X \leq c_2), (d_1 < Y \leq d_2)) = 0, \tag{17.3}$$

for any four real numbers c_1, c_2, d_1, d_2. In case of finite sample spaces, these considerations of infinity can be ignored, and as before, it is then easy to see that for X and Y to be independent it is enough that

$$Cov((X = c), (Y = d)) = 0 \tag{17.4}$$

for any pair of real numbers c and d.

If W is a random variable on the sample space $\mathbb{R}$, then for any random variable X we can form $W(X)$ and get a new random variable. That is, the observed value of $W(X)$ is obtained by getting the observed value of X and then putting this value into W to obtain a new value. For instance, if W is the random variable such that $W(x) = x^3$, then the observed value of $W(X)$ is gotten by raising the observed value of X to the third power. It is important to realize that in general $E(W(X))$ and $W(E(X))$ will be quite different. For instance, if X has nonzero standard deviation, then by [5.7] we have $E(X^2) \neq (E(X))^2$, which is to say that if F is the random variable on $\mathbb{R}$ such that $F(x) = x^2$, then $E(F(X)) \neq F(E(X))$. Notice that if W_1 and W_2 are both random variables on $\mathbb{R}$, then

$$(W_1 + W_2)(X) = W_1(X) + W_2(X) \tag{17.5}$$

$$(W_1 W_2)(X) = W_1(X) W_2(X). \tag{17.6}$$

We can form simple random variables on $\mathbb{R}$ that are useful by rounding off. To be more specific, let us suppose we are working to say n decimal place accuracy. Let L_n be the random variable on $\mathbb{R}$ so that $L_n(x)$ is the result of rounding down to n decimal places, so, for instance, with $n = 5$, then $L_5(3.4268794) = 3.42687$. Let U_n be the corresponding random variable that rounds up to five decimal places, so, for instance, $U_5(3.4268721) = 3.42688$. Let R_n be the ordinary rounding to n decimal place accuracy. Clearly we have $L_n \leq R_n \leq U_n$, and therefore for any random variable X we have $L_n(X) \leq R_n(X) \leq U_n(X)$. In fact, $L_n(X), R_n(X)$, and $U_n(X)$ are exactly the same as in [2.1] and [2.2]. Recall then from [3.12]

$$|E(X) - E(R_n(X))| \leq (10)^{-n}. \tag{17.7}$$

We can apply this to the random variables on $\mathbb{R}$ itself. This means that if W is any random variable on $\mathbb{R}$, then to any required degree of accuracy, we can use sufficiently many decimal place rounding off accuracy so that W and $R_n(W)$ agree to the required level of accuracy. That is, we can replace W by $R_n(W)$ in the calculations. Now $R_n(W)$ is easily seen to be a discrete random variable, consequently, $R_n(W)$ is a linear combination of events in $\mathbb{R}$. This means $R_n(W(X))$ is a linear combination of events which are each simple statements of the form $X\ in\ J$, where J is an interval in $\mathbb{R}$. This means that for X and Y to be independent, by [5.10] and [6.12] it is necessary and sufficient that

$$Cov(F(X), G(Y)) = 0, \tag{17.8}$$

for any discrete random variables F, G on $\mathbb{R}$. Hence by [17.7], we see that X and Y are independent if and only if [17.8] holds for all random variables F, G on $\mathbb{R}$. Notice this is the same as saying $E(F(X)G(Y)) = E(F(X))E(G(Y))$ for any random variables F, G on $\mathbb{R}$. In particular, this means that for any positive real number t, we have

$$E(t^{X+Y}) = E(t^X t^Y) = E(t^X)E(t^Y),$$

whenever X and Y are independent random variables. This equation can be usefully used in calculating distributions. That is, we define the MOMENT GENERATING FUNCTION of X to be

$$\psi_X(t) = E(t^X).$$

Then we have from the previous equations for any independent random variables X, Y that $\psi_{X+Y} = \psi_X \psi_Y$. Notice that as the moment generating function is computed through expectation, the moment generating function for X can only depend on the distribution of X. This means that if X and Y have the same distribution, then $\psi_X = \psi_Y$. Thus, if X is any random variable, and T_n is the total of n independent observations of X, then $\psi_{T_n} = \psi_X^n$. Because if X_k is the $k-$th observation of X, then X_k and X have the same distribution, but $X_1, X_2, X_3, ..., X_n$ are all mutually independent and therefore as $T_n = X_1 + X_2 + X_3 + ... + X_n$, it follows that

$$\psi_{T_n} = \psi_{X_1}\psi_{X_2}\psi_{X_3}...\psi_{X_n} = \psi_X^n.$$

Consider the case of a simple random variable, say $X = 2A + 3B + 5C$, where A, B, C, D are mutually exclusive events whose union is S. Then as $t^0 = 1$, we get

$$\psi_X(t) = t^2 P(A) + t^3 P(B) + t^5 P(C) + P(D).$$

Notice in this case we get a polynomial in t as the moment generating function of X, and that the terms are encoding the distribution by having the coefficient of t^v being $P(X = v)$ for any value v that is possible for X. That is to find the probability of v, we just look for the term $\psi_X(t)$ having exponent v, and then its coefficient is the probability that $X = v$. Thus, if X and Y are independent simple random variables, then we can compute the distribution of $X + Y$ by computing $\psi_{X+Y} = \psi_X \psi_Y$ and then reading off the distribution from the coefficients after the product of the polynomials is computed. A useful special case of these methods is the case of an event. Let $P(A) = p$ and $P(notA) = q$. Then $\psi_A(t) = pt + q$, and T_n is now the random variable which counts the number of occurrences of A in n independent trials, so by the preceding we have $\psi_{T_n} = (pt + q)^n$, giving the usual binomial distribution when the binomial theorem is applied to expand the right side in terms of powers of t and the distribution is decoded. That is, applying[11.14] gives

$$\psi_{T_n} = (pt+q)^n = \sum_{k=0}^{n} C(n,k)(pt)^k q^{n-k} \tag{17.9}$$

$$= \sum_{k=0}^{n} C(n,k)p^k q^{n-k} t^k,$$

and this decodes as

$$P(T_n = k) = C(n,k)p^k q^{n-k}, \tag{17.10}$$

which is the formula for the binomial distribution, [11.15].

18. APPENDIX: THE GEOMETRY OF RANDOM VARIABLES

It is often useful to have a picture in mind when dealing with mathematical problems. In the case of random variables, given the sample space S we have in general an ALGEBRA OF RANDOM VARIABLES ON S by which we mean the set of all random variables defined on S with the structure provided by addition and multiplication of random variables. Moreover, given an expectation model E for random variables on S, we can use it to define what is called an inner product for random variables on S. In more detail, if X and Y are two such random variables, then we define their inner product as $E(XY)$ which is just a number. We often write this inner product as $\langle X|Y\rangle_E = E(XY)$ or simply $\langle X|Y\rangle = E(XY)$, when there is no confusion as to E. We then have in case c is a number or constant

$$\langle cX|Y\rangle = c\langle X|Y\rangle, \tag{18.1}$$

$$\langle X+W|Y\rangle = \langle X|Y\rangle + \langle W|Y\rangle, \tag{18.2}$$

and

$$\langle X|Y\rangle = \langle Y|X\rangle, \tag{18.3}$$

which are sort of like the ordinary associative, distributive, and commutative laws of multiplication. We should also point out that for any random variable X we have

$$\langle X|X\rangle \geq 0 \tag{18.4}$$

which is crucial for defining length. Using the inner product and its positivity property [18.4], we define the length or 2-norm of a random variable as

$$\|X\|_2 = \sqrt{\langle X|X\rangle} = \sqrt{E(X^2)}. \tag{18.5}$$

More generally, for any positive number p we could define the $p-$norm by

$$\|X\|_p = (E(X^p))^{1/p}, \tag{18.6}$$

but there is no inner product for the general $p-$norm for $p \neq 2$. Consequently, we will restrict attention to the case $p = 2$ and drop the subscript from the 2-norm and simply refer to it as the norm or length of a random variable, so $\|X\| = \|X\|_2$ for any random variable X. Now to see there is geometry in this, let us think for the moment of the example of arrows in ordinary three dimensional space. We will consider that the only relevant properties of such arrows are length and direction. That is if the arrow is moved without changing its length or direction, we consider it to be the same arrow we started with. If v denotes an arrow in space, then we let $|v|$ denote its length. Such imaginary arrows are useful for representing all kinds of geometric information. For instance, a movement from one place to another can be represented as the arrow pointing in the direction from the initial point to the final point and whose length specifies the final distance actually moved-that is, it specifies the actual displacement resulting from the move and not all the motion used to accomplish the move. Such displacements are easily seen to have a natural addition known as HEAD TO TAIL ADDITION OF ARROWS. That is, since we can move arrows as long as we do not stretch them or change their direction, to form $v+w$ we move w so that its tail is at the head of v and then the sum is the arrow with tail at the tail of v and head at the head of w. Arrows can be useful in representing forces in physics and velocity and acceleration in motion. For instance when driving a car, your velocity arrow points straight out in the direction of motion and has length equal to the speed registered on the speedometer if the wheels are not slipping on the road surface. The head to tail addition works in these examples as well. We can multiply arrows by ordinary numbers. If v is an arrow and c is a number, then cv is the arrow whose length is $|cv| = |c||v|$ and whose direction is the same as that of v in case c is positive and is the opposite direction of v if c is negative. To form the inner product of the arrows v and w we imagine a number line through the arrow v and we project perpendicularly onto this line from the head and tail of w. We put a scale on the number line so that the tail of v is at 0 and the head of v is at $|v|$. If x denotes the numerical value on this number line of the projection of the tail of w and if y denotes the value on this number line of the projection of the head of w, then we define the inner product $\langle v|w\rangle = |v|(y - x)$. If w is the head to tail sum of to arrows, then the projection of the triangle picture of the three arrows in the head to tail sum leads to the distributive law for this inner product, and reversing the roles of v and w in this process can be seen to give the same result by similar triangles. We therefore have the same rules for arrows as for random variables, and in fact the geometry is the same. To see the geometric meaning of the inner product, notice that if v and w are perpendicular, then $\langle v|w\rangle = |v|(y-x) = 0$, because both the head and tail of w project to the same point on the number line through v so that $x = y$ and $y-x = 0$. Thus the inner product is measuring something that takes into account the angle between the arrows. Since the rules for inner product are the same for random variables and arrows, anything we deduce from these basic rules for random variables will be true for the arrows, and vis-versa. For instance, from the basic rules, [18.2] and [18.3], we see that

$$\|X + Y\|^2 = \|X\|^2 + \|Y\|^2 + 2\langle X|Y\rangle. \tag{18.7}$$

Now, we can notice that [5.6] can be rewritten as

$$\langle X|Y\rangle = Cov(X, Y) + \mu_X\mu_Y \tag{18.8}$$

and in particular [5.7] can be rewritten as

$$\|X\|^2 = Var(X) + \mu_X^2. \tag{18.9}$$

In fact, because $(\mu_X + \mu_Y)^2 = \mu_X^2 + \mu_Y^2 + 2\mu_X\mu_Y$, we see that by [18.8] and [18.9] it is the case that [18.7] and [5.12] are actually equivalent equations. Better still, by [5.10] and [5.9] we see that $Cov(X, Y)$ also satisfies the basic rules of inner product given by [18.1], [18.2], and [18.3], and so we could say that [5.12] is actually a special case of [18.7]. Alternately, we can notice that by definition, [5.3], if X and Y are random variables, then with $D_X = X - \mu_X$ and $D_Y = y = \mu_Y$ we have

$$Cov(X, Y) = E(D_XD_Y) = \langle D_X|D_Y\rangle,$$

and by [5.4],

$$Var(X) = E(D_X^2) = \|D_X\|^2,$$

so that [5.12] results from [18.7] and the fact that $D_{X+Y} = D_X + D_Y$ as pointed out in [5.1]. Thinking in terms of the arrow picture, if $\langle X|Y\rangle = 0$, then somehow X and Y should be thought of as perpendicular. In fact, if we think of the addition of random variables as being geometrically a triangle as in head to tail addition of arrows, then for the case where $\langle X|Y\rangle = 0$ we should think of X and Y as perpendicular, so the head to tail addition picture becomes a picture of a right triangle and in this case [18.7] simplifies to $\|X+Y\|^2 = \|X\|^2 + \|Y\|^2$ which amounts to the Pythagorean Theorem for random variables. Somehow then, $\langle X|Y\rangle$ must be related to an angle between the random variables X and Y.

Suppose X and Y are any random variables which we fix for the following argument. Then for any number t we can form the random variable $X+tY$. By [18.4], we have $\|X+tY\|^2 \geq 0$ for any value of t and so by [18.7] we find that for every value of t we have

$$\|X + tY\|^2 = \|X\|^2 + t^2\|Y\|^2 + 2t\langle X|Y\rangle \geq 0. \tag{18.10}$$

If $\langle X|Y\rangle \neq 0$ and $\|Y\| = 0$, then by choosing t to be very large in absolute value and with opposite sign to $\langle X|Y\rangle$, we could contradict [18.10] which is not possible. That is to say, we must conclude from this that $\langle X|Y\rangle = 0$ if $\|Y\| = 0$. But, this conclusion must apply equally well to X meaning that $\langle X|Y\rangle = 0$ if either $\|X\|$ or $\|Y\|$ is 0. In particular, if either σ_X or σ_Y is zero, then $Cov(X, Y) = 0$, which we see on replacing X and Y by D_X and D_Y, respectively. Now, suppose then that Y is not 0, and as [18.10] holds for all values of t, let us set

$$t = \frac{-\langle X|Y\rangle}{\|Y\|^2}. \tag{18.11}$$

This results in the inequality

$$\|X\|^2 + \frac{\langle X|Y\rangle^2}{\|Y\|^2} - 2\frac{\langle X|Y\rangle^2}{\|Y\|^2} \geq 0. \tag{18.12}$$

Multiplying throughout by $\|Y\|^2$ and then cancelling this reduces to the general inequality

$$\|X\|^2\|Y\|^2 - \langle X|Y\rangle^2 \geq 0, \tag{18.13}$$

which is the same as

$$|\langle X|Y\rangle| \leq \|X\|\|Y\|, \tag{18.14}$$

an inequality known as the CAUCHY-SCHWARZ INEQUALITY. In view of our remarks about the case where either $\|X\|$ or $\|Y\|$ is zero, we see that this inequality [18.14] holds with no restriction-that is to say it holds for any pair of random variables. It gives [6.12] on replacing X and Y by their deviations from their means, D_X and D_Y, respectively. It is the Cauchy-Schwarz inequality that allows us to get directly at the angle between two random variables. We define the angle θ between X and Y to be the angle between 0 and π radians with the property that $\cos\theta = \langle X|Y\rangle/\|X\|\|Y\|$. Because $\cos(\pi/2) = 0$, it then follows that when the inner product is zero the angle between the two random variables is a right angle as before. Moreover, in the head to tail addition rule we see that for arrows v and w with their tails together we have $|v - w|$ is the distance between their tips. If we think of the random variables as points in space by placing all their tails at the 0 random variable, then we are thinking of their tips as being the points in space and $\|X - Y\|$ becomes the distance from X to Y. For instance, if c is any constant and X any random variable, then

$\|X-c\|^2 = Var(X-c) + (E(X-c))^2 = Var(X) + (\mu_X - c)^2$. Now from this we see that the smallest the distance from X to any constant c is obtained by choosing that constant to be $c = \mu_X$. That is, for each random variable, the constant closest to it is its mean. Notice also that for any constant c, we have $\langle c|D_X\rangle = cE(D_X) = 0$, so any deviation random variable is perpendicular to any constant. Since $D_X = X - \mu_X$, we see that $X = D_X + \mu_X$ and thinking in terms of head to tail addition, the triangle is a right triangle. Now, the constants form a line through the constant 1, since every constant is just a multiple of 1, (remember multiplying an arrow by a number just stretches it) so this shows that all the deviation variables are in a subspace perpendicular to the line of constants. In this picture then, the deviation D_X is just the perpendicular projection of X onto this subspace. We see then that the correlation coefficient ρ between X and Y is just the cosine of the angle between their deviations, D_X and D_Y. When we look for the optimal regression line $W = mX + c$ for guessing Y when the value of X is given to us, we are minimizing the mean squared residual $E(R^2)$, where $R = Y - W$ with optimal choice of m and c. In our picture then we are really just trying to find m and c so as to make $W = mX + c$ come as close as possible to Y in the space of random variables. Since $D_W = mD_X$, when the choice is optimal, then D_Y will be closest to D_W and $E(Y) = E(W)$.

References

[1] Dupré, M. J., Unknowns and Guessing Expectation and Probability, on my website.

[2] Dupré, M. J. and Tipler, F. T., The Cox theorem, unknowns and plausible value, LANL arXiv, 2006.

[3] E. T. Jaynes, *Probability Theory-The Logic of Science*, Cambridge University Press, 1999.

Department of Mathematics, Tulane University, New Orleans, LA 70118 USA
E-mail address, M. J. Dupré: mdupre@tulane.edu