MATH-111 (DUPRÉ) SPRING 2010 LECTURES

1. **LECTURE** MONDAY 11 JANUARY 2010

We discussed the syllabus and course policy as well as the location of my office and my office hours. I will generally be in my office Monday, Wednesday, and Friday from 9 to 10 AM and from 1 to 2 PM. My office is Gibson Hall Room 309A. I will often be in my office at other times, so if you need to see me come up to the office and if I am not there, check the fourth floor. The syllabus and work schedule are on my Tulane website and you will receive a link in an email from blackboard.

We briefly discussed the idea that guessing must have certain logical constraints. We discussed the idea that when we guess we must base our guess on our knowledge of some factual information which we assume to be true, so our guess depends ont this information. But, given our information for instance if we guess one unknown quantity to be 10 and another unknown quantity to be 20, then logical consistency would dictate that we should guess their sum to be 30. In mathematics, the technical terminology for guessing is **EXPECTED VALUE**.

2. **LECTURE** WEDNESDAY 13 JANUARY 2010

We discussed the logic of guessing in some detail. We use capital letters near the end of the alphabet to denote **UNKNOWNS** which are to be thought of as numerical quantities whose values are certain but are to some extent unknown to us, whereas we use lower case letters for numbers that are definitely known to us. For instance we discussed the example of the outside air temperature in degrees Fahrenheit and its relation to the temperature in degrees Celsius. If the temperature in degrees Celsius is denoted by X and the temperature in degrees Fahrenheit is denoted by Y, then we know the conversion formula $Y = (9/5)X + 32$ so even if we do not know the exact temperature, if we must guess values for X and for Y, then our guesses are constrained by the conversion formula. If we guess the outside temperature in degrees Celsius is 10, then we are guessing the value of X is 10. Then our guess for Y must be $(9/5)10+32 = 50$. The formula which relates Y to X here has the general form $Y = aX + b$ where a and b are definite numbers. Since in general any equation of this form can be viewed as simply specifying a way to change units, this means that if we guess X has value 10, then we should guess that $aX + b$ has value $10a + b$ no matter how a and b are chosen. Remember, choices of units are in fact somewhat arbitrary.

In general, we deal with unknowns and statements of factual nature. Thus we will not use subjective statements of opinion. The statements we deal with are restricted to be statements of fact which are either true or false, even though in certain situations we may not know whether a given statement is true or false. We will use capital letters near the beginning of the alphabet for statements. Thus, if A is a statement, then A is either true or false.

In dealing with guessing values for unknowns, it is useful to have a notation which contains the symbols indicating what we are guessing and the factual statements we are using to arrive at our guess. We use

$$E(X|K)$$

to denote the guess for the unknown X when K is the statement of the factual information we are using to arrive at our guess. In words, $E(X|K)$ is the *Expected Value of X given K*.

We noted that because units can always be changed in many arbitrary ways, it should always be the case that for any specified numbers a and b we should have

$$E(aX + b|K) = aE(X|K) + b.$$

This in some sense merely expresses the fact that numerical quantities have an intrinsic value that goes beyond the choice of units, even though numerical specifications require units to have meaning. There is something about knowing the outside temperature that is beyond the choice of units used to specify it.

If we take the case where we chose $a = 0$, then the preceding equation says for any definite number b it must be the case that

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

More generally, if

K implies that the value of X is b,

then,

$E(X|K) = b$.

This is just the most basic expression of logical consistency for guessing-whenever your information tells you the value of the unknown you are trying to guess, then you do not really have to guess.

More generally, if we have any two unknowns X and Y then we can form their **Sum** $X + Y$ and their **Product** XY which are new unknowns. For instance, if we have a basket of oranges and a basket of apples and if X is the number or oranges in the basket of oranges and Y is the number of apples in the basket of apples, then $X + Y$ makes perfectly good sense. Likewise, if X is the number of apples in the basket of apples and Y is the outside temperature in degrees Celsius, then $X + Y$ is a perfectly specified unknown. If, based on information K, I guess the value 12 for X and the value 10 for Y, then to be logically consistent, based on K, I should guess 22 for the value of $X + Y$. Using our notation, this leads to the general rule

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

called the **Additive Property of Expectation**, and which is simply expressing the logical consistency of addition with guessing.

We can also notice that as far as logic is concerned, if our information K tells us $X \leq Y$, even without knowing either, as far as guessing is concerned, to be logically consistent with our information we must choose a guess for X that is no more than our guess for Y. That is it must generally be true that

if K implies $X \leq Y$,

then,

$E(X|K) \leq E(Y|K)$.

This is known as the **Order Property** of Expectation or the **Positivity Property** of Expectation.

We need to keep in mind that our guessing procedure is to be though of as mechanized so as to be performed by a robot. No subjective information is allowed, only factual statements whose truth value we may not know.

We noted that if we have any statement A, then we can form an unknown called the **Indicator** of A, denoted I_A, whose value is either zero or one according to whether A is false or true. Thus if we think A is true, then we would guess the unknown I_A has value one, whereas if we think A is false, then we should guess the unknown I_A has value zero. Thus the statement "$I_A = 1$" is logically equivalent to the statement "A *is true*", they both have the same meaning.

We can notice that no matter what,

$$0 \leq I_A \leq 1,$$

and therefore

$$0 = E(0|K) \leq E(I_A|K) \leq E(1|K) = 1,$$

so always

$$0 \leq E(I_A|K) \leq 1,$$

no matter what. We must keep in mind that our general rules so far do not tell us to guess either a zero or a one for the value of the indicator. More generally if X is an unknown whose possible values are the whole numbers one through six, then the only thing we can say for sure is that

$$1 \leq E(X|K) \leq 6,$$

but logical consistency may not allow us to guess one of the whole numbers. So far, we do not know how to guess here, but we will soon. For the example here, think of a dice in a box which we cannot see and X is the number on the top face. Our rules so far tell us our guess should be a real number between one and six, but our rules do not tell us that our guess should be a whole number.

We noted that we can use logic to combine statements and algebra to combine indicators, and this gives a useful way to turn logic into algebra. For instance, it is easy to see that if A and B are any statements, then since $A\&B$ is the statement that both A and B are true, it must be the case that

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

that is to say, the indicator of $A\&B$ is simply the product of the indicator of A with the indicator of B. We can also easily see that the negation of A, denoted *not* A, has indicator

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

Finally, we noted that the statement A *or* B is the statement that at least one of the statements A and B is true, possibly both. The only way that A *or* B can be false is for both A to be false and B to be false. Since combining statements with "&" involved multiplication of the indicators, we might try addition for combining statements with "*or*". Unfortunately, if A and B are both true, then $I_A + I_B$ has the value two which is not allowed for an indicator. But this is easily fixed by subtracting one exactly when (and only when) both are true, that is, just subtract the indicator of $A\&B$. Thus,

$$I_{(A\ or\ B)} = I_A + I_B - I_{A\&B} = I_A + I_B - I_A I_B.$$

We will use these facts next time to see how our rules for guessing force us to guess certain values for indicators in many situations. In other words, in many situations we find that there is a unique solution to the problem of guessing-we have no "leeway".

3. **LECTURE** FRIDAY 15 JANUARY 2010

Today we began by reviewing the basic rules for guessing dictated by logic, that is to say the rules for expectation. We generally use capital letters near the end of the alphabet for unknowns and capital letters near the beginning of the alphabet for statements. Lower case letters are symbols for numbers which are known to us. Thus, if X is any unknown, then $E(X|K)$ is the expected value of X given K is the statement of the information used to evaluate how we should guess the value of X. It is useful to also have another symbol for $E(X|K)$ and that is the Greek letter μ which you will see in your text book. Thus we can write μ in place of $E(X|K)$ when X and K are understood. Sometimes we tag tag the μ with a subscript to indicate information if necessary. Thus we can write

$$\mu = \mu_X = \mu_{(X|K)} = E(X|K).$$

So, in brief, our basic rules are now

$E(aX + b|K) = aE(X|K) + b$,

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

if K implies that $X = c$, then $E(X|K) = c$,

if K implies that $X \leq Y$, then $E(X|K) \leq E(Y|K)$.

We also created an unknown for each statement called its indicator. For any statement A, we denote its indicator by I_A, and remember it is simply the unknown whose value is one if A is true and zero otherwise. Thus knowing the value of I_A is exactly the same as knowing whether or not A is actually true.

We observed that for the indicator of any statement A, we know $0 \leq I_A \leq 1$, so

$0 \leq E(I_A|K) \leq 1$, for any statement A.

Moreover, for any statements A and B, it is easy to check that

$$I_{(A\&B)} = I_A I_B,$$

$$I_{not\ A} = 1 - I_A,$$

and

$$I_{(A\ or\ B)} = I_A + I_B - I_{(A\&B)}.$$

We ended last time thinking about how a robot programmed to follow our rules would evaluate $E(I_A|K)$ in a simple example: the dice in the box. The statement K says that no value for the number up on the dice should be given any preference. We considered the statement A which says "the number up is even". The robot knows that he must guess a number for I_A that is between zero and one from our rules above, but because of K he cannot make any preference of A over $B = not\ A$. Thus he must accept that $E(I_A|K) + E(I_B|K)$. On the other hand, by our addition rule, since we have $I_A + I_B = 1$, it follows that the robot's guesses must satisfy

$$E(I_A|K) + E(I_B|K) = 1.$$

If we set $E(I_A|K) = \mu = E(I_B|K)$, which we can do as the robot must guess equal values for these two indicators in order to follow the dictates of K and our rules, then this tells us

$$\mu + \mu = 1,$$

or

$$2\mu = 1,$$

giving

$$\mu = \frac{1}{2},$$

which is to say finally,

$$E(I_A|K) = \frac{1}{2} = E(I_B|K).$$

If A is the statement that the number up is 1 or 2, if B is the statement that the number up is 3 or 4, if C is the statement that the number up is 5 or 6, then we know one and only one of these statements is true, so

$$I_A + I_B + I_C = 1,$$

and again, the robot cannot make any preference over any one of these statements so must guess the same value for all three, but the sum of the guesses must equal 1, so he must guess each to be one third. That is

$$E(I_A|K) = E(I_B|K) = E(I_C|K) = \frac{1}{3}.$$

On the other hand, suppose that for real number r, we let (r) denote the statement "the number up on the dice is the number r," so exactly one of the statements (1),(2),(3),(4),(5),(6) is certainly true for the dice in the box. Let N be the statement that the number up on the box is even. Then making the robot use the statement $N\&K$, we see that

$K\&N$ implies that $I_{(2)} + I_{(4)} + I_{(6)} = 1$,

but none of these three statements can be preferred, so he must guess the same number for all three again, and by our rules all the guesses must add up to one, so again, the only possibility is to guess each to be one third, which is now

$$E(I_{(2)}|N\&K) = E(I_{(4)}|N\&K) = E(I_{(4)}|N\&K) = \frac{1}{3}.$$

Notice that these calculations are giving exactly what we would calculate if we were calculating probabilities. Thus, if you have no idea what number on the dice is up, then there is a fifty percent chance it is even, whereas if you know that the number up is even but have no idea which even number it is, then each of the numbers 2,4,6 has a one third chance of being the number up. For the number up itself, we notice that using K, the robot cannot prefer any of the statements (1),(2),(3),(4),(5),(6), so he must guess the same number for all their indicators, but again, exactly one of these six statements is true and all the others are false,

$$I_{(1)} + I_{(2)} + I_{(3)} + I_{(4)} + I_{(5)} + I_{(6)} = 1,$$

and this means

$$E(I_{(1)}|K) = E(I_{(2)}|K) = E(I_{(3)}|K) = E(I_{(4)}|K) = E(I_{(5)}|K) = E(I_{(6)}|K) = \frac{1}{6}.$$

Again, this is obviously the probability of each statement. We are seeing that our robot is actually using probabilities for guesses for values of indicators. With this as motivation, we will make the general definition of **PROBABILITY**. For any statements A and B whatsoever, we **DEFINE** the probability of A given B, denoted $P(A|B)$ by the equation

$$P(A|B) = E(I_A|B).$$

From the rules for expectation, and our algebraic rules for indicators, we immediately obtain basic rules of probability:

$$0 \leq P(A|B) \leq 1,$$

$$P(A \text{ or } B|K) = P(A|K) + P(B|K) - P(A\&B|K).$$

We also see right away, that any time we have n statements one of which must be true and all the others false, if K tells us none of the statements are preferred, then each must have probability $1/n$. Thus, our rules are giving us the probabilities in many situations. What about the unknown X itself? To see how to guess it involves a new rule. First, we take the equation

$$1 = I_{(1)} + I_{(2)} + I_{(3)} + I_{(4)} + I_{(5)} + I_{(6)}$$

and multiply both sides by X getting

$$X = XI_{(1)} + XI_{(2)} + XI_{(3)} + XI_{(4)} + XI_{(5)} + XI_{(6)},$$

which now by our rules give us

$$E(X|K) = E(XI_{(1)}|K) + E(XI_{(2)}|K + E(XI_{(3)}|K) + E(XI_{(4)}|K) + E(XI_{(5)}|K) + E(XI_{(6)}|K),$$

Notice we must calculate a sum of terms and each has the form

$$E(XI_N|K)$$

where N is a statement. Notice that we have no general rule for guess the product of two unknowns in terms of the guesses for the factors, so let us look back at our examples. Take the case where $X = I_{(2)}$ and N is the statement that the number up is even. Notice that $(2)\&N = (2)$, as saying the number up is two and its even is redundant, its just the same as saying the number up is two. Thus

$$E(I_{(2)}I_N|K) = P((2)\&N|K) = P((2)|K) = \frac{1}{6},$$

whereas earlier we calculated

$$P((2)|N\&K) = E(I_{(2)}|N\&K) = \frac{1}{3}$$

and

$$P(N|K) = \frac{1}{2}.$$

Notice that the product of the last two numbers gives the $1/6$ we need. This would lead us to suspect the general **MULTIPLICATION RULE**:

$$E(XI_N|K) = E(X|N\&K)P(N|K).$$

In fact this is the rule we will use to calculate $E(X|K)$ for the dice. Notice that for any number r we have

$$E(X|(r)\&K) = r,$$

since $(r)\&K$ implies the number up is r. For each whole number one through six, the probability is exactly one sixth, so our previous equation expressing $E(X|K)$ as a sum of expected values of products of X with indicators when combined with the multiplication rule gives

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

Notice that the guess dictated by these rules is **not a possible value**. This is a feature of guessing which may at first be counter intuitive, but we must accept it as it is a consequence of the rules. One way to think of this at first is to think that the robot is trying to guess a number which is simultaneously as close as possible to all the possible values.

4. **LECTURE** MONDAY 18 JANUARY 2010

NO CLASS FOR HOLIDAY (MARTIN LUTHER KING JR).

5. **LECTURE** WEDNESDAY 20 JANUARY 2010

Today we discussed the **MULTIPLICATION RULE:**

$E(XI_N|K) = E(X|N\&K)P(N|K)$, for any unknown X and any statements N and K.

We began by noting that most of our rules follow from one single rule and the mere assumption that some form of rule exists. The single most fundamental rule of guessing is the **RETRACTION RULE:**

if K implies that $X = c$, then $E(X|K) = c$.

For instance, if your information tells you that the value of X is 7 then your guess based on that information must be 7. Remember, lower case letters stand for definite numbers-they are unknowns whose description actually tells their value, so for them there is really no guessing. In fact, the retraction rule in particular tells us that if c is any number, then

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

This is because K certainly implies that $c = c$ because $c = c$ is a true statement and therefore by the retraction rule, $E(c|K) = c$. As a point of logic, remember that the only way the implication "P implies Q" can be false is for P to be true and Q to be false. Thus any time Q is true then "P implies Q" is true". Since $c = c$ is always true, this means that for any statement K it is true that "K implies $c = c$" is a true statement and therefore by the retraction rule we have $E(c|K) = c$, always, no matter what K is and no matter what c is.

Now, the retraction rule puts an enormous constraint on what rules can be. For instance, if we consider the rule for changing units or the **GENERAL RESCALE RULE:**

$E(aX + b|K) = aE(X|K) + b$, for any numbers a, b and any unknown X,

we can see that the rule tells us that as soon as the numerical value of $E(X|K)$ has been worked out, to calculate $E(aX+b)$ we need only add to b the number a multiplied by the number already found for $E(X|K)$. We do not need to examine the unknown $aX + b$ as an unknown itself in order to find what we guess for its value, if we have already guessed a value for X. In particular, we see that if X and Y are any two unknowns such that $E(X|K) = E(Y|K)$, then $E(aX + b|K) = E(aY + b|K)$, since the resulting guess only depends on the numerical value of $E(X|K)$ and this is the same as $E(Y|K)$. But if we put $E(X|K) = c$, where c is the definite number we have decided to guess for the value of X on the basis of assuming K, then in particular by the retraction rule,

$$E(X|K) = c = E(c|K)$$

and therefore

$$E(aX + b|K) = E(ac + b|K) = ac + b = aE(X|K) + b,$$

so the general rescaling rule follows from the retraction rule and the assumption that some form of rescaling rule exists.

As another example, consider the **ADDITION RULE:**

$E(X + Y|K) = E(X|K) + E(Y|K)$, for any unknowns X and Y.

In particular, this rule implies that merely knowing the numerical values you guess for X and for Y is enough to somehow determine what must be guessed for $X + Y$. Put another way, this says that if U and V is any other pair of unknowns and if $E(X|K) = E(U|K)$ and if $E(Y|K) = E(V|K)$, then

$$E(X+Y|K) = E(U+V|K),$$

which is to say that our guess for the value of $X+Y$ has to be the same as our guess for $U+V$. Again, if we write $E(X|K) = a$ and $E(Y|K) = b$, then by the retraction rule, we have

$$E(X|K) = a = E(a|K)$$

and

$$E(Y|K) = b = E(b|K),$$

and therefore by our previous observation it must be the case that we guess the same thing for $X+Y$ as we guess for $a+b$, that is to say

$$E(X+Y|K) = E(a+b|K) = a+b = E(X|K) + E(Y|K),$$

and this is the addition rule. That is the addition rule follows from the retraction rule and the mere assumption that our guess for $X+Y$ can always be determined somehow using only the pair of numerical values from our guesses for X and for Y. The assumption of the existence of some form of rule gives the exact rule using the retraction rule.

Now, notice that we do NOT have a rule for guessing products of unknowns in general. If there were such a rule for finding how to guess XY as soon as you decided what to guess for X and for Y, then the only possible rule would be to multiply your guesses together, by the same reasoning as we used for the addition rule. The argument would go like this. Assuming that somehow the two numbers resulting from guessing X and Y determine what should be guessed for XY would mean that if $E(X|K) = E(U|K)$ and $E(Y|K) = E(V|K)$, then always $E(XY|K) = E(UV|K)$. But putting $E(X|K) = a$ and $E(Y|K) = b$ we have here with a playing the role of U and b playing the role of V, that we could conclude

$$E(XY|K) = E(ab|K) = ab = E(X|K)E(Y|K),$$

and we arrive at the conclusion that we have to simply multiply our guesses together to arrive at the guess for XY. We will see that as a result of the retraction rule and the addition rule that in fact this multiplication WOULD GIVE THE WRONG ANSWER. In particular, this means the underlying assumption is wrong. To determine the guess for XY must and will turn out to involve more than simply knowing what you guessed for X and what you guessed for Y.

Returning now to the **MULTIPLICATION RULE** that we will prove using the retraction rule is that

$$E(XI_N|K) = E(X|N\&K)E(I_N|K) = E(X|N\&K)P(N|K).$$

Notice that it is a restricted form of product whose guess we can determine, we can determine XY is one of the factors is an indicator. Remember for Y to be an indicator is the same as having $Y^2 = Y$, since the only numbers which equal their own square are zero and one. Also notice that to determine the guess for XY when $Y = I_N$ is an indicator we need to know $E(X|N\&K)$, the guess for X assuming that N is true, and also we need $E(I_N|K)$ which by definition is $P(N|K)$. Thus, if $Y = I_N$ is an indicator, then N is logically equivalent to the statement $Y = 1$, so we can also say that the multiplication rule says that

$E(XY|K) = E(X|(Y=1)\&K)E(Y|K)$, for any unknowns X and Y, provided that $Y^2 = Y$.

To prove the multiplication rule, we first observe that we must assume that some form of rule exists. That is we assume that given merely the numerical value we guess for X assuming N true, that is the number $E(X|N\&K)$, and the number $E(I_N|K)$, we can determine $E(XI_N|K)$. This means that we assume that if Y is any other unknowns for which we happen to make the same guess as for X under the assumption that N is true, that is if

$$E(X|N\&K) = E(Y|N\&K),$$

then we must arrive at the same value for our guesses for the unknowns XI_N and YI_N, which is to say we must assume here it is valid to conclude that

$$E(XI_N|K) = E(YI_N|K).$$

But now, we can work out the multiplication rule using the retraction property. Suppose that $E(X|N\&K) = b$. By the retraction rule, $E(X|N\&K) = b = E(b|N\&K)$, so letting b play the role of Y, we conclude that

$$E(XI_N|K) = E(bI_N|K).$$

But now by the general rescaling rule we know

$$E(bI_N|K) = bE(I_N|K) = bP(N|K) = E(X|N\&K)P(N|K),$$

and combining this equation with the previous equation gives

$$E(XI_N|K) = E(X|N\&K)P(N|K),$$

which is the multiplication rule.

As a consequence of the multiplication rule we can prove the **SAVAGE SURE THING PRINCIPLE:**

(SSTP) for any statement M, if $E(X|M\&K) = c = E(X|(not\ M)\&K)$, then $E(X|K) = c$.

That is if M is some statement and if we find we should guess X has value c assuming M is true, and if we find we should guess X has the value c assuming M is not true, then we should guess X has value c whether or not M is true. To see this, notice that if we put $N = not\ M$, then $I_M + I_N = 1$, so as

$$X = X(I_M + I_N) = XI_M + XI_N,$$

we have

$$1 = E(I_M + I_N|K) = E(I_M|K) + E(I_N|K) = P(M|K) + P(N|K),$$

and therefore by the multiplication rule,

$$E(X|K) = E(XI_M|K) + E(XI_N|K) = E(X|M\&K)P(M|K) + E(X|N\&K)P(N|K)$$
$$= cP(M|K) + cP(N|K) = c[P(M|K) + P(N|K)] = c1 = c.$$

On the other hand, if we do not assume the multiplication rule, but instead assume the Savage sure thing principle, then for any two unknowns X and Y, if it is the case that $E(X|N\&K) = E(Y|N\&K)$, then this is the same as saying

$$E(XI_N|N\&K) = EYI_N|N\&K)$$

which is in turn the same as saying

$$E([X - Y]I_N|N\&K) = 0.$$

Then we note that certainly, since $I_N = 0$ if N is false, that is to say

$$E([X - Y]I_N|N\&K) = E([X - Y]0|(not\ N)\&K) = E(0|(not\ N), \&K) = 0,$$

so our guess for $[X - Y]I_N$ is the same, namely zero whether or not N is assumed true or false, so by the Savage sure thing principle, we must have

$$E([X - Y]I_N|K) = 0,$$

and this means we have shown that

$$E(XI_N|K) = E(YI_N|K).$$

That is we have shown on the basis of the Savage Sure Thing Principle, that if $E(X|N\&K) = E(Y|N\&K)$, then $E(XI_N|K) = E(YI_N|K)$, which means some form of rule must hold for

determining $E(XI_N|K)$ from $E(X|N\&K)$ if the Savage sure thing principle holds. But our arguments show that if such a rule holds in some form, then it must be the multiplication rule

$$E(XI_N|K) = E(X|N\&K) \cdot E(I_N|K).$$

We finished the class by noting that if A, B, C are any statements of which exactly one is true and the others are false, then

$$1 = I_A + I_B + I_C,$$

so

$$X = XI_A + XI_B + XI_C,$$

and therefore

$$\begin{aligned} E(X|K) &= E(XI_A|K) + E(XI_B|K) + E(XI_C) \\ &= E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K). \end{aligned}$$

The end result is that in this situation

$$E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K).$$

We finished by using the calculator's statistical computation feature to calculate the preceding formula in an example. In applications, instead of just three statements there may be hundreds, so calculators are necessary.

6. **LECTURE** FRIDAY 22 JANUARY 2010

We have been writing $E(X|K)$ for the expected value or optimal guess for X given we assume K is true, and when new information is assumed, say statement N, then we have $E(X|N\&K)$ for the expected value of X given we assume both N and K are true. We usually have K as the statement of our background information which in any problem stays the same throughout, so we will simplify our notation by writing

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

and

$$P(A|N\&K) = P(A|N).$$

Thus, we do not write down the K explicitly when it is reasonably understood, but rather only write in the given information if something new is given. Thus, for the general multiplication rule, we have

$$E(XI_N) = E(X|N)P(N),$$

instead of the more lengthy

$$E(XI_N|K) = E(X|N\&K)P(N|K).$$

They both say the same thing provided that we understand K is given throughout.

In the last lecture, using the multiplication rule, we worked out the formula for expectation that results whenever we have several statements of which exactly one is true. For instance if A, B, C are statements for which we know exactly one is true and all the others are false, then we have

$$SURE = A \text{ or } B \text{ or } C$$

which is the same as

$$1 = I_A + I_B + I_C,$$

and this means in particular that

$$1 = P(A) + P(B) + P(C),$$

and for any unknown X,

$$X = XI_A + XI_B + XI_C.$$

Applying the addition rule for expectation gives

$$E(X) = E(XI_A) + E(XI_B) + E(XI_C),$$

and applying the multiplication rule to each term then gives

$$E(X) = E(X|A)P(A) + E(X|B)P(B) + E(X|C)P(C).$$

There is nothing special here about the fact that there are three statements of which exactly one is true, and in some applications there could be thousands. To make this useful requires that we can actually work out the conditional expected values $E(X|A),\ E(X|B),\ E(X|C)$, and as well find the probabilities $P(A),\ P(B),\ P(C)$. Obviously this could be laborious, but for the expected values, often we choose the statements so that in each case, the value of X is completely determined. Thus for instance, suppose that we have five statements of which exactly one is true, denoted A, B, C, D, F. Suppose that if A is true, then we know X definitely has the value v_A, if B is true, then we know X definitely has the value v_B, if C is true, then X definitely has the value v_C, if D is true, then X definitely has the value v_D, and if F is true, then X definitely has the value v_F. In this case, we know

$$E(X|A) = v_A,$$
$$E(X|B) = v_B,$$
$$E(X|C) = v_C,$$
$$E(X|D) = v_D,$$
$$E(X|F) = v_F.$$

Then our formula says

$$E(X) = v_A P(A) + v_B P(B) + v_C P(C) + v_D P(D) + v_F P(F).$$

Thus, knowing the values v_A, v_B, v_C, v_D, v_F completely reduces the problem of computing $E(X)$ to the problem of calculating the probabilities

$$P(A),$$
$$P(B),$$
$$P(C),$$
$$P(D),$$
$$P(F).$$

In short, the addition rule and the multiplication rule together reduce the problem of calculating expectation to the problem of calculating probability.

As an example, suppose we have the table below giving the information concerning the values of two unknowns X and Y in case of of the events is true and as well the probabilities for the events.

TABLE	**PROB**	**Val(X)**	**Val(Y)**	**Val(XY)**
A	.15	2	3	6
B	.2	−3	4	−12
C	.3	5	4	20
D	.2	5	2	10
F	.15	2	3	6

Thus, the table gives the event names in the first column, in the next column is the column of probabilities, so $P(A) = .15$ and $P(C) = .3$, for instance. The next column gives the values for X in case of each event in the list, and so on. Thus if B is true, then $X = -3$ and $Y = 4$ which means that $XY = -12$ if B is true. Once the information is tabulated like this it is a routine matter to put the columns into lists in the calculator. If we enter the probabilities in L_a, the values of X in L_b, the values of Y in L_c, in the TI calculator, then we can put the values of XY in L_d almost instantly by using the store command $L_b L_c \to L_d$. The calculator does all the multiplications for us all at once. Thus, in the table we do not need the values of XY, they are merely there so you can easily see them. Also, to calculate the expected value of X now just go to the STAT CALC menu and select the "1-Var Stats" and when it comes up on the screen, type after it L_b, L_a since the probability list must always go last. To calculate the expected values of X, Y and XY all at once, select the "2-Var Stats" and when it comes up on the screen, after it type L_b, L_c, L_a, and hit the ENTER button. Notice the probability list again goes last, the list of X values goes first and the list of Y values goes second. The readout first gives the expected value of X and the other statistical calculations for X followed by those for Y and finally you see Σxy which here will be the expected value of XY. I used $a = 1, b = 2, c = 3$, so the probabilities were in L_1, the values of X in L_2, and the values of Y, in L_3. The readout gives the

$$E(X) = \Sigma x = \bar{x} = 2.5,$$
$$E(Y) = \Sigma y = \bar{y} = 3.3,$$

and

$$E(XY) = \Sigma xy = 7.4.$$

Notice that $E(X)$ multiplied by $E(Y)$ is NOT $E(XY)$, since

$$(2.5)(3.3) = 8.25 \neq 7.4.$$

This is an example showing that the general multiplication rule that we might suspect is in fact FALSE-you cannot compute the expected value of a product by multiplying expected values, and as we showed in the last lecture, this means that you cannot compute the expected value of XY only using the expected values of X and of Y alone-more information will be required, since if not, the retraction rule would imply that the expected value of the product is always just the product of expected values. We will next see what this extra information is.

To deal with two unknowns, in general we have to allow for the possibility that one of them actually contains information about the other. As an example, if we have a pond full of trout, and if one of them is pulled from the pond let X be his length in inches and Y be his weight in pounds. If you are given the information that the trout is longer than average for these trout, then you would guess his weight is likely to be above average for these trout. Since we know now that the expected value is the average as we see from examples, this leads us to think of the relationship between two such unknowns as partially captured by how much knowing one is above or below average influences us to to think the other is above or below average.

To examine this in more detail, we first point out that from our examples so far, we begin to realize that if you know the actual average of population, then that is what you should guess for something taken from the population. For instance, assume that the overall average of the lengths of all the trout in the pond is 14 inches. If X is the length of a trout taken from the pond which you cannot see, then your best guess as to the length of this trout is 14 inches. That is to say, $E(X) = 14$. On the other hand, $X - 14$ is then your error and in this situation, we call $D_X = X - 14$ the deviation in length for this trout. Notice if the trout is of above average length, then it has a positive deviation, whereas if it has below average length, then it has a negative deviation. If we try to guess our error, we would try to guess the value of $D_X = X - 14$. But, from our rules of expectation, we know that

$$E(D_X) = E(X - 14) = E(X) - 14 = 14 - 14 = 0.$$

This means that we would guess our error is zero. But remember that the optimal guess is the average, so what is happening is that the positive deviations are canceling the negative deviations-all the errors are averaging out. To stop this from happening and thereby get a better handle on the error of our guess, we should square the deviation and guess that. Precisely, we should use

$$Var(X) = E(D_X^2) = E((X - 14)^2)$$

to gauge our guess of our error in this situation. We call $Var(X)$ the **VARIANCE** of X. To make up for having squared the deviations, we define the **STANDARD DEVIATION**, denoted σ_X to be the square root of the variance of X, so

$$\sigma_X = \sqrt{Var(X)}.$$

The standard deviation is really our best reasonable way to gauge our error.

To get a handle on the way length and weight relate to each other, we can likewise consider the product of deviations. Suppose that the overall average weight of the fish in the pond is 4 pounds and Y is the weight of the fish pulled from the pond. The product of the deviations is then

$$D_X D_Y = (X - 14)(Y - 4).$$

If you are told that the fish from the pond has above average length, then you would reasonably think that it is more likely than not that this fish is above average in weight. Of course it is not certain, since there may be some long skinny fish in the pond, but these skinny fish are likely to be rare, that is fish which are longer than average but weighing less than average. This means that if D_X is positive, then the same is likely to be true for D_Y. Thus, in this case, the product $D_X D_Y$ is also positive. On the other hand, if you are told the fish from the pond is shorter than average, so D_X is negative, then you would be reasonable to guess that it weighs less than average so that D_Y is also negative. Notice when both deviations are negative their product is again positive as the product of two negative numbers is positive. We have realized now that $D_X D_Y$ should very likely tend to be positive in this situation, so that if you had to guess a value for this product your should guess a positive number. That is here it is reasonable that

$$E(D_X D_Y) > 0.$$

If this expected product of deviations $E(D_X D_Y)$ is very large then we would think there is a close relationship between length and weight for this population of fish, whereas if $E(D_X D_Y)$ is close to zero, then we might think there is not a close relationship between length and weight here. If the expected product of deviations is negative but large, then there is still a close relationship between X and Y, but we call it a negative relationship in this case. In any case, we call this the **COVARIANCE** of X and Y, denoted $Cov(X, Y)$. This means by definition

$$Cov(X, Y) = E(D_X D_Y),$$

so we see right away, that $Var(X) = Cov(X, X)$, that is, the variance of X is just its covariance with itself. Of course X should be very well related to itself, since if you are told X is above average, then you know Y is above average in case $Y = X$ is the same unknown. However, we could be fooled in this situation if there is a lot of variation in the population of fish. In other words, if the relationship is not very strong but there are enormous deviations, then the products will be enormous leading to a large expected product or covariance. To compensate for large deviations, we divide by the standard deviation of each unknown and call the result the **CORRELATION COEFFICIENT** of X and Y which is denoted by ρ, the Greek letter *rho*. Thus, by definition,

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

The correlation coefficient gives the true measure of how well X and Y relate. If it is near zero, then there is not much relationship whereas if it is near ± 1, then there is a strong relationship. In fact, we will see that we can use regression analysis to guess a best prediction of Y when we have the value of X, and when we do, the square of the correlation coefficient, ρ^2, tells us the amount of variation in Y that X accounts for, and this is called the **COEFFICIENT OF DETERMINATION**.

To calculate ρ and ρ^2 quickly with the calculator, make sure you have "diagnostics on" by going to the catalogue button (second function of a button on the bottom row), and scroll down until you see "diagnostics on" in the long alphabetical list. With the cursor on "diagnostics on", hit the enter button several times until you see "done" appear on the screen. Then go to the "linreg(ax+b)" in the CALC menu followed by L_b, L_c, L_a, with the data from the above table, and pushing the enter button gives a readout with the table data above gives

$$a = -.0918918919$$

$$b = 3.52972973$$

$$r^2 = .1280460789$$

$$r = -.3578352678.$$

In the readout here, the reported value of r is the correlation coefficient ρ, so the reported value of r^2 is the coefficient of determination. This coefficient of determination is fairly small and tells us that there is not a lot of use here in trying to predict Y from X in the tabulated example.

7. LECTURE MONDAY 25 JANUARY 2010

Today we reviewed expectation, covariance, standard deviation, and correlation, and looked at how they all relate in situations where there are two related unknowns, such as length and weight of fish. The basic formulas are, when we have ANY two unknowns X and Y,

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

$$D_X = X - \mu_X,$$

$$Cov(X,Y) = E(D_X D_Y),$$

$$\sigma_X^2 = Var(X) = Cov(X,X),$$

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

$$\rho^2 = \text{ coefficient of determination.}$$

Remember μ_X then is just another symbol for the best guess for the unknown, it tends to be referred to as the **MEAN**, but it is just the same as the expected value, $E(X)$. It is convenient to have this other expression for the mean or expected value, as always using $E(X)$ in equations and expressions can sometimes be hard to read. The second equation simply says the D_X is the deviation from the mean. Since your optimal guess is μ_X, we see that D_X is just the error you make.

Since

$$E(D_X) = E(X - \mu_X) = E(X) - \mu_X = \mu_X - \mu_X = 0,$$

we conclude that if you have to guess your error, you would guess zero, but we can also conclude from our rules and some algebra, that we also have

$$Cov(X,Y) = E(D_X D_Y) = E[(X - \mu_X)(X - \mu_Y)] = E(XY) - \mu_X \mu_Y,$$

$$\sigma_X^2 = Var(X) = Cov(X,X) = E(D_X^2) = E(X^2) - (\mu_X)^2,$$

$$\rho = \frac{E(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y},$$

$$Cov(X,Y) = \rho \sigma_X \sigma_Y,$$

where the last expressions in each of these first two equations results from multiplying out the the expression inside the expectation and applying the expectation rules to the result, keeping in mind that $E(X - \mu_X) = 0$. When turned around, the first two equations also tell us that

$$E(XY) = \mu_X \mu_Y + Cov(X,Y) = \mu_X \mu_Y + \rho \sigma_X \sigma_Y,$$

and

$$E(X^2) = \mu_X^2 + \sigma_X^2.$$

These last two equations tell us that when it comes to logical consistency in guessing, if you cannot be certain of the values for X and Y, then you cannot simply multiply your guesses to get the best guess for the product of two unknowns.

It is not immediately obvious, but in fact, using only our rules and some algebra, it can be shown that

$$\rho^2 \leq 1,$$

and therefore

$$-1 \leq \rho \leq 1.$$

A case where $\rho = 1$ would be a case where there is perfect positive correlation between X and Y, and the case where $\rho = -1$ would be a case where there is perfect negative correlation between X and Y. A case where $\rho = 0$ would be a case where X and Y are completely uncorrelated. In general, ρ is a number between negative one and one, the closer to one the better the positive correlation, and the closer to negative one, the better the negative correlation. The coefficient of determination is really the final arbiter or how well the two variables relate.

Typically, the numbers you would want to find in a situation with two related unknowns X and Y, are the five numbers: μ_X, μ_Y, σ_X, σ_Y, and ρ. For the case of the fish in the pond, where X is length and Y is weight, if you took a large sample, and calculated the means and standard deviations and correlation in the data, these would estimate the true population values which have the Greek symbols.

For now, we will assume we are given these values of means, standard deviations, and correlation to see how it can be used to better your guess for Y when you are given the value of X. To begin, let's take an example with specific numbers. Suppose that, thinking of length of fish in inches and weight of fish in pounds, for fish from a pond,

$$\mu_X = 14,$$

$$\mu_Y = 5,$$

$$\sigma_X = 3,$$

$$\sigma_Y = 1.7,$$

and

$$\rho = .7.$$

Remember, the mean or expected value of an unknown is always the best guess unless there is additional information beyond the mere background information. Thus, if you are told a fish has been taken from the pond and you need to guess its length, then you should guess 14 whereas if you need to guess its weight, you would guess 5. Of course, since ρ is $7/10$, which is positive, if you are told the fish is 20 inches long, then you would want to increase your guess for the weight substantially above 5. Notice that the 20 is 6 units above 14, which is two standard deviations above the mean, as the standard deviation in length is 3. Thus if the correlation of length and weight were perfect ($\rho = 1$), you would guess that the weight is also two standard deviations above the mean. Now the weight standard deviation is 1.7, so two standard deviations is 3.4, and therefore with perfect correlation, the best guess for the weight of a 20 inch fish is 8.4 pounds. However, we only have $\rho = .7$, so this in fact turns out to mean that we should not increase our weight guess by two standard deviations, but only by (2)(.7) standard deviations which is $(.7)(3.4) = 2.18$, and that results in a guess of only 7.18 for the twenty inch fish. Notice that even with the correlation not being perfect, if you are told that the fish has length 14 inches, then you should guess his weight is exactly 5 pounds.

In general with two unknowns we look for a simple linear relationship, and this means that if we graph the value y we guess for Y when the value of X is given to be the number x, this straight line must go right through the point (μ_X, μ_Y), reflecting the fact that no matter what the correlation, we have μ_Y is still the best guess for Y when we are told that X actually has the value μ_X. If the line has slope α, then the line must have equation

$$y = \beta(x - \mu_X) + \mu_Y.$$

Notice that this can also be written as

$$y = \alpha + \beta x$$

where

$$\alpha = \mu_Y - \alpha\mu_X.$$

From our example, we can see that the general formula for α is

$$\alpha = \rho\frac{\sigma_Y}{\sigma_X}.$$

To understand something about why these equations are used, we must realize that to say a guess is best in some sense means that in some sense we are trying to minimize our error. But, we obviously cannot know what our error will be until we actually can see what the true value is, and that may even be impossible, so we even have to guess our error. We have already seen that our actual error in the sense of the simple deviation is expected to be zero. This means that when we guess, we also guess our error is zero. but remember that the guess for a square is not just the square of the guess, so our guess for the squared deviation will usually be positive and not zero. In fact, that is the variance, σ^2. You might ask why the squared error is the variance-maybe if you guess differently you could have an even smaller guess for the squared error. Suppose that instead of μ_X you guess the number c. Your error is now

$$X - c = (X - \mu) + (\mu_X - c),$$

so using some algebra,

$$(X - c)^2 = (X - \mu_X)^2 + (\mu_X - c)^2 + 2(\mu_x - c)(X - \mu_X).$$

Remember now that $\mu_X - c$ is just a number, say d. We have then

$$d = \mu_X - c,$$

$$(X - c)^2 = (X - \mu)^2 + d^2 + 2d(X - \mu_X) = D_X^2 + d^2 = 2dD_X.$$

But, $E(D_X) = 0$ and $E(D_X^2) = \sigma_X^2$, so applying expectation to the previous equation we have

$$E[(X - c)^2] = \sigma_X^2 + d^2 + 2d(0) = \sigma_X^2 + d^2.$$

Now, the best we can do to minimize this squared error is to make d^2 as small as possible, since any square is non-negative. Obviously, then the best one can do to minimize the guess for the squared error is to set $d = 0$. Since $d = \mu_x - c$, this means the best you can do to minimize your guess for what the squared error will be is to choose $c = \mu_X$.

In the case of two unknowns X and Y, if we try to use an equation of the form $y = \alpha + \beta x$ to guess the value y of Y when we are given the value x of X, then we have the problem of choosing the coefficients α and β. The criterion we use here is to try to minimize our guess for the squared error again. Precisely, if we use the value a for α and b for β, then our guess for the squared error is

$$E[(Y - [a + bX])^2]$$

and in a similar fashion, we can rearrange algebraically to see that this can be expressed as a sum of squared terms so the minimum is attained by choosing each to be zero. The result is

$$\beta = \rho \frac{\sigma_Y}{\sigma_X},$$

and

$$\alpha = \mu_Y - \beta \mu_X.$$

You use the standard deviations and correlation coefficient to find β, and you use the means and β to find α. In the TI calculator, when you use the "linreg(a+bx)", the value of a in the readout gives the value of α, and the value of b in the readout gives the value for β. When sample data is used, the value for a reported is then an approximation of the true α, and the value of b reported is an approximation for the value of β.

8. **LECTURE** WEDNESDAY 27 JANUARY 2010

Today we discussed the difference between population data and sample data and the use of sample data to estimate the various population parameters such as the mean, standard deviation, variance, covariance, and correlation coefficient. In particular, we noted that in several formulas a non-obvious correction factor

$$\frac{n}{n-1}$$

shows up which gives best estimates in the case of the variance and covariance rather than blindly using the sample data as if it were the whole population, which would lead to underestimates. We also noted that this factor completely cancels out when calculating the correlation coefficient, so when using the calculator to calculate linear regression coefficients and correlation coefficients, and coefficients of determination, we do not need to worry about the distinction between sample and population. Thus, if data for whole populations or unknowns is entered for the unknowns X and Y, then the reported value of r is in fact the value of ρ, the true correlation coefficient, whereas if the data is merely sample data, the reported value of r is denoted r in your text book and is merely an estimate of ρ, called the sample correlation coefficient. Likewise, in the case of sample data, the reported value of r^2 is merely an estimate of the true coefficient of determination, ρ^2 which tells how well the regression analysis actually works. In many applications, a coefficient of determination as low as one tenth could be useful. However, with such a low coefficient of determination, we should suspect other unknowns are lurking and which should be searched for so as to improve guess work.

We noted that if we have a sample of values for an unknown X, then the sample mean $\bar{x}$ gives an estimate of the true mean $\mu_X = E(X)$. If Y is another unknown and we have sample data for Y, then its sample mean $\bar{y}$ gives and estimate of $\mu_Y = E(Y)$. For instance, if we have a sample of size n for X with sample mean $\bar{x} = 7$, then we would have

$$E(X|\bar{x} = 7) = 7,$$

since if the only information you have about X is the sample mean with value 7, then that should be your guess for the value of X. You certainly have no basis for guessing a higher value than 7 and you have no basis for guessing a lower value than 7. On the other hand, if the sample size n is just $n = 3$, then you would not think of this guess as very reliable, whereas if your sample size is $n = 1000$, then you would probably think this gives a fairly reliable guess.

Recall, that we have the useful computation formula for variance (recall the useful notation $D_X = X - \mu_X$)

$$E(D_X^2) = E((X - \mu_X)^2) = Var(X) = E(X^2) - (\mu_X^2).$$

We notice that as a consequence, whenever the mean of the squares equals the square of the mean, then the variance and therefore the standard deviation must be zero. On the other hand, every squared deviation is non negative, so if the variance is zero, then all deviations must be zero, and this means the only possible value of X is μ. Thus, for the case of the dice in a box, as there is more than one possible value, the standard deviation and variance are not zero, and the mean of the square is more than the square of the mean. To see this easily in the simplest example, suppose that $X = I_A$ is an indicator of event A. In this case, $X^2 = X$, since the only values an indicator can have are zero and one and these numbers each equal their own square. We then have

$$E(X^2) = E(X) = E(I_A) = P(A),$$

whereas, since

$$E(X) = E(I_A) = P(A),$$

we have

$$[E(X)]^2 = [P(A)]^2.$$

Obviously, in this case, as a probability is between zero and one, the only way that $P(A)$ can equal its own square is for $P(A)$ to be zero or one, meaning that A is either something we know is certainly true or A is something we know is certainly false. Any time

$$0 < P(A) < 1,$$

we have $P(A)$ is not equal to its square and thus the indicator of A is an unknown for which the expected value of the square is different than the square of its expected value. Thus in general, we cannot simply multiply our guess for X and Y in order to get the best guess for the product XY.

This means that usually we have $E(XY)$ is different from $E(X)E(Y) = \mu_X\mu_Y$. Remember we also have the formula

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

This shows that whenever the covariance is not zero, the mean of the product is different than the product of the means. From sample data, you might suspect that just as $\bar{x}$ and $\bar{y}$ are the best estimates of μ_X and μ_Y, that $\bar{x}\bar{y}$ gives the best estimate of the mean of XY and that therefore

$$\overline{xy} - \bar{x}\bar{y}$$

gives the best estimate of $Cov(X,Y)$. Unfortunately, it can be shown using our rules that in fact the best estimate of $Cov(X,Y)$ from the sample data for a sample of size n is the *sample covariance*

$$cov(x,y) = \frac{n}{n-1}[\overline{xy} - \bar{x}\bar{y}].$$

In particular, the best estimate for $Var(X) = \sigma_X^2$ is the *sample variance*

$$s_x^2 = \frac{n}{n-1}[(\overline{x^2}) - (\bar{x})^2].$$

The sample correlation coefficient r is the estimate of ρ which is formed by putting these estimates together, and so as

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

this leads to the *sample correlation coefficient*, denoted by r where

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\frac{n}{n-1}[\overline{xy} - \bar{x}\bar{y}]}{\sqrt{\frac{n}{n-1}[(\overline{x^2}) - (\bar{x})^2]}\sqrt{\frac{n}{n-1}[(\overline{y^2}) - (\bar{y})^2}}.$$

You can now see that all the correction factors $n/(n-1)$ in the formula for r simply cancel giving the simpler calculation formula

$$r = \frac{[\overline{xy} - \bar{x}\bar{y}]}{\sqrt{[(\overline{x^2}) - (\bar{x})^2]}\sqrt{[(\overline{y^2}) - (\bar{y})^2}}.$$

Always keep in mind that when using regression analysis, any sample data will give estimates for the regression line in the calculator readout. Using the values of a and b in "linreg(ax+b)" to form the equation $y = ax + b$ to use for guessing Y when X is given to you can always be done, but you must look at the value of ρ^2 to know if it has any reliability. For ρ^2 near one the linear regression is very reliable, whereas for ρ^2 near zero, the linear regression is virtually useless. We noted that if there is enough data, a value for r^2 small but positive can sometimes be evidence that there is correlation in situations where correlation is to be avoided.

We pointed out that the sample data can be plotted in a scatter plot and if the picture shows a reasonable trend, drawing the regression line totally by sight without looking at any numbers can often lead to reasonable results. In many situations, it is actually the value of r^2 that is the only thing that needs to be calculated from the data so you can tell if the eyeball work has any validity.

Using sample data examples, we worked some problems using regression and observed that the calculations always alow the regression analysis to give a guess, but depending on the sample data, the guess can be reasonable or not.

9. **LECTURE** FRIDAY 29 JANUARY 2010

Today we discussed deviations, covariance, and variance. We noted that as $D_X = X - \mu_X = X - E(X)$ is the deviation of X from its mean for any unknown X, then certain things done to unknowns are reflected in their deviations. For instance, if we double X, then all deviations are doubled, which is to say that

$$D_{(2X)} = 2D_X.$$

For instance if A makes 10 thousand dollars more than B in annual salary, and all salaries are doubled, then A makes 20 thousand dollars more than B. On the other hand, if all salaries are simply increased by a fixed amount, say 5 thousand dollars, then A still makes exactly 10 thousand dollars more than B. In general then, for any constant c, we have

$$D_{(cX)} = cD_X.$$

If we have two unknowns X and Y we can form their total T, and notice that

$$D_T = D_X + D_Y.$$

For instance, if X is the daily balance in credit account A, and Y is the daily balance in credit account B, then the deviation D_X is the amount that the balance exceeds what is expected and likewise for D_Y. If the daily balance for account A is on average one thousand dollars, and the daily balance for account B is on average two thousand dollars, then if I have both accounts, then my total daily balance for the two accounts together is $T = X + Y$ and the average daily balance is three thousand dollars. If today, account A is 500 dollars over the average for account A, then its balance is 15 hundred dollars, and the deviation is 5 hundred dollars. If account B today is 2 hundred dollars over average, then today its balance is 22 hundred dollars and the deviation is two hundred dollars. Notice the total T is today 37 hundred dollars, which is 7 hundred dollars above what is average for T.

The simplest deviation to deal with is that for a constant, c. This is because $E(c) = c$, so

$$D_c = c - \mu_c = c - E(c) = c - c = 0,$$

and therefore

$$D_c = 0,\ c\ constant.$$

That is, a constant certainly never deviates from its value-its value is certain. In particular, we have

$$D_{X+c} = D_X + D_c = D_X,$$

or

$$D_{X+c} = D_X,\ c\ constant,$$

which goes along with our prior observation that adding a fixed amount to every persons salary does nothing to the differences between various salaries.

Now, we need to use these observations to work out some simple rules for dealing with variance and covariance. Remember that we have

$$\rho\sigma_X\sigma_Y = Cov(X,Y) = E(D_XD_Y).$$

The right hand side can be used to work out some simple properties of covariance whereas in problems with two unknowns you will be given means, standard deviations, and correlation to calculate basic covariance. Thus, as multiplication is commutative, $D_XD_Y = D_YD_X$, it follows that

$$Cov(X,Y) = Cov(Y,X).$$

Compare this with $ab = ba$ for multiplication of ordinary numbers. For dealing with a constant, as $D_c = 0$, it follows that

$$Cov(X, c) = 0, \ c \ constant.$$

In particular, as $Var(X) = Cov(X, X)$, it must be the case that

$$Var(c) = 0, \ c \ constant.$$

Since

$$D_{bX} = bD_X,$$

we must have

$$Cov(bX, Y) = bCov(X, Y), \ b \ constant.$$

Thus

$$Cov(bX, cY) = bcCov(X, Y),$$

and as $Var(X) = Cov(X, X)$, this also means we have

$$Var(cX) = c^2 Var(X), \ c \ constant$$

so

$$\sigma_{cX} = |c|\sigma_X, \ c \ constant.$$

For the total of two unknowns, we have for $T = X + Y$,

$$D_T = D_X + D_Y,$$

so for any unknown W, we have

$$D_W D_T = D_W D_X + D_W D_Y,$$

and therefore

$$Cov(W, T) = E(D_W D_T) = E(D_W D_X + D_W D_Y) = E(D_W D_X) + E(D_W D_Y) = Cov(W, X) + Cov(W, Y).$$

This shows that in general,

$$Cov(W, X + Y) = Cov(W, X) + Cov(W, Y).$$

Again there is a similarity here with the ordinary distributive law for multiplication of numbers

$$a(b + c) = ab + ac.$$

Now, with ordinary numbers, we have

$$(a + b)(c + d) = ac + ad + bc + bd.$$

This can be pictured with areas of rectangles by taking a horizontal side of length $a + b$ and a vertical side of length $c + d$. Marking off a length a from the lower left corner on the horizontal edge and marking off a length c on the vertical edge we see four rectangles. The lower left has area ac, the upper left has area ad, the lower right has area bc, and the upper right has area cd. The total area is the sum of these terms and illustrates the above equation. We can do the same schematically with covariance. Mark the horizontal edge and label the two parts as U and W and then do the same with the vertical edge marking with X and Y. Then the lower left rectangle inside is marked $Cov(U, X)$, the upper left rectangle is marked $Cov(U, Y)$, the lower right rectangle is marked $Cov(W, X)$ and the upper right rectangle marked $Cov(W, Y)$. We then see that

$$Cov(U + W, X + Y) = Cov(U, X) + Cov(U, Y) + Cov(W, X) + Cov(W, Y).$$

We can apply this to the variance of a total, since $Var(X) = Cov(X, X)$. This is similar to the special case for numbers of squaring a binomial

$$(a + b)^2 = a^2 + b^2 + 2ab.$$

With covariance, it takes the form

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y).$$

Since

$$Cov(X, -Y) = Cov(X, (-1)Y) = (-1)Cov(X, Y) = -Cov(X, Y),$$

we can see that

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

in similar fashion to

$$(a - b)^2 = a^2 + b^2 - 2ab.$$

For instance, if the covariance and X and Y happens to be zero then both $Var(X + Y)$ and $Var(X-Y)$ will simply be $Var(X)+Var(Y)$, that is in the case of two uncorrelated unknowns, the variance of the *sum or difference* is simply the *sum* of the variances, you do not subtract variances for the case of a difference of two unknowns, the variance terms will always add. You should try to use these little geometric pictures to help visualize how to deal with the various terms.

As an example, suppose that we have two unknowns X and Y, with

$$\mu_X = 70,\ \mu_Y = 80,$$
$$\sigma_X = 5,\ \sigma_Y = 7,$$
$$\rho = .8.$$

To calculate the variance of $X + Y$, we must go through covariance, as

$$Cov(X, Y) = \rho\sigma_X\sigma_Y = (.8)(5)(7) = (4)(7) = 28,$$

so

$$Var(X + Y) = Var(x) + Var(Y) + 2Cov(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2[\rho\sigma_X\sigma_Y]$$
$$= (5)^2 + (7)^2 + 2[(.8)(5)(7)] = 25 + 49 + (2)(28) = 25 + 49 + 56 = 130.$$

We now know that

$$Var(X + Y) = 130,$$

so the standard deviation of the total $T = X + Y$ is

$$\sigma_T = \sqrt{130}.$$

Notice that to find the standard deviation of the total you must work with variances, to find variance of the total requires working with covariances, so finally in the end once you find the variance, you can take the square root to find the standard deviation of the total if it is needed.

We also reviewed the formula

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

and its special case using $X = Y$,

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

Notice that this last equation is sort of like the Pythagorean Theorem. In fact it is closely related to the Pythagorean Theorem. To remember it, think of a right triangle with the horizontal and vertical sides marked as μ_X and σ_X respectively, then the hypotenuse is marked $\sqrt{E(X^2)}$. We often need to deal with expected squares as when dealing with errors, we generally want to minimize the expected squared error.

For instance, when dealing with two unknowns X and Y, to try to use X values to guess Y values, we start by forming $W = a + bX$, and try to choose values for a and b so that a value for X gives a value for W which gives a good guess for Y. Notice that

$$R = Y - W$$

is the error when using W to guess Y, so we want to choose values for a and b for which $E(R^2)$ is minimum. Notice that

$$E(R^2) = \mu_R^2 + \sigma_R^2 = \mu_R^2 + Var(R) = \mu_R^2 + Var(Y - W)$$
$$= \mu_R^2 + Var(Y) + Var(W) - 2Cov(W, Y) = \mu_R^2 + \sigma_Y^2 + Var(W) - 2Cov(W, Y),$$

or simply

$$E(R^2) = \mu_R^2 + \sigma_Y^2 + Var(W) - 2Cov(W, Y).$$

Now, we have from our covariance rules

$$Cov(W, Y) = Cov(a + bX, Y) = bCov(X, Y) = b\rho\sigma_X\sigma_Y$$

and

$$Var(W) = Var(a + bX) = Var(bX) = b^2Var(X) = b^2\sigma_X^2.$$

When we substitute this into the previous equation, we have

$$E(R^2) = \mu_R^2 + \sigma_Y^2 + b^2\sigma_X^2 - 2b\rho\sigma_X\sigma_Y.$$

A little trick can be used to simplify this last equation. If we add and subtract the same thing from σ_Y^2,

$$\sigma_Y^2 = (1 - \rho^2)\sigma_Y^2 + \rho^2\sigma_Y^2,$$

and substitute this into our equation for $E(R^2)$, we have

$$E(R^2) = \mu_R^2 + (1 - \rho^2)\sigma_Y^2 + \rho^2\sigma_Y + b^2\sigma_X^2 - 2b\rho\sigma_X\sigma_Y.$$

Concentrate now on the last three terms and make a slight rearrangement:

$$\rho^2\sigma_Y^2 + b^2\sigma_X^2 - 2b\rho\sigma_X\sigma_Y = (\rho\sigma_Y)^2 + (b\sigma_X)^2 - 2(\rho\sigma_Y)(b\sigma_X).$$

This last expression reminds us of

$$(c - d)^2 = c^2 + d^2 - 2cd,$$

where $c = \rho\sigma_Y$ and $d = b\sigma_X$. Thus, the last three terms are just $(\rho\sigma_Y - b\sigma_X)^2$, which when substituted into the previous equation for $E(R^2)$ now gives

$$E(R^2) = \mu_R^2 + (1 - \rho^2)\sigma_Y^2 + (\rho\sigma_Y - b\sigma_X)^2.$$

The thing to notice here is that two of the three terms are squares so the smallest they can possibly be is zero. Remember, we still have not chosen a and b, and we want to choose them to give the smallest possible value to $E(R^2)$. When we look at the last squared term, then we can make it zero by simply solving the equation

$$\rho\sigma_Y - b\sigma_X = 0,$$

for b, since we have no control of the standard deviations, they are simply given. On the other hand, this gives the simple equation for b,

$$b = \rho\frac{\sigma_Y}{\sigma_X}.$$

To make the first square term zero, keep in mind we have now already chosen b, so we only need a. But to make the first square term zero is simply to set $\mu_R = 0$. Now, this gives

$$0 = \mu_R = E(R) = E(Y-W) = E(Y)-E(W) = \mu_Y - E(a+bX) = \mu_Y - (a+b\mu_X) = [\mu_Y - b\mu_X] - a.$$

This means, as b is already chosen, we have using that value of b, we should chose a to be

$$a = \mu_Y - b\mu_X.$$

Thus, the correlation coefficient, ρ, and the standard deviations for X and for Y are sued to determine b, and once b is found, we use the expected values of X and Y to determine a. Finally, notice that with these optimal values for a and b, usually denoted by α and β, respectively, we have both the first and last terms in the equation for $E(R^2)$ are zero and the result is

$$E(R^2) = (1-\rho^2)\sigma_Y^2.$$

This fact is useful because as any square is non negative, we must have

$$0 \leq E(R^2).$$

But also $\sigma_Y^2 \geq 0$, so therefore it must be true that

$$1 - \rho^2 \geq 0,$$

and therefore

$$\rho^2 \leq 1.$$

This forces us to realize that

$$-1 \leq \rho \leq 1,$$

for any two unknowns X and Y. These equations of course depend on choosing the optimal values for a and b denoted α and β, respectively which we found above to be

$$\beta = \rho\frac{\sigma_Y}{\sigma_X},$$

and

$$\alpha = \mu_Y - \beta\mu_X.$$

For instance for the example,

$$\beta = \rho\frac{\sigma_Y}{\sigma_X} = (.8)\frac{7}{5} = \frac{(4)(7)}{5^2} = \frac{28}{25} = 1.12,$$

and

$$\alpha = 80 - \beta(70) = 80 - (\frac{28}{25})(70) = 80 - (28)\frac{14}{5} = 80 - (28)\frac{28}{10} = 80 - 78.4 = 1.6.$$

This means that to use X to guess values for Y, we would use

$$W = 1.6 + (1.12)X,$$

so whenever we have a value x for X we have the best guess

$$y = 1.6 + (1.12)x$$

is the value for Y. That is, in general,

$$E(Y|X = x) = \alpha + \beta x,$$

and in the example

$$E(Y|X = x) = 1.6 + (1.12)x.$$

So, if we are find out the value of X is 75, then our best guess for the value of Y is

$$E(Y|X = 75) = 1.6 + (1.12)(75) = 85.6.$$

In general, if we have data on two related unknowns concerning a population such as length and weight in a population of fish, then using the data to calculate values of r and a and b gives approximate values for ρ and α and β, respectively.

10. **LECTURE** MONDAY 1 FEBRUARY 2010

Today we reviewed for TEST 1 to be given in class Wednesday 4 February 2010. We went over PRACTICE TEST 1 ANSWERS SPRING 2010. Make sure that the version of the practice test you use for review is the correct practice test and especially NOT practice test 1 from last spring. The version for SPRING 2010 has been posted online since early Friday morning of last week.

We reviewed the basic rules for mean, variance, covariance, standard deviation, correlation, linear regression, probability, conditional probability, conditional expectation, and their applications to various problems.

11. **LECTURE** WEDNESDAY 3 FEBRUARY 2010

TEST 1 IN LECTURE CLASS.

12. **LECTURE** FRIDAY 5 FEBRUARY 2010

We reviewed conditional expectation and probability and worked problems with Bayes' Theorem.

13. **LECTURE** MONDAY 8 FEBRUARY 2010

We discussed the general rules for counting. We denote by $n(A)$ the number of things in the set A. We noted the ADDITION RULE FOR COUNTING:

$$n(A \cup B) = n(A) + n(B),\ A,\ B,\ disjoint.$$

We noted that a useful way to denote the results of stepwise procedures is with a sequence where the k^{th} entry in the sequence denotes the result of the k^{th} step of the process. In particular, if we have a sequence of sets $A_1, A_2, A_3, ..., A_n$ and if the process consists of choosing one item from each set, then the set of all sequences $(a_1, a_2, a_3, ..., a_n)$ gives the set of all possible outcomes for the stepwise process, and this is also the **CARTESIAN PRODUCT**, P, of all these sets, denoted

$$P = A_1 \times A_2 \times A_3 \times ... \times A_n = \{(a_1, a_2, a_3, ..., a_n)\ such\ that\ a_k\ in\ A_k,\ each\ k \leq n\}.$$

We noted that in this case, we have a simple multiplication result for $n(P)$, namely,

$$n(P) = n(A_1) \cdot n(A_2) \cdot n(A_3) \cdot \cdots \cdot n(A_n).$$

We also see that in this situation, as a stepwise process, all the various steps are independent of each other. But, in fact that is not necessary for the multiplication rule. Notice that $n(A_k)$ is the number of ways to perform the k^{th} step of the process of choosing one item from each of these sets. More generally, we can consider processes where the set of possibilities for the outcome at step k may depend on the history of what the results were for the previous steps but where the number of those possibilities does not depend on the history of the previous steps. For instance, when you deal from a standard deck of cards, you have 52 possibilities for the first card dealt, but what is possible for the second card depends on what was dealt on the first card. On the other hand, no matter what card was dealt for the first card, there are only 51 possibilities for the second card, and no matter what cards are dealt for the first two cards, there are 50 possibilities for the third card, and so on.

Thus, in general, if the number of ways to perform each step is independent of the particular results on the previous steps, then the multiplication rule still works. We still use P to denote the set of outcomes for the n step process, but it is no longer a cartesian product. That is, if for each k, the number of ways to perform step k is m_k independent of what came before in the process, then for all n steps, the total number of outcomes is $n(P)$ where

$$n(P) = m_1 \cdot m_2 \cdot m_3 \cdot \cdots \cdot m_n.$$

We call the the **MULTIPLICATION PRINCIPLE FOR COUNTING**.

As a particular example, we noted that the number of ways to arrange r things from a set of n things is nPr which is calculated in the PRB menu of the MATH menu in the calculator. To arrange r things from a set of n things, is for instance to place r letters into r given mailboxes, when there are n letters in all to chose from. If the mailboxes are $B_1, B_2, B_3, ..., B_n$, then considering step k as the act of choosing one of the letters to go in B_k, there are n choices for a letter to go in B_1, so step one has n ways to be done, whereas then step 2 has only $n-1$ ways

to be done, step three likewise has only $n-2$ ways to be done, and so on, until finally step r has only $n-(r-1)=n-r+1$ ways to be done. This means

$$nPr = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-r+1).$$

To arrange all n things then there are $n!$ ways where

$$n! = nPn = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1.$$

We read $n!$ as "n factorial".

Now, we can express the nPr in terms of factorials using the formulas so far, but it is easier to simply apply the multiplication principle for counting. We can view the problem of arranging all n things as done in two steps. Step 1 arrange r of them and step 2 arrange the rest, that is the remaining $n-r$ things. According to the multiplication principle for counting, this means that

$$n! = nPn = (nPr)([n-r]P[n-r] = (nPr)[(n-r)!],$$

and therefore, on solving this equation for nPr we find

$$nPr = \frac{n!}{(n-r)!},$$

a useful formula expressing nPr in terms of factorials.

For instance, using your calculator, you can calculate 52 nPr 5, which is the number of ways to deal out, in order, 5 cards from a standard 52 card deck, and you find it is 311875200, a fairly large number. Thus, there is a first card, a second card, a third card, a fourth card, and a fifth card, here, and it matters say whether you got the ace of diamonds first or second. In many card games, it does not matter in which order you are dealt your cards, it only matters which cards you get. For instance, when we speak of the number of 5 card hands from a standard deck of cards, we do not care in which order the cards are dealt, we only care about which cards we get as a final result of the deal. In general, we are often interested in the number of ways to choose r things from a set of n things, and this number is denoted nCr in your calculator. To calculate the number of 5 card hands from a standard deck, we need to calculate 52 nCr 5 = 2598960, a much smaller number than 52 nPr 5 = 311875200.

We can express nCr in terms of nPr using the multiplication principle for counting. Imagine arranging r things from a set of n things done as a two step procedure. Step 1 choose the r things to be arranged, and then Step 2, arrange all r of the things chosen in step 1. Notice that Step 1 can be done in nCr ways and Step 2 can be done in $rPr = r!$ ways, so by the multiplication principle for counting,

$$nPr = (nCr)(r!)$$

which can be solved for nCr giving

$$nCr = \frac{nPr}{r!} = \frac{n!}{(n-r)!r!}.$$

We noted that this formula shows that the number of ways to chose r things from n things must be the same as the number of ways to chose $n-r$ things from a set of n things. This may seem strange at first, but when you stop to think about it a minute, you can notice that deciding which r things to include in what you chose is exactly the same as deciding which $n-r$ things not to include in what you chose. That is choosing what to include is the same as choosing what to exclude.

A more useful formula relating these numbers is a result of the addition principle for counting which have not really applied yet. This is the formula for Pascal's Triangle

$$[n+1]C[r+1] = nCr + nC[r+1].$$

To see why this is true, imagine a box containing $n+1$ blocks of which one is red and all the rest are white. The outcomes for choosing $r+1$ blocks can be separated into two sets of outcomes by simply noticing that any time you chose $r+1$ blocks, you either did or did not get the red block. Let A be the set of all possible choices where you do get the red block and B be the set of all possible choices where you did not get the red block. To create an outcome in A, you first get the red block (there is only one way to do that as there is only one red block), and next choose the remaining r blocks you need from the n white blocks in the box. Thus,

$$n(A) = 1 \cdot (nCr) = nCr.$$

To create an outcome in B, you cannot use the red block, so you must chose all $r+1$ blocks from the n white blocks in the box, so

$$n(B) = nC[r+1].$$

Since every outcome of choosing $r+1$ blocks must fall in either A or B, we must have

$$n(A \cup B) = [n+1]C[r+1].$$

Since A and B are clearly disjoint, by the addition principle for counting we have

$$[n+1]C[r+1] = nCr + nC[r+1].$$

We finished the class by using these counting methods to calculate the probability of being dealt all five cards of the same suit when being dealt 5 cards from a standard deck of cards.

14. **LECTURE** WEDNESDAY 10 FEBRUARY 2010

CLASS DID NOT MEET

15. **LECTURE** FRIDAY 12 FEBRUARY 2010

CLASS DID NOT MEET

16. **LECTURE** MONDAY 15 FEBRUARY 2010

CLASS DID NOT MEET

17. **LECTURE** WEDNESDAY 17 FEBRUARY 2010

CLASS DID NOT MEET

18. **LECTURE** FRIDAY 19 FEBRUARY 2010

CLASS DID NOT MEET

19. **LECTURE** MONDAY 22 FEBRUARY 2010

Today we discussed distributions for unknowns, the cumulative distribution function for an unknown, and the probability density function or probability distribution function for an unknown. In particular, we discussed the binomial distribution and how to calculate it with the calculator as well as the hypergeometric distribution and its calculation.

The **CUMULATIVE DISTRIBUTION FUNCTION** of the unknown X is usually denoted F_X and is the real valued function of the real variable x whose rule is

$$F_X(x) = P(X \leq x).$$

We observed that if b is a specific number for which $P(X = b)$ is positive, then the graph of F_X has a jump at $x = b$. If F_X does not have a jump at $x = b$, then it must be that $P(X = b) = 0$. This does not mean that X cannot equal b.

We defined the **PROBABILITY DENSITY FUNCTION** for X, denoted f_X, as the function whose graph has the property that for any two numbers a, b with $a < b$, the area under the graph between $x = a$ and $x = b$ gives $P(a < X \leq b)$. We observed that if F_X does not have a jump at $x = b$, then $f_X(b)$ is the slope of the tangent line to the graph of F_X at the point $(b, F_X(b))$. On the other hand, we observed that if F_X does have a jump at $x = b$, which means that $P(X = b) > 0$, then the graph of f_X would have to have "infinite height" at $x = b$, which seems contradictory. We resolve this issue pictorially by using spikes. If $P(X = b) > 0$, we put a spike of height $P(X = b)$ right over the point $x = b$ on the graph. Thus, $P(a < X \leq b)$ is the area under the graph of f_X between $x = a$ and $x = b$ plus the heights of all the spikes over points x with $a < x \leq b$. For instance, if X is the number up on a loaded dice, then the graph of f_X consists of six spikes over the points $x = 1, 2, 3, 4, 5, 6$ whose TOTAL height is one. On the other hand, if X is the length of a fish to be selected from a population of fish, then f_X would be a curve without spikes.

In case we have a population of fish with mean length$\mu = 40$ and standard deviation $\sigma = 5$, then the length of a fish, X, from this population, given we know nothing else about this fish, would be governed by the normal distribution, so f_X is a "bell curve'. Generally, when a population is normal, this is stated as an assumption, but in fact, it can be proven mathematically, that whenever the only thing you know is μ_X and σ_X, then f_X is the normal distribution.

20. **LECTURE** WEDNESDAY 24 FEBRUARY 2010

Today we discussed the binomial and hypergeometric distributions and some of their applications.

21. **LECTURE** FRIDAY 26 FEBRUARY 2010

Today we reviewed the binomial and hypergeometric distributions for counting successes and then discussed the Poisson distribution for counting successes. The situation for the Poisson distribution is that there is no longer a number of trials, but rather an amount that is examined. For instance, you might be counting the number of tadpoles in 5.23 gallons of pond water or the number of bears in 24.6 square miles of forrest. Here, the only number that characterizes the distribution is the number expected. If you expect 43.8 bears in the 24.6 square mile of forrest, then you have the numerical information required to calculate the probability that you will actually find 42 bears in that region of forrest. But in order for the calculation we do to be valid here, we need an assumption. For the bears in the forrest the assumption is that bear counts for disjoint parts of the forrest are independent of each other. Likewise for the tadpole count the assumption would be that tadpole counts for disjoint regions of the pond water are independent of each other. In that case, the Poisson distribution applies, and in your calculator, the format is simple. If μ is what you expect, and X is the actual count, then

$$P(X = k) = e^{-\mu}\frac{\mu^k}{k!} = poissonpdf(\mu, k).$$

We worked several examples using the Poisson distribution in the calculator and observed that in a sense, the Poisson distribution is a certain "limit" of the binomial distribution. We also observed that if W is the amount you must examine before finding the first success, then

$$P(W > x) = poissonpdf(\mu, 0),$$

where μ is the expected count for the amount x. For instance, if we expect 6 trolleys per hour, then we expect 1.5 trolleys in $x = 15$ minutes, so the probability we must wait more than 15 minutes for a trolley when on average they arrive at 6 per hour is

$$poissonpdf(1.5, 0) = .2231301601.$$

22. **LECTURE** MONDAY 1 MARCH 2010

Today we began by reviewing the three counting distributions: binomial, hypergeometric, Poisson. In each there is a sample size, but in the first two (binomial and hypergeometric), the sample size is itself a count of the number of trials or observations made. In the case of the Poisson distribution, the sample size must be measured. The only parameter you must know of figure out for the Poisson distribution is the expected value, and is often expressed as what is expected for a sample of one unit size. For instance, if I examine pond water and expect to fine ten tadpoles per gallon of pond water, then in 5.3 gallons I expect to find 53 tadpoles. If I am standing on the corner watching buses go by, and if the go by on average at ten per hour, then I expect to see five buses go by in half an hour and fifteen buses go by in an hour and a half. What distinguishes the binomial from the hypergeometric is that in the binomial we must assume that all the trials are independent, so if the population is finite, we must be drawing with replacement for the binomial to apply, whereas if we are tossing a coin over and over to see if heads comes up, then the successive trials are reasonably assumed independent. If a policeman is observing traffic with a radar gun to see which cars are speeding, and if he watches the next one hundred cars go by, it is reasonable to assume that the different drivers are independent of each other in their decision whether to speed or not, so it would be reasonable to assume that the results of different observations as to whether a car is speeding or not are independent of each other. Now, one could argue that in heavy traffic, if one car is speeding then they are probably all moving together roughly so as not to bump into one another, since no driver wants to have his trunk smashed for only going the speed limit. The assumption of independence in practice needs to be examined carefully. For instance on a lonely stretch of rural highway where cars are separated by large distances, it would be reasonable to assume that the different cars are independent of each other as to whether or not they are speeding. For the hypergeometric distribution we have the situation where we are drawing from a small finite population without replacement. For instance, if we are going to draw twenty cards from a standard deck of cards (without replacement) and count the number of diamonds we get, then we would use the hypergeometric distribution if we can assume a "fair deal" that is all sets of twenty cards which could possibly be taken from the deck are all equally likely to be the one we get. We reviewed the formulas and methods of calculating probabilities with these distributions. We also discussed the distribution for the waiting "time" W in the setting of the Poisson distribution. Thus, if you expect to find μ per unit, then in time t you expect to find μt and therefore if you had to wait more than t, which means $W > t$, then in time t you actually saw none. Thus,

$$P(W > t) = poissonpdf(\mu t, 0).$$

Now as we discussed before, the Poisson distribution is actually simple to compute, and if the count X is governed by the Poisson distribution with mean μ, then

$$P(X = k) = \frac{e^{-\mu}(\mu)^k}{k!} = poissonpdf(\mu, k),$$

so putting $k = 0$ and replacing μ by μt we see that

$$P(W > t) = e^{\mu t}.$$

But this means we can actually easily calculate the cumulative distribution function for the waiting time W, since

$$F_W(t) = P(W \leq t) = 1 - P(W > t) = 1 - e^{-\mu t}.$$

The waiting time is an example of a continuous variable, since it must be measured, we cannot simply count. All of the three counting distributions however are examples of sampling. In the case of the binomial and hypergeometric, the sample size n is the number of trials.

We need to go on now to discuss sampling in general terms. After all, the main use of sampling is in trying to find characteristics or parameters of populations that are effectively infinite or much to large to possibly observe all the members of the population. If we want to know the average length of all the salmon in the Pacific Ocean, it is not practical or desirable to actually try to measure the length of every such salmon. We need to estimate with a a large sample, but as soon as we admit that we must settle for such an estimate, the question becomes what do we expect to get for the sample mean and how far off could it be from the population mean. In order to answer questions such as these, we must consider sampling in general somewhat theoretical terms-it is unavoidable.

To begin, suppose that X is a random variable, which means that X is the result of some observation that can be repeated, such as measuring the length of a salmon caught in the Pacific. Now to deal with sampling in general theoretical terms, we simply imagine that we want to try to guess the result of sampling before actually doing it. Thus we will have a sequence of observed values to guess

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

where n is the sample size. Of course, if we knew the true mean of X denoted

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

then you would certainly guess that the value of each observation will also be μ_X, which is to say that

$$E(X_k) = \mu_X, \ for\ all\ k,\ 1 \leq k \leq n.$$

For instance, if I said we are going to take a sample of one hundred salmon from the Pacific, and if you happen to know that the average length of all salmon in the Pacific is 31 inches, then you would certainly guess in advance that the length of the fifth fish in the sample is going to be 31 inches. After all, you certainly would not guess a number less than 31 and you certainly would not guess a number more than 31.

Now, the first step in calculating the sample mean is to total all the observed values, so that total in advance is also unknown to us, so we designate it as T_n, and this means

$$T_n = X_1 + X_2 + X_3 + ... + X_n.$$

Since we just saw that our guess for each of these unknowns in the sum is μ_X and since there are n such terms, it follows from the rules of expectation that

$$E(T_n) = n\mu_X.$$

For instance, if the salmon in the Pacific average 31 inches in length, and if I catch one hundred of them a lay them nose to tail on the dock and use a tape measure to measure the distance from the tail of the first salmon to the nose of the last, then I will expect in advance to find the length to be $(100)(31) = 3100$ inches.

The next step in calculating the sample mean is to take the sample total and divide by the number of observations which is n, the sample size. Therefore, if we designate the sample mean we will get as the unknown $\bar{X}_n$, which is of course also unknown to us before we actually take the sample and compute it, then

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

This means that by our rules of expectation

$$E(\bar{X}_n) = E(\frac{1}{n}T_n) = \frac{1}{n}E(T_n) = \frac{1}{n}n\mu_X = \mu_X.$$

That is we finally arrive at a fairly remarkable result:

$$E(\bar{X}_n) = \mu_X.$$

That means that whenever we take a sample, in advance of actually taking the measurements we are expecting to get the true population mean or true expected value μ_X as the result for the sample mean. Of course, we have already observed that in expectation theory, you do not always get what you expect, and in fact in many situations you almost never get what you expect. For instance, recall the dice in the box where X is the number up, so $\mu_X = 7/2$ which is not even possible to get for the number up. This means that when sampling, even though we always expect the sample mean to be the true mean, in fact it most likely will not be, and we need to know how far off it might be. That is, we need to know the standard deviations for T_n and $\bar{X}_n$. Up until now, we have not said anything about the actual method of sampling, but to calculate standard deviations, some assumptions must be made. A simple assumption to make is that all observations are uncorrelated. Recall that if U and W are any two uncorrelated unknowns, then

$$Var(U + W) = Var(U) + Var(W),$$

so we can apply this to calculate $Var(T_n)$ if all observations are uncorrelated with each other. In particular this is the case if all observations are independent of each other such as in repeated independent trials governed by the binomial distribution. Thus, assuming sampling with uncorrelated observations (SUO), we have

$$Var(T_n) = Var(X_1) + Var(X_2) + Var(X_3) + ... + Var(X_n).$$

Now the next thing we notice is that all the observations have the same distribution as X and therefore all have the same variance. For instance, if you know that the chance of a Pacific salmon being under 40 inches in length is 75 percent, and I ask what is the chance the fifth salmon I catch will have length under 40 inches, then that is obviously 75 percent. Any probability question about the length of the fifth salmon has the same answer as for the first salmon. This means all the observations X_k have the same distribution as X and therefore the same variance as X, namely σ_X^2. Therefore,

$$Var(T_n) = n\sigma_X^2, \ using \ SUO.$$

Taking square roots then gives us the standard deviation for T_n, namely

$$\sigma_{T_n} = [\sqrt{n}]\sigma_X, \ using \ SUO.$$

Thus, even though a sample total of one hundred salmon is expected to be one hundred times as long as the average salmon length, the standard deviation of the total is only ten times as big as the standard deviation in length for a single salmon. When we apply this to the sample mean, we find

$$\sigma_{\bar{X}_n} = \sigma_{(1/n)T_n} = \frac{1}{n}\sigma_{T_n} = \frac{1}{n}[\sqrt{n}]\sigma_X = \frac{\sigma_X}{\sqrt{n}},$$

or simply

$$\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}, \ using \ SUO.$$

This is a truly remarkable result and it explains exactly why large samples should produce sample means close to the true mean. For instance, we will see that there is always a large probability that observations fall within a few standard deviations of the true mean, say within

ten standard deviations of the true mean to be extremely conservative. If the standard deviation for the salmon is say 8 inches, and if we have a sample of size $n = 10000$, then $\sqrt{n} = 100$, so the standard deviation of the sample mean $\bar{X}$ is $8/100 = .08$, so ten standard deviations is now less than an inch. This means such a large sample is almost certain to give the true mean to within an inch. We will soon see that in fact it is very likely to be much more accurate. However, you can see from this that we can get as accurate as we like by taking samples sufficiently large and with as high a probability as we like of being within the stated accuracy. Exactly how likely we are to achieve a given level of accuracy with a given sample size will be the subject of most of the rest of this course.

23. **LECTURE** WEDNESDAY 3 MARCH 2010

Today we reviewed the basic facts about the sampling distribution, that is the the distribution for T_n and for $\bar{X}_n$ for sampling X with samples of size n. Remember, before we take a sample, the observations are unknowns $X_1, X_2, X_3, ..., X_n$ whose total T_n is called the sample total and whose average $\bar{X}_n$ is called the sample mean. We observed that X_k has the same distribution as X and therefore

$$E(X_k) = \mu_X, \ all \ k \leq n,$$

and

$$\sigma_{X_k} = \sigma_X, \ all \ k \leq n.$$

From these facts and the rules of expectation we find

$$E(T_n) = n\mu_X,$$

and

$$E(\bar{X}_n) = \mu_X.$$

Thus any time we take a sample, we always expect the sample mean to be the true mean μ_X, before we actually look at the observation results. Of course, we know we usually do not get what we expect. How far off the sample mean is from the true mean depends on the standard deviation of $\bar{X}_n$.

If all observations are uncorrelated, we showed last time that

$$Var(T_n) = n\sigma_X^2,$$

so

$$\sigma_{T_n} = (\sqrt{n})\sigma_X.$$

Since $\bar{X}_n = (1/n)T_n$, it follows that

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

In particular, if all the observations are independent of each other, then they are all uncorrelated, so these equations for standard deviations hold. When this is the case that all observations are independent of each other, we say that we are doing **INDEPENDENT RANDOM SAMPLING (IRS)**. This means that for large samples we should be very likely to get a sample mean near the true mean. Thus

$$\sigma_{T_n}(IRS) = (\sqrt{n})\sigma_X$$

and

$$\sigma_{\bar{X}_n}(IRS) = \frac{\sigma_X}{\sqrt{n}}.$$

To see how likely the sample mean is to be near the true mean for large samples we can begin with a simple inequality known as **Tchebeychev's Inequality**. For this inequality, we begin by asking how likely it is that the distance from X to μ_X is at least $k\sigma_X$, where k is some positive number that we will choose for convenience later. That is, what is

$$P(|X - \mu_X| \geq k\sigma_X) = ?$$

This at first certainly looks like a difficult problem, but we have the tools to solve it easily to our satisfaction. First, let A be the statement that $|X - \mu_X| \geq k\sigma_X$, so A is a factual statement which is either true or false,

$$A: \ |X - \mu_X| \geq k\sigma_X.$$

Thus, the probability we seek is simply $P(A)$. Next, since two non-negative numbers stand in a given size relation if and only if their squares are in that same size relation, we see A is the same as

$$A: \ (X - \mu_X)^2 \geq k^2 \cdot \sigma_X^2.$$

Next, we perform a little mathematical trick. Consider the general inequality:

$$(X - \mu_X)^2 \geq k^2 \cdot \sigma_X^2 \cdot I_A,$$

which is certainly either true or false, depending on the value of I_A, that is depending on whether A is true or false. Notice if A is true, then $I_A = 1$, and the inequality just restates A so the inequality is true if A is true. On the other hand, if A is false, then $I_A = 0$, so the left side of the inequality is zero, and as the right hand side is something squared, it is at least zero, so the inequality is actually true even when A is false. Thus the inequality is generally true, and we can therefore apply the expectation to both sides to get

$$E((X - \mu_X)^2) \geq k^2 \cdot \sigma_X^2 \cdot E(I_A).$$

But now, just recall that by definition

$$E(I_A) = P(A),$$

and also by definition,

$$E((X - \mu_X)^2) = Var(X) = \sigma_X^2.$$

When we substitute in these facts, we find simply

$$\sigma_X^2 \geq k^2 \cdot \sigma_X^2 \cdot P(A).$$

Now we can cancel the σ_X factors on both sides to find simply

$$1 \geq k^2 \cdot P(A),$$

or in other words,

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

If we use the fact that $P(not\ A) = 1 - P(A)$, here, we can also conclude that

$$P(not\ A) \geq i - \frac{1}{k^2}.$$

Replacing A by the actual statement it is standing for now gives two equivalent inequalities (either one is) called Tchebeychev's inequality:

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

and

$$P(|X - \mu_X| < k\sigma_X) \geq 1 - \frac{1}{k^2}.$$

For instance, if we take $k = 1$, we find no useful information, since

$$1 - \frac{1}{1^2} = 0.$$

If $k = 2$, then we get

$$1 - \frac{1}{2^2} = \frac{3}{4},$$

which says that we always have at least a seventy five percent chance of being within two standard deviations of the true mean. If we take $k = 3$, we find we always have at least 8/9 of a chance of being within three standard deviations of the true mean. If we take $k = 4$, we find we always have at least 96 percent chance of being within four standard deviations of the true mean. If we take $k = 10$, we find we always have at least a 99 percent chance of being within ten standard deviations of the true mean.

When we combine Tchebeychev's inequality with the results for the standard deviation of the sample mean $\bar{X}_n$, the results become very striking. For instance, if we have $\sigma_X = 10$, then for $n = 10000$, we have $\sqrt{n} = 100$, so

$$\sigma_{\bar{X}_n} = \frac{1}{10}.$$

Thus ten standard deviations for $\bar{X}$ here is only one unit. Therefore, by Tchebeychev's inequality, any sample of this size has a 99 percent chance of giving a sample mean within one unit of the true mean.

Obviously, with larger samples we can get as close as we want to the true mean with as much certainty as we want, but the sample sizes would not be practical. Fortunately for statisticians, there is a much more powerful theorem which comes to the rescue.

CENTRAL LIMIT THEOREM: As n tends to infinity, both T_n and $\bar{X}_n$ become normally distributed. In fact in practical terms, we will consider T_n and $\bar{X}_n$ to be normal whenever $n \geq 30$.

We did calculations which showed with the normal distribution built into the calculator we find easily that with a sample of only size 900 we are in fact 99.7 percent sure to have the sample mean be within one unit of the true mean.

24. **LECTURE** FRIDAY 5 MARCH 2010

Today we discussed the application of our sampling theory results to the counting distributions, the binomial, the hypergeometric, and the Poisson. We recalled that when sampling any X, with the sample total denoted T_n and the sample mean denoted $\bar{X}_n$, we have

$$E(T_n) = n \cdot \mu_X,$$

and

$$E(\bar{X}_n) = \mu_X,$$

but for standard deviations we must assume something about the sampling method. For Independent Random Sampling (IRS), that is under the assumption all observations are independent of each other, we have

$$\sigma_{T_n}(IRS) = \sqrt{n} \cdot \sigma_X,$$

and

$$\sigma_{\bar{X}_n}(IRS) = \frac{\sigma_X}{\sqrt{n}}.$$

Often, in sampling we do not use IRS, since that allows the possibility of measuring or observing the same member of the population twice. Thus, if we are sampling to determine the mean blood pressure of a population, we would probably not want to accidentally measure the same person's blood pressure more than once. We say that we have a **SIMPLE RANDOM SAMPLE(SRS)** if all sample of size n are equally likely to be the sample we actually choose with our sampling method. For instance, when dealing cards, say a five card hand, we think of the deal as being fair of all possible five card hands are equally likely to be the hand we get and as well for our opponents in the game. However, it is clear that a SRS is not an IRS. Obviously for instance, if you are dealt an ace on your first card, you know you are less likely to get another ace than you were to receive the first ace. It turns out that there is a simple correction factor which gives the variances and therefore the standard deviations when using SRS. Thus,

$$Var(T_n(SRS)) = \frac{N-n}{N-1} \cdot Var(T_n(IRS)) = \frac{N-n}{N-1} \cdot n \cdot Var(X),$$

$$\sigma_{T_n}(SRS) = \sqrt{\frac{N-n}{N-1}} \cdot \sigma_{T_n}(IRS) = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{n} \cdot \sigma_X,$$

$$Var(\bar{X}_n(SRS)) = \frac{N-n}{N-1} \cdot Var(\bar{X}_n(IRS)) = \frac{N-n}{N-1} \cdot \frac{Var(X)}{n},$$

$$\sigma_{\bar{X}_n}(SRS) = \sqrt{\frac{N-n}{N-1}} \cdot \sigma_{\bar{X}_n}(IRS) = \frac{N-n}{N-1} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

We can therefore simply say that when using SRS, to calculate the standard deviations for sampling, you must multiply the IRS results by the correction factor c_{SRS}, where

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

Lets put these results to work on the counting distributions. For instance when counting, we are dealing with the basic variable $X = I_A$, where A is the statement that you got a success (one trial). Thus, you could be counting red blocks when drawing blocks from a box with or without replacement. In this case, we have

$$\mu_X = E(X) = E(I_A) = P(A) = p,$$

where $p = P(A)$, the probability of success on a single trial is called the success rate. Thus, if you draw 20 blocks from a box containing 30 red blocks and 70 white blocks and you are counting the number of red blocks you get, every time you go to draw a block you have a 30 percent chance of success (getting a red block), whether or not you draw with or without replacement. Drawing with replacement gives IRS, whereas without replacement gives SRS. Either way, we see that T_n now gives the success count, so T_n is binomial if we use IRS and hypergeometric if we use SRS. In either case, the expected value of T_n is the same

$$\mu_{T_n} = n \cdot \mu_X = n \cdot p.$$

Thus, if we draw 20 blocks from the box of blocks, we expect to get 6 red blocks and whether we are using IRS (drawing with replacement) or SRS (drawing without replacement) is irrelevant to this fact. Notice that

$$P(T_n(IRS) = k) = binompdf(n, p, k) = (n\ [nCr]\ k)p^k(1-p)^{n-k},$$

so the mean could be calculated directly from the distribution by summing values times probabilities from 0 to n, but this would be a lot of work, and our sampling theory makes it obvious that the result is simply np. Moreover, the calculation would be just as much or more work using SRS, as

$$P(T_n(SRS)) = k) = \frac{(R\ nCr\ k)([N-R]\ nCr\ [n-k])}{(N\ nCr\ n)},$$

so all these probabilities for $k = 0$ to $k = n$ would have to be calculated, all values multiplied by their probabilities and added, and we see from our theory that the result is guaranteed to again be simply np. Of course, calculating variances and standard deviations directly from these distributions would be even harder, but our sampling theory now gives the results very easily as soon as we recall the standard deviation of $X = I_A$. Here, we have $X^2 = X$ as $X = I_A$ only takes values zero or one, so

$$Var(X) = E(X^2) - [E(X)]^2 = E(X) - [E(X)]^2 = p - p^2 = p(1-p),$$

which is success rate multiplied by failure rate for the variance of an indicator of an event. Therefore the standard deviation of $X = I_A$ is simply

$$\sigma_{I_A} = \sqrt{p(1-p)}.$$

As a result, we have

$$\sigma_{T_n}(BINOMIAL) = \sqrt{n}\sqrt{p(1-p)} = \sqrt{np(1-p)} = \sqrt{\mu_{T_n}(1-p)},$$

and

$$\sigma_{T_n}(HYPERGEOMETRIC) = \sqrt{\frac{N-n}{N-1}}\sqrt{np(1-p)} = \sqrt{\frac{N-n}{N-1}}\sqrt{\mu_{T_n}(1-p)}.$$

For the Poisson distribution, we can recall that it can be considered as a limit of the binomial distribution as the number of trials $n \to \infty$, in a controlled way. That is, if T is the success count here, and we are given the expected value $\mu = E(T)$ as the count expected in a unit size sample, then for very very large n, in a sample of size $1/n$ to expect to count $\mu_n = \mu/n$, so as μ/n becomes very much less than one, we can ignore the possibility of ever having a count above one. For instance, if we expect 6 buses per hour at our stop, then in one second we expect 1/600 of a bus, and as the chance of two buses in any one second period is essentially zero, we can think of either one or zero buses as the only possibility during any given second. Thus,

the count for one second is simply an indicator of whether a single bus comes or not. Thus, as n becomes very very large, the success count for a sample of size $1/n$ can be considered to be an indicator of whether a single success happens in that little sample, so if I_A denotes this indicator, then

$$T_{1/n} = I_A,$$

and

$$\frac{\mu}{n} = E(T_{1/n}) = E(I_A) = P(A) = p_n,$$

so we regard

$$p_n = \mu/n$$

now as the success rate, and as there are n of these disjoint small samples of size $1/n$ in the unit, and all are independent, we should have

$$poisson(\mu, k) = \lim_{n\to\infty} binompdf(n, \mu/n, k)$$

and that is just what we find experimentally with the calculator. In particular, as the success rate p_n tends to zero, the failure rate tends to one, so in the formula for the standard deviation of the binomial we find

$$\sigma_T = \lim_{n\to\infty} \sigma_{T_n}(BINOMIAL(n, \mu/n) = \sqrt{\mu_T}.$$

That is, the standard deviation of the Poisson distribution is always just the square root of the expected value.

25. **LECTURE** MONDAY 8 MARCH 2010

Today we reviewed for TEST 2 in class.

26. **LECTURE** WEDNESDAY 10 MARCH 2010

Today we had TEST 2 in class.

27. **LECTURE** FRIDAY 12 MARCH 2010

Today we discussed continuous distributions in general, as well as the uniform and normal distributions. For a continuous unknown recall that we picture its distribution as given by a curve of the form $y = f(x)$ whose graph is never below the horizontal axis and where the total area under the graph is one. Then, the probability $P(a < X < b)$ is the area under the graph of f between the limits $x = a$ and $x = b$. Notice that as the probability of X being exactly a or exactly b is zero, we have

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b).$$

Thus, we can approximate probabilities by simply looking at the picture of the distribution curve. Notice that where the curve is high we are more likely to find values than where the curve is low, for a given range length. That is, if the curve is lower from 8 to 18 than from 20 to 28, then the probability of X being between 8 and 18 is less than the probability of X being between 20 and 28, even though we are considering the same length of a range of values, namely ten units.

In general, when we look at the distribution of the continuous unknown X, the mean $\mu - E(X)$ is the **Balance Point** of the distribution, which can be shown using the laws of physics. That is if we assume the region under the distribution curve is made of uniform metal of uniform thickness, then where that balances is the true mean μ.

The fact that $P(X = c) = 0$ for every value of c may take a little getting used to at first. To see that it must be true, notice that we can make a narrow rectangle of height h and width d where h is so high that the region under the graph of f from $x = c - (d/2)$ to $x = c + (d/2)$ is entirely contained in the rectangle and therefore,

$$P(X = c) \leq P(c - [d/2] < X < c + [d/2]) \leq h \cdot d.$$

But if you think that $P(X = c) = p > 0$, then I can choose d so small that $hd < p$ and conclude that $P(X = c) < p$, which would be a contradiction. This at first appears to be a paradox, since when we measure a continuous variable, we always get some value. The resolution to this conundrum is that in any practical setting, observing a continuous unknown always involves a measurement. We discussed the fact that if X is a continuous unknown, then in practice that means a measuring device is involved in observing X and that means a level of measurement accuracy must always be considered. As already noted, for continuous X

$$P(X = c) = 0,$$

when c is any definite real number. On the other hand, if we measure X to two decimal place accuracy, then the statement that X is observed to have value 7.34 really means that X is somewhere between 7.335 and 7.345, which is a range of values, and the area under the distribution curve for X between limits $x = 7.335$ and $x = 7.345$ can very well have a positive value.

In general, if X is an unknown, we can consider $R_n(X)$ as the result of rounding off X to n decimal place accuracy. For instance, we can model counting unknowns with continuous

unknowns, by simply rounding off to the nearest whole number. That is, if X is a continuous unknown, and if $X \geq -(.5)$, then when we round off X to the nearest whole number to get $R_0(X)$, then we can think of the value as the result of a count.

The simplest continuous distribution is the **UNIFORM DISTRIBUTION**. Here, we have an unknown, and the only thing we know is the minimum value Min and the maximum value Max. Since that is all we know, there is no basis for thinking any range of values more likely than another if the lengths of the ranges are the same. Thus, all points of the distribution curve must have the same height, h. Since the total area under the curve is one, it follows that

$$h \cdot [Max - Min] = 1,$$

and therefore

$$h = \frac{1}{Max - Min}.$$

This means that the distribution curve is simply a horizontal line segment extending from the point (Min, h) to the point (Max, h) in the coordinate plane. Then, the probability of a range is simply

$$P(a < X < b) = \frac{b - a}{Max - Min}.$$

Also, the balance point is obviously the average of the minimum and maximum values, so

$$\mu_X = \frac{Min + Max}{2},\ X\ uniform.$$

Because of the Central Limit Theorem, the normal distribution is one of the most important distributions in applications. For the normal distribution you only need the mean μ and standard deviation σ. In fact, whenever you only know the mean and standard deviation of X, then the distribution must be normal purely from the standpoint of your information.

We reviewed the Central Limit Theorem and used the calculator to show that the binomial distribution for $n = 40$ and $p = .45$ is very well approximated by the normal distribution with

$$\mu = np$$

and standard deviation

$$\sigma = \sqrt{np(1-p)}.$$

28. **LECTURE** MONDAY 15 MARCH 2010

Today we reviewed continuous distributions and particularly the normal distribution. We began by observing that if X is any unknown and if R_nX is the unknown whose value is the result of rounding off X to n decimal places, then RnX is a discrete unknown and in fact $Y_n = (10^n)R_nX$ has only integer values. Thus, if $X \geq 0$, then Y_n is like a counting unknown. We noted that we can form the lower round L_nX and the upper round U_nX, where for L_nX we simply replace all decimal places after the n^{th} with zeroes, and for U_nX, we raise the n^{th} decimal place by one and replace all decimal places after the n^{th} with zeroes. Obviously, both

$$L_nX \leq X \leq U_nX,$$

and

$$L_nX \leq R_nX \leq U_nX.$$

It follows that

$$|R_nX - X| \leq U_nX - L_nX.$$

We can apply the expectation to these first two inequalities and find as well that

$$E(L_nX) \leq E(X) \leq E(U_nX)$$

and

$$E(L_nX) \leq E(R_nX) \leq E(U_nX),$$

so it likewise follows that

$$|E(R_nX) - E(X)| \leq E(U_nX - L_nX).$$

On the other hand, it is also obvious that

$$U_nX - L_nX = \frac{1}{10^n}.$$

Therefore, we have both

$$|R_nX - X| \leq \frac{1}{10^n}$$

and

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

The significance of these inequalities is that if we are working to n decimal place accuracy with our measurements and using the results to compute expectations, then our resulting expectation calculations will also have n decimal place accuracy.

We observed that we can picture the result of the calculation of area under a distribution curve as likewise approximately the result of using the distribution for the discrete unknown R_nX. The distribution of R_nX is best pictured as a series of spikes, and the corresponding area under the distribution curve for X is formed by a large number of approximating rectangles. We can see this more clearly. Suppose f_X is the probability density function for X, and x is an n decimal place number. We can consider the interval whose left edge is

$$x_- = x - \frac{1}{2(10^n)}$$

and whose right edge is

$$x_+ = x + \frac{1}{2(10^n)}.$$

The area $A(x)$ under f_X between these two limits is

$$A(x) = P(x_- \leq X \leq x_+).$$

But, then also

$$P(R_nX = x) = A(x),$$

by definition of R_nX. In the distribution picture for R_nX, we think of the spike on x as representing the probability that R_nX has value x, so it has height $p_x = A(x)$. Now notice

$$\Delta x = x_+ - x_- = \frac{1}{10^n},$$

it follows that

$$A(x) = (10^n) \cdot p_x \cdot \frac{1}{10^n} = (10^n) \cdot p_x \cdot \Delta x.$$

This is the area of a rectangle whose base has length $1/(10^n)$ and whose height is $(10^n)p_x$. On the other hand, as $A(x)$ is the area under the graph of f_X between limits $x = x_-$ and $x = x_+$, and since this region under the curve is very thin (for large n), it follows that to very good approximation

$$A(x) =^{approx}= f_X(x) \cdot \frac{1}{10^n} = f_X(x) \cdot \Delta x.$$

Therefore we approximately have

$$f_X(x) \cdot \Delta x = A(x) = (10^n) \cdot p_x \cdot \Delta x,$$

and therefore it must be that

$$f_X(x) =^{approx}= (10^n) \cdot p_x,$$

or equivalently,

$$p_x =^{approx}= \frac{f_X(x)}{10^n} = f_X(x) \cdot \frac{1}{10^n} = f_X(x) \cdot \Delta x,$$

gives very approximately the formula for the spike heights for the distribution of R_nX. Alternately, we can say for large n,

$$p_x = P(R_nX = x) =^{approx}= f_X(x) \cdot \Delta x.$$

This means that if $a < b$ are two n decimal place numbers, then

$$P(a \leq R_nX \leq b) = \sum_{a \leq x \leq b} p_x =^{approx}= \sum_{a \leq x \leq b} f_X(x) \cdot \Delta x.$$

For instance, since the expected value of R_nX is the sum of values multiplied by their probabilities,

$$E(R_nX) = \sum x \cdot p_x =^{approx}= \sum x \cdot f_X(x) \cdot \Delta x.$$

If g is any real valued function on the real line which is reasonable enough that it can be graphed with paper and pencil, then $g(X)$ is the unknown whose value is $g(x)$ if X has value x. Then

$$E(g(R_nX)) = \sum g(x) \cdot p_x =^{approx}= \sum g(x) \cdot f_X(x) \cdot \Delta x.$$

In the limit as $n \to \infty$, the term on the right "converges" to what is called the Riemann integral of $g \cdot f_X$ which is denoted

$$\int g(x)f_X(x)dx = \lim_{n\to\infty} \sum g(x)f_X(x) \cdot \Delta x,$$

whereas the term on the left converges to

$$E(g(X)) = \lim_{n\to\infty} E(g(R_n X)).$$

We can therefore say that

$$E(g(X)) = \int g(x)f_X(x)dx.$$

For instance, if B is a set of real numbers and if I_B, denotes the indicator of the statement that X is in B, assuming B is reasonably well behaved as a subset of the real line, then with $g = I_B$, the integral as well as the sums only use terms for which x is in B and as well, for those values of x we have $I_B(x) = 1$. For instance, if B is the set of all numbers between a and b, then

$$\int I_B(x)f_X(x)dx = \lim_{n\to\infty} \sum_{a\leq x\leq b} f_X(x)\Delta x = P(X \ in \ B).$$

It is customary to denote

$$\int_B g(x)dx = \int I_B(x) \cdot g(x)dx,$$

for any function g, so we have

$$P(X \ in \ B) = \int_B f_X(x)dx.$$

This is the precise mathematical statement that probability is given by area under the density curve.

29. **LECTURE** WEDNESDAY 17 MARCH 2010

Today we discussed the normal distribution and how to calculate a **CONFIDENCE INTERVAL** and its **MARGIN OF ERROR**. We denote the margin of error in a confidence interval as ME. Thus the purpose of a confidence interval is to give an estimate for the true mean μ_X, for the unknown X based on sample data. If $\bar{x}$ is the mean of our sample, then obviously $\bar{x}$ is the best guess for μ_X, based only on that sample data, however, there may be error, and when we are required to have a certain confidence C in our statement including the margin of error, then we will generally state that

$$\mu_X = \bar{x} \pm ME \; with \; confidence \; C$$

to mean that

$$P(\bar{x} - ME \leq \mu_X \leq \bar{x} + ME) = C.$$

Here, we call C the **LEVEL OF CONFIDENCE**. We noted that when $\bar{X}_n$ is normally distributed we can always express

$$ME = z_C \cdot \frac{\sigma_X}{\sqrt{n}}.$$

Let Z denote the standard normal random variable, so

$$\mu_Z = 0$$

and

$$\sigma_Z = 1.$$

The number z_C is chosen so that

$$P(-z_C \leq Z \leq z_C) = C.$$

Notice that $-z_C \leq Z \leq z_C$ is the "middle region" which leaves out two tails of total area $1 - C$ so by symmetry, each tail has area $(1 - C)/2$. Therefore, the area to the left of z_C is A_C where

$$A_C = C + \frac{1 - C}{2} = \frac{2C + 1 - C}{2} = \frac{1 + C}{2},$$

which of course is simply the average of C and one. But, using the inverse normal in the calculator, we have

$$z_C = invNorm((1 + C)/2, 0, 1).$$

We worked examples of calculating confidence intervals and their margins of error using the calculator. We also noted that usually we are given the sample data and the level of confidence, and then we compute the margin of error, but we also need the standard deviation of X to use this method. In the calculator go to the test menu and go down to the "zInterval" and follow the dialogue. If you do not have σ_X, then you must use s, the sample standard deviation in its place, so you cannot use the "zInterval" but instead must use the "tInterval".

We also noted that if you know σ_X, and if you are given C and an allowed margin of error ME, then solving the margin of error formula for n we find

$$n = (\frac{z_C \cdot \sigma_X}{ME})^2,$$

where we always must round up to get our whole number value for the sample size. Thus, this formula tells us how big the sample will have to be in order to achieve a desired level of accuracy with our desired level of confidence.

30. **LECTURE** FRIDAY 19 MARCH 2010

Today we had a review of confidence intervals and the normal distribution.

31. **LECTURE** MONDAY 22 MARCH 2010

Today we discussed the $t-$distribution and its use for calculating margins of error for confidence intervals for means of normal populations using independent random sampling (IRS) in the case where the population standard deviation is unknown.

The $t-$ distribution is really a whole family of distributions parametrized by positive whole numbers called the number of **DEGREES OF FREEDOM**, which we will denote by df. When you have a single sample of size n, the number of degrees of freedom is simply

$$df = n - 1.$$

The $t-$ distribution has the shape roughly of a standard normal bell curve but is not quite as tightly centered about zero. As the df increases the distribution is more and more tightly centered about zero until in the limit as $df \to \infty$ the $t-$distribution becomes the standard normal or $z-$distribution. Because of this we can think of the standard normal as being the $t-$distribution for an infinite number of degrees of freedom.

The use of the $t-$distribution comes from the fact that when using observed values of $\bar{X}$ to estimate μ_X, if we know σ_X, then the margin of error for confidence level C is

$$ME = z_C \cdot \frac{\sigma_X}{\sqrt{n}},$$

so if σ_X is not known to us, then we need to use the sample data to estimate σ_X. Of course, the sample standard deviation is designed precisely to do that. In fact, as an unknown, if $X_1, X_2, X_3, ..., X_n$ are the sample observations to be made (so we do not know yet what they will turn out to be), then

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

where T_n is the sample total

$$T_n = X_1 + X_2 + X_3 + ... + X_n,$$

so we have

$$T_n = n \cdot \bar{X}_n.$$

The sample deviations from the sample mean are the n numbers

$$X_1 - \bar{X}_n,\ X_2 - \bar{X}_n,\ ...,\ X_n - \bar{X}_n,$$

and notice that if we add them all up we have

$$T_n - n \cdot \bar{X}_n = 0.$$

This means we have n deviations but they are always related by an equation and this means that the freedom of the deviations to vary themselves is cut from n dimensions down to $n-1$ dimensions. This is also the reason for the fact that when calculating s^2, the sample standard deviation, we divide the sum of squared deviations by $n-1$ instead of n. More specifically, it is because of this that we can show using our rules of expectation and variance that

$$E[(X_1-\bar{X})^2+(X_2-\bar{X})^2+...+(X_n-\bar{X})^2]=(n-1)\cdot\sigma_X^2.$$

From this it follows that the sample variance is expected to turn out to be the true population variance, but of course, as usual, you rarely get what you expect. We can define the new unknown

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

so that the value of S^2 is the observed sample variance, so S_n^2 is the sample variance unknown and S_n is the sample standard deviation unknown. Then

$$E(S_n^2)=\sigma_X^2.$$

Recall that when we calculate the margin of error in a confidence interval when σ_X is known we have the formula for the margin of error, denoted ME, which is simply

$$ME=z_C\cdot\frac{\sigma_X}{\sqrt{n}}.$$

Here, z_C is the point on the standard normal Z for which

$$P(|Z|\leq z_C)=C.$$

Keep in mind that

$$|Z|\leq z_C$$

is equivalent to

$$-z_C\leq Z\leq z_C.$$

To have error no more than $x>0$ with probability C is to say that

$$P(|X-\mu_X|\leq x)=C$$

but the inequality

$$|X-\mu_X|\leq x$$

is equivalent to

$$\frac{|X-\mu_X|}{\sigma_X/\sqrt{n}}\leq\frac{x}{\sigma_X/\sqrt{n}}.$$

But,

$$\frac{|X-\mu_X|}{\sigma_X/\sqrt{n}}=Z_{\bar{X}_n}$$

is the standardization of $\bar{X}_n$, so that means the last inequality is the same as saying

$$|Z|\leq\frac{x}{\sigma_X/\sqrt{n}}.$$

When we compare this with

$$|Z|\leq z_C,$$

we see they have the same probability which is C precisely with

$$z_C = \frac{x}{\sigma_X/\sqrt{n}},$$

and therefore

$$ME = x = z_C \cdot \frac{\sigma_X}{\sqrt{n}}.$$

Thus our margin of error formula is really coming from the standardization formula

$$Z = Z_{\bar{X}_n} = \frac{\bar{X}_n - \mu_X}{\sqrt{n}}.$$

If we replace σ_X by an observed value of S_n in our calculations, then we are using a new unknown we call t instead of Z given by

$$t = \frac{\bar{X}_n - \mu_X}{S_n/\sqrt{n}}.$$

It can be shown using the rules of expectation, that if X is normal, then assuming IRS, the unknown t has the $t-$distribution for $n-1$ degrees of freedom here. Therefore, we replace z_C by t_C calculated with the $t-$distribution instead of the $Z-$distribution or standard normal distribution.

We worked examples with the calculator and saw how to compute the margin of error using the tInterval in the calculator's TEST MENU.

32. **LECTURE** WEDNESDAY 24 MARCH 2010

We reviewed confidence intervals for means and proportions.

33. **LECTURE** FRIDAY 26 MARCH 2010

Answered questions about confidence intervals.

34. **LECTURE** MONDAY 29 MARCH 2010

NO CLASS-SPRING BREAK

35. **LECTURE** WEDNESDAY 31 MARCH 2010

NO CLASS-SPRING BREAK

36. **LECTURE** FRIDAY 2 APRIL 2010

NO CLASS-SPRING BREAK

37. **LECTURE** MONDAY 5 APRIL 2010

NO CLASS-SPRING BREAK

38. **LECTURE** WEDNESDAY 7 APRIL 2010

39. **LECTURE** FRIDAY 9 APRIL 2010

40. **LECTURE** MONDAY 12 APRIL 2010

41. **LECTURE** WEDNESDAY 14 APRIL 2010

42. **LECTURE** FRIDAY 16 APRIL 2010

43. **LECTURE** MONDAY 19 APRIL 2010

44. **LECTURE** WEDNESDAY 21 APRIL 2010

45. **LECTURE** FRIDAY 23 APRIL 2010

46. **LECTURE** MONDAY 26 APRIL 2010

47. **LECTURE** WEDNESDAY 28 APRIL 2010

48. **LECTURE** FRIDAY 30 APRIL 2010

49. **LECTURE** MONDAY 24 AUGUST 2009

We discussed the general rules for guessing unknown quantities so as to maintain logical consistency. We use capital letters to denote unknown quantities and statements of unknown truth value. In a given situation, we generally have some background information to start with, which we denote by K. If X is an unknown quantity, then $E(X|K)$ is the notation we use to designate our guess for the numerical value of X given that we assume the statement K is true. In a situation where K is well understood, we may drop it from the notation and write simply $E(X)$ for short to designate $E(X|K)$, but we should keep in mind that there is always a background information statement we are using to make our guess. Another notation which we will some times use is the Greek letter μ which we tag with subscript X if necessary. Thus for notation,

$$E(X|K) = E(X) = \mu_X = \mu, \tag{49.1}$$

all indicate the same thing, namely our guess, with various symbols included in the notation when necessary to avoid confusion. This will become clearer as you begin to use the notation in problems.

We assume that our unknowns such as X are described in a way which makes it clear that there is a value for the unknown, but we may have incomplete information about what that value is. For instance, as a beginning example, suppose that we have a box sitting on the table and inside, where we cannot see, is a single standard dice as used in the game of craps at the casino. We could use X to denote the number (of spots) on the top face of the dice in the box. Our background information K states that there is a definite face on top and it is in the box where we cannot see inside. We know that there are six possibilities for the number on top, but how should we choose a number for our guess. When we only consider a single such problem, there does not seem any clear way to proceed. It is when we begin to consider several problems and their relationships that we begin to realize that there should be some logical constraints on how to guess in order to maintain logical consistency. We will use capital letters to denote unknowns and statements, and lower case letters to denote numbers which we actually know. For instance, the most obvious rule should be that if our information happens to tell us the value of X, then that is the value we should guess. For instance if K says the dice is in the box and the face with two spots is on top, then $E(X|K) = 2$ is the only thing that makes sense. In this case, we observe that K implies the statement $X = 2$ and so if we base our guess on K, then it only makes sense to guess $E(X|K) = 2$, that is to say, it only makes sense to guess that 2 is the value of X given we assume K to be true. More generally, if c is any definite number and if K implies that $X = c$, so K tells us the value of X is c, then we should guess that $X = c$ if we are basing our guess on K. This gives our first rule.

NORMALIZATION RULE: If K implies that $X = c$, then

$$E(X|K) = c. \tag{49.2}$$

More generally, instead of telling us the exact value of X our information K might only tell us an inequality restricting possibilities for the value of X. For instance, in the dice example, our background information telling us that the dice in the box is a standard dice as used in the casino actually implies that $1 \leq X \leq 6$. Thus, it certainly would not make sense to guess 8 is the value of X in this example. In fact we should also have $1 \leq E(X|K) \leq 6$ in the dice example. More generally, when dealing with any unknown, if a and b are definite numbers and our statement K implies that $a \leq X \leq b$, then we should definitely restrict our guess to be a number between a and b. To be precise, we will always assume the next rule is enforced.

POSITIVITY RULE: If K implies that $a \leq X \leq b$, then

$$a \leq E(X|K) \leq b. \tag{49.3}$$

In general, an unknown numerical quantity has only a numerical value as we will restrict the units to be part of the description. For instance, suppose that K is the information that outside there is a fish in an ice chest and X is the weight of the fish in pounds, then the value of X is simply a number. This means that we can add unknowns. For instance if Y is the height in feet of a specific tree outside which we can see off in the distance, then $X + Y$ is defined to be the result of adding the weight of the fish in pounds to the height of the tree in feet. You may protest that it makes no sense to add those two numbers together, but there are many cases where it does make sense, and it is simplest not to have to worry about the units as they are built into the unknowns. If you have guessed the weight of the fish to be 30 pounds and the height of the tree to be 60 feet, then it only makes sense to guess that the sum of the two numbers is 90. Now suppose that we have two boxes on the table in front of us and in each box is a bank book for a savings account, but we cannot see the balance on either bank book. Suppose that we know the owner of each bank book and have some information about what the balance of each might be. Suppose that X is the value of the bank book on the left and Y is the value of the bank book on the right, both in US dollars. If K is the statement of what I know about the owners of the bank books and the information describing the physical setup here, and if I have already guessed that the bank book in the box on the left is in dollars worth 3000 and if I have already decided to guess the one on the right in dollars is worth 4000, then it only makes sense that I should guess 7000 for the value of $X + Y$. That is, in any situation where unknowns are added to form new unknowns, if I can guess each summand, then I just add my guesses up to get my guess for the value of the sum of the unknowns. This is our next rule which we will assume to be always true.

ADDITIVITY: If X and Y are any unknowns, then $X + Y$ denotes the unknown whose value is the sum of the individual numerical values, and with any background information K we have

$$E(X + Y|K) = E(X|K) + E(Y|K). \tag{49.4}$$

Suppose that the box on the table contains a gold nugget which we cannot see. It might be very small or it might fill up the whole box. Let X be the weight in ounces of the nugget. Let Y be the value of the nugget in dollars. Suppose that our background information tells us that gold is worth 800 dollars an ounce. If we have guessed that the weight of the nugget is 3 ounces, that means we have determined $E(X|K) = 3$, then we should guess the value of the nugget in dollars to be 2400. Here K implies we have $Y = 800X$ is true, and thus $E(Y|K) = 2400$, or $E(800X|K) = 800E(X|K)$. This gives us our final rule for the day.

HOMOGENEITY: If K implies that $Y = cX$, then

$$E(Y|K) = E(cX|K) = cE(X|K). \tag{49.5}$$

To summarize, we have our four basic rules for guessing in order to maintain logical consistency:

If K implies $X = c$, then $E(X|K) = c$.
If K implies that $a \leq X \leq b$, then $a \leq E(X|K) \leq b$.
$E(X + Y|K) = E(X|K) + E(Y|K)$.
If K implies that $Y = cX$, then $E(Y|K) = E(cX|K) = cE(X|K)$.

50. **LECTURE** WEDNESDAY 26 AUGUST 2009

We began by reviewing the four basic rules of guessing.

NORMALIZATION RULE: If K implies that $X = c$, then

$$E(X|K) = c. \tag{50.1}$$

POSITIVITY RULE: If K implies that $a \leq X \leq b$, then [50.1]

$$a \leq E(X|K) \leq b. \tag{50.2}$$

ADDITIVITY: If X and Y are any unknowns, then $X + Y$ denotes the unknown whose value is the sum of the individual numerical values, and with any background information K we have [49.4]

$$E(X + Y|K) = E(X|K) + E(Y|K). \tag{50.3}$$

HOMOGENEITY: If K implies that $Y = cX$, then [49.5]

$$E(Y|K) = E(cX|K) = cE(X|K). \tag{50.4}$$

We can use the guessing procedure on statements to evaluate how likely a statement is to be true. The only type statements we consider are statements which are either true or false. We do not deal with statements such as "Mozart's music is better than Bach's", in other words, the statements we deal with are factual statements which are clearly either true or false. The truth value of a factual statement we deal with may not be known to us from our background information statement K. But, based on K we want to guess how likely a new statement is to be true. The result is called probability. Suppose that K is our background information statement and that N is a new statement. Suppose that K tells us something about N but does not tell us the truth value of N. We use N to define a very special unknown called the *Indicator* of N denoted by I_N, with the provision that I_N can only have value 0 or 1 according to whether N is false or true. That is if we know N is true, then we know that $I_N = 1$. If we know that N is false, then we know that $I_N = 0$. That is, knowing the value of I_N is the same as knowing whether N is true or false, the truth value of N. Now our rules for guessing do not tell us how to proceed here if we do not know whether N is true or false. But, since I_N is an unknown, we will go ahead and define what we will call the *Probability of N given K*, denoted $P(N|K)$, by the following formula.

DEFINITION OF PROBABILITY

$$P(N|K) = E(I_N|K). \tag{50.5}$$

Just as with the $E(X|K)$ notation, we drop the $|K$ from the notation if no confusion results. That is, if we are calculating several probabilities all with the same given information K, then we would simply write $P(N)$ for $P(N|K)$. In other words, when we understand we are basing our calculations on K, we often find it simpler to write $P(N)$ and just keep in mind that actually $P(N) = P(N|K)$. We can be sure that $0 \leq I_N \leq 1$, since I_N can only be either 0 or 1, and therefore by the Positivity Rule, [50.1], we know that

$$0 \leq E(I_N|K) \leq 1, \tag{50.6}$$

and therefore,

$$0 \leq P(N|K) \leq 1. \tag{50.7}$$

If K implies that N is true, then this is the same as saying K implies that $I_N = 1$, and therefore by the Normalization Rule, [50.1], we must have $P(N|K) = E(I_N|K) = 1$. If K implies that N is false, then this is the same as saying K implies $I_N = 0$, and again using the Normalization Rule, in this case we find $P(N|K) = E(I_N|K) = 0$. Thus, if K tells us N is true, then $P(N|K) = 1$, whereas if K tells us N is false, then $P(N|K) = 0$. By [50.7], we might say that if K does not tell us whether N is true or false, then $P(N|K)$ should be a number somewhere strictly between 0 and 1, and we can begin by thinking that the closer the probability is to 1, the more likely N is to be true as judged with the background information K.

Using logic we can combine statements using the logical connectives "$\&, or, not$". Thus, $notN$ is the negation of statement N, so $notN$ is true exactly if N is false. It is easy to see here that in terms of indicators we can write

$$I_{notN} = 1 - I_N. \tag{50.8}$$

It then follows immediately from the Additivity and Homogeneity Rules that

$$P(notN|K) = 1 - P(N|K). \tag{50.9}$$

Thus, when the weatherman says there is 30% chance of rain, that is the same as saying there is a 70% chance it will not rain.

In case we have two statements, say statement A and statement B, then we can form the statement $A\&B$ which to be true requires that both of these individual statements be true. We then easily check that

$$I_{A\&B} = I_A I_B, \tag{50.10}$$

so to get the indicator of $A\&B$ we simply multiply their individual indicator unknowns together.

Since "&" goes with multiplication, we might guess that "or" goes with addition, so we might be tempted to guess that I_{AorB} is the same as $I_A + I_B$. Here we have to keep in mind that in logic, "or" does not mean the exclusive "or" of everyday talk. For $AorB$ to be a true statement, it only need be the case that at least one of these statements is true, but that allows the possibility that they are both true. If both are true, then the value of $I_A + I_B$ would be 2 and that is not allowed for an indicator. We need to subtract 1 exactly in the case they are both true and subtract zero otherwise, that is we need to subtract $I_{A\&B}$. The result you can easily check is that

$$I_{AorB} = I_A + I_B - I_{A\&B}. \tag{50.11}$$

It now follows immediately from our Addition and Homogeneity Rules that

$$P(AorB|K) = P(A|K) + P(B|K) - P(A\&B). \tag{50.12}$$

We use S to denote a statement which is true for sure such as "1=1," and we use Φ to denote a statement which is false for sure such as $1 \neq 1$. Notice that $I_S = 1$ and $I_\Phi = 0$. We then must have $P(S|K) = P(Sure|K) = 1$ and $P(\Phi|K) = 0$.

The fundamental rules of probability are simply [50.7], [50.12], and $P(Sure|K) = 1$.

In many situations we have some finite number of statements of which exactly one is true and all the others are false, but K does not tell us which of these statements is the one that is true. In this case their indicators must add up to 1 and hence their probabilities must add up to 1. For instance, if we have statements A, B, C and exactly one is true and the other two are false, then $I_A + I_B + I_C = 1$, so when the Additive Rule and the definition of probability is applied, we find that $P(A|K) + P(B|K) + P(C|K) = 1$. In any situation where our information K does

not tell us any of these three statements is more likely true than another, we must accept all three probabilities are the same, and as they add up to 1 we see each of these statements has probability 1/3. The same would apply if there were 6 different statements and K does not allow us to conclude any one more likely than another, then each of the 6 statements must have probability 1/6 given K. For instance, for the case of the dice in the box where we cannot see it, if that is the extent of our information, then we conclude that all faces are equally likely to be the one on top so each has probability 1/6 of being the one on top. We call this the Principle of Indifference. In general, if there are n statements and K tells us exactly one is true but gives no information allowing us to judge any being more likely than the others, we conclude they all have the same probability, $1/n$. We generally refer to this as the *Model of Equally Likely Outcomes*. In gambling situations, we generally say a game is *FAIR* when the model of equally likely outcomes is in effect. Thus we speak of a fair pair of dice or a fair roulette wheel or a fair lottery. For instance, if a box contains 3 red blocks and 2 blue blocks, and one block is removed and we do not see which one has been removed, then with K the statement of these facts, if R is the statement that the removed block is red, then $P(R|K) = 3/5$, since each block has probability 1/5 of being the block that was removed, and three of these are red. Try making up symbols for the statements that each of the 5 blocks is the one removed on the first draw, and then assuming each has probability 1/5 demonstrate that $P(R|K) = 3/5$.

Returning to the equation $1 = I_A + I_B + I_C$ when K tells us exactly one of the statements A, B, C is true, if we multiply through each side by X, we arrive at the equation

$$X = XI_A + XI_B + XI_C,$$

and our Addition Rule then says

$$E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K).$$

This means that if we can figure out how to deal with the computation of $E(XI_N|K)$ when N is some new information, then it can be applied to each term of the preceding equation to calculate the value $E(X|K)$. The problem of how to compute $E(XI_N|K)$ leads to a new rule called the Multiplication Rule which is our final fundamental rule. As this rule is more complicated than the four basic rules, we will deal with this in the next lecture. But in a sense it is the most important rule because it allows us to determine $E(X|K)$ in many situations. That is finally, our guess will be completely determined by our rules, so in a sense, it is not really just guessing.

51. **LECTURE** FRIDAY 28 AUGUST 2009

We have previously discussed four basic rules for guessing and defined the notation $E(X|K)$ for our guess of the value of the unknown X based on the information in the statement K. The technical term mathematicians and statisticians use here is *Expectation*. Thus we refer to $E(X|K)$ as the *Expected Value* of X given K. We saw that the four basic properties of $E(X|K)$ are dictated by the requirement that guessing should be at least logically consistent and consistent with addition of numbers. We also previously used these rules to determine the rules of probability. But there is a final fundamental rule which we call the *Multiplication Rule* which is more difficult than the four basic rules, and which is necessary for the determination of expectation. The multiplication rule will in fact allow us to determine all expected values from probabilities and those in turn can often be determined by the model of equally likely outcomes. To get an idea of what is needed, recall that when we worked out the rules of probability from the rules of expectation, we also pointed out that in many problems we are presented with the situation of having some finite number of statements and K tells us exactly one of them is true but does not tell us which one is true. In this case, recall, we know that their indicators must add up to 1 and hence their probabilities do also. For instance, if there are three statements A, B, C of which according to K exactly one is true but K does not say which of the three is the one that is true, then we know

$$1 = I_A + I_B + I_C, \tag{51.1}$$

and thus we have $1 = P(A|K) + P(B|K) + P(C|K)$. But we can multiply both sides of [51.1] by X and now arrive at the equation

$$X = XI_A + XI_B + XI_C. \tag{51.2}$$

We then find by the Addition Rule that

$$E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K). \tag{51.3}$$

Notice that we could apply this same method even if there were thousands of these statements instead of only three. We can use computers to do the addition. But we still need to know each term in the sum. This is the general problem. If N is some new statement, how do we determine $E(XI_N|K)$??? Well, remember that I_N is 0 if N is false and 1 if N is true, and this means that if N is false then XI_N has the value 0 but if N is true then XI_N simply has the value of X itself. To determine $E(XI_N|K)$, we there for have to modify our guess for X based on K to include two things: (1) the way to modify our guess due to the fact that the possible values of X may be different if N is assumed to be true and (2) the way to modify our guess due to the fact that K may not tell us the actual truth value of N, that is whether N is true or false. We can see that if N is true, then we should begin by figuring out $E(X|N\&K)$ in order to deal with (1). As far as (2) is concerned, the best that K can do is to tell us $P(N|K)$. We therefore have two numbers to begin with here, first the expected value of X given that both N and K are actually true and second the probability of N given that K is true. The first is an expected value and the second is a probability. We are looking for a way to combine these two numbers to arrive at $E(XI_N|K)$, and which will always remain consistent with the four basic rules. We will begin by assuming that there is some general rule here which is consistent with the four basic rules. Suppose we imagine that there is such a general rule which is known to an oracle, say the Oracle at Delphi, the voice of the God Apollo. The oracle knows the rule and today is the day it is dealing with questioners who have questions about application of this rule to their problems of guessing unknowns. In order to save time, since his calculation only depends on the expected value and the probability, the oracle asks that each questioner not bother him with the details of his specific unknown, but rather simply tell the oracle the two

numbers, first the expected value and second the probability for his problem and present his required offering of gold and then the oracle will announce the value of the result which is the expected value of the unknown multiplied by the indicator of the new information statement. Imagine you are in a long line before the oracle and you hear the person behind you talking to the person behind him and you realize that you both have the same expected value to report to the oracle, namely 8. This sparks your interest to listen further and you realize his unknown is entirely different from yours but miraculously, his probability is the same number as yours, namely .3, even though the new statement he is dealing with and his background statement are both entirely different from yours. Notice that if we were dealing with the problem of finding the expected value of a sum of unknowns, the process could work the same way and you, knowing the Addition Rule could easily play the role of the oracle. But, we can outsmart the oracle and save our gold. First, we realize that since we will both be reporting the same pair of numbers to the oracle, the oracle will have to give the same answer in both cases. That would allow us to split the cost and save half of our gold. But we can do even better. Suppose that we think of the case where our background information K tells us that the value of X is exactly 8. The oracle must give the same answer in this case as well. But in this case we have

$$E(XI_N|K) = E(8I_N|K) = 8E(I_N|K) = 8P(N|K) = 8 * .3 = 2.4$$

which means the final answer the oracle must give is simply 2.4, the product of our two numbers. In fact, we see that the only way the oracle can operate and remain consistent with our four basic rules is to simple multiply each pair of numbers it is presented with. This finally gives us our *Multiplication Rule*, and obviously we see why it is so named.

MULTIPLICATION RULE:

$$E(XI_N|K) = E(X|N\&K)P(N|K). \tag{51.4}$$

Mathematically, the multiplication rule is really the most fundamental rule, as it has so many applications and can be used to give the addition rule for probability, even though we will not demonstrate this here.

Returning to the situation where we have the three statements of which exactly one is true, from [51.3] and the Multiplication Rule we have

$$E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K),$$

so

$$E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K).$$

This result gives us the general rule due to Bayes in the case of probability which allows us to reduce the problem of guessing to the problem of calculating probabilities.

GENERAL BAYES RULE FOR EXPECTATION:

Any time we have a finite sequence of statements $A, B, C, ...$ and our background information tells us exactly one is true (so all others are false), then

$$E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K) + ... \tag{51.5}$$

can be used to determine the guess for the value of X once we have *determined all the probabilities* of the statements $A, B, C...$ *and the guesses we would make in each case.*

For instance, for the dice in the box, we can take statement A_1 to be the statement that 1 is on the top face, and likewise define A_2, A_3, A_4, A_5, A_6. If K says we cannot see in the box, then each of these statements has probability 1/6 given K but obviously $E(X|A_1\&K) = 1$ and $E(X|A_2\&K) = 2$ and so on, so our previous results on probability tell us here that each of these statements has probability 1/6, whereas the General Bayes Rule for Expectation tells us that $E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$. Notice that the multiplication rule together

with the four basic rules has now determined what we should guess. In some sense, we have removed the guesswork in guessing or in another sense, we have turned guesswork into an actual process which leads to a definite result. Any two people following these basic logical rules must arrive at the same result or else one of them has violated a rule of logical consistency or the multiplication rule. We are often presented with a type of problem where we have a table giving a list of all possible values of an unknown together with their probabilities. Then we know the probabilities must add up to one, so if one of the entries in the probability list is missing we can easily figure it out. To find the guess for the unknown, we simply multiply each value by its probability and add up all the products, as that is what [51.5] is saying to do. In the TI-83/4 calculator, simply put the values in a list and the corresponding probabilities in another list so that each value is on the same list level as it's probability and then do the 1-var stat L_v, L_p, where v is the list number of the value list and p is the list number of the probability list.

The multiplication rule can immediately be applied to give the rule for conditional probability for calculating $P(A\&B|K)$. We just use the fact that by [50.10] we know $I_{A\&B} = I_A I_B$, so using $X = I_A$ and $N = B$ in the Multiplication Rule [51.4] we get

$$P(A\&B|K) = E(I_{A\&B}|K) = E(I_A I_B|K) = E(I_A|B\&K)P(B|K) = P(A|B\&K)P(B|K),$$

so finally we have the simple result.

CONDITIONAL PROBABILITY RULE:

$$P(A\&B|K) = P(A|B\&K)P(B|K). \tag{51.6}$$

LAST SPRING THIS LECTURE ALSO INCLUDED THE FOLLOWING APPLICATIONS WHICH WE WILL GET TO IN THE NEAR FUTURE.

The conditional probability rule has many applications, and we begin by reconsidering the blocks in the box problem, where blocks are being drawn successively from a box one after another without replacement. Suppose there are 3 red and 2 blue blocks in the box for a total of 5 blocks and this is our background information together with the statement that we cannot see what is in the box or tell by feel what color a block is. We must reach into the box and grab a block and pull it out without seeing which block we have until we have already chosen it. What is the chance that of the first two blocks drawn both are red? We will use R to denote that the block is red. So we are asking, what is $P(bothR)$? We can notice that "both R" is the same statement as "$1^{st}R\&2^{nd}R$" and then apply the Conditional Probability Rule to find easily

$$P(bothR) = P(2^{nd}R|1^{st}R)P(1^{st}R) = (2/4)(3/5) = .3$$

Recall that in many situations we have some finite number of statements of which exactly one is true and all the others are false, but K does not tell us which of these statements is the one that is true. In this case their indicators must add up to 1 and hence their probabilities must add up to 1. For instance, if we have statements A, B, C and exactly one is true and the other two are false, then $I_A + I_B + I_C = 1$, so when the Additive Rule and the definition of probability is applied, we find that $P(A|K) + P(B|K) + P(C|K) = 1$. In any situation where our information K does not tell us any of these three statements is more likely true than another, we must accept all three probabilities are the same, and as they add up to 1 we see each of these statements has probability 1/3. The same would apply if there were 6 different statements and K does not allow us to conclude any one more likely than another, then each of the 6 statements must have probability 1/6 given K. For instance, for the case of the dice in the box where we cannot see it, if that is the extent of our information, then we conclude that all faces are equally likely to be the one on top so each has probability 1/6 of being the one on top. We call this the Principle of Indifference. In general, if there are n statements and K

tells us exactly one is true but gives no information allowing us to judge any being more likely than the others, we conclude they all have the same probability, $1/n$. We generally refer to this as the *Model of Equally Likely Outcomes*. In gambling situations, we generally say a game is *FAIR* when the model of equally likely outcomes is in effect. Thus we speak of a fair pair of dice or a fair roulette wheel or a fair lottery. For instance, if a box contains 3 red blocks and 2 blue blocks, and one block is removed and we do not see which one has been removed, then with K the statement of these facts, if R is the statement that the removed block is red, then $P(R|K) = 3/5$, since each block has probability $1/5$ of being the block that was removed, and three of these are red. Try making up symbols for the statements that each of the 5 blocks is the one removed on the first draw, and then assuming each has probability $1/5$ demonstrate that $P(R|K) = 3/5$. If a second block is removed and we are told that the first block removed was red, then $P(2^{nd}R|1^{st}R\&K) = 2/4$. Try to figure out the meaning of $P(1^{st}R|2^{nd}R\&K)$. Does this make sense?

The General Bayes Rule for Expectation can be used to formulate a useful rule for probability by taking X in [51.5] to be the indicator of a statement. Can you figure out what this formula should be?

52. **LECTURE** MONDAY 31 AUGUST 2009

The teaching assistant Ms Xu discussed calculations with data and the use of the calculator.

53. **LECTURE** WEDNESDAY 2 SEPTEMBER 2009

Today we discussed the use of the TI-83/4(+) calculator to calculate one variable statistics on sample and population data. To do the calculations, the data must first be entered into lists in the calculator. To do that, one must first turn ON the calculator (lower left hand corner button, its second function is OFF). The data lists are accessed by pressing the stat button which across the top shows the menu of menus EDIT CALC TEST. At the bottom of the EDIT screen menu will appear "SetUpEditor". Putting the cursor on "SetUpEditor" and pressing ENTER will put the lists into the standard factory format. With the cursor on "Edit" at the top of the EDIT screen menu, when ENTER is pressed the lists will appear in a grid. At the top on the left is the first list L_1, in the top middle is L_2 and on the top right is L_3. Scrolling to the right will make more lists appear. The standard factory setting gives 6 lists, but more lists can be created and named if necessary. To clear a list, put the cursor on the list name at the top and then press the CLEAR button. This at first sight appears to do nothing, but if the cursor is moved into the box below the list name, all the data in that list immediately disappears and new data can be entered. To enter data into a list, move the cursor into the box below the list name, type the numerical value of a data score and press enter. Repeating this one can enter more possibly different numbers in a list than we will ever need in this course. If a score has been left out, put the cursor on the score in its place and press "2nd" followed by "DEL", because the second function of the delete button is the insert operation. This will push all those scores down and create a zero in the location where you wish to enter the score that was left out. You then simply type the left out score and hit the ENTER key. To delete a score, simply put the cursor on that score and press the DELETE button, "DEL". If our data is just a list of numbers, we simply enter it into a list and then to calculate the statistics for that data, we press the STAT button again and put the cursor on CALC at the top. This immediately gives us the entire menu of statistical calculations the calculator can perform on data. At the top of the list is "1:1-Var Stats" and it is the default if you simply press ENTER. To call another calculation instead, either move the cursor to that line in the menu or type the symbol (number or letter as the case may be) for that line (it appears followed by a colon). The list of symbols for the lines of any menu is 1,2,3,4,5,6,7,8,9,0,A,B,C,D,E,... For instance in the CALC menu the fourth line is "4:LinReg(ax+b)", and to call up LinReg(ax+b) we can either move the cursor onto that line and hit ENTER, or simply type "4" followed by ENTER. When 1-Var Stats is called, one sees a blinking cursor. In TI calculators, the blinking cursor represents the calculator asking you for information. In this case, the calculator needs to know which list contains the data you want to have the calculator use for the statistical calculations. The default is list 1 or "L_1". Hitting ENTER will cause the calculator to calculate the statistics on the data in "L_1". If you want the calculations done on the data in list 2, you respond to the blinking cursor by typing "L_2" which is accomplished by pressing "2nd" followed by "2". When you press ENTER, the statistical calculations will then be done on the data in list 2.

When you look at the readout of the statistical calculations, you see beneath the "1-Var Stats" the actual values calculated for the statistics: $\bar{x}$ = mean of all scores, Σx = sum of all scores, Σx^2 = the sum of squares of all the scores (each score is squared and all those resulting numbers are added up), S_x =the sample standard deviation, σ_x = the population standard deviation, n = the data size (the number of scores in the list), and here you see preceding the symbol n is a downward pointing arrow. In any readout, the downward pointing arrow signals you that by scrolling down, you will be able to see more results than can be displayed on a single screen. Scrolling down then shows: $\min X$ = the minimum score, Q_1 = the first quartile score (the score which separates the bottom 25 percent from the top 75 percent), Med= the

median of the data (the score which separates the bottom 50 percent from the top 50 percent), Q_3 = the third quartile score (the score which separates the bottom 75 percent from the top 25 percent, and finally, at the very bottom of the readout appears max$X$ = the maximum score.

The mean, sum, data size, n, and the two standard deviations, S_x and σ_x are our first concern. If the data represents all the blood pressure scores in an entire population, then if John Doe is in the population, and X is John Doe's blood pressure, our best guess for that is $\mu_X = E(X|K) = \bar{x}$. That is, the population mean is our best guess, and if our data is for the whole population, then the calculator symbol $\bar{x}$ is actually representing the population mean $\mu_X = E(X|K)$. Also, as S_x is the sample standard deviation and our data is for a whole population, then we should ignore it and use the population standard deviation, $\sigma_X = \sigma_x$. (notice the calculator calls the variable or unknown x instead of X in this case. If we want to guess what our error will be when we use μ_X as our guess for John Doe's blood pressure, then σ_X is our best bet. We will see that in general, if X is any unknown, when using $E(X|K)$ as our guess for the value of X our best bet is that the error will be σ_X where

$$\sigma_X^2 = E((X - \mu_X)^2|K)$$

(more about this formula later). If we guess anything else, it will also appear that our error will be more.

Notice that our error will depend on how much variation there is in the blood pressure scores throughout the population. If everyone in the population has blood pressure 120, then the mean is 120, and when we guess that, our error is zero. The more the variation, the larger our error is likely to be, and also the larger the standard deviation computed for the data.

Suppose instead our data represents the blood pressure scores from a sample taken from the population we are interested in. In this case, we will see that the sample mean is the best guess for John Doe's blood pressure, which is to say, if our background information is K and if $\bar{x}$ is the sample mean, then

$$E(X|\ \bar{x}\ \&\ K) = \bar{x}.$$

However, the σ_x calculated by the computer is no longer the population standard deviation. It is merely the standard deviation for the sample considered as itself a population, and that is not what we want. For instance, if there is only one score in the sample, then there is certainly no deviation or variation in the data and the standard deviation would be zero, even though the actual population might have plenty of variation. This points to the fact that since there are fewer scores in a sample than in the whole population, generally, the sample data will have less variation than the population as a whole. The sample standard deviation S_x is meant to compensate for this fact of small variation due to a small number of scores in the sample (small n). In the calculator, the relation between the σ_x and the S_x is simple. If you know the sample size, n, then always

$$S_x = \sigma_x \sqrt{\frac{n}{n-1}}.$$

Why this is the case will become clearer much later in the course, but suffice it to say, that the choice of this "fudge factor" to modify the standard deviation from the sample data to get a good estimate of the population standard deviation is chosen in an optimal sense, in fact, in the sense that its square is expected to be the square of the population standard deviation in the ordinary sense of expectation:

$$E(S_x^2|K) = \sigma_X^2.$$

The square of standard deviation is called VARIANCE. Thus the equation says that the sample variance is expected to be the population variance.

Often, for population data we have scores and there probabilities or a list of scores and for each score in the list we know the percentage of the population or sample having that score. In this case, we enter the probabilities or percentages in another list, say L_3, and call up the 1-Var Stats and enter L_1, L_3 as answer to the blinking cursor. Each percentage or probability must

be on the same level in the list as the score for which it is the probability or percentage. In this case, the calculator does not know the actual number of scores or there is no such number, and the calculator adds up the probabilities and gets $n = 1$. Here, the formula for S_x would have zero in the denominator, so the calculator leaves out any value for S_x. It only reports σ_x as far as standard deviation is concerned. If we have $n = 100$ for a sample, then $100L_3 \to L_2$ (here the little arrow is the symbol for the store button on the lower left of the calculator-it has STO and a little arrow head on it) stores the score frequencies in list 2, since the frequency of each score is simply its probability (in the data) or percentage multiplied by 100. In this case, 1-Var Stats L_1, L_2 will have the same values for $\bar{x}$ and σ_x, but the S_x appears and you can check that the formula above for S_x in terms of n and σ_x actually works.

54. **LECTURE** FRIDAY 4 SEPTEMBER 2009

The teaching assistant Ms Xu discussed calculations with data and the use of the calculator.

55. **LECTURE** MONDAY 7 SEPTEMBER 2009

LABOR DAY. NO LECTURE.

56. **LECTURE** WEDNESDAY 9 SEPTEMBER 2009

Today we will begin learning how to deal with 2 related unknowns. In many situations there are several unknowns to deal with, some are easy to measure and some are more difficult to measure. In such a situation, we can imagine trying to develop a method for guessing the value of the unknowns that are difficult to measure by simply using the ones that are easy to measure. Let's begin with a simple example. Suppose we have a tuna fish pulled out of the Pacific Ocean. Let X be its length in feet and Y be its weight in pounds. For instance, suppose that $E(X) = \mu_X = 4$ and $E(Y) = \mu_Y = 300$. Of course, either μ_X or $E(X) = \mu_X$ here is short hand for $E(X|K)$ where K is the statement of what we know about Pacific tuna fish and contains the fact that the fish in question was pulled out of the Pacific and is in fact a tuna fish. Now, it is generally easier to measure the length of a fish than its weight, since measuring the length only requires a simple tape measure, whereas the weight requires a scale, and moreover, if the fish weighs hundreds of pounds it has to be hoisted up onto the scale which could be difficult. If we are given the additional information that the fish is 6 feet long, we would probably be inclined to increase our guess for the weight of the fish beyond a mere 300 pounds. On the other hand, if the fish is only 2 feet long, we would likely want to guess his weight to be less than 300 pounds. This merely expresses the idea that longer fish tend to weigh more. But this is certainly not a hard and fast rule. We might sometimes find a long skinny fish whose length is actually above average but whose weight is below average. This is clearly not a common thing for tuna fish. Let us use the symbol $D_X = X - \mu_X = X - 4$ to denote the deviation in length for a fish and so likewise $D_Y = Y - \mu_Y = Y - 300$. Notice that if $X = 6$, then $D_X = 2$, whereas if $X = 3$, then $D_X = -1$.

To summarize, we are thinking here that a fish which is longer than average should tend weigh more than average whereas a fish which is shorter than average should tend to have weight below average. But these are only tendencies, not hard and fast rules which are certain. In terms of deviations, we can say that if $D_X \geq 0$ then D_Y should tend to be greater than or equal to zero, and therefore the product of deviations $D_X D_Y$ should tend to be greater than or equal to zero. On the other hand, if $D_X \leq 0$, which means the fish is shorter than average, then D_Y should tend to also be less than or equal to zero. Since negative times negative is positive, in this case again we conclude that the product of deviations should tend to be greater than equal to zero. Overall, we can assess this by evaluating $E(D_X D_Y)$. We define the **Covariance** of X with Y, denoted $Cov(X, Y)$ by the formula

$$Cov(X, Y) = E(D_X D_Y).$$

We therefore have

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

One problem with the use of covariance to get an idea of how well two unknowns relate is that the number could be large just because there is a lot of variation in the variables themselves. For instance, if it is the case that all tuna fish are very close to 4 feet long and weigh close to 300 pounds, then all the deviations will be small, so the products will be small and thus the expected product will be small and that is the covariance. It might be the case that for such a population of tuna the relation between length and weight is very good. In another population of tuna, there might be enormous variation in length as well as in weight giving a

large covariance when the relation between length and weight is not so good. To get a better idea of the relationship, we need to compensate for the deviations within each unknown. To do this, we simply use the covariance of each unknown *with itself*. This is called the **Variance** of the unknown, that is, precisely, we denote by $Var(X)$ the variance of the unknown X which is defined to be its covariance with itself:

$$Var(X) = Cov(X, X).$$

It follows that

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

We define the **Standard Deviation** of X, denoted by σ_X, by the formula

$$\sigma_X = \sqrt{Var(X)}.$$

Remember, for populations or for samples, the square of the standard deviation is the variance and the square root of the variance is the standard deviation. Do not get that backwards. In the variance, we can notice that $D_X^2 \geq 0$, and therefore

$$Var(X) = E(D_X^2) \geq 0.$$

Because of this, the square root of the variance will always make sense, since variance can never be negative.

We can now make use of the standard deviation to standardize the covariance to give an intrinsic measure of the relationship between two variables. To do this we form the **Correlation Coefficient**, denoted by the Greek letter ρ and defined by the formula

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

Therefore, we can see that

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = E(\frac{D_X}{\sigma_X}\frac{D_Y}{\sigma_Y}) = E(Z_X Z_Y).$$

We have in the last expression introduced the notation for the **Standardization** of an unknown. The standardization of X is denoted Z_X and is the new unknown defined by

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

The standardization merely converts to units of standard deviation. For instance, in our example, if we have $\sigma_X = .25$, then a 3.5 foot fish has deviation $-.5$ and therefore standardized length score of $-.5/(.25) = -2$. A standardized score of -2 just means here that the fish is 2 standard deviations below the mean which is .5 feet below the mean, and as the mean is 4, this means $4 - .5 = 3.5$. When we look at the formula

$$\rho = E(Z_X Z_Y)$$

keeping in mind that the Z stands for standardizing the unknown, we can see that the correlation coefficient gives a standard measure of the relationship between two unknowns. On the other hand, from the formula defining ρ, we see that we can calculate the covariance when we know the correlation coefficient and the standard deviations simply using the formula

$$Cov(X, Y) = \rho \sigma_X \sigma_Y,$$

just multiply the three numbers when you have them all.

Now, clearly a large value for ρ indicates a strong positive relationship between the two unknowns, whereas a value near zero would indicate their relationship is not of much use. To see how to use this information, let us imagine that we want to find a simple linear expression of the form $a + bx$ so that when we know the length x of a particular fish, we calculate $a + bx$ to get our guess for the weight of the fish. We would have to chose the numbers a and b in advance somehow. What we would be doing is trying to use the unknown $a + bX$ to guess Y.

For simplicity, put $W = a + bX$. If we know X, then we know W exactly through the equation $W = a+bX$, and we use that as our guess for the value of Y. That is in symbols, more precisely,

$$E(Y|X = x) = a + bx.$$

For instance, if we have decided in advance that we should use $a = 375$ and $b = 3$, then when we see fish that is 6 feet long we guess the weight to be $175 + 3 * 6 = 375 + 18 = 493$. Of course, I have just pulled the values $a = 375$ and $b = 3$ out of thin air here, so this method is so far fairly worthless. We need to have a criterion for picking the numbers a and b.

To optimize our choice of a and b, we need to consider that what we need to do is minimize our error in guessing somehow. To do this, we notice that our error or **Residual** denoted R is given by

$$R = W - Y = (a + bX) - Y.$$

Overall, the negative errors may appear to balance the positive errors, so to get rid of that possibility, we take the attitude that we want to *overall* minimize the squared residuals which means, we want to minimize $E(R^2)$. Now, using our basic rules for expectation and a lot of algebra which would probably put you to sleep, we can get the following useful equation for the expected squared residual, which we can call the **regression residual equation**:

$$E(R^2) = [E(R)]^2 + (1 - \rho^2)\sigma_Y^2 + (\rho\sigma_Y - b\sigma_X)^2.$$

At first, this equation probably looks like a nightmare, but it has a couple of simple features which make the solution to our problem fall right out. Because, we can notice that two of the terms on the right side are squares. A term which is a square can never be negative, so the smallest it can possibly be is zero. For instance, setting $\rho\sigma_Y - b\sigma_X = 0$ makes the last term zero, and this equation is easily solved for b giving

$$b = \frac{\rho\sigma_Y}{\sigma_X}$$

as the optimal choice for b, the number we call the **Regression Slope**. On the other hand, if we set $E(R) = 0$, then the first term on the right vanishes. Since $R = (a + bX) - Y$, we have from our rules for expectation that $E(R) = a + b\mu_X - \mu_Y$, and therefore setting $E(R) = 0$ gives the equation $a + b\mu_X - \mu_Y = 0$ and this means

$$a = \mu_Y - b\mu_X.$$

We therefore have

$$b = \frac{\rho\sigma_Y}{\sigma_X}$$

and

$$a = \mu_Y - b\mu_X = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$$

as fairly simple equations giving the optimal choice for the numbers a and b which we use for our regression equation. Finally, notice that when we use the optimal choices for a and b, then both squared terms in the residual regression equation vanish and it simplifies to

$$E(R^2) = (1 - \rho^2)\sigma_Y^2.$$

This is a very useful fact, since first of all, it tells us that $\rho^2 \leq 1$, and therefore, $-1 \leq \rho \leq 1$. This is because as $R^2 \geq 0$, it follows that $E(R^2) \geq 0$, whereas we already know that $\sigma_Y^2 \geq 0$. This forces $1 - \rho^2 \geq 0$. The second thing this equation tells us is that the fraction $1 - \rho^2$ of the variance of Y is in the squared error, so it must be that ρ^2 is the fraction of the variance of Y that is being accounted for with the regression equation being used to guess Y. That is, if $\rho^2 = .8$, then our expected squared error will be only 80 percent of what it would have been if we did not use the regression equation. If I do not measure the fish, I will guess his weight as 300 pounds, and will expect a certain squared error in this guess. If I measure the fish and use the optimal regression equation to guess the weight of the fish, I cut my expected squared error down to only 80 percent of what it would have been.

For more details on these facts or to see the residual regression equation proven using the basic rules for expectation and algebra, you can go to THE EXPECTATION PRIMER down near the bottom of the MATH-111 page on my website. This is not rocket science, it merely requires the basic rules of expectation we worked out in class the first week of class and high school algebra. The interesting thing here is that the residual regression equation by this method is completely general, whereas in typical statistics classes only special cases are proven using calculus of several variables, which puts the foundation of regression analysis even beyond a student who has had the typical first year freshman calculus course.

The equations require that we know the means and standard deviations for each of our two unknowns and as well we need to know the correlation coefficient relating the two unknowns. If we have sample data for the pair of unknowns, then we can enter the data in two lists in the calculator as paired data. For instance in case of length and weight of fish, we enter the lengths in one list and the weights in another list in such a way that the length of each fish is on the same line as its weight. The 2-Var Stats in the calculator will give the sample means and sample standard deviations. The LinReg(a+bx) will give the sample estimates for a, b, ρ, and ρ^2. The sample estimate of ρ is called the **Sample Correlation Coefficient** and has the symbol r. Thus in the LinReg readout in the calculator you will only see the values of a, b, r and r^2 reported. These then are used to work out the optimal guess given the sample data. We can say that if N is the statement of our sample data which results in specific values a and b for the regression equation, then

$$E(Y|(X = x)\&N) = a + bx.$$

If at first you only see a and b reported in the readout of LinReg and you need r and r^2, you must turn on the diagnostics in your calculator. Once this is done, it will stay on until you turn them off, so just leave them on for the duration of this course. To turn on the diagnostics in your calculator, notice that the second function of the zero button is CATALOG. When you go to the catalog you find the entire listing of all functions in the calculator which is fairly enormous, but all are arranged alphabetically. Go down the list until you see the line "diagnosticOn" and with the little arrow head beside this line press the enter button a few times until you see "done". Henceforth you will always see both r and r^2 right below a and b in the readout of the LinReg. Notice there are both LinReg(ax+b) and LinReg(a+bx). I have chosen to use the later form which is further down in the list of the Stat CALC menu because if is more in line with symbols used in your textbook. If you use the wrong one it will simply give the of the values for a and b. For instance if you use the expression $a + bx$ for regression but use LinReg(ax+b) in the calculator, your reported readout value for a will be what you should use for b in the regression expression and your reported readout value for b will be what you should use for a in the regression expression.

57. **LECTURE** FRIDAY 11 SEPTEMBER 2009

Today we discussed the properties of covariance and variance which follow from the rules for expectation and a little algebra. The first useful fact is that

$$Cov(X,Y) = Cov(Y,X)$$

because if D_X is the deviation of X from its mean, then $Cov(X,Y) = E(D_X D_Y)$ and multiplication does not depend on order: $D_X D_Y = D_Y D_X$. Also, the additivity of expectation means that if we have any three unknowns, say W, X, Y, then

$$Cov(W, X+Y) = Cov(W,X) + Cov(W,Y).$$

This is just like the distributive law

$$a(b+c) = ab + ac,$$

is we think of $Cov(X,Y)$ as some kind of multiplication of X and Y. For instance, just as in algebra we know

$$(a+b)^2 = a^2 + b^2 + 2ab,$$

here with covariance, we find

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

Here you should keep in mind that

$$\sigma_X^2 = Var(X) = Cov(X,X).$$

Therefore we can alternately write the formula for $Var(X+Y)$ as

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y.$$

In this last equation, we have also used the formula $Cov(X,Y) = \rho\sigma_X\sigma_Y$ which gives the covariance in terms of the correlation and standard deviations. From these equations, we see that the variance and covariance must come into play when we want to find the standard deviation for $X+Y$ when we only know σ_X, σ_Y, and ρ.

Another useful equation which follows from the rules of expectation and the definition of covariance is that

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

In particular, this shows that if X and Y are correlated, then the expected value of the product will be different than the product of expected values. In particular, applying this to variance, as $Var(X) = Cov(X,X)$, we see that

$$\sigma_X^2 = Var(X) = E(X^2) - \mu_X^2.$$

From this we see that the mean of the square is always at least as big as the square of the mean. Moreover, the only way the two can be equal is for the variance to be zero. For an unknown to have zero variance means the unknown can only have a single value-it is a constant. Put another way, any time there is more than one possible value for something, if we want to guess its square we should guess more than the square of our guess for the thing itself. How much more is exactly the variance, according to these equations.

If c is a constant, then its deviation from its mean is zero, since $\mu_c = c$ and therefore $D_c = 0$. Thus $D_c^2 = 0$ and therefore $\sigma_c = 0$. If we add a constant c to an unknown, it does nothing to the variance, since in the formula for $Var(X+c)$ all the terms containing σ_c are zero. Thus,

$$Var(X+c) = Var(X).$$

For instance, if X is the salary of a worker in Duckburg and if everyone gets a five thousand dollar raise, then all the relative differences in salary stay the same. If you made 7 thousand dollars more than your neighbor before the raise, you still do after the raise. On the other hand, if all salaries are doubled, then all the differences between salaries double as well which doubles the standard deviation. If c is any constant, then $Var(cX) = c^2 Var(X)$, and $\sigma_{(cX)} = |c|\sigma_X$,

since the square root of c^2 is the absolute value of c, which we donote by $|c|$. Thus, for any number c, the equation

$$|c| = \sqrt{(c^2)}$$

defines its absolute value. For instance, $|-4| = 4$ whereas $|4| = 4$. In general, if we rescale X to form the new unknown $Y = bX + c$, then

$$E(Y) = bE(X) + c,$$

but for standard deviation we have only

$$\sigma_Y = |b|\sigma_X,$$

since the addition of the constant c to bX has no effect on standard deviation,

$$\sigma_Y = \sigma_{(bX)} = |b|\sigma_X.$$

As an example, suppose that the mean salary in Duckburg is 50K dollars with standard deviation 10K dollars. If everyone gets a 5K raise plus 20 percent of their original salary, then for a person with salary X, his new salary is $Y = 5 + X + .2X = (1.2)X + 5$. This means the new means salary will be $(1.2)50 + 5 = 65$ thousand dollars whereas the new standard deviation will be $(1.2)10 = 12$ thousand dollars.

Let us return now to probability and some simple typical quiz problems about drawing blocks from a box. These problems illustrate the utility of thinking about probability in terms of information. Any time two set-ups are the same as far as the information is concerned, all the probabilities will be the same. Consider a box containing 5 blocks of which 3 are red and 2 are blue. Suppose blocks are drawn one after another from the box. Let R be the statement that a given drawn block is red and B the statement that a given drawn block is blue. For instance, "$2^{nd}R$" is the statement that the second block drawn is red. Obviously, $P(1^{st}R) = 3/5$. If we ask for $P(2^{nd}R|1^{st}R)$, we can realize that for the draw of the second block, it is the case that we know there are two red blocks and two blue blocks in the box, so $P(2^{nd}R|1^{st}R) = 2/4 = 1/2$. On the other hand, there seems to be a problem about asking for $P(1^{st}R|2^{nd}R)$. Does this make sense? Actually, we can make sense out of this. The question is asking you to give the probability of drawing a red block on the the first draw if you can somehow know that you will get a red block on the second draw. How could this possibly be? Let us suppose that instead of blocks in the box we have playing cards in the box. Let's say a whole deck of cards in the box. Drawing cards from the box is just like being dealt a hand of cards. When we play cards, we do not insist that the dealer put all the cards in a box and shake it up and then draw randomly from the box of cards. We generally shuffle the deck and then simply deal cards from the top of the deck one after another. Notice, that as far as information is concerned, the situation is the same. You do not know which card is where in the stack, so any card can be anywhere. For instance, one fourth of the cards are spades, so if we ask for the probability that the first card dealt is a spade, then it is 1/4. However, the probability the second is a spade given the first was is 12/51. If I tell you that the card underneath the top card is a spade, then you know the second card dealt will be a spade. The probability that the first is a spade given that the second is a spade is therefore again 12/51. The same is the case with the blocks. Just imagine the blocks are stacked but you cannot see the stack. If I ask you the chance the top block is red given that the block right beneath it is red, the probability is obviously simply 2/4=1/2. But, informationwise, the situation is the same whether we are drawing blocks from the box or drawing blocks off the top of a stack one after another. The time factor is an illusion here, you should think of the future as already having been determined, you are just not aware of what it is. Where you know the future, it works for you just the same as knowing the past, as far as probability is concerned.

58. **LECTURE** MONDAY 14 SEPTEMBER 2009

Today we reviewed for TEST 1 scheduled for Wednesday 16 September 2009.

When reviewing the formulas for linear regression, we noted that from the formula for the optimal a and b giving the linear regression $W = a + bX$ for Y on X, since

$$b = \rho\frac{\sigma_Y}{\sigma_X},$$

gives the regression slope which we think of as rise over run for the regression line, and as

$$a = \mu_Y - b\mu_X,$$

we can see that the equation

$$E(Y|X = x) = a + bx$$

for computing the best guess for Y when given the value of X, dictates that

$$E(Y|X = \mu_X) = \mu_Y.$$

Indeed,

$$E(Y|X = \mu_X) = a + b\mu_X = (\mu_Y - b\mu_X) + b\mu_X = \mu_Y.$$

Thus if we graph the equation $y = a + bx$, whose graph we call the regression line, then we see it must always pass though the point (μ_X, μ_Y). This means that we can alternately compute $E(Y|X = x)$ by simply using the regression slope b and the deviations from the mean. For instance, suppose

$$\mu_X = 100,\ \mu_Y = 200,\ \sigma_X = 4,\ \sigma_Y = 20,$$

and $\rho = .5$. Since the regression line passes right through the point $(100, 200)$, if we are told that X actually has the value 100, then we should guess 200 for Y. Notice that the rise over the run or regression slope is just

$$b = \frac{(.5)20}{5} = 2.$$

If we know that X is 110, then we would increase our guess for Y above 200 by the amount $(2)(10) = 20$, so our guess for the value of Y should be 220. If we know that X is 115, this is 15 above the mean for X and therefore we should increase our guess for Y by the amount $(2)(15) = 30$ and therefore our guess should be 230.

Lets for the moment imagine a situation of perfect correlation, so $\rho = 1$. The our regression slope is just

$$b = \frac{\sigma_Y}{\sigma_X}.$$

This means that if X is known to be one standard deviation above the mean, then we should guess that Y is also one standard deviation above the mean. If we know X is 2 standard deviations below the mean, then we should guess that Y is two standard deviations below the mean. In the previous example, if we change the value of ρ to 1, then the regression slope is $b = 20/4 = 5$. If we know X is 108, then we would know that X is two standard deviations above the mean so we should guess Y is two standard deviations above the mean which is $2(20) = 40$ above the mean for Y so

$$E(Y|X = 108) = 200 + 2 * 20 = 200 + 40 = 240.$$

With perfect correlation, the number of standard deviations that X is from the mean dictates the number of standard deviations from the mean for Y we should guess, which we then simply add to $\mu_Y = 200$. Of course the same applies if X is below the mean, but we are then subtracting instead of adding. Thus, if we know X is 88, then that is 12 below the mean, and since the standard deviation for X is 4, this means that X is 3 standard deviations below the mean, so we should guess Y is also 3 standard deviations below the mean. As the standard deviation for Y is 20, three standard deviations is 60, so we guess Y is 60 below its mean of 200 or $200 - 3(20) = 200 - 60 = 140$. Now, lets go back to the case where $\rho = .5$. Then we see that when the run is $\sigma_X = 2$, then the rise is $\rho\sigma_Y = .5(20) = 10$. That is if we know X is one standard

deviation above its mean, then now instead of guessing Y to be one standard deviation above its mean, we would correct for the correlation coefficient of $\rho = .5$ by guessing Y to be only $1/2$ a standard deviation above its mean, which is therefore $(1/2)20 = 10$, so we guess the value of Y is $200 + 10 = 210$. In general, we can standardize X and Y by setting

$$Z_X = \frac{X - \mu_X}{\sigma_X}$$

and likewise for Y. Thus in our example we have

$$Z_X = \frac{X - \mu_X}{\sigma_X} = \frac{X - 100}{2}$$

and

$$Z_Y = \frac{Y - \mu_Y}{\sigma_Y} = \frac{Y - 200}{20}.$$

Thus a value of $x = 108$ for X is the same as a value

$$z_X = \frac{108 - 100}{2} = \frac{8}{2} = 4$$

which means we should guess Z_Y has value $z_Y = \rho z_X = (1/2)4 = 2$. Notice that

$$Y = \mu_Y + \sigma_Y Z_Y,$$

so that if we know the value of Z_Y, then we know the value of Y. For instance, in the case at hand, we have Y should be guessed as having value

$$y = 200 + 20(2) = 240.$$

59. **LECTURE** FRIDAY 18 SEPTEMBER 2009

Today we discussed Bayes' Rule for computing probabilities in situations which can be broken down into all the various alternatives. If for instance, we know exactly one of the statements A, B, C is true, then for their indicators we know that

$$1 = I_A + I_B + I_C,$$

so on multiplying through by any unknown X, we have

$$X = XI_A + XI_B + XI_C,$$

and therefore

$$E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K).$$

But to each term on the right hand side we can apply the multiplication rule

$$E(XI_N|K) = E(X|N\&K)P(N|K).$$

The result is sometimes called Bayes' Rule:

$$E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K).$$

The expectation calculation is thereby broken down into the calculation of expected values in each of the separate cases. If you know the probability of each case being the one to happen, then the expected value is the sum of products, each term the product of the expected value in one case multiplied by the probability that case actually is the case.

In particular, we can take the unknown X to itself be the indicator of a statement, say D. Remembering that

$$P(N|K) = E(I_N|K)$$

for any statement N, by definition, and recalling that

$$I_{M\&N} = I_M I_N,$$

the expectation equation for Bayes' Rule becomes a probability equation, also sometimes called Bayes' Rule:

$$P(D|K) = P(D|A\&K)P(A|K) + P(D|B\&K)P(B|K) + P(D|C\&K)P(C|K).$$

The multiplication rule can be turned around to give

$$E(X|N\&K) = \frac{E(XI_N|K)}{P(N|K)}$$

and in case of probability

$$P(M|N\&K) = \frac{P(M\&N|K)}{P(N|K)}.$$

Dropping the background information K, we have for short,

$$P(M|N) = \frac{P(M\&N)}{P(N)} = \frac{P(BOTH)}{P(GIVEN)}.$$

Combining this with Bayes' Rule we could for instance calculate the conditional probabilities $P(A|D),\ P(B|D),\ P(C|D)$.

In an applied situation, we dealt with the example of the Acme Widget Corporation which has three widget factories, A, B, and C. If we know the percentage of defective widgets coming from each factory, if we know the percentage of total widget production coming from each factory, then we can calculate the probability a widget is defective. Then, using the formulas above, if we have a defective widget, we can calculate the probability it came from factory A. Likewise, we can do the same for factories B and C.

In case there are only two categories A and B, then B =not$A$. In such a situation, a problem will be dealing with just two statements, so it is sometimes not immediately clear which is to play the role of the different cases and which is the statement playing the role of D above. But

if you are confused as to which way to proceed here, just pick a way to proceed, and if it is the wrong way, you will quickly find that the problem information does not give you what you need, so then go back and try the other way. Specifically, if we are dealing with say M and N, if we try to use the cases M and not$M$, then we will need the conditional probabilities $P(N|M)$ and $P(N|\text{not}M)$. Thus, if we realize that we are given $P(N|M)$, then the cases must be M and not$M$, whereas if we see we are given $P(M|N)$, then the cases must be N and not$N$.

As an example, suppose we have a student, Sam, taking a multiple choice test which is machine graded. Suppose that it is the case that each question has 5 possible answers to choose from and the answer must be marked on an answer sheet. If Sam knows the answer, we suppose he has a 90 percent chance of marking correctly, whereas if he does not know the answer, he has only a 20 percent chance of marking correctly. Suppose he knows the answers to 70 percent of the questions on the test. We want to know the percentage of questions he marks correctly and if we see a question marked correctly, we want to know the chance he actually knew the answer. We begin by choosing symbols. Let's use K for the statement he knows the answer to the question and C for the statement the answer to the question is marked correctly. Do we have $P(K|C)$ given to us in the problem information? The answer is no. How about $P(C|K)$. That is the chance he marks correctly given that he knows the answer, which is given to us as 90 percent. Thus the two cases we consider are C and not$C$. Using Bayes' Rule, we then find that

$$P(C) = P(C|K)P(K) + P(C|not\ K)P(not\ K) = (.9)(.7) + (.2)(.3) = .63 + .06 = .69.$$

We therefore know that the probability he marks a question correctly is 69 percent. If we see a correctly marked question the chance he actually knew the answer is $P(K|C)$, and this is

$$P(K|C) = \frac{P(K\&C)}{P(C)} = \frac{P(C|K)P(K)}{P(C)} = \frac{.63}{.69} = \frac{63}{69} = \frac{21}{23}.$$

This is about 91.3 percent (to three significant figures). Notice that the numerator is one of the terms in the calculation of $P(C)$. This will always be true in these Bayesian analysis problems, so when you do the calculation for the probability in the denominator, keep track of the individual terms before adding them all up, so you can then simply look at the total and pick out the correct term for the numerator. This will save you time and effort.

60. **LECTURE** MONDAY 21 SEPTEMBER 2009

Today, we began by demonstrating that Bayes'Rule and Baysian analysis can be used to calculate probabilities that we have already calculated using the information view of probability theory, which at first appear paradoxical because of time ordering. Specifically, consider a box containing 3 red blocks and 2 blue blocks. The experiment is to draw the blocks one after another from the box without replacement. As far as information is concerned, this is the same as having a stack of blocks and the experiment is to draw the blocks from the stack from the top down one after another, such as is done when dealing cards from a deck (stack of cards). If the two processes were not the same as regards the state of our information, then card dealers in casinos would have to draw cards from a box that has been shaken instead of dealing from a shuffled deck. If I ask for the probability that the second block is blue given that the first is red, there is no problem in visualizing this to be 2/4, since we can imagine that there has been one red block removed as we start to make the second draw, so the box contains 4 blocks of which 2 are red and 2 are blue. If I ask for the probability that the first block is blue given that the second block is red, there seems to be a problem visualizing what this means from the standpoint of drawing the blocks one after another from the box, but it does not seem to be a problem at all in the second way of viewing the experiment in terms of a stack of blocks. In the second view, the given information merely states that the block in the second position from the top is red, so for the positions unknown to us there are 4 blocks of which 2 are red and 2 are blue, so the chance a blue block is on top is simply 2/4. This is the same result as in the reverse time order. In symbols, let B be "blue" and R be "red". Accepting this view that the two experimental setups are equivalent, we can quickly figure complicated conditional probabilities such as

$$P(3^{rd}B|1^{st}R \ \& \ 2^{nd}R) = \frac{2}{3}.$$

In terms of the stack, telling you the top two blocks are red lets you know that for the bottom three positions there are 2 blue blocks and only one red block, so the chance the third position contains a blue block is simply 2/3. It is also clear in the stack picture, that without any information about what block is where, if I ask for the probability that the third block is blue, it is 2/5, whereas in the block drawing picture, it sometimes appears confusing as to what is meant by the probability that the third is blue. You must keep in mind that probability is about the state of your information. If I ask what is the probability that the third block drawn is blue, you must imagine that you have no way of knowing what the results of the first two draws were. If someone else drew the blocks from the box where you could not see the results, and after all the blocks are drawn we ask what is the chance that when we ask the experimenter the result of the third draw we will be told that it was blue, then that probability is what we are seeking. We see that the experimenter could be arranging the blocks in a row as he takes them out of the box in order to keep track of the results. A row of blocks is clearly equivalent to a stack of blocks as far is the information is concerned.

Now let's use Bayesian analysis to compute these probabilities from the rules of probability directly, without using the information view. For instance, we can say that either the second block will be red or it will not be red for sure, so we can write using Bayes' Rule:

$$P(2^{nd}\ B) = P(2^{nd}\ B|1^{st}\ R)P(1^{st}\ R) + P(2^{nd}\ B|1^{st}\ not\ R)P(1^{st}\ not\ R)$$

$$= (\frac{2}{4})(\frac{3}{5}) + (\frac{1}{4})(\frac{2}{5}) = \frac{(2)(3)+(1)(2)}{(4)(5)} = (\frac{2}{5})(\frac{3+1}{4}) = \frac{2}{5}.$$

The final result here is the same as we found quickly by viewing the blocks as being stacked. Now lets work out the probability that the first is red given that the second is blue using the

Bayesian analysis. We have

$$P(1^{st}\ R|2^{nd}\ B) = \frac{P(2^{nd}\ B\ \&\ 1^{st}\ R)}{P(2^{nd}\ B)} = \frac{P(2^{nd}\ B|1^{st}\ R)P(1^{st}\ R)}{P(2^{nd}\ B)} = \frac{(\frac{2}{4})(\frac{3}{5})}{(\frac{2}{5})} = \frac{3}{4}.$$

Notice, how much quicker it is to view the blocks as stacked and think of the given information as telling us there is a blue block in the second position in the stack, so the chance the top block is red is simply 3 out of 4, or 3/4.

We next began the theory of counting which is necessary for dealing with similar problems involving many more objects. For instance, if we are dealing cards from a standard deck of 52 cards, we might ask the probability that if we deal five cards we get two of the same denomination which is called a pair. You should keep in mind that a standard deck of cards has 4 suits: Spades, Hearts, Diamonds, Clubs, and 13 denominations: Ace, 2,3,4,5,6,7,8,9,10, Jack, Queen, King. Here, the problems in merely counting up the different possibilities become substantial, and it is useful to keep in mind a few simple counting principles. The first is so obvious, that it might seem too simple to state, but sometimes, it is crucial. To begin, suppose we have any two sets A and B, say both contained in the set S. By

$$A \cup B = \{x \in S :\ x \in A\ or\ x \in B\}$$

we mean the set theoretic union of A and B which is a new set consisting of the members of A and B all thrown in together to form a single set, whereas by

$$A \cap B = \{x \in S : x \in A\ \&\ x \in B\}$$

we mean the set theoretic intersection, that is the set consisting of exactly what the two sets have in common, or their overlap. if F is any finite set, we denote by $n(F)$ the number of members of F. Next, suppose that the sets A and B have nothing in common or no overlap, that is in set language they are disjoint, expressed in symbols as

$$A \cap B = \emptyset.$$

For instance, suppose I have two boxes, one red and one blue, each having some finite number of things inside. Suppose I dump the contents of the two boxes in a third box which is green. If A is the set of original contents of the red box, if B is the set of original contents of the blue box, and if C is the set of contents ending up in the green box, then

$$C = A \cup B,$$

and obviously as $A \cap B = \emptyset$, we must also have

$$n(C) = n(A) + n(B),$$

or generally,

$$n(A \cup B) = n(A) + n(B),\ A \cap B = \emptyset.$$

This simple fact sort of expresses the idea of conservation of "stuff" in some sense. It is called the Addition Rule for counting. For instance suppose I ask how many 5 card hands have all hearts or at least 4 diamonds and one club. Let A be the set of all possible 5 card hands having all hearts and let B be the set of all 5 card hands having 4 diamonds and one club. Then since a hand cannot have both all hearts and 4 diamonds and a club, the two sets are disjoint. It follows that if C is the set of all 5 card hands having either all hearts or 4 diamonds and a club, then $C = A \cup B$, and therefore,

$$n(C) = n(A) + n(B).$$

We see the Addition Rule allows us to break down the counting problem into parts.

The next rule is called the Multiplication Rule and is the rule for counting the number of distinguishable outcomes in a stepwise procedure. Suppose we consider a procedure which takes several steps. What you do in life could be considered a sequence of steps. Starting from this morning, for instance, Step 1: you got out of bed, Step 2: you got dressed, Step 3: you gathered up your laundry and put it all in the washing machine,..., and obviously, at each stage of this

process you have many alternatives. In general, the alternatives you have available at each step of your life depends heavily on the choices you have previously made. However, when it comes to counting problems, there are many situations where the number of options you have available at each stage of a process will NOT depend on the particular choices which preceded. The card dealing experiment is a prime example. Suppose you are dealt the first 5 cards from the top of a shuffled deck of cards, turned up one after another in a row. The order in which cards appear is important here. This game is called 5 card showdown and in Poker one bets after each card is turned up, so the order in which the cards appear is important to the players. The number of cards you have available for the first card dealt to you is 52, since you have no idea of where any particular cards are positioned in the stack. Once you are dealt a card, you only have 51 cards available for the second card dealt to you, independent of what you got on your first card. Likewise, when you consider what is available for the third card, you know there are 50 possibilities no matter which two cards you received for the first two dealt. The number of possibilities for each step is totally independent of what came before-totally independent of the prior history. This is certainly not the situation in life in general, but it is here. Now, keep in mind that the cards available at each stage DO depend on what came before, it is just that the NUMBER of cards available does not depend on what came before. If I get the Ace of Hearts as the first card dealt, then it is not possible to get the Ace of Hears as the third card dealt. We want to count the number of possible outcomes of dealing out 5 cards from a shuffled deck. Well we have 52 possibilities for the first step, and for each of those 52 possibilities there 51 possibilities for the second step resulting in (52)(51) possibilities for the outcome of the first two steps. Notice the outcome of the first two steps is a two card hand with an ordering of the cards-there is a first card and a second card. Getting the Ace of Hearts followed by the Ace of Clubs is a different outcome than getting the Ace of Clubs followed by the Ace of Hearts. To specify an outcome of this process, we need to specify what happens on each draw. Thus to actually specify an outcome we need to write down a sequence of 5 cards which are all different. The answer to the question of the number of outcomes is to count the number of all such sequences. Clearly there are 52 possibilities for the first card or first entry of the sequence, there are for each of these 52 possibilities then 51 possibilities for the second card, so (52)(51) possibilities for the first two cards or first two entries in the sequence, and for each of these (52)(51) possibilities for the first two cards, there are 50 possibilities for the third card for a total of (52)(51)(50) sequences of three cards, and so on. Clearly there are

$$(52)(51)(50)(49)(48)$$

possible outcomes for the game of 5 card showdown simply simply as far as the cards dealt are concerned. For instance, if we ask how many ways are there to arrange all the cards in the deck, or how many results are possible for shuffling the whole deck, that would be

$$(52)(51)(50)...(3)(2)(1).$$

which is an astronomical number-in fact it is astronomical in comparison to astronomical numbers. In fact,

$$(52)(51)(50)...(3)(2)(1) > 8 * 10^{67}.$$

This is an 8 followed by 67 zeroes. For short we denote this product of all the consecutive integers in the sequence $1, 2, 3, ..., r$ by the symbol

$$r! = (1)(2)(3)...(r),$$

which is read "r *factorial*". Obviously these numbers get big very fast as r increases. Now, back to the problem of computing (52)(51)(50)(49)(48). We could do this easily with the calculator, but if we were going to deal out 13 cards, it would be tedious. Fortunately the PRB menu if the MATH menu has the factorial on line 4 of the menu, so factorials can be quickly calculated. Notice also, that we can express the number of arrangements of 5 cards taken from the deck in

terms of factorials as

$$(52)(51)(50)(49)(48) = \frac{52!}{48!} = \frac{52!}{(52-5)!}.$$

This last expression makes it clear what the answer is if we deal out 13 cards in order instead of only 5. We just replace the 5 with 13 in the expression. This means there are

$$(52)(51)(50)(49)(48) = \frac{52!}{48!} = \frac{52!}{(52-5)!}$$

ways to arrange the top 13 cards in the deck. This type of counting arrangements comes up so often that it is useful to have a notation for it. We denote by $P(n,r)$ the number of ways to arrange r things chosen from a set of n things. Thus

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

Here the capital P stands for *permutation* which is the mathematical word for arrangement. In the calculator, to calculate the number of arrangements look for the symbol "nPr" in the PRB menu of the MATH menu. Then to calculate $P(52,5)$, you begin by typing 52, then hit the MATH button and put the cursor on PRB, and then type the number of the line on which you see "nPr" or else put the cursor directly on it and press the enter button. At this point you should see "52 nPr" on your screen so then type the 5 and you should then see "52 nPr 5" on your screen. At that point, you hit the enter button and the calculator gives you the answer.

These counting formulas for permutations are an illustration of the general counting principle we call the Multiplication Rule. In a stepwise procedure, if the number of options at each stage is independent of the history, and if m_k denotes the number of options available for the k^{th} step, $k = 1, 2, 3, ..., n$, then the total number of outcomes or sequences for all n steps is the number N which is the product of all the numbers available for each step:

$$N = m_1 m_2 m_3 ... m_n.$$

61. **LECTURE** WEDNESDAY 23 SEPTEMBER 2009

Today we continued our discussion of methods of counting. We reviewed the Addition Rule and the Multiplication Rules given in the previous lecture. We observed that

$$n(A \cup B) = n(A) + n(B) - n(A \cap B),$$

when counting finite sets. We reviewed the use of the Multiplication Rule for counting arrangements, and the formula for $P(n,k)$, which gives the number of ways of arranging k things taken from a set of n things. Specifically, we have

$$P(n,n) = n! = (1)(2)(3)...(n)$$

and more generally, we have

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

giving the number of ways to arrange k things chosen from a set of n things. We discussed the fact that we define $0! = 1$ in order to make the two formulas consistent in case $k = 0$.

Next, we dealt with the number of WORDS which are obtained by rearranging symbols. The number of ways to (re)arrange the letters $ABCD$ is obviously $(4!)$. In case of a word like $MISSISSIPPI$, we see that as some of the letters are alike, we cannot distinguish for instance which S went where in the rearrangement. Each arrangement is called a WORD even though it may not appear in any dictionary. For a mathematician, a word is simply an arrangement of symbols, which we may as well take to be a string of symbols. Let us call this number of arrangements x as we do not actually know what it is. To count the number of ways in this case, we form new distinguishable symbols. From the M tag it to get M_1, the first new symbol. From the four I's we get the new symbols I_1, I_2, I_3, I_4, from the four S's we get S_1, S_2, S_3, S_4, and from the two P's we get the two symbols P_1, P_2. Notice these new symbols are just the result of putting tags on the original symbols so as to make them all different. We can easily count the number of ways to arrange the eleven new symbols as they are all different, so it is $(11!)$. The key to finding x, the number of ways to arrange the untagged symbols is to realize that the job of arranging the tagged symbols can be accomplished as a two step procedure where Step 1 is to arrange the untagged symbols. After all, if you see an arrangement of the tagged symbols, you cannot tell whether it was accomplished by actually rearranging the tagged symbols or accomplished by rearranging the untagged symbols and then attaching the tags as the last step. Thus there are

$$(1!)(4!)(4!)(2!)$$

ways to put the tags on the untagged symbols once they have been arranged, so we must have, by the Multiplication Rule

$$x[(1!)(4!)(4!)(2!)] = 11!,$$

and therefore

$$x = \frac{11!}{(1!)(4!)(4!)(2!)}$$

gives the number of ways to arrange the letters $MISSISSIPPI$. It is useful to have a symbol for this. We will write

$$C(11; 1, 4, 4, 2) = \frac{11!}{(1!)(4!)(4!)(2!)}.$$

Notice that the sum of the numbers after the semi-colon is the number before the semi-colon. The order of the numbers after the semi-colon clearly does not matter. Thus

$$C(11; 1, 4, 4, 2) = C(11; 2, 4, 1, 4).$$

Moreover, if the numbers after the semi-colon do not add up to the number before the semi-colon, we assume that the last number was just left out. Thus,

$$C(11; 1, 4, 4) = C(11; 1, 4, 4, 2)$$

whereas

$$C(11;4,1,2) = C(11;4,1,2,4).$$

More generally, if we have a set n objects of which k_1 are all alike and considered indistinguishable among themselves, of which another subset of k_2 are considered all alike and indistinguishable among themselves, and so on, say we have m different types of objects which can be distinguished, k_1 of the first type, k_2 of the second type and so on and finally k_m of the last type, then we must have

$$k_1 + k_2 + ... + k_m = n$$

and the number of distinguishable arrangements of the objects is $C(n; k_1, k_2, ..., k_m)$ which is given by the formula

$$C(n; k_1, k_2, ..., k_m) = \frac{n!}{(k_1!)(k_2!)...(k_m!)} = C(n; k_1, k_2, ..., k_{m-1}).$$

As a special case, suppose that there only two types, type A and type B. For instance we could be asking for the number of words which can be formed with the letters $AAABBBB$. Such a word has 7 positions in which letters must be filled in, and notice as soon as we fill in the three A's, the whole word is determined as all the rest of the positions will be filled in with B's. Thus, the job of making a word here is equivalent to the job of choosing the 3 positions from the 7 available in which to put the A's. Notice also, that it does not matter in which order we decide which of the positions will get an A. It only matters which of the seven positions are chosen. Thus,

$$C(7;3,4) = C(7;3) = C(7;4)$$

gives the number of ways to choose 3 things from a set of 7 things, and we see that that is the same as the number of ways to choose 4 things from a set of 7 things.

In general, then

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

gives the number of ways to choose r things from a set of n things. Clearly, 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 a set of n things, as determining which r things to choose is the same job as determining which $n-r$ things not to choose. You can form a club by either choosing who will be in the club or deciding who is not in the club.

To see the relationship between the number of ways to arrange r things chosen from n, notice that such an arrangement can be accomplished in two steps where the first step is to choose the r things from the n things and the second step is to arrange the r things you have chosen on the first step. This means we must have

$$P(n,r) = C(n;r)P(r,r)$$

and therefore

$$C(n;r) = \frac{P(n,r)}{P(r,r)}$$

and of course, when we use the formulas

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

and

$$P(r,r) = r!,$$

we see that

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

just as before.

We computed the number of ways to play 5 card showdown-dealing out 5 cards one after another where the order in which the cards appear is important. It is

$$P(52,5) = 311,875,200$$

which is a very big number. When we talk about a 5 card hand, we do NOT care about the order in which the cards are dealt, we just care about which 5 cards we received in the deal. the number of 5 card hands is therefore

$$C(52;5) = 2,598,960.$$

clearly a very big number, but small in comparison to the number of things that can happen in dealing out 5 card showdown. In the game of Poker, there are only 5 card hands even if the game is played with more than 5 cards. Different hands are of different value, and if you are playing a game with more than 5 cards, then you use the best 5 to "play". The best hand is a Royal Flush which consists of the Ace, King, Queen, Jack, and Ten, all of the same suit. For the 4 suits, let us use the symbol H for Heart, D for Diamond, C for Club, and S for Spade. For the different denominations, let us use the symbol A for Ace, K for King, Q for Queen, J for Jack, T for Ten, 9 for 9, 8 for 8, and so on. The a Heart Royal Flush is AH, KH, QH, JH, TH. There is only one such hand out of all 2598960 possible Poker hands, so your chance of getting the Heart Royal Flush when dealt 5 cards from the standard 52 card deck is

$$\frac{1}{2598960},$$

a very small probability.

In general, for calculating the probabilities of the various Poker hands, it is useful to key in on the number of denominations in the type of hand. As an example, if you want to calculate the probability of getting a hand with two pair when dealt 5 cards, you can begin by noticing that such a hand can only have 3 different denominations, so step 1 in forming such a hand would be to choose the 3 denominations from the 13 available. That can be done in $C(13;3)$ ways. The next step would be to decide which one of the three chosen denominations is not to be paired or else choose which two of the denominations are to be paired. This can be done in $3 = C(3;1) = C(3;2)$ ways, and next, for the denomination not to be paired, choose one of that denomination. This can be done in $C(4;1) = 4$ ways. Next for each of the two denominations to be paired, choose two cards of that denomination. That can be done in $C(4;2) = 6$ ways for each of those two denominations, and it does not matter the order in which one does these last two choices. Multiplying all the numbers of ways for each of the steps, we find that there are

$$C(13;3)C(3;1)C(4;1)C(4;2)C(4;2) = (286)(3)(4)(6)(6) = 123,552$$

ways to form a 5 card hand containing two pair, no more and no less. For instance, a hand with 4 of a kind is not counted here as such a hand only has 2 denominations. this means that the probability of being dealt a hand containing exactly two pair is

$$\frac{123552}{2598960} = .0475390156,$$

or just under 5 percent.

62. **LECTURE** FRIDAY 25 SEPTEMBER 2009

Today we calculated the number of 5 card Poker hands of each type. Remember there are

$$C(52;5) = (52\ nCr\ 5) = 2,598,960$$

possible 5 card Poker hands. For each type of hand, we imagine a stepwise process for actually making such a hand which involves therefore a sequence of decisions. You want to make sure that any possible hand of the given type could be the result, and also you should make sure the process you are thinking of has the property that any change in any decision at any step will definitely change the final outcome. As last time, we note that for many types of Poker hand, a key characteristic is the number of different denominations that type of hand has. The different types of hands (followed by their number) in decreasing order of power are:

ROYAL FLUSH: $C(4;1) = 4$

STRAIGHT FLUSH (INCLUDING ROYAL FLUSH):

$$C(4;1) \cdot C(10;1) = 4 * 10 = 40$$

STRAIGHT FLUSH BUT NOT ROYAL FLUSH:

$$40 - 4 = 36$$

FOUR OF A KIND:

$$C(13;2) \cdot C(2;1) \cdot C(4;4) \cdot C(4;1) = 78 * 2 * 1 * 4 = 624$$

or alternately a slight variation in process:

$$P(13;2) \cdot C(4;4) \cdot C(4;1) = 156 * 1 * 4 = 624$$

FULL HOUSE (PAIR & 3 OF A KIND):

$$C(13;2) \cdot C(2;1) \cdot C(4;3) \cdot C(4;2) = 78 * 2 * 4 * 6 = 3744$$

or alternately a slight variation in process:

$$P(13;2) \cdot C(4;3) \cdot C(4;2) = 156 * 4 * 6 = 3744$$

FLUSH (ALL OF THE SAME SUIT INCLUDING STRAIGHT FLUSH):

$$C(4;1) \cdot C(13;5) = 4 * 1287 = 5148$$

FLUSH BUT NOT STRAIGHT FLUSH:

$$5148 - 40 = 5108$$

STRAIGHT (INCLUDING STRAIGHT FLUSH):

$$C(10;1) \cdot [C(4;1)^5] = 10 * (4^5) = 10 * (2^{10}) = 10 * 1024 = 10,240$$

STRAIGHT BUT NOT STRAIGHT FLUSH:

$$10240 - 40 = 10200$$

THREE OF A KIND:

$$C(13;3) \cdot C(3;1) \cdot C(4;3) \cdot C(4;1) \cdot C(4;1) = 286 * 3 * 4 * 4 * 4 = 54,912$$

TWO PAIR:

$$C(13;3) \cdot C(3;2) \cdot C(4;2) \cdot C(4;2) \cdot C(4;1) = 286 * 3 * 6 * 6 * 4 = 123,552$$

ONE PAIR:

$$C(13;4) \cdot C(4;1) \cdot C(4;2) \cdot C(4;1) \cdot C(4;1) \cdot C(4;1) = 715 * 4 * 6 * 4 * 4 * 4 = 1,098,240$$

NOTHING (NO PAIR NOR STRAIGHT NOR FLUSH):

$$C(13;5)\cdot C(4;1)\cdot C(4;1)\cdot C(4;1)\cdot C(4;1)\cdot C(4;1) = 1287 * (4^5) - 10200 - 5108 - 40$$

$$= 1317888 - 10200 - 5108 - 40 = 1,302,540$$

Thus, for the different hands in non-overlapping types we have

ROYAL FLUSH:.....4
NON-ROYAL STRAIGHT FLUSH:.....36
FULL HOUSE:.....624
FOUR OF A KIND:.....3744
FLUSH BUT NOT STRAIGHT:.....5108
STRAIGHT BUT NOT FLUSH:.....10200
THREE OF A KIND:.....54,912
TWO PAIR:.....123,552
ONE PAIR:.....1,098,240
NOTHING:.....1,302,540

GRAND TOTAL:.....2,598,960

Of course to calculate the probability of being dealt any of these types of hands, you simply divide the number for that type in the table by the grand total of $C(52;5) = 2,598,960$, which is the total number of possible 5 card Poker hands. In particular, the probability of nothing is

$$P(NOTHING) = \frac{1302540}{2598960} = .501177394,$$

which means there is a roughly 50 percent chance of getting nothing and a 50 percent chance of getting something whenever you are dealt 5 cards from a standard 52 card deck, the chance of getting nothing being only very slightly better than the chance of getting something.

63. **LECTURE** MONDAY 28 SEPTEMBER 2009

NO CLASS TODAY BECAUSE OF YOM KIPPUR.

64. **LECTURE** WEDNESDAY 30 SEPTEMBER 2009

Today we discussed independence for pairs of statements or events, sequences of independent trials for a repeatable experiment, and the binomial and multinomial distributions.

We say that statement A is *Independent* of statement B, provided that knowing B is true does not effect the probability that A is true, that is more precisely, A is independent of B if and only if

$$P(A|B) = P(A).$$

Since the multiplication rule tells us that always

$$P(A|B)P(B) = P(A \ \& \ B),$$

we see that A is independent of B if and only if

$$P(A \ \& \ B) = P(A)P(B).$$

But this last condition would be true if and only if B is independent of A, so we see that the following five statements all say the same thing:

$$A \textit{ is independent of } B,$$
$$P(A|B) = P(A),$$
$$P(A \ \& \ B) = P(A)P(B),$$
$$P(B|A) = P(B),$$
$$B \textit{ is independent of } A,$$

that is, if any one of the above five statements are true, then all are true. If any one of the above five statements is true for A and B, we say that A and B are *mutually independent*, and thus we know all five are true.

Suppose that we have a repeatable experiment and that we have say 3 mutually exclusive statements about the outcome for any one single trial of the experiment. For instance, suppose we are tossing a dice over and over, and

$$A = \{1, 2\}$$
$$B = \{3, 4\}$$
$$C = \{5, 6\},$$

so we know that each time the dice is tossed, the outcome is A or B or C. When we toss the dice 4 times, the outcome can be specified by a sequence of symbols or a "word". For instance, the word

$$ABAC$$

means that on for the first toss A is true, on the second toss B is true, on the third toss, A is true, and on the fourth toss C is true. Thus all the outcomes for the sequence of four tosses are four letter words using the alphabet A, B, C. Notice that the word $ABAC$ indicates in particular that A happened twice, but that is not the only outcome where A happens twice. If the outcome (sequence of outcomes) had been instead $AABC$ or $AABB$, it would again be the case that A happened twice. But if we ask for all outcomes where A happens twice and B happens once, we would have all possible four letter words obtained by rearranging the letters of the word $AABC$. We know that number is

$$C(4; 2, 1, 1) = C(4; 2, 1) = \frac{4!}{2!1!1!},$$

using the same counting method we used to count the number of ways to rearrange the letters in the word $MISSISSIPPI$. On the other hand, assuming the successive tosses are independent, we know that

$$P(AABC) = P(A)P(A)P(B)P(C) = [P(A)]^2[P(B)]^1[P(C)]^1.$$

In fact if $WXYZ$ is any word which results from rearranging the letters $AABC$, then

$$P(WXYZ) = P(W)P(X)P(Y)P(Z) = [P(A)]^2[P(B)]^1[P(C)]^1.$$

As a result, if T is the statement that A happened twice and B happened only once, which means that C must happen once, then

$$P(T) = C(4; 2, 1, 1)[P(A)]^2[P(B)]^1[P(C)]^1.$$

For more detail, let S be the set of all words which result from rearranging the letters in the word $AABC$. We then have

$$n(S) = C(4; 2, 1, 1).$$

For each $W \in S$, we have

$$P(W) = P(A)P(A)P(B)P(C) = [P(A)]^2[P(B)]^1[P(C)]^1,$$

whereas these words in S are all mutually exclusive outcomes. We therefore have

$$I_T = \Sigma_{W \in S} I_W,$$

so

$$\begin{aligned} P(T) = E(I_T) = \Sigma_{W \in S} E(I_W) = \Sigma_{W \in S} P(W) \\ = \Sigma_{W \in S}[P(A)]^2[P(B)]^1[P(C)]^1 \\ = C(4; 2, 1, 1)[P(A)]^2[P(B)]^1[P(C)]^1. \end{aligned}$$

In these situations of dealing with words, it is convenient to use the expression A^k to be the word formed by simply repeating the letter A sequentially k times, that is the $k-$letter word with only the letter A. Then we can include such an expression in a word, so $AABC = A^2BC$. Thus, all words in S are rearrangements of A^2BC.

More generally, if we have a repeatable experiment with possible outcomes of types $A_1, A_2, , , , A_m$ and if we perform n independent trials, we can ask for the probability that we have r_1 results of type A_1, and r_2 results of type A_2, and so on, so r_m results of type A_m. Let us call this outcome the statement $T(r_1, r_2, ..., r_m)$. Notice we must have

$$r_1 + r_2 + ... + r_m = n.$$

All the words representing outcomes for which $T(r_1, r_2, ..., r_m)$ is true are the words gotten by rearranging the letters of the single word

$$A_1^{r_1} A_2^{r_2} ... A_m^{r_m}.$$

We then have, using the same reasoning as before, that in general,

$$P(T(r_1, r_2, ..., r_m)) = C(n; r_1, r_2, ..., r_m)[P(A_1)]^{r_1}[P(A_2)]^{r_2}...[P(A_m))]^{r_m}.$$

Again, this is because the statement $T(r_1, r_2, ..., r_m)$ is simply that the outcome is any one of the $C(n; r_1, r_2, ..., r_m)$ words which can be obtained by rearranging the letters of $A_1^{r_1} A_2^{r_2} ... A_m^{r_m}$ and each has the same probability, namely

$$P(A_1^{r_1} A_2^{r_2} ... A_m^{r_m}) = [P(A_1)]^{r_1}[P(A_2)]^{r_2}...[P(A_m))]^{r_m}.$$

Of particular interest is the case where $n = 2$ and on each trial either A happens or it does not. In this case, the distribution giving the probability that A happens exactly k times in n trials is called the binomial distribution. We worked an example with the binomial distribution using the calculator's distribution menu.

For instance, if a policeman is hiding behind a tree with a radar gun watching 100 cars go by, and if in general 30 percent of the cars speed, then he should expect to find 30 speeders, but the chance of that is actually very small. There are 101 possibilities here for the number of cars he finds speeding: 0,1,2,3,...,100. If we ask for the chance that he finds 27 cars speeding, it would be

$$binompdf(100, .3, 27) = C(100; 27)(.3)^{27}(.7)^{73}.$$

The binomial distribution in the calculator is in the distribution menu which is the second function of the "VARS" button. Here is is useful to think of the number T that he will find speeding as an unknown, so we have $E(T) = 100(.3)$ for the expected value of T, and we say that T has the binomial distribution $binomial(100, .3)$ and loosely speaking, we say that T is binomially distributed here. More generally, if we have n independent trials with success probability p, on any one individual trial, then we say T, the total number of successes in n trials is binomially distributed or more precisely, has the distribution $binomial(n, p)$. In that case,

$$P(T = k) = binompdf(n, p, k) = C(n; k)p^k(1 - p)^{n-k}.$$

65. **LECTURE** FRIDAY 2 OCTOBER 2009

Today we discussed the hypergeometric and binomial distributions and how to tell which to use.

In general, to say we know the distribution of the unknown X is to say that no matter what two numbers a and b are given us, we can always give

$$P(a < X \leq b).$$

If X is a *count* such as the number of times heads comes up when a coin is tossed 30 times, then we know that X can only have whole number values and in this case, we know the distribution as soon as we know

$$P(X = k)$$

for every possible whole number k. For instance, in this case,

$$P(2 < X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5).$$

For the case of unknowns which are counts, we will deal mainly with the hypergeometric, the binomial, and the Poisson distributions. The Poisson distribution will be discussed after Test 2. For Test 2, you will need to know the hypergeometric and binomial distributions.

In general, the situation for these two counting distributions (binomial or hypergeometric) is the situation of repeated trials. There are two cases which predominate.

CASE A: you have some finite population (say a deck of cards or a box of blocks or a room full of people) and you are drawing repeatedly from the population WITHOUT REPLACEMENT. In this case, the successive draws are NOT INDEPENDENT OF EACH OTHER. If some property is counted out of what you drew (the number of diamonds, the number of red blocks, the number with high blood pressure), then the count has the HYPERGEOMETRIC DISTRIBUTION.

CASE B: you are drawing WITH REPLACEMENT from a finite population, or you have an effectively infinite population-a situation where you can repeat the experiment ad infinitum and on each trial probability of success is the same no matter what you know about the previous results. For example, tossing a dice over and over an counting the number of times an even number comes up, or tossing a coin over and over and counting the number of times heads comes up.

For instance, suppose that you have a box containing 20 blocks of which exactly 5 are RED. If we draw 10 blocks one after another, then the number R of red blocks we will get is an unknown before we actually draw the 10 blocks. Let R_k be the statement that k^{th} draw results in a red block. Clearly we know in either CASE A or CASE B, each time we draw a block we have a 25 percent chance it will be red,

$$P(R_k) = P(\textit{get red block}) = \frac{1}{4}.$$

Thus in either case we know that

$$E(R) = (10)(\frac{1}{4}) = \frac{5}{2} = 2.5.$$

However, it is obvious that in CASE A

$$P(2^{nd}Red|1^{st}Red) \neq P(2^{nd}Red) = \frac{1}{4},$$

whereas in CASE B it is obvious that

$$P(2^{nd}R|1^{st}) = P(2^{nd}R) = \frac{1}{4}.$$

That is we have

$$P(R_2|R_1) \neq P(R_2) = .25$$

in CASE A, whereas in CASE B we have

$$P(R_2|R_1) = P(R_2).$$

In CASE A, the result of the second draw is DEPENDENT on the result of the first draw, red blocks are getting use up each time one is drawn. If we see that the first 5 blocks drawn are red in CASE A, then we know it is impossible to ever draw another red block as they have all been used up. In CASE B, since blocks are being replaced after each draw, we could in fact end up with $R = 10$ even though there are only 5 blocks in the box. If we are tossing a dice and counting the number of times a 6 comes up, there is no limit to the number, then no matter how many times we get a 6, we know it is always possible to get a 6 on the next toss with the same probability as on the first toss. Of course, if we toss a dice 99 times and get 6 on every toss, then we are pretty sure the next toss will result in a 6, whereas if we had never tossed the dice before, we would assume a probability of 1/6 for the chance of getting a 6. The point is, that after we know the probability of tossing a 6 and are sure of it, then the probabilities do not change as we continue to toss the dice. The successive tosses are *independent* in this situation. In either the dice toss or the box of blocks situation, the essential property of CASE B is that as we are about to do each trial, the experimental set up is back to its original state *as far as our information is concerned*. To make this clearer, consider the experiment of drawing cards successively from a standard deck. In CASE A, the deck is initially shuffled, so we have no idea where any specific cards are in the stack. In CASE B, each time a card is drawn, we record the result, put the card back in the deck and reshuffle the deck. Now, the deck changes its physical state after each reshuffle, but our information is the same after each reshuffle-we have no idea where any specific card is.

If you understood the method of calculating probabilities for Poker hands which we went over in a previous lecture, then calculating the probabilities in the case of the hypergeometric distribution is easy. For instance, in the case of the box of 20 blocks of which 5 are red, if we draw 10 without replacement, we know we expect 2.5 red blocks. If we ask for the probability that $R = 3$, we have to calculate the total number of ways to choose 10 blocks from the box, $C(20; 10)$, and calculate the total number of ways to draw 10 blocks so as to get exactly 3 red blocks, which is $C(5; 3)C(15; 7)$, since to get exactly 3 red means to choose three red and then choose 7 not red. Thus, in CASE A,

$$P(R = 3) = \frac{C(5; 3)C(15; 7)}{C(20; 10)}.$$

More generally, suppose we have a finite population of size N from which we draw n things, and suppose we count the number X which has some specified property S. Suppose that we know that there are exactly M of the things in the whole population which have the property S. In CASE A, we draw without replacement and

$$P(X = k) = \frac{C(M; k)C(N - M; n - k)}{C(N; n)}.$$

We call this the formula for the HYPERGEOMETRIC DISTRIBUTION. Notice the numerator is the number of ways to actually choose the n objects so as to get exactly k with property S. For to do this, you must step one choose k of the M things which have property S and step two choose the remaining $n - k$ things from the $N - M$ things in the population which do not have property S. In general, the things with the property we count are called successes. Thus M here is the size of the population of successes and $N - M$ is therefore the size of the population of failures.

Now lets consider CASE B. The successive draws are done with replacement, or there are so many blocks in the box and so many red blocks in the box that for all practical purposes no matter how many red blocks we draw, there is no significant difference in the probability the next will be red, or the situation is like tossing the dice or flipping a coin. All trials are

independent of each other, so each outcome can be specified as a word using two symbols S and F, here S stands for success and F for failure. Thus,

$$SSFFSSFFFF$$

in the case of the block drawing with CASE B, that the first two blocks were red, the next two were not red, the next two were red and the last four were not red. No matter how these symbols are ordered, the probability of such an outcome is simply

$$P(S)^4P(F)^6.$$

But if we ask for $P(X = 4)$, then we have to realize that there are many such sequences which end up with the result that we got exactly 4 successes. The number of such sequences is simply the number of ways to arrange these symbols to make a 10 letter word with exactly 4 S's. This we know is simply $C(10; 4)$. Thus,

$$P(X = 4) = C(10; 4)P(S)^4P(F)^6.$$

In general, we write

$$p = P(S)$$

and call this the *success probability* or the *success rate*. It is the probability of success on a single trial, in either CASE A or CASE B. However, we now see that if we put

$$q = P(F) = 1 - P(S) = 1 - p,$$

then q is the *failure rate* or *failure probability* and in n trials,

$$P(X = k) = C(n; k)p^kq^{n-k}.$$

This is the formula for the BINOMIAL DISTRIBUTION. This distribution is in your calculator in the distribution menu as discussed in the last lecture. When you see *pdf* on the end of the name of the distribution for a counting unknown, in your calculator, it stands for *probability distribution function* and means that it gives the probabilities $P(X = k)$ for any value of k you enter in proper format. When you see *cdf* it stands for *cumulative distribution function* and gives the probabilities $P(X \leq k)$ for any value of k you enter in proper format. Thus, in CASE B, the format is

$$P(X = k) = \mathit{binompdf}(n, p, k)$$

and

$$P(X \leq k) = \mathit{binomcdf}(n, p, k).$$

Be careful to notice that for the case of a counting unknown,

$$P(X < k) = P(X \leq k - 1) = \mathit{binomcdf}(n, p, k - 1).$$

For counting unknowns, "less than" and "less than or equal to" are very different.

Finally, we discussed the example of airline overbooking. If an airline knows that 90 percent of the people show up for their reservations and they do not want a lot of empty seats flying around, then they will overbook and hope that there are enough seats for those that show up. If there are 300 seats on a plane and 325 reservations, then there is a problem if more than 300 show up. We want to know the probability of this problem. Let X be the number of people who show up. It is reasonable that as far as the airline knows, all the reservations are independent of each other in the sense that if the twenty third person on the list shows up, we still no nothing more about the others on the list. Therefore the probability of a problem can be calculated in two ways and either gives the same result-just stay consistent with your point of view. The first way is the probability we want is

$$P(X > 300) = 1 - P(X \leq 300) = 1 - \mathit{binomcdf}(325, .9, 300).$$

The other way is to think in terms of the failures. If Y is the count of failures, then the failure rate is $q = .1$, that is alternately, we think of failure as success and success as failure. We are now interested in the probability

$$P(Y < 25) = P(Y \leq 29) = \mathit{binomcdf}(325, .1, 29).$$

If you do the calculations, you find the same answer either way, which is of course merely a reflection of the fact that probability is a logically consistent theory.

66. **LECTURE** MONDAY 5 OCTOBER 2009

Today we discussed problems on the practice test related to counting and reviewed for Test 2. We noted that using Pascal's Triangle, the numbers $C(n;r)$ for small n can be quickly calculated because of the formula

$$C(n;r) + C(n;r+1) = C(n+1;r+1).$$

To see why this must always be true, imagine that we have a box containing 101 blocks of which 100 are white blocks and just one block is red. Imagine we want to draw 38 blocks without replacement from this box. Notice that if you draw 38 blocks from this box, then either you did on did not get the red block. Thus the total number of ways to do this job is the number of ways where you do get the red block added to the number of ways where you do not get the red block. To get 38 blocks so as to get the red block, step 1 choose the red block (only one way-there is only one red block, $C(1;1) = 1$) and step two, choose the other 37 blocks from the remaining 100 white blocks, $C(100;37)$ ways. Thus there are

$$C(1;1)C(100;37) = C(100;37)$$

ways to get 38 blocks out of the box so that one of them is the red block. On the other hand, if asked to get 38 blocks from the box so as to not get the red block, then all 38 blocks must be taken from the 100 white blocks. Thus

$$C(100;37) + C(100;38) = C(101;38).$$

There is obviously nothing special about the numbers 100 and 37 here, so in general, if we have $n+1$ blocks in the box of which n are white and one is red, and if we draw $r+1$ blocks without replacement, then we either do or do not get the red block, so

$$C(n;r) + C(n;r+1) = C(n+1;r+1).$$

Obviously, the number of ways to choose nothing from the empty set is 1, so $C(0;0) = 1$. In fact, it is obvious that for any n we must have

$$C(n;0) = 1 = C(n;n),$$

since there is only one way to choose nothing and there is only one way to take everything. We also know that

$$C(n;1) = n = C(n;n-1),$$

since there are n ways to choose a single thing from a set of n things, and there are n ways to leave one thing behind taking all but one thing which means n ways of choosing $n-1$ things from the n things. Thus, if we arrange the numbers $C(n;r)$ for fixed n all in a horizontal row, as

$$C(n;0) \;\; C(n;1) \;\; C(n;2) \;\; ... \;\; C(n;n-2) \;\; C(n;n-1) \;\; C(n;n),$$

then for sure we will see

$$1 \;\; n \;\; C(n;2) \;\; ... \;\; C(n;n-2) \;\; n \;\; 1.$$

We can then use the formula for Pascal's Triangle

$$C(n;r) + C(n;r+1) = C(n+1;r+1)$$

to fill in the terms in the line for choosing from $n+1$ things giving the numbers $C(n+1;r)$ from the line for choosing from n things.

We start off with the line for choosing from nothing, $n = 0$. Of course it only has the single number $1 = C(0;0)$. We then fill in below the lines for higher values of n using the formula for Pascal's Triangle:

$$
\begin{array}{c}
1 \\
1\ \ 1 \\
1\ \ 2\ \ 1 \\
1\ \ 3\ \ 3\ \ 1 \\
1\ \ 4\ \ 6\ \ 4\ \ 1 \\
1\ \ 5\ \ 10\ 10\ \ 5\ \ 1 \\
1\ \ 6\ \ 15\ \ 20\ \ 15\ \ 6\ \ 1 \\
1\ \ 7\ \ 21\ \ 35\ 35\ \ 21\ \ 7\ \ 1 \\
1\ \ 8\ \ 28\ \ 56\ 70\ 56\ \ 28\ \ 8\ \ 1 \\
1\ \ 9\ \ 36\ 84\ 126\ 126\ 84\ 36\ \ 9\ \ 1
\end{array}
$$

Each number in the array is the sum of the two numbers above closest to it, so it can easily be written down as shown, and it provides the numbers $C(n;r)$ quickly for $n \leq 9$.

67. **LECTURE** WEDNESDAY 7 OCTOBER 2009

Today we had TEST 2 in lecture class.

68. **LECTURE** FRIDAY 9 OCTOBER 2009

Today we discussed the Poisson distribution. We began by discussing the difference between discrete and continuous quantities. For us, the difference is simple. If you are dealing with something you count, then its a discrete quantity, whereas if you are dealing with something you have to measure with some form of measuring device, then it is a continuous quantity. Thus, height, weight, blood pressure, volume, area, length, time, are all continuous quantities.

We have already dealt with continuous unknowns, but we have not dealt with the specific technicalities of their distributions, whereas we have done so for two discrete counting distributions: the hypergeometric and the binomial, which in reality are whole infinite families of distributions. In particular, an important parameter of these distributions is the sample size, which is the number of trials, denoted n. For these distributions, it is obviously a discrete quantity. It tells the number of things examined where we were counting successes. However, sometimes we count successes when the amount we examine must be specified by a continuous quantity. For instance, we could stand at a trolley stop for two hours and count the number of trolleys that arrive during that time. The unknown we are observing here is discrete, but the sample size is now a continuous quantity. This tells us we cannot be dealing with either the binomial or hypergeometric distributions. In this situation, we are dealing with the Poisson distribution. Here, we must assume that disjoint intervals of time are independent of one another. That is, we assume that if we watch from 5pm to 7pm, and if we see 3 trolleys arrive between 5pm and 5:15 pm, then that does not give us any help in guessing what will happen between 5:15pm and 5:45pm, as the two intervals of time are disjoint. Notice the two time intervals actually do have 5:15pm in common, but we shall see that for continuous quantities, a single point is negligible.

For the Poisson distribution, the only parameter is the expected value, μ. Thus, if X is the number of tadpoles in a gallon of water taken from a pond in the swamp, and if we assume that the numbers of tadpoles in disjoint volumes of water are independent of each other, and if we expect 7.3 tadpoles per gallon "on average", then the entire Poisson distribution is determined by the number 7.3. If we examine a particular gallon of pond water here, the probability of finding 6 tadpoles is

$$P(X = 6) = \frac{(7.3)^6 e^{-7.3}}{6!} = poissonpdf(7.3, 6).$$

In general, if X is any unknown count having the Poisson distribution with mean μ, then

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!} = poisonpdf(\mu, k).$$

We note the fact that we can make sense of adding up an infinite number of numbers here (summing a series) and that

$$e^{\mu} = \sum_{k}^{\infty} \frac{\mu^k}{k!}.$$

Notice that for the Poisson unknown with mean μ, the probability $P(X = k)$ is simply the result of dividing the k^{th} term of the series by e^{μ}. Since the sum of the series is e^{μ}, this means that the sum of the probabilities must be 1. That is,

$$\sum_{k}^{\infty} P(X = k) = 1.$$

Of course, you can notice right away that if we have any sequence of positive numbers

$$b_1, b_2, b_3, ...$$

and if the series has a sum, say

$$S = \sum_k^\infty b_k,$$

then we can form a distribution for a counting unknown by setting

$$P(X = k) = \frac{b_k}{S},$$

since then all the probabilities are guaranteed to add up to 1. However, this procedure usually does not produce anything practical.

If we need to know $P(X \leq k)$, then we calculate using the cdf in our calculator. Thus, for the Poisson distribution,

$$P(X \leq k) = \sum_{i \leq k} P(X = i) = \mathit{poissoncdf}(\mu, k).$$

We also noticed that if X is governed by the Poisson distribution with mean μ, then it can be easily rescaled to deal with different sample sizes. For instance, is we expect $\mu = 6$ trolleys per hour on average at the trolley stop, then we expect 3 in a half hour and twelve during two hours. Thus if asked for the probability that 4 trolleys arrive between 5pm and 5:30pm, the answer is

$$P(X = 4) = \mathit{poissonpdf}(3, 4).$$

If asked the probability that 2 trolleys arrive between 5pm and 5:15pm, we note that during one quarter hour we expect $6/4 = 1.5$ trolleys to arrive, so the answer is

$$P(X = 2) = \mathit{poissonpdf}(1.5, 2).$$

We can notice in the trolley example, that during a single minute we expect $\mu = 1/10$ trolleys to arrive. We can view this as being equivalent to having a 10 percent chance of seeing a whole trolley during any given single minute. Since all the successive single minutes are independent of each other, we might guess that it is reasonable to try using the binomial distribution. To do this, we regard the hour as consisting of 60 independent trials, each having a ten percent chance of success. In fact, if we do this we find that for k near 6,

$$\mathit{binompdf}(60, .1, k) =^{ap}= \mathit{poissonpdf}(6, k).$$

Here I use

$$=^{ap}=$$

to denote "approximately equal". If we replace the one minute intervals by six second intervals, then there are 600 disjoint 6 second intervals making up the full hour, and in each we expect $\mu = 1/100$ trolleys to arrive or alternately, we can think that there is a one percent chance of a whole trolley arriving in a give 6 second interval. In this case, we find that the approximation

$$\mathit{poissonpdf}(6, k) =^{ap}= \mathit{binompdf}(600, 1/100, k)$$

is much more accurate. In fact, for any Poisson unknown X with mean μ is what is expected per unit size sample, we can regard the unit as a disjoint union of n smaller samples of size $1/n$. Then in each we expect the count to be μ/n. When we choose n so large that μ/n is small and much less than 1, we can regard this as saying that μ/n is the probability the count will be exactly 1 for that small sample examined, and the unit sample becomes n independent trials for getting a whole to appear. We then find that as n gets very very large, the approximation can be made extremely accurate

$$P(X = k) = \mathit{poisson}(\mu, k) = \mathit{binompdf}(n, \mu/n, k),$$

for k near μ. How near k has to be to μ will depend on how big n is taken to be. In fact, it can be shown precisely that for any k, we have

$$\lim_{n\to\infty} poisson(\mu, k) = binompdf(n, \mu/n, k).$$

The last topic of the lecture was the use of the Poisson distribution to determine probabilities for a continuous unknown, the *waiting time*. For instance in the trolley example, we can ask what is the probability that I have to wait more than 15 minutes for a trolley. Notice that is exactly the same as saying the for the first fifteen minutes zero trolleys arrive. If trolleys arrive at 6 per hour, then in fifteen minutes I expect 1.5 trolleys, so if W is the waiting time, we have

$$P(W > 15) = poisson(1.5, 0).$$

In general, if $t \geq 0$, then during t hours, we expect μt trolleys to arrive, so the probability we must wait longer than than time t is

$$P(W > t) = poisson(\mu t, 0).$$

For instance, in the example we have $t = 1/4$ hour, so $\mu t = 6 * 1/4 = 1.5$. We can also recall here that

$$poissonpdf(\mu, k) = \frac{\mu^k e^{-\mu}}{k!},$$

so this means that

$$P(W > t) = \frac{(\mu t)^0 e^{-\mu t}}{0!} = e^{-\mu t}.$$

Thus,

$$P(W \leq t) = 1 - P(W > t) = 1 - e^{-\mu t}.$$

69. **LECTURE** MONDAY 12 OCTOBER 2009

Today we began by reviewing the difference between continuous and discrete quantities. In practical terms, to determine the value of a **discrete quantity** we generally have to **count**, whereas to determine the value of a **continuous quantity** we generally have to **measure**. We then reviewed the counting distributions including the Poisson distribution from the last lecture. We noted that of the three counting distributions, only the Poisson distribution has the sample size specified as a continuous quantity. The amount of "stuff" examined to determine the count has a continuous measure. For instance, if we watch a trolley stop for 2.3 hours and count the number of trolleys that arrive, then even though the unknown here is discrete, the sample size is the amount of time we watch which must be measured. For the hypergeometric and binomial distributions, there is always a number of trials, denoted by n since it is the sample size. Thus, to determine which distribution to use, if the question asks the probability of a certain number of successes, and if there is a continuous amount examined, then we would check that the assumptions of the Poisson distribution are in effect. If the sample size or amount examined for successes must be a whole number of trials, then we either have the hypergeometric or binomial distributions. In that case, if we are drawing without replacement from a finite population, then the distribution is hypergeometric, otherwise, we would check to see that the trials are independent, in which case the distribution of the count is binomial.

We reviewed the computations of Poisson probabilities and reviewed the fact that the waiting time distribution can be calculated using the Poisson distribution. That is, if we expect μ per unit time, and if W is the time we will have to wait, then

$$P(W > t) = poissonpdf(\mu t, 0) = e^{-\mu t},$$

and therefore

$$P(W \leq t) = 1 - e^{-\mu t}.$$

We discussed the idea of representing distributions with pictures. In the case of a discrete counting distribution, the picture is naturally formed by putting a horizontal axis with the possible values of the count equally spaced on the horizontal axis, and then using a chosen vertical scale, we draw a spike over each possible value whose height is the probability of that value.

In the case of a continuous unknown X, it is generally the case that there are whole continuous ranges of possible values. For instance, if we deal with blood pressure, and if you expect my blood pressure is 120, and if x is some number between 100 and 140, you probably would not be able to rule that out as being my blood pressure before you measure my blood pressure. Again, we represent the possible values on a horizontal axis, but now we cannot restrict to whole numbers. We deal with this pictorially by imagining that there is a continuous curve above the horizontal axis which has the property that the area under the curve between limits a and b, that is the area trapped under the curve and above the horizontal axis between the vertical lines through $x = a$ and $x = b$, actually gives the probability that X is between a and b. We call the function f_X whose graph gives this curve the **Probability Density Function** for X. Thus, in the case of my blood pressure, if you had a picture of the curve representing the probability density function for my blood pressure, then the area under that curve between $x = 100$ and $x = 140$ would give the probability that my blood pressure will turn out to be in that range. In general here, then the total area under the whole probability density curve between limits $-\infty$ and ∞ must be equal to 1, since whatever the value of X is, we know it satisfies

$$-\infty < X < \infty,$$

for sure.

In general, if X is any unknown, it could be a mixture of continuous and discrete. Its picture would have a density function as well as a possible countable infinity of spikes. If A is the total

area under the density curve and T is the total of the heights of all the spikes, then we have the constraint

$$A + T = 1.$$

Here, if two numbers a and b are given, then to find $P(a < X \leq b)$, we compute the area under the density curve between these two limits and as well we add up the total heights of all the spikes we find in that range, and the sum is the probability we are looking for. However, in practical problems, we almost never run into the situation where we have this mixture of continuous and discrete. Thus, in practical terms, unknowns are usually continuous (no spikes in the picture) or else discrete (only spikes in the picture).

The first continuous distribution we encountered above was that for the waiting time. We saw that if W is the waiting time, then when we expect μ successes in a sample size of one unit (of time or volume or length or area and so on), we can calculate using the equation

$$P(W > t) = 1 - e^{-\mu t}$$

and

$$P(W \leq t) = 1 - e^{-\mu t}.$$

If we look at the graph of $y = e^{-\mu t}$ as t varies from 0 to ∞, we see that it is continually decreasing, getting to zero at ∞, so the probability of waiting forever is zero. Likewise, we see that as t increases therefore the probability of waiting no more than time t is increasing, getting to 1 at ∞. If we ask for the probability that your waiting time is between the limits a and b, with $a < b$, then that must be

$$P(a < W \leq b) = P(W \leq b) - P(W \leq a) = 1 - e^{-\mu b} - [1 - e^{-\mu a}] = e^{-\mu a} - e^{\mu b}.$$

We therefore have a way of calculating all probabilities for the waiting time unknown, so we have the entire distribution.

The first thing to notice about a continuous unknown X is that if v is any real number, the probability that X takes the exact value v must be zero. For this would be the area between the limits $x = v$ and $x = v$. There is no area between. It is as if our method is telling us to find the area of the spike whose height is the height of the density curve over the precise point v on the horizontal axis-there is no area of in a geometric line segment such as a spike. Thus, the probability my blood pressure is exactly 120 is zero!! How can we reconcile this fact with the idea that we expected my blood pressure to be 120, and when the measurement is carried out, the result is actually a number. The answer here is that whenever we observe a continuous unknown, we must always measure and measurement always has limits to accuracy. Thus, in any application of a measuring device there is always a specified level of accuracy for the measurement. If we want to measure my blood pressure to 2 decimal place accuracy and if the value reported is 123.46, then that really means that my blood pressure was somewhere between 123.455 and 123.465, and that is a continuous range of possibilities. So, there will be positive area trapped under the density curve between these two limits, and thus there will be a positive probability of such an observed value. In general then if X is any unknown and x is any real number,

$$P(X = x,\ \textit{exactly}) = 0,$$

but if x is a two decimal place number, then very possibly

$$P(X = x,\ \textit{to two decimal place accuracy}) > 0.$$

We finally discussed the normal distribution. If X is normal with mean μ and standard deviation σ, then

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

This probably looks very mysterious at first glance, but the fraction in front of the e is just a normalization factor to make the total area under the density curve equal to 1, which we know

must be the case for any density curve. We can next notice that

$$\frac{x-\mu}{\sigma} = z,$$

the standard score. Thus, if we put $A = \sigma\sqrt{2\pi}$, then the density for the normal distribution becomes simply

$$f_X(x) = \frac{1}{A}e^{-z^2/2}.$$

Evidently, A must be the area under the curve

$$y = e^{-z^2/2}$$

where z is dependent on x. This means that the probabilities are really only depending on the standard scores, so we could standardize everything and then calculate probabilities. If Z is a standard unknown, it has mean zero and standard deviation one. Thus we have

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} = \frac{1}{A}e^{-z^2/2},$$

that is here the area under the curve

$$y = e^{-z^2/2}$$

as a function of z is just $A = \sqrt{2\pi}$, and the probabilities are determined by simply converting everything to standard scores. Thus, if $\mu_X = 120$ and $\sigma_X = 15$, then the probability that X is within one standard deviation of the mean is $P(105 < X < 135)$ whereas the probability that X is within two standard deviations of the mean is $P(90 < X < 150)$. We then have

$$P(105 < X < 135) = P(-1 < Z < 1)$$

and

$$P(90 < X < 150) = P(-2 < Z < 2).$$

If X is normally distributed, then we can calculate the probability density curve using the distribution menu

$$f_X(x) = \mathit{normalpdf}(x, \mu, \sigma).$$

To actually calculate a probability, we must calculate an area under the curve between two given limits. Thus,

$$P(a < X \leq b) = P(\leq X \leq b) = \mathit{normalcdf}(a, b, \mu, \sigma).$$

You can check for instance that

$$\mathit{normalcdf}(105, 135, 120, 15) = \mathit{normalcdf}(-1, 1, 0, 1)$$

with a value of about 68 percent. This tells us that about 68 percent of any normal population is within one standard deviation from the mean. You can also check that

$$\mathit{normalcdf}(90, 150, 120, 15) = \mathit{nomalcdf}(-2, 2, 0, 1)$$

with a value of about 95 percent. This tells us that about 95 percent of any normal population is within 2 standard deviations of the mean. Likewise,

$$\mathit{normalcdf}(75, 165, 120, 15) = \mathit{normalcdf}(-3, 3, 0, 1)$$

with a value of about 99.7 percent, which tells us that about 99.7 percent of any normal population is within 99.7 percent of the mean. If you are only dealing with percentages to the nearest whole percent, then 99.7 rounds off to 100 percent. Thus, in many practical situations we can think of the whole normal population as lying within 3 standard deviations of the true mean. When we come across an observation that is more than 3 standard deviations from the true mean, we should either think of it as remarkable or possibly caused by some problem. Thus, if someone's blood pressure is three standard deviations above the mean, they probably need immediate medical attention.

70. **LECTURE** WEDNESDAY 14 OCTOBER 2009

Today we continued discussing the normal distribution. We began recalling the distribution $f = f_X$ for a normal unknown with mean μ and standard deviation σ is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2},\ z = \frac{x-\mu}{\sigma},$$

so z is simply the standard score for x. We looked at the graph of

$$Y_1 = e^{-.5X^2}$$

using the graphing capability of the calculator and noted that geometrically it has the same shape as the normal distribution, or what in everyday language is called the "bell curve".

Next, we did some calculations of probabilities with the normal distribution. Remember,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = normalcdf(a, b, \mu, \sigma),$$

for any two numbers a and b with $a \leq b$. Also remember, for any a we have

$$P(X = a\ exactly) = 0$$

and therefore also

$$P(X = b\ exactly) = 0,$$

which accounts for why all the different inequality conditions above have the same probability.

As a matter of fact, since the calculator calculates areas under the distribution by integration, and as integration from right to left is the negative of integration from left to right, we can find that always, for any two numbers a and b it is true that

$$normalcdf(a, b, \mu, \sigma) = -normalcdf(b, a, \mu, \sigma).$$

This means that if you accidentally put the larger number in before the smaller number, the answer will be a negative number. But we know probability can never be negative, so if you get a negative answer, the first place to look for your mistake here is to check you put the numbers in correctly. We will see that in certain situations, it is useful to put the larger number first.

Suppose we ask for the probability of being below average in a normal population. By symmetry of the distribution, we see that must be 50 percent, no calculation required. What is the chance of being exactly average? That has to be ZERO! Remember, the probability of any value happening exactly is zero. Likewise, then, the probability of being above average is exactly 50 percent.

Sometimes we want to know a probability such as the probability of being less than a specific number x. This is apparently the area under the distribution curve from x all the way left to $-\infty$. We call such a region under the curve a *tail* and its area is called a *tail area*. More specifically, since the region extends to the left all the way to negative infinity, we call it a *left tail*. There is no way to enter $-\infty$ into the calculator. Using any very very large number with a negative sign in front will give an approximate answer. To get the most accurate answer using the calculator, we can notice that in case $\mu \leq x$, we just have

$$P(X \leq x) = P(X \leq \mu) + P(\mu < X \leq x) = .5 + normalcdf(\mu, x, \mu, \sigma).$$

In case $x < \mu$ we want to subtract the the area between x and μ from $1/2$ and this is the same as adding with the limits entered in reverse order, so we have

$$P(X < x) = P(X \leq x) = P(X \leq \mu) - P(x < X \leq \mu)$$

$$= .5 - normalcdf(x, \mu, \mu, \sigma) = .5 + nomralcdf(\mu, x, \mu, \sigma).$$

Notice this means that we have in either case, no matter whether or not x exceeds μ that we can always calculate the left tail area as

$$P(X \leq x) = .5 + normalcdf(\mu, x, \mu, \sigma).$$

You can think of this as being the result of getting from $-\infty$ to x by going through μ. The $1/2$ gets you from $-\infty$ to μ and the $normalcdf(\mu, x, \mu, \sigma)$ gets you from μ to x and the computer does not care where x is in relation to μ.

In case of a right tail, that is a region described by $X \geq x$, we calculate its area in case $x < \mu$ as

$$P(X > x) = P(X \geq x) = P(X > \mu) + P(\mu \leq X \leq x)$$
$$= .5 + normalcdf(x, \mu, \mu, \sigma) = .5 - normalcdf(\mu, x, \mu, \sigma).$$

If $x \geq \mu$, then we want to subtract the area between μ and x from $1/2$, so we get

$$P(X \geq x) = .5 - normalcdf(\mu, x, \mu, \sigma),$$

which is again the same result. Thus, if we call The right tail area A_+ and the left tail area A_-, then we have

$$A_\pm = .5 \mp normalcdf(\mu, x, \mu, \sigma),$$

in general.

Another type of problem we often encounter is the problem of calculating a probability of being within a certain distance d from a specific number c. To say the distance from X to c is less than d is the same as saying

$$|X - c| < d$$

which is also the same as saying

$$c - d < X < c + d.$$

Therefore,

$$P(|X - c| < d) = normalcdf(c - d, c + d, \mu, \sigma).$$

Finally we have problems where the information effectively gives us an area under the distribution curve and we want the boundaries of the region. For instance, if we know μ and σ for a specific normal population, we might want to know the score x for which 90 percent is below x and only 10 percent is above x. This means we want to solve the equation

$$P(X < x) = .9.$$

Notice this is going backwards, we have the probability, we want to know the boundary score. Whenever we go backwards in mathematics we call it inversion, and here we use the inverse normal in the calculator. Thus, we have

$$x = invNorm(.9, \mu, \sigma).$$

In general, if we want to know the cut off score for a given left tail area A, the score is x given by

$$x = invNorm(A, \mu, \sigma).$$

We refer to these scores as centile scores in every day language. Thus, $invNorm(.9, \mu, \sigma)$ is the 90^{th} percentile score. If you beat this score, you are in the top 10 percent. If you do not beat this score, you are in the bottom 90 percent. We would thus also call this score the upper 10 percentile score.

If we want to know the score x for which $P(-x < X < x) = .8$, then we notice that from symmetry

$$.2 = P(not\ [-x < X < x]) = P(X < -x) + P(X > x) = 2P(X < -x) = 2P(X > x).$$

Thus $P(X > x) = .1$ and $P(X \leq x) = 1 - .1 = .9$. We therefore have again $x = invNorm(.9, \mu, \sigma)$. Notice that .9 is half way from .8 to 1. In general, if we want x so that

$$P(-x < X < x) = A,$$

then

$$P(X < -x) = P(X > x) = \frac{1 - A}{2},$$

so

$$P(X < x) = 1 - \frac{1-A}{2} = A + \frac{1-A}{2} = \frac{1+A}{2}.$$

This means

$$x = invNorm(\frac{1+A}{2}, \mu, \sigma).$$

Here, notice that $(1 + A)/2$ is the average of A with 1, or what is the same thing, we can see that $(1 + A)/2$ is half way from A to 1.

71. **LECTURE** FRIDAY 16 OCTOBER 2009

NO LECTURE TODAY BECAUSE OF FALL BREAK

72. **LECTURE** MONDAY 19 OCTOBER 2009

Today we discussed Thebeychev's inequality, sampling distributions, and the *CENTRAL LIMIT THEOREM*. We noticed that if we are sampling the unknown X, then we are actually creating a sequence of new unknowns

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

where n is the sample size and X_k is the future value of the k^{th} observation. Then we form the *sample total* denoted T_n given by

$$T_n = X_1 + X_2 + ... + X_n$$

and the *sample mean* denoted $\bar{X}_n$ and given by

$$\bar{X_n} = \frac{1}{n} T_n.$$

We showed that

$$E(T_n) = n\mu_X$$

and

$$E(\bar{X}_n) = \mu_X,$$

so in particular, whenever we take a sample, our best guess in advance of looking at the data is that the sample mean will turn out to be the true mean.

Notice that as the sample observations are all really observations of X, they all have the same distribution as X and therefore in particular, they all have the same mean and standard deviation. For instance, just to take an example, suppose we have a population of tuna fish. Let X be the weight of a tuna fish selected from this population and which is lying out in the parking lot, where you cannot see it. If you think there is a 30 percent chance that this tuna fish weighs between 300 and 325 pounds, then the same could be said of the weight of the third tuna fish in our sample, which is X_3. After all, the only thing we know about these tuna fish is that they both come from the same population. Thus, whatever we know about probabilities of various values of X can equally well be applied to X_3 or any other X_k, with $1 \leq k \leq n$.

This means all the sample observations $X_1, X_2, ..., X_n$ must all have the same distribution as X and therefore in particular, they must all have the same mean as X, which is μ_X, and they must all have the same standard deviation as X, namely σ_X. That is, to emphasize the point, for every k,

$$\mu_{X_k} = E(X_k) = E(X) = \mu_X,$$

and, as well,

$$\sigma_{X_k} = \sigma_X,$$

or what is the same thing, by squaring both sides of this last equation,

$$Var(X_k) = Var(X).$$

To see why these equations for the expected value of the sample total and the expected value of the sample mean are true, we first use the Addition Rule of expectation as applied to T_n. We have

$$E(T_n) = E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n) = \mu_{X_1} + \mu_{X_2} + ... + \mu_{X_n} = n\mu_X,$$

and therefore, as far as the sample total is concerned,

$$E(T_n) = n\mu_X.$$

But then for the sample mean $\bar{X}_n$ we have

$$E(\bar{X}_n) = E(\frac{1}{n} T_n) = \frac{1}{n} E(T_n) = \frac{1}{n} n\mu_X = \mu_X,$$

and therefore,

$$E(\bar{X}_n) = \mu_X.$$

Notice these equations say for instance, that if the mean weight of tuna fish is 300 pounds, and if I know this then I would guess any sample of these tuna fish would have mean weight 300 pounds, before I actually see the data. If I know 10 of these fish are in the back of my friends pickup truck, then I would guess the load is 3000 pounds. Of course, typically we know that when it comes to unknowns, we often do not get what we guess or expect, and to guess how far off our guess would be from reality depends on standard deviations. Thus we need to also find the standard deviations for T_n and $\bar{X}_n$. To do this we need assumptions on covariances of the various observations. The simplest assumption is that they are all zero which is the case if they are all independent of each other.

We observed that for this simplest assumption, *Independent Random Sampling* (IRS), that is if all the sample observation unknowns $X_1, ..., X_n$ are pairwise independent of each other, then

$$\sigma_{T_n} = \sqrt{n}\ \sigma_X,$$

and

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

To see why the equations for standard deviation are true, under the assumption of independent random sampling, we know that all the covariances between the various observations are zero. Recall that if X and Y are independent variables, then $Cov(X, Y) = 0$ and therefore

$$Var(X + Y) = Var(X) + Var(Y).$$

Applied to our sample total, the assumption that all these sample observations are independent means that

$$Var(T_n) = Var(X_1) + Var(X_2) + ... + Var(X_n) = nVar(X)$$

or

$$Var(T_n) = nVar(X),$$

so taking square roots of both sides gives

$$\sigma_{T_n} = \sqrt{n}\sigma_X.$$

As far as the standard deviation of the sample mean $\bar{X}_n$ is concerned, it is often referred to as the *Standard Error of the Mean*, and we have

$$\sigma_{\bar{X}_n} = \sigma_{(1/n)T_n} = \frac{1}{n}\sigma_{T_n} = \frac{1}{n}\sqrt{n}\sigma_X = \frac{1}{\sqrt{n}}\sigma_X,$$

so finally we have

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

The Central Limit Theorem says that as n, the sample size, tends to infinity, the distributions for T_n and for $\bar{X}_n$ become normal. In practice, this means that if $n \geq 30$, we will assume these distributions are normal.

Tchebeychev's inequality says that for any unknown X and for any positive number k, it is always true that

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}.$$

The amazing thing about this inequality is that it is always true, we need not make any assumptions about X and combined with our above facts about sampling distributions with IRS, it tells us that if we take a large enough sample we will be very likely to find our sample mean is very close to the true population mean. In fact, it means that no matter how close we need to be to the true population mean, and no matter how sure we need to be of the result, then we can insure this by making the sample size large enough. However, the Central Limit Theorem comes in to play in sampling situations much more strongly than Tchebeychev's inequality, and guarantees that sample sizes much smaller than required by Tchebeychev's inequality will suffice

to give accurate estimates of population means. However, the proof of central limit theorem, though possible with our level of theory is beyond the level of difficulty we will deal with.

The proof of Tchebeychev's Inequality, on the other hand, is actually very easy and uses only the basic definitions, so we will give that proof here. Remember that

$$\sigma_X^2 = E((X - \mu_X)^2),$$

by definition, and that

$$|X - \mu_X| \geq k\sigma_X$$

if and only if

$$(X - \mu_X)^2 \geq k^2\sigma_X^2.$$

Let A stand for the statement that $|X - \mu_X| \geq k\sigma_X$. Thus, in terms of A, Tchebeychev's Inequality says merely

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

Now A is equivalent to

$$(X - \mu_X)^2 \geq k^2\sigma_X^2.$$

And as A is a statement, it is either true or false, so let I_A be the indicator of A. Thus, I_A is an unknown, and it is either one or zero. It is one if A is true and zero if A is false. Now, we can form the unknown

$$k^2\sigma_X^2 I_A,$$

and we can ask how it compares to

$$(X - \mu_X)^2.$$

That is, we consider the inequality between unknowns

$$(X - \mu_X)^2 \geq k^2\sigma_X^2 I_A.$$

We will now see that the above inequality is actually true. After all, if A is false, then the right hand side is zero and the square of any number is at least zero, whereas if A is true, then the right hand side of the inequality is $k^2\sigma_X^2$ but when A is true, this is less than or equal to the left hand side, by definition of what statement A says. The inequality has to be true whether or not A is!! We therefore know from our basic rules of expectation, that

$$E((X - \mu_X)^2) \geq E(k^2\sigma_X^2 I_A) = k^2\sigma_X^2 E(I_A).$$

But by definition of probability,

$$P(A) = E(I_A),$$

and also by definition,

$$\sigma_X^2 = E((X - \mu_X)^2),$$

so

$$\sigma_X^2 \geq k^2\sigma_X^2 P(A),$$

and after canceling, we have

$$k^2 P(A) \leq 1,$$

so finally

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

which is Tchebeychev's Inequality.

73. **LECTURE** WEDNESDAY 21 OCTOBER 2009

Today we discussed the application of the formulas for mean and standard deviation of the sampling distributions for sample total and sample mean. If X is any unknown for which repeated observations can be made, and if we take a sample, the sample data before we see the data is a sequence of unknowns $X_1, X_2, ..., X_n$ all having the same distribution as X and therefore the same mean and standard deviation as X. We denote the sample total by T_n and the sample mean by $\bar{X}_n$. Notice they are capitalized as they are new unknowns-before we look at the sample data, we do not know their values. We found that

$$E(T_n) = n\mu_X$$

and

$$E(\bar{X}_n) = \mu_X.$$

Moreover, if we use independent random sampling (IRS) so all observations are independent of each other, then

$$\sigma_{T_n} = \sqrt{n}\sigma_X,$$

and

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

We also discussed the Central Limit Theorem which says as n tends to ∞ that T_n and $\bar{X}_n$ become normally distributed. Moreover, in practical applications, we assume these two unknowns are normal whenever $n \geq 30$.

We observed that if we are trying to keep a lamp lit with a bunch of light bulbs, then the amount of time we keep the lamp lit is an example of a sample total, and so if X is the life of a light bulb from a certain population of light bulbs, and if we choose n bulbs from this population, then the time we expect to keep our lamp lit is $n\mu_X$. For instance, if a typical bulb is expected to last 750 hours, then we expect for 10 of these bulbs to be able to provide 7500 hours of light and 100 of these bulbs to provide 75000 hours of light. If $\sigma_X = 25$ hours, then to calculate the probability that we get at least 7000 hours from 10 bulbs, we must know the distribution of T_{10}. If we know the population of bulbs is normal, then so are T_{10} and $\bar{X}_{10}$. Thus we can say

$$P(get\ 7000\ hours\ of\ light) = P(T_n \geq 7000)$$
$$= .5 + normalcdf(7000, 7500, 7500, 25 * \sqrt{10}) =^{ap}= 1,$$

so we can be virtually certain to be able to have at least 7000 hours of light from our 10 bulbs if we assume that X is normal. If we have 100 bulbs, then we do not need to assume that X is normal to use the normal distribution in the probability computation. The Central Limit Theorem allows us to assume that T_n and $\bar{X}_n$ are both normal as $n \geq 30$, if $n = 100$. Thus, the probability that 100 bulbs will keep our lamp lit for at least 74000 hours is

$$P(T_{100} \geq 74000) = .5 + normalcdf(74000, 750 * 100, 750 * 100, 25 * \sqrt{100}) =^{ap}= .999999,$$

which means again that we are virtually certain to have at least 74000 hours of light from the 100 bulbs.

In case of average life for 10 bulbs, if we ask for the probability that 10 bulbs have an average life between 753 and 757 hours, then assuming X is normal, the so is $\bar{X}_{10}$, and therefore

$$P(753 \leq \bar{X}_{10} \leq 757) = normalcdf(753, 757, 750, 25/\sqrt{10}).$$

For 100 bulbs we do not need to assume normality as the Central Limit Theorem gives it to us, and automatically

$$P((753 \leq \bar{X}_{100} \leq 757) = normalcdf(753, 757, 750, 25/\sqrt{100}).$$

Notice in all these examples how the formulas for mean and standard deviation come in.

74. **LECTURE** FRIDAY 23 OCTOBER 2009

Today we discussed the sampling distribution and how the distributions for sample mean $\bar{X}_n$ and sample total T_n depend on whether we are doing independent random sampling (IRS), or *Simple Random Sampling* (SRS). For sampling the unknown X remember we are creating new unknowns $X_1, X_2, ..., X_n$ for a sample of size n, and all these unknowns have the same distribution as X so in particular they all have mean μ_X and all have standard deviation σ_X. We can refer to these new unknowns $X_1, X_2, ..., X_n$ as the *sample observations*. When we use IRS, all the sample observations are independent of each other. For instance, in a finite population, to have IRS, we must sample with replacement. For SRS, we sample without replacement in such a way that all subsets of n things from the population are equally likely. For instance, when you deal a hand of 5 cards from a standard deck of cards, if the deal is fair, then any 5 cards are just as likely as any other 5 cards-all possible 5 card hands are equally likely and each has probability one out of 2598960. If we replace each card and shuffle after dealing it, then we would have an IRS.

Recall that we always have

$$E(\bar{X}_n) = \mu_X$$

and

$$E(T_n) = n\mu_X,$$

but for standard deviations of these unknowns we assumed IRS, and the result was

$$\sigma_{\bar{X}_n}(IRS) = \frac{1}{\sqrt{n}}\sigma_X,$$

and

$$\sigma_{T_n}(IRS) = \sqrt{n}\sigma_X.$$

To begin understanding how SRS works, we recall the problem of determining probabilities when drawing colored blocks from a box. If the box contains 3 yellow blocks, 3 red blocks, and 2 blue blocks, then we know that to analyze the probabilities for SRS, that is drawing blocks without replacement, we can think of the blocks as stacked (like a deck of cards), and the successive draws are simply performed by repeatedly selecting the block at the top of the stack. We recall that for instance, to calculate the probability that the second block is red given that the first is red and the third is yellow, we simply imagine we are told the top block in the stack is red and the third block in the stack is yellow, and we see immediately that this conditional probability is 2/6. But also, it is just as clear that the probability the fourth block is red given the first is red and the third is yellow is also 2/6. In fact, we see that however the second draw result is dependent on our knowledge of the results of the first and third draws, the fourth draw result has the exact same dependence on the first and third draws. But also notice, that if we ask for the probability that the second is red given that the first yellow and the fourth is red, we still get 2/6. In fact, we are seeing that the result of each draw has the same dependence on another draw as any other draw depends on any other draw. This would also be the case if the blocks were numbered instead of colored, and for observing any numerical unknown in a finite population, it is the same as drawing numbered blocks from a box. That is to say in mathematical terms, *for SRS, all the sample observations have exactly the same correlation or covariance with each other:*

$$Cov(X_k, X_l) = b_n,$$

for some number b_n which does not depend on the particular pair (k, l) if $k \neq l$. Moreover, these sample observation correlations cannot have any dependence on the sample size n, since when you are making the k^{th} and l^{th} observations, it is not even necessary to know how many observations you will go on to make. Thus, we have

$$b_n = b$$

independent of n. To look at this another way, since the correlation is the same for any pair of observations in the sample, it is the same as the correlation of the first observation with the second observation, but this only depends on the results of the first two draws no matter what the sample size n as long as $n \geq 2$. Of course, if $k = l$, then

$$Cov(X_k, X_l) = Cov(X, X) = \sigma_X^2.$$

We therefore see that for a sample of size n we must have

$$Cov(X_k, T_n) = (n-1)b + \sigma_X^2,$$

because when we expand the expression on the left side using

$$T_n = X_1 + X_2 + ... + X_n,$$

we have

$$Cov(X_k, T_n) = Cov(X_k, X_1) + Cov(X_k, X_2) + ... + Cov(X_k, X_k) + ... + Cov(X_k, X_n).$$

Notice each term $Cov(X_k, X_l)$ where $k \neq l$ has the value b but the term where $k = l$ has the value σ_X^2. Since there are $n-1$ terms where $k \neq l$, it follows that

$$Cov(X_k, T_n) = (n-1)b + \sigma_X^2,$$

and also we see this result does not depend on k, since k does not appear in the expression on the right hand side here. That is to say, we have discovered that in SRS, all the sample observations have exactly the same correlation with the sample total. For SRS of size n, let us denote this by $c_n(SRS)$, so

$$c_n(SRS) = (n-1)b + \sigma_X^2.$$

Then we must have

$$Var(T_n) = Cov(T_n, T_n) = Cov(X_1, T_n) + Cov(X_2, T_n) + ... + Cov(X_n, T_n) = nc_n(SRS).$$

This means that

$$\sigma_{T_n}^2 = nc_n(SRS).$$

Next, we can observe that if we take an SRS of size $n = N$, where as usual N is the population size, then the sample total is fixed, it cannot change from sample to sample, since *there is only one SRS of size $n = N$, namely the sample consisting of the whole population.* Thus T_N is actually a constant, and therefore its variance is zero, and therefore

$$0 = Var(T_N) = Nc_N(SRS),$$

so as $N \neq 0$,

$$c_N(SRS) = 0.$$

Since

$$c_N(SRS) = (N-1)b + \sigma_X^2,$$

this means

$$b = -\frac{\sigma_X^2}{N-1}.$$

Now, we know b, so we can calculate $c_n(SRS)$ and find

$$c_n(SRS) = (n-1)b + \sigma_X^2 = \sigma_X^2 - \frac{n-1}{N-1}\sigma_X^2 = [1 - \frac{n-1}{N-1}]\sigma_X^2$$

$$= \frac{(N-1)-(n-1)}{N-1}\sigma_X^2 = \frac{N-n}{N-1}\sigma_X^2,$$

so we have

$$c_n(SRS) = \frac{N-n}{N-1}\sigma_X^2,$$

and

$$Cov(X_k, X_l) = -\frac{1}{N-1}\sigma_X^2.$$

Since

$$\sigma_{T_n}^2 = Var(T_n) = nc_n(SRS),$$

we now have

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

Taking square roots we get

$$\sigma_{T_n}(SRS) = \sqrt{\frac{N-n}{N-1}}\sqrt{n}\sigma_X.$$

But, we know already that

$$\sigma_{T_n}(IRS) = \sqrt{n}\sigma_X,$$

so this means that

$$\sigma_{T_n}(SRS) = [\sigma_{T_n}(IRS)]\sqrt{\frac{N-n}{N-1}}.$$

We also have $\bar{X}_n = T_n/n$, so

$$\sigma_{\bar{X}_n}(SRS) = \frac{1}{n}\sigma_{T_n}(SRS)$$

$$= [\frac{1}{n}][\sigma_{T_n}(IRS)]\sqrt{\frac{N-n}{N-1}} = [\sigma_{\bar{X}_n}(IRS)]\sqrt{\frac{N-n}{N-1}},$$

so

$$\sigma_{\bar{X}_n}(SRS) = [\sigma_{\bar{X}_n}(IRS)]\sqrt{\frac{N-n}{N-1}},$$

as well. That is to say, the expression

$$\sqrt{\frac{N-n}{N-1}}$$

should be simply viewed as the correction factor for converting from IRS to SRS when calculating standard deviations for sampling distributions. We can notice that this correction factor is obviously 1 for all practical purposes when the population is enormous in comparison to the sample size. When the sample size is about 10 percent of the population size, then the factor is close .95, which means that using the IRS values of standard deviation introduces about a 5 percent error. If the sample size is about 5 percent of the population size, then the correction factor for SRS is about .975, so the error is about 2.5 percent. If the sample size is about one percent of the population size, then the error is only about a half of a percent. If the sample size is only about one tenth of one percent of the population, then the correction factor is about .9995, so the error is only about one twentieth of one percent. You can see here that for populations that are very large in comparison to the sample size, the difference between SRS and IRS is negligible.

Another thing to notice here is that as

$$Cov(X_k, X_l) = \frac{-\sigma_X^2}{N-1},$$

that for large populations the correlation between observations is negligible, but in any case it is *always negative*, no matter the population size. You can think about that as being reasonable because any time you see something larger than average when sampling, that means the other observations are less likely to be above average, if you sample without replacement.

Our main application of this correction factor will be in computing standard deviations for counting unknowns. The difference between a binomial counting unknown and a hypergeometric counting unknown is just the difference between IRS and SRS, so we must have

$$\sigma_{binomial} = \sigma_{hypergeo}\sqrt{\frac{N-n}{N-1}}.$$

Since we will find it is easy to see that

$$\sigma_{binomial} = \sqrt{np(1-p)}$$

gives the standard deviation of a binomial count with success rate p and n independent trials, this means that

$$\sigma_{hypergeo} = \sqrt{np(1-p)}\sqrt{\frac{N-n}{N-1}}.$$

Of course if R is the size of the population of successes, then $p = R/N$ in the case of the hypergeometric count.

75. **LECTURE** MONDAY 26 OCTOBER 2009

Today we reviewed the results from last week for the sampling distributions for T_n, and $\bar{X}_n = T_n/n$, the sample total and sample mean unknowns. We had seen that when sampling the random variable X, that is, an unknown which can be repeatedly observed, then we are actually creating a whole sequence $X_1, X_2, ...X_n$, of observation unknowns all having the same distribution as X and therefore having the same mean μ_X and standard deviation σ_X. We saw that

$$E(T_n) = n\mu_X,$$

and

$$E(\bar{X}_n) = \mu_X.$$

However, as usual, we don't often get what we expect, so we must also be interested in finding the standard deviations for these unknowns. To do that, we began by assuming Independent Random Sampling (IRS) which means we assume all the observation unknowns are independent of each other. We found then

$$\sigma_{T_n} =^{IRS}= \sqrt{n}\sigma_X$$

and

$$\sigma_{\bar{X}_n} =^{IRS}= \frac{\sigma_X}{\sqrt{n}}.$$

Alternately, we can write

$$[\sigma_{T_n}]_{(IRS)} = \sqrt{n}\sigma_X$$

and

$$[\sigma_{\bar{X}_n}]_{(IRS)} = \frac{\sigma_X}{\sqrt{n}}.$$

We had also observed last time that when doing Simple Random Sampling (drawing randomly from a finite population without replacement), then the observation unknowns are no longer independent and that in fact

$$Cov(X_k, X_l) = -\frac{\sigma_X^2}{N-1},$$

for each pair $k \neq l$, and this means all observations have the same correlation coefficient ρ given by

$$\rho = \frac{Cov(X_k, X_l)}{\sigma_X \sigma_X} = -\frac{\sigma_X^2}{(N-1)\sigma_X^2} = -\frac{1}{N-1},$$

or simply

$$\rho = -\frac{1}{N-1}.$$

Notice that this tells us in particular that all the sample observations are negatively correlated, which when you think about it is intuitively clear-if you see one observation is larger than average, you know the other observations are less likely to be as large. Thus you can think that having one observation larger than expected makes it more likely that the others are smaller than that observation.

As a consequence of this correlation formula we had then found that for using SRS we have

$$[\sigma_{T_n}]_{(SRS)} = \sqrt{\frac{N-n}{N-1}}\ \sqrt{n}\sigma_X = \sqrt{\frac{N-n}{N-1}}\ [\sigma_{T_n}]_{(IRS)}$$

and

$$[\sigma_{\bar{X}_n}]_{(SRS)} = \sqrt{\frac{N-n}{N-1}}\ \frac{\sigma_X}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}}\ [\sigma_{\bar{X}_n}]_{(IRS)}.$$

Thus, we can think of the factor $c_{(SRS)}$ given by

$$c_{(SRS)} = \sqrt{\frac{N-n}{N-1}}$$

as a correction factor for standard deviations required whenever we are using SRS instead of IRS, or in other words, the correction for sampling without replacement instead of sampling with replacement.

Today we next went on to discuss the particular case that X is simply the indicator of an event A, so $X = I_A$ and is one if A happens and zero if not. We noticed that in this case, the sample total T_n is simply the success count if we regard each occurrence of A as a success. This is simply because the only possible value of an indicator is zero or one, one for success and zero for failure. Thus, all the observations in this case form a sequence of zeroes and ones, so the total is simply the number of ones, which is of course the number of successes. But this means that T_n is binomially distributed when using IRS whereas is hypergeometric when using SRS. We know that by definition here,

$$\mu_X = E(I_A) = P(A) = p$$

is the success rate, so now our formulas for expected sample total and expected sample mean give us

$$E(T_n) = np$$

and

$$E(\bar{X}_n) = p.$$

Notice that $\bar{X}_n = T/n$ is just the proportion of successes in the sample, and our formula tells us the same thing our common sense tells us. Namely that the proportion in the sample is expected to be the success rate and so the success count should be np. However, our common sense will not tell us very much about the standard deviation, but our theory here gives it to us. Thus, if T has the binomial distribution for n independent trials with success rate p, then we have $T = T_n$ where $X = I_A$ for A the event of success on a single trial, and $P(A) = p$. This means that

$$\mu_T = E(T_n) = n\mu_X = np$$

and

$$[\sigma_T]_{(binom)} = [\sigma_{T_n}]_{(IRS)} = \sqrt{n}\sigma_{I_A},$$

and for the sample proportion of successes, T/n, we have

$$\sigma_{(T/n)} = [\sigma_{\bar{X}_n}]_{(IRS)} = \frac{\sigma_{I_A}}{\sqrt{n}}.$$

We see now that the only thing we need to calculate to have the standard deviations for the binomial and hypergeometric distributions is the standard deviation of an indicator $X = I_A$. Now the crucial property of an indicator is simply $X^2 = X$, since the only numbers which equal their squares are zero and one. In general, for any unknown, we know

$$\sigma_X^2 = E(X^2) - \mu_X^2,$$

or

$$E(X^2) = \mu_X^2 + \sigma_X^2.$$

We can think of the hypotenuse of a right triangle with short sides μ_X and σ_X has hypotenuse of length H given by

$$H = \sqrt{E(X^2)}.$$

In our case, we have $X^2 = X$, and

$$E(X) = E(I_A) = P(A) = p,$$

so the equation says

$$\sigma_X^2 = E(X^2) - \mu_X^2 = E(X) - p^2 = p - p^2 = p(1-p).$$

Thus for any event A the standard deviation of its indicator is

$$\sigma_{I_A} = \sqrt{p(1-p)}.$$

When we use this in our standard deviation formulas for success total and success proportion, we have

$$[\sigma_T]_{(binom)} = [\sigma_{T_n}]_{(IRS)} = \sqrt{n}\sigma_{I_A} = \sqrt{n}\sqrt{p(1-p)} = \sqrt{np(1-p)} = \sqrt{\mu_T(1-p)},$$

and

$$[\sigma_{T/n}]_{(binom)} = [\sigma_{\bar{X}_n}]_{(IRS)} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

In case T is the total success count when sampling without replacement in a finite population, then we know that T has the hypergeometric distribution, so if N is the population size and R is the size of the population of successes, then $p = P(A) = R/N$, and

$$E(T) = np = \frac{nR}{N},$$

and as we are now doing SRS,

$$[\sigma_T]_{(hypergeo)} = [\sigma_{T_n}]_{(SRS)} = [\sigma_T]_{(binom)}\sqrt{\frac{N-n}{N-1}}$$
$$= \sqrt{np(1-p)}\sqrt{\frac{N-n}{N-1}}.$$

Thus, in a finite population, the only difference between the hypergeometric distribution and the binomial distribution is the difference between sampling without replacement or SRS and sampling with replacement or IRS, so to get the standard deviation for the hypergeometric distribution, we simply multiply that for the binomial distribution (with the same number of trials and success rate) by the SRS correction factor.

Finally today, we reviewed the Central Limit Theorem which says that as n tends to infinity, both T_n and $\bar{X}_n$ become normal. In practical terms we will always take these to be normal when $n \geq 30$. We illustrated the Central Limit Theorem by picturing the binomial distribution for $n = 13$ and $p = .45$ and noticing that when plotting the tops of the spikes for the distribution picture, a bell curve fits nicely through all the points.

76. **LECTURE** WEDNESDAY 28 OCTOBER 2009

Today we began with a discussion of how the accuracy of measurement relates to the accuracy of computed expectation values when dealing with continuous unknowns.

Suppose that X is any unknown and we decide to observe X to n decimal place accuracy. We have really then replaced X by a new unknown, the *Round* of X to n decimal places, which we can denote by $R_n(X)$ and which, of course, is the result of rounding off X to n decimal places. For instance, if we observe the value for X is 32.3789, then this means that $R_2(X)$ is observed to be 32.38, or

$$R_2(32.3789) = 32.38.$$

Now, we could instead just round X down automatically to get what we call the *Lower Round* of X denoted $L_n(X)$, so

$$L_2(32.3789) = 32.37.$$

On the other hand, we could just always round up automatically to get what we call the *Upper Round* if X denoted $U_n(X)$. Thus

$$U_2(32.3789) = 32.38 = R_2(X),$$

but

$$U_2(32.3719) = 32.38 \neq R_2(32.3719) = 32.37 = L_2(32.3719).$$

Obviously we always have no matter what the value of X that $R_n(X)$ will be the same as one or the other of the two unknowns $L_n(X)$ and $R_n(X)$. That is, in any case, we can say that

$$L_n(X) \leq R_n(X) \leq U_n(X).$$

Also it is obvious that

$$L_n(X) \leq X \leq U_n(X).$$

Now if a and b are any two numbers, and x and y are also numbers, and if we know both inequalities

$$a \leq x \leq b$$

and

$$a \leq y \leq b$$

are true, then clearly the distance from x to y cannot exceed $b - a$, so we must have

$$|x - y| \leq b - a.$$

Now for sure

$$U_n(X) - L_n(X) = \frac{1}{10^n},$$

so therefore

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

But, we can apply the expectation to both the inequalities

$$L_n(X) \leq X \leq U_n(X)$$

and

$$L_n(X) \leq R_n(X) \leq U_n(X),$$

so from the order preserving property of expectation, we also know both inequalities

$$E(L_n(X)) \leq E(X) \leq E(U_n(X))$$

and

$$E(L_n(X)) \leq E(R_n(X)) \leq E(U_n(X)).$$

Therefore we also must have

$$|E(R_n(X)) - E(X)| \leq E(U_n(X)) - E(L_n(X)) = E(U_n(X) - L_n(X)) = E(10^{-n}) = \frac{1}{10^n},$$

so in fact,

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

Notice that this means if we measure or observe to n decimal place accuracy, then our expected values are correct to n decimal place accuracy. To be precise here on what we are saying about the accuracy of our expected values, we must say that if we have the exact probability for each possible value of $R_n(X)$, then using these probabilities to calculate the expected value of $R_n(X)$ will give us the expected value of X itself to n decimal place accuracy. If there are inaccuracies in the various probabilities of values for $R_n(X)$, then these inaccuracies would cause further inaccuracies in our expected value calculations.

In any case, the main thing to notice here is that

$$L_n(X),\ R_n(X),\ \textit{and}\ U_n(X)$$

are *all discrete unknowns*. In fact, the unknowns

$$10^n L_n(X),\ 10^n R_n(X),\ \textit{and}\ 10^n U_n(X)$$

have only integer values, that is values in the set

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

For unknowns which are non-negative, this means that all calculations can be reduced to dealing with "counting" unknowns, those with values in the set of whole numbers

$$\mathbb{W} = \{0, 1, 2, 3, ...\}.$$

Finally here, we can use these considerations to see what is going on with the calculation of covariance and expected values of products of unknowns. If X and Y are unknowns,

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

so the calculation of covariance one way or another involves calculating the expected value of a product of unknowns. Our previous considerations tell us that for all practical purposes, it is enough to understand how to do this for discrete unknowns (and therefore, actually, it is enough to deal with those having only integer values). If X is discrete, let V_X be the set of possible values of X. For each possible value v in V_X, let A_v be the event that X has the value v, which is $(X = v)$, and let I_v be the indicator of A_v. Then

$$X = \sum_{v \in V_X} vI_v = \sum_{v \in V_X} vI_{(X=v)},$$

and we know

$$E(X) = \sum_{v \in V_X} vE(I_v) = \sum_{v \in V_X} vP(A_v) = \sum_{v \in V_X} vP(X = v).$$

If we use w to denote values of Y, we likewise have

$$Y = \sum_{w \in V_Y} wI_{(Y=w)}$$

and

$$E(Y) = \sum_{w \in V_Y} vP(Y = w).$$

Notice that

$$XY = \sum_{v \in V_X} \sum_{w \in V_Y} vwI_{(X=v)} I_{(Y=w)},$$

but we can use that fact that

$$I_A I_B = I_{A\&B}$$

here to tell us always

$$I_{(X=v)} I_{(Y=w)} = I_{(X=v)\&(Y=w)}.$$

Therefore

$$XY = \sum_{v \in V_X} \sum_{w \in V_Y} vwI_{(X=v)\&(Y=w)},$$

so

$$E(XY) = \sum_{v \in V_X} \sum_{w \in V_Y} vwP((X = v)\&(Y = w)).$$

Thus, to compute the expected for the product XY, it is not enough to know just the values and probabilities for each of the unknowns separately, we must also know their *Joint Distribution*, that is the probabilities for all the combinations of possible values of the two unknowns considered together. We can also see this using the multiplication rule for expectation. For we have

$$XY = \sum_{w \in V_Y} XwI_{(Y=w)},$$

so

$$E(XY) = \sum_{w \in V_Y} wE(XI_{(Y=w)}) = \sum_{w \in V_Y} wE(X|Y = w)P(Y = w)$$

$$= \sum_{w \in V_Y} \sum_{v \in V_X} vwP(X = v|Y = w)P(Y = w) = \sum_{v \in V_X} \sum_{w \in V_Y} vwP((X = v)\&(Y = w)).$$

Finally today, we began the discussion of confidence intervals. If we have a population of tuna fish under consideration and we want to know the true population mean weight, μ, then we can take a large sample to get an idea, and by Tchebeychev's Inequality, we know that for large enough samples the sample mean is very likely to be close the the true mean. If we have a sample mean of 325 pounds and that is our only information, then we should obviously guess $\mu = 325$, which means more precisely, if X is the weight of a tuna fish, then

$$E(X|\bar{x} = 325) = 325.$$

To know what our error would likely be, we must use the standard deviation.

77. **LECTURE** FRIDAY 30 OCTOBER 2009

Today we discussed the computation of confidence intervals. In general, any non-trivial statement about a population for which we have incomplete knowledge cannot be made with absolute certainty. If A is any statement about such a population, we call $C = P(A)$ our *Confidence* in the statement. We often begin by choosing C and call it the it Level of Confidence at which we work. Most population parameters such as mean and standard deviation must be estimated from sample data, and as a result there is likely to be error in our results. Thus, we express our estimate as a *Point Estimate* plus or minus a *Margin of Error* which we denote by ME. We will only be concerned with the simplest case, namely estimations of the true population mean μ_X. To do this, we begin by standardizing $\bar{X}_n$, to form the standard unknown Z, where

$$Z = \frac{\bar{X}_n - \mu_X}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_n - \mu_X}{(\sigma_X/\sqrt{n})}.$$

Then we have $\mu_Z = 0$ and $\sigma_Z = 1$. Notice that if we use the mean of a sample to guess the true mean, then our error is the numerator of Z here. Also, if $\bar{X}_n$ is normal, then so is Z. To be able to have normality of Z here, it therefore is enough that either $n \geq 30$ or X is itself normal. To have confidence C that

$$|\bar{X}_n - \mu_X| \leq M,$$

is the same as saying

$$P(|\bar{X}_n - \mu_X| \leq M) = C.$$

On the other hand, as Z is now assumed normal, we know that if

$$z_C = invNorm(\frac{1+C}{2}, 0, 1),$$

then

$$P(|Z| \leq z_C) = C.$$

Thus, as

$$|Z| \leq z_C$$

if and only if

$$|\bar{X}_n - \mu_X| \leq z_C \frac{\sigma_X}{\sqrt{n}},$$

it follows that our margin of error with confidence C is

$$ME = z_C \frac{\sigma_X}{\sqrt{n}}.$$

Of course this all depends on knowing the population standard deviation σ_X to start with, and that is usually not reasonable if you do not even know the true mean, even though there are applications where this is the case. If you do not know σ_X, then as the sample standard deviation s_x from your sample data is an estimate of σ_X, we could try replacing σ_X by s_x in our previous calculations. The problem is that now, the sample standard deviation s_x is just an observed value of the random variable $S = S_n(X)$ which is the sample standard deviation of our sample as an unknown number. Notice that for instance,

$$S^2 = \frac{n}{n-1}[(X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + ... + (X_n - \bar{X}_n)^2],$$

where $X_1, X_2, ..., X_n$ are our sample observation unknowns. In fact, it can be shown that

$$E(S^2) = \sigma_X^2,$$

and it is in this sense that we say s_x is an estimate of σ_X. It is really the sample variance which is expected to be the true variance. When we standardize using S in place of σ_X in the preceding formula we do not get the standard Z but rather we get a new unknown which has a distribution slightly different from the standard normal and it is called the *Student t-distribution*, or simply the t-distribution for short. It is actually a whole family of distributions

parametrized by a number we call *Degrees of Freedom*. In our case, the number of degrees of freedom which we denote by df is $df = n - 1$. For $df = d$, we denote by t_d the unknown with the $t-$distribution for d degrees of freedom. It can then be shown with our rules of expectation and a lot of algebra (see for instance the Expectation Primer on my website), that

$$t_d = \frac{\bar{X}_n - \mu_X}{S/\sqrt{n}}$$

does indeed have the $t-$distribution for $df = n - 1$, provided that X itself is a normal random variable. Thus, if σ_X is not known, then our method proceeds with the assumption that X is a normal random variable. In that case, the z_C is replaced by t_C for the given level of confidence C. All the reasoning is exactly the same because the $t-$distribution is also a bell shaped distribution which is symmetric with mean zero-it is just a little flatter than the standard normal distribution. Thus, t_C is chosen so that

$$P(|t_d| \leq t_C) = C,$$

which can be calculated as

$$t_C = invt(\frac{1+C}{2}, d),\ d = n - 1,$$

if you have the inverse $t-$distribution in your calculator.

To actually calculate a confidence interval using the TEST menu in the statistical menu of the TI calculator, if population standard deviation is known use the $z-$interval, whereas if it is not known, use the $t-$interval in the menu. The format for entering the information is self explanatory in each case, but you should try and practice with it. If you only want the margin of error as opposed to the confidence interval, simply enter $\bar{x} = 0$.

Finally, we discussed the fact that there is only one criterion for choosing between z and t for confidence intervals and that is simply whether or not you know σ_X. If you know σ_X you use the $z-$interval and if you do not know σ_X, you use the $t-$interval. Often textbooks mistakenly give the impression that the $t-$interval is for small samples, but if you know σ_X, and you want to use the $t-$distribution, then X must be assumed normal so $\bar{X}_n$ is normal no matter how small the sample size and therefore the $z-$interval should be used instead.

78. **LECTURE** MONDAY 2 NOVEMBER 2009

Today we reviewed the basics of confidence intervals for the true mean μ_X for an unknown population or random variable X. Remember, the confidence interval has the form

$$\mu_X = a \pm b.$$

We would read this in "everyday language" or "common parlance" as "a give or take b", and it means that

$$a - b \leq \mu_X \leq a + b.$$

Keep in mind that generally in statistical settings, nothing can be said about an unknown population with absolute certainty, so

The number a is our *Point Estimate* whereas the number b is our *Margin of Error*. If you have the sample mean $\bar{x}$ of a sample, then the point estimate is that sample mean. We denote the margin of error by ME. For a given *Level of Confidence* denoted C, to have confidence C in statement A simply means $P(A) = C$. Thus, to say

$$\mu_X = \bar{x} \pm ME$$

with confidence C is the same as saying

$$P(\bar{x} - ME \leq \mu_X \leq \bar{x} + ME) = C,$$

or equivalently,

$$P(|\bar{x} - \mu_X|) = C.$$

Obviously the margin of error depends on the level of confidence, and if you think carefully, you will realize that for given data, if you raise the confidence level, then you must allow a larger margin of error. If I say I am 90 percent certain my error is less than 5, then I would not be 99 percent certain of that the error is less than 5. This means the size of the error must now possibly be larger than 5 if I require 99 percent certainty. Keep in mind, the distribution of $\bar{X}_n$ is normal, so the higher the area we want in the middle of the distribution, the farther out we must go.

Our formula for the margin of error for confidence level C is

$$ME_C = z_C \frac{\sigma_X}{\sqrt{n}},\ \sigma_X \textit{ given},$$

or

$$ME_C = t_C \frac{s_x}{\sqrt{n}},\ \sigma_X \textit{ unknown}.$$

To calculate a confidence interval given the statistical information such as data or statistics, we do not need the margin of error formulas above, but to solve the problem of figuring out how large a sample to use, we must use these formulas. To calculate a confidence interval simply press the stat button on your calculator and go to the test menu. Next, in case you know σ_X, scroll down to the $z-$interval and call it up, and enter your information. If you do not know σ_X, then use the $t-$interval instead. After entering the $\bar{x}$, the n, the level of confidence C, and the σ_X or s_x as the case may be, then press enter, and the readout gives two numbers in parenthesis separated by a comma, followed by the value of $\bar{x}$ and the value of n. This is the format for the readout of all confidence intervals in the calculator. Here, the first number is $\bar{x} - ME$, and the second number is $\bar{x} + ME$, Clearly, if you just want the margin of error ME you can just enter $\bar{x} = 0$ in the calculator.

If we are given a required level of confidence C and told to make sure we are that confident that our margin of error is does not exceed the number E, then that is simply requiring

$$z_C \frac{\sigma_X}{\sqrt{n}} = ME_C \leq E.$$

Obviously, we want to make our sample no bigger than necessary, and clearly increasing n always makes the margin of error smaller, so we would take the smallest sample size that satisfies the above inequality. A simple way to find it is to turn the inequality into an equality and solve it

for the sample size n. Of course the solution n_C will generally not be a whole number, so our required sample size is $n \geq n_C$, that is we must round up to the nearest whole number. Solving the equation is simple. The equation

$$z_C \frac{\sigma_X}{\sqrt{n_C}} = E$$

is equivalent to

$$\frac{z_C \sigma_X}{E} = \sqrt{n_C},$$

so squaring both sides gives

$$n_C = \left(\frac{z_C \sigma_X}{E}\right)^2.$$

Obviously, in real applications, we do not know σ_X, but often we can find a number B for which we are reasonably certain that $\sigma_X \leq B$. In that case, we simply replace σ_X by B in the formula so

$$n_C = \left(\frac{z_C B}{E}\right)^2.$$

Finally, we discussed the problem of determining a probability of *True Proportion*. For instance, if A is a statement about the outcome of a repeatable experiment, such as tossing a dice, then we can try to estimate $P(A)$ by making n trials of the experiment and using the proportion of successes in our data as an estimate of $p = P(A)$. We are sampling the indicator of A which we denote by I_A. In this case, the sample mean $\bar{x}$ is the sample proportion, as all the values of the indicator are zeroes and ones, so the sample total is the number of successes $T_n = x$, and the sample mean is then just $\bar{x} = T_n/n = x/n$. But in order not to confuse this use of x with the situation with a general continuous random variable, we use $\hat{p}$ to denote the sample proportion $\bar{x} = x/n$. Thus, $\hat{p}$ is the estimate of the true proportion p. Since the distribution of T_n here is actually binomial, we need the binomial distribution to be approximately normal which it will be provided that $n \geq 30$ and p not close to either zero or one. This type of confidence interval is called a 1-Prop $z-$interval in the calculator. You only need the level of confidence C the number of trials n and the number of successes x. Once these are entered, you press enter and the calculator gives the answer in the standard format.

Clearly, when estimating a proportions or probability, we need to have small ME as the number we are estimating is between zero and one. With the estimate expressed as a percentage, we want our ME to be only a few percentage points, so as a decimal fraction, we can begin with looking at requiring $ME \leq .01$. The unknown here is an indicator, and we know that the standard deviation of an indicator I is

$$\sigma_I = \sqrt{p(1-p)} = \sqrt{p - p^2}.$$

We can recognize that if we graph the equation $y = p - p^2$, its graph is a downward parabola crossing the horizontal $p-$axis at $p = 0$ and $p = 1$. The maximum is clearly right between zero and one at $p = 1/2$. Using that value of p gives

$$\sigma_{max} = \sqrt{(1/2)(1-(1/2))} = \sqrt{(1/2)(1/2)} = 1/2.$$

This means no matter what, we are sure that

$$\sigma_I \leq \frac{1}{2}, \ I \text{ an indicator.}$$

This means that when estimating true proportions, to find the required sample size for given confidence, we use

$$n_C = \left(\frac{z_C(1/2)}{E}\right)^2, \ \text{to make } ME \leq E.$$

We discussed some elementary applications of these formulas for finding required sample sizes. Notice that if you double the allowed margin of error the sample size required goes down by a factor of four, whereas if you triple the allowed margin of error, then the required sample size

goes down by a factor of nine. Thus, for 99 percent confidence and only one percentage point error, we find

$$n_C = 16587.2415,$$

so

$$n \geq 16588.$$

If we allow 2 percentage points in the margin of error, then the sample size required is

$$n \geq n_C$$

where

$$n_C = 4146.810374$$

and therefore the required sample size is only $n \geq 4147$. If we allow three percentage points in the margin of error, we find

$$n_C = 1843.026833,$$

so we need $n \geq 1844$. Thus, to keep things in round numbers, if we take a sample of size $n = 2000$, then our proportion estimates will be accurate to within three percentage points with 99 percent confidence.

79. **LECTURE** WEDNESDAY 4 NOVEMBER 2009

Today we discussed confidence intervals and reviewed what had been done in the last few lectures. In particular, we discussed the problem of finding the required sample size to make the margin of error ME_C with confidence C at most E, that is

$$ME_C \leq E.$$

Here we need an upper bound B on the standard deviation σ_X, meaning that B is a number for which we are sure

$$\sigma_X \leq B.$$

We know the sample size must be at least n_C where

$$n_C = \left(\frac{z_C \sigma_X}{\sqrt{n}}\right)^2$$

which obviously increases as σ_X increases. Thus, if we do not know σ_X but so know $\sigma_X \leq B$, then we would, to be safe, use B instead of σ_X. This means that the required sample size is $n \geq n_C$ where

$$n_C = \left(\frac{z_C B}{\sqrt{n}}\right)^2, \ \sigma_X \ unknown, \ \sigma_X \leq B.$$

In case of proportions, we demonstrated that $\sigma \leq 1/2$, and consequently $B = 1/2$, and the required sample sizes are easy to compute, but the result is in the thousands. This means that anything which can be done to lower the value of B can result in a reduction in sample cost. We know that $0 \leq p \leq 1$, and that given the value of p, we have $\sigma = \sqrt{p(1-p)}$. Since the worst case is $p = .5$, as that leads to the largest value for σ, we saw from the parabola graph of the equation $v = p(1-p) = p - p^2$, that as p gets closer to $1/2$, the value of σ^2 and likewise the value of σ increases. Thus, if J is the set of numbers between zero and one, and if K is any subset of J, then the worst possible value of p in K is the value closest to $1/2$. If we are sure that p is in the subset K, then we can use this to lessen the required sample size. If we call p_K the number in K closest to $1/2$, then we can take $B = B_K$, where

$$B_K = \sqrt{p_K(1-p_K)}.$$

As then $B < 1/2$, the required sample size is now less in this case. Notice how the required sample size depends on all the inputs z_C, B, and E. If we double E, the sample size is cut to one fourth of what would be required for the originally given E. In general, as E goes up, the required sample size goes down, but much faster because of the squaring effect in the formula. If the allowed error E is tripled, the required sample size is cut to only one ninth of the original required sample size. If we increase our required level of confidence C, then z_C increases and therefore so does the required sample size. If we increase the bound B on the standard deviation, again the required sample size goes up, and likewise anything we can do to decrease B will make the required sample size go down, and go down fast, because of the squaring effect in the formula for n_C. Thus, if we can decrease B by ten percent, its new value is $(.9)B$ and squared is $(.81)B^2$ which means the sample size is cut by 19 percent of what was originally required. We call these effects *Scaling Effects*. When dealing with sample sizes in the thousands, a slight decrease in B can lead to a substantial reduction in the required sample size. For instance, if we know that of all the possible values for the true proportion the one closest to $1/2$ is $6/10$, then the value of B is cut from $1/2$ down to

$$B = \sqrt{(.6)(.4)} = \sqrt{.24},$$

which squared is $.24 = 24/100$. As $(1/2)^2 = .25$, we have decreased the value of B by a factor of

$$\frac{.24}{.25} = \frac{24}{25} = .96,$$

and this means that we can cut the sample size by 4 percent. When dealing with sample sizes in the thousands, a four percent reduction in the required sample size could be a substantial cost saving.

One simple way to approach this is to purely use scaling effects to figure required sample sizes. We can begin by noticing that for popular values of the level of confidence C all the corresponding values of z_C are near 2. As for 95 percent confidence we know that z_C is approximately 1.960, this means for $z_C = 2$, we have over 95 percent confidence. So imagine to start that the C has been chosen so as to make $z_C = 2$. Take $B = 1$ and $E = 1$ to start. The required sample size for this case is then simply $n_C = 2^2 = 4$. If we are dealing with proportions, then we want E to be only a few percentage points, so as a next step, take $E = .01 = 1/100$. This multiplies the required sample size by $100^2 = 10,000$, so now the required sample size is 40,000. But in general, for proportions, we can take $B = 1/2$ which squared gives $1/4$, so now the required sample size is down to 10,000. If we want 99 percent confidence, then the value of z_C is approximately 2.576, which multiplies the required sample size by a factor of $(2.576/2)^2 = 1.658944$, so multiplying by 10,000 gives a required sample size of at least $n_C = 16,589.44$. Thus, if we allow the error E to double from one percentage point to two percentage points, then the required sample size is cut from 16,590 down to $n_C = 16,589.44/4 = 4147.36$. This seems to take longer than simply using the calculator, but can often be estimated in your head this way without the aid of a calculator.

80. **LECTURE** FRIDAY 6 NOVEMBER 2009

Today we discussed hypothesis testing and its comparison to a criminal trial. We discussed the significance of data which is also called the P-value of data.

To summarize, in a **HYPOTHESIS TEST**, we deal with two competing exclusive hypotheses, called the **NULL HYPOTHESIS** and the **ALTERNATE HYPOTHESIS**. However, these hypotheses are not treated equally. Rather, in a hypothesis test we try to **DISPROVE** the Null Hypothesis by **PROVING** the Alternate Hypothesis. The proof of the alternate hypothesis must be based on **EVIDENCE** which statisticians refer to as **DATA**.

It is customary to denote the Null Hypothesis by H_0 and the Alternate Hypothesis by H_1 or H_a or H_{alt}. For instance, most useful example to keep in mind is the example of a **Criminal Trial**. In this case, the Null Hypothesis is that the accused person is innocent. The Alternate Hypothesis is that the accused person is guilty. We could summarize this as

$$H_0 : \ Accused\ is\ Innocent$$

versus

$$H_{alt} : \ Accused\ is\ GUILTY$$

and keep in mind that the *Burden of Proof* is on the Prosecutor. The jury will in fact be instructed that the accused is to be thought of as innocent throughout the trial and that the evidence presented by the prosecutor must be evaluated under that assumption.

In any hypothesis test, we assume that H_0 is true for purposes of argument, and any statement which is a purely logical consequence of H_0 is also then assumed true. We then look at evidence or data, and evaluate how contradictory our data is of H_0. We are dealing with a type of "proof by contradiction".

As an example, suppose that we have a large box full of tiny colored beads, millions of them. Suppose that H_0 is the statement that there are no red beads in the box. To try to disprove this hypothesis, we take a bead out of the box and examine it. This bead is data or evidence. If it is red, it is a perfect contradiction of H_0. That is, we can notice that this would be impossible if we assume H_0. In probabilistic terms, we can notice this is the same as

$$P(get\ a\ red\ bead |\ H_0) = 0.$$

Suppose now instead H_0 only says that the true proportion of red beads in the box does not exceed ten percent, or in symbols

$$H_0 : \ p_R \leq .1$$

versus

$$H_1 : \ p_R > .1.$$

Notice that getting a single red bead from the box no longer gives a perfect contradiction of H_0, and in fact no sample of beads from the box can result in a perfect contradiction of this H_0. However, if we draw 10 beads from the box and 4 are red, then we might tend to be suspicious that H_0 is wrong. This is now because if we calculate

$$P(get\ 4\ out\ of\ 10\ red |\ H_0),$$

we find a small number. Clearly, if we take a bigger sample here, our evidence should be stronger if this high percentage of red beads happens again. On the other hand, if we draw 100 beads and get exactly 10 red beads we have no contradiction at all, and yet the calculated probability of getting 10 out of 100 red beads is small. Thus having evidence which has a small probability of happening is not of itself proof of the alternate hypothesis. Somehow, the **RELEVANCE** of out data or evidence must be considered. Surely getting 10 out of 100 red beads in a sample of beads from the box is somewhat irrelevant ("so what") as far as trying to prove the alternate hypothesis here.

These considerations lead us to adopt the following numerical measure of how contradictory our data is of H_0, and we call it the $P-$**Value** or **SIGNIFICANCE** of our data or evidence, defined as

$$P-Value\ of\ our\ data\ = P(data\ as\ or\ more\ contradictory\ of\ H_0\ than\ our\ data|\ H_0).$$

When this number is small, then we would regard our data as strong evidence in favor of the alternate hypothesis. How small this must be, is an issue will defer to later.

Let us now consider the problem of proving that the box of beads has more than 10 percent red beads when we find 10 out of 100 beads drawn to be red. If we calculate the probability of getting exactly 10 out of 100 beads, the probability that this could happen under H_0 is just the binomial probability if 10 successes out of 100 trials with a true success rate $p_0 = .1$, and where we have used the subscript zero on our true success rate to remind us that that is due to the assumption that H_0 is true here, since we really do not know the true proportion of red beads in the box. If x denotes the number of red beads in a sample of size n, this probability is

$$P(x = 10\ red\ out\ of\ n = 100\ drawn\ |\ H_0) = binompdf(100, .1, 10) = .1318653468,$$

which is not a very big number. On the other hand, if we compute the significance of this data or its $P-$value, then we must ask what is more contradictory of the results of our data. That would certainly be drawing 100 beads and finding $x > 10$ that is more than 10 red beads in the sample. Notice the correspondence between the inequality $x > 10$ and the alternate hypothesis $H_1 : p_R > .1$. With the symbols on the same side (here on the right side) and with the numbers on the same side (here on the left side), the inequality symbols go exactly the same way. This means that the $P-$value or significance of the data is according to the above definition

$$P(X_R \geq 10|X_R\ binomial\ n = 100,\ p_0 = .1) = 1 - binomialcdf(100, .1, 9) = .5487098154.$$

This is certainly not near zero, which means the result of finding 10 red beads out of 100 should be ignored as it obviously should be. But what if we find 23 red beads our of 100. Now the $P-$value computation gives the very small number .0001141563199. We therefore reasonably conclude that the box has more than 10 percent red beads on the basis of such a sample result.

It is important to realize the logic here of computing with the assumed true proportion $p_0 = .1$ when H_0 only says that $p_R \leq .1$. This is because we certainly would not use a value of p_0 larger than .1, but if we use a smaller number than .1, then any argument we give could be attacked as making an unfair assumption. The hardest thing to disprove here is that $p_R = .1$ among all the possible values allowed under H_0. You should think about this until you understand it. All the logic involved in hypothesis testing is actually very simple, but just complicated enough to be confusing when first encountered. It is like learning to ride a bicycle. Once you "get it", it is easy, but to do this you need to spend a little time seriously thinking about what is going on. If you do not do this, it will always be confusing and you will make mistakes and get things mixed up.

To do the preceding calculations directly in the calculator, we would press the "stat" button and go to the TEST menu and scroll down to the "1-prop-Z-test". Here you would enter the value p_0 assumed for the true proportion under H_0, and then we would enter the sample information asked for, and finally choose the form of the alternate hypothesis and choose "calculate". When we press the "enter" button, the $P-$value is given as p in the readout.

81. **LECTURE** MONDAY 9 NOVEMBER 2009

We reviewed the basics of Hypothesis Testing presented in the last lecture. The main example is the criminal trial where H_0 is that the accused is innocent. Remember, the accused need not present any evidence, the burden of proof is on the prosecutor. If the evidence presented is not sufficiently contradictory of H_0, then the accused is set free. Notice, that nothing has been proven when the accused gets to go free. It is merely the result of insufficient evidence of guilt. Last time we defined the Significance of the data or evidence as

$$P - Value\ of\ our\ data\ = P(data\ as\ or\ more\ contradictory\ of\ H_0\ than\ our\ data|\ H_0).$$

We can consider how this definition selects for relevance of evidence as opposed to simply looking to see if the evidence itself is of low probability. For if we simply think that whenever our evidence has low probability we have diss-proven H_0, then we can prove a person guilty whenever we find irrelevant rare evidence. For instance, in a murder trial, if the prosecutor notices that the eye color of the accused is a very rare shade of blue-green, then that evidence given H_0 is very small, because it is independent of H_0, and of low probability in its own right. On the other hand, if we consider the above definition of significance of the evidence, all evidence is more contradictory than irrelevant evidence, and consequently the $P-$value of the eye color evidence is near one which is not near zero.

We went on to relate this to the problem of beads in a box discussed in the previous lecture. Remember, in any hypothesis testing situation, in order to convince others that your data is good evidence in favor of the alternate hypothesis and thus very contradictory of the null hypothesis, you must have a $P-$value near zero. How close to zero is a matter of judgement in actual practice, and would depend on the seriousness of the situation. Often, a value denoted α is chosen in advance which is near zero, and then we regard the $P-$value as small enough if it does not exceed α. Such a value α is called a **LEVEL OF SIGNIFICANCE**. It plays a similar role to the level of confidence when dealing with confidence intervals, but significance and confidence should never be confused-they are two different things. For instance, clearly we should use a value of α which is near zero, whereas we should use a value of confidence which is near one. Typical or popular levels of significance are $\alpha = .05, .01, .001, .1$. Smaller values of significance level are set in serious situations. In general, nobody pays much attention to data until its significance gets down around $\alpha = .05$.

82. **LECTURE** WEDNESDAY 11 NOVEMBER 2009

Today we continued the discussion of Hypothesis Testing and used the example of testing rope for mountain climbing as an example. If you need to know that the mean rope breaking strength of your climbing rope is $\mu > 1000$ in order to be safe, then you should take a sample of the rope you intend to use and test it. If the data proves $\mu > 1000$ to your satisfaction, then it is reasonable to use the rope. Obviously if a sample mean for $n = 36$ pieces of rope is only $\bar{x} = 998$, then you would not think the rope is safe, but if the sample mean is $\bar{x} = 1002$, you still might not think the rope is safe, because as $\bar{X}$ is normally distributed, there is always a chance that the sample mean could turn out slightly larger than 1000 even though the true mean is maybe even smaller than 1000. You should be careful here, because your life is at stake. To know how good or bad a sample mean is, we need the standard deviation for the rope breaking strength, σ. Say we know $\sigma = 50$. We evaluate the significance or P-value of the data for a sample mean of 1002 by calculating

$$P(\bar{X} \geq 1002 | \mu = 1000,\ \bar{X}\ normal\ , \sigma = 50,\ n = 36).$$

This number is the P-value or significance of the data. It is given the symbol p in your calculator readout for any hypothesis test in the test menu. If your data were in perfect contradiction of the null hypothesis giving a perfect proof of the alternate hypothesis, then $P - value = p = 0$. This means that in statistical situations, we want to see a very small number for the P-value, and if we do not see a small enough number, then we will not think of the data as proving the alternate hypothesis. Typically remember, we want $p \leq .05$ before we pay much attention to the data, but in serious situations, such as in the mountain climbing rope example, we probably want a much lower P-value before we will trust the rope for mountain climbing. In general, we often set a small number value in advance called a Level of Significance and denoted α and require that $P - value \leq \alpha$ in order to establish the alternate hypothesis. If the P-value of the data turns out larger than α, then we simply say the data is inconclusive-it does not prove anything.

Remember that in any hypothesis test, you are trying to prove the Alternate Hypothesis and trying to disprove the Null Hypothesis. So in the rope climbing example, I want to prove that $\mu > 1000$, so that is the alternate hypothesis and $\mu \leq 1000$ is the null hypothesis. Notice that we always put the possibility of equality in the null hypothesis. When we calculate, we will in fact use the null hypothesis by taking that equality to give a hypothetical value to the true mean which we denote by μ_0. Thus, in our example we have

$$\mu_0 = 1000.$$

In tests in the test menu of your calculator, symbols with subscript zero refer to values under the null hypothesis. You pick the appropriate test from the test menu, for our example the $z-$test, and enter the statistical information and most important, you must choose the correct alternate hypothesis. Our alternate hypothesis is $\mu > 1000$ but in the calculator you will have entered $\mu_0 = 1000$, so your alternate hypothesis is

$$\mu > \mu_0.$$

If the Consumer Protection Agency suspects that Acme Corporation's 1000 pound test rope is not a s strong as advertised, then the Consumer Protection Agency will take a sample of Acme's rope and test it to see if the data will prove that the rope is not as strong as advertised. Thus, for the Consumer Protection Agency, the alternate hypothesis they will try to prove with the data is $\mu < 1000$, so their choice in the test menu for alternate hypothesis will be

$$\mu < \mu_0.$$

From their point of view, if the sample mean is only $\bar{x} = 998$, then they might not think this is strong enough evidence to prosecute Acme. There is too big a chance that this sample mean

could have happened even though $\mu = 1000$. For instance for a sample of size $n = 100$, the P-value is not low enough to impress a judge or jury. However, if their is strong enough suspicion against Acme rope, the Agency could take a much larger sample. For a sample of 100000 pieces of rope, if the sample mean is 998, the P-value of the data is zero to the level of accuracy in the calculator. For such a large sample, the sample mean will be a very very accurate estimate of the true mean, so even a difference of 2 pounds is too much and Acme would be found guilty.

Try thinking about these issues using your common sense. Hypothesis testing is confusing until you understand it, but then it is easy. It is like learning to ride a bicycle. You have to keep trying examples until you "get it".