

MATH-111 (DUPRÉ) SPRING 2009 LECTURES

1. LECTURE MONDAY 12 JANUARY 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,

$$(1.1) \quad E(X|K) = E(X) = \mu_X = \mu,$$

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

$$(1.2) \quad E(X|K) = c.$$

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

$$(1.3) \quad a \leq E(X|K) \leq b.$$

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

$$(1.4) \quad E(X + Y|K) = E(X|K) + E(Y|K).$$

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

$$(1.5) \quad E(Y|K) = E(cX|K) = cE(X|K).$$

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 $E(Y|K) = E(cX|K) = cE(X|K)$.

2. LECTURE WEDNESDAY 14 JANUARY 2009

We began by reviewing the four basic rules of guessing.

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

$$(2.1) \quad E(X|K) = c.$$

POSITIVITY RULE: If K implies that $a \leq X \leq b$, then [2.1]

$$(2.2) \quad a \leq E(X|K) \leq b.$$

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 [1.4]

$$(2.3) \quad E(X + Y|K) = E(X|K) + E(Y|K).$$

HOMOGENEITY: If K implies that $Y = cX$, then [1.5]

$$(2.4) \quad E(Y|K) = E(cX|K) = cE(X|K).$$

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

$$(2.5) \quad P(N|K) = E(I_N|K).$$

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, [2.1], we know that

$$(2.6) \quad 0 \leq E(I_N|K) \leq 1,$$

and therefore,

$$(2.7) \quad 0 \leq P(N|K) \leq 1.$$

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, [2.1], 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 [2.7], 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, $\text{not}N$ is the negation of statement N , so $\text{not}N$ is true exactly if N is false. It is easy to see here that in terms of indicators we can write

$$(2.8) \quad I_{\text{not}N} = 1 - I_N.$$

It then follows immediately from the Additivity and Homogeneity Rules that

$$(2.9) \quad P(\text{not}N|K) = 1 - P(N|K).$$

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

$$(2.10) \quad I_{A\&B} = I_A I_B,$$

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 $I_{A\text{or}B}$ 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 $A\text{or}B$ 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

$$(2.11) \quad I_{A\text{or}B} = I_A + I_B - I_{A\&B}.$$

It now follows immediately from our Addition and Homogeneity Rules that

$$(2.12) \quad P(A\text{or}B|K) = P(A|K) + P(B|K) - P(A\&B).$$

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 $P(S|K) = P(\text{Sure}|K) = 1$ and $P(\Phi|K) = 0$.

The fundamental rules of probability are simply [2.7], [2.12], and $P(\text{Sure}|K) = 1$.

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.

3. LECTURE FRIDAY 16 JANUARY 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

$$(3.1) \quad 1 = I_A + I_B + I_C,$$

and thus we have $1 = P(A|K) + P(B|K) + P(C|K)$. But we can multiply both sides of [3.1] by X and now arrive at the equation

$$(3.2) \quad X = XI_A + XI_B + XI_C.$$

We then find by the Addition Rule that

$$(3.3) \quad E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K).$$

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 therefore 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 simply 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:

$$(3.4) \quad E(XI_N|K) = E(X|N\&K)P(N|K).$$

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 [3.3] 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, \dots and our background information tells us exactly one is true (so all others are false), then

$$(3.5) \quad E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K) + \dots$$

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, \dots 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 [3.5] 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 [2.10] we know $I_{A\&B} = I_A I_B$, so using $X = I_A$ and $N = B$ in the Multiplication Rule [3.4] 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:

$$(3.6) \quad P(A\&B|K) = P(A|B\&K)P(B|K).$$

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 $P(\text{both } R)$? We can notice that "both R" is the same statement as " $1^{\text{st}} R \& 2^{\text{nd}} R$ " and then apply the Conditional Probability Rule to find easily

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

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 [3.5] to be the indicator of a statement. Can you figure out what this formula should be?

4. LECTURE MONDAY 19 JANUARY 2009

No lecture today because of Martin Luther King Day.

5. LECTURE WEDNESDAY 21 JANUARY 2009

What we have been calling our guess for the value of the unknown X given the statement K goes by two technical names among mathematicians and statisticians. The first of these is the *Expected Value* of X given the information in the statement K . The second is *Mean* of X given K . We began by reviewing the four basic rules and the Multiplication Rule ([1.4],[1.5],[1.3],[2.1],[3.4]), which now become the **RULES FOR CONDITIONAL EXPECTATION**.

NORMALIZATION RULE: If K implies that $X = c$, then [2.1]

$$(5.1) \quad E(X|K) = c.$$

POSITIVITY RULE: If K implies that $a \leq X \leq b$, then [1.3]

$$(5.2) \quad a \leq E(X|K) \leq b.$$

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 [1.4]

$$(5.3) \quad E(X + Y|K) = E(X|K) + E(Y|K).$$

HOMOGENEITY: If K implies that $Y = cX$, then [1.5]

$$(5.4) \quad E(Y|K) = E(cX|K) = cE(X|K).$$

MULTIPLICATION RULE:

For any unknown X and any statement N and any background statement K , we have [3.4]

$$(5.5) \quad E(XI_N|K) = E(X|N \& K)P(N|K).$$

The operator E which turns unknowns into expected values is then called the **CONDITIONAL EXPECTATION**. The word conditional is often left out here, but we should keep in mind that all expectations and probabilities are conditional on the information we assume as background, and the results depend heavily on the information assumed. We also reviewed the General Bayes Rule for Conditional Expectation [3.5] and its use in computations with the TI-83/4. We briefly discussed the meanings of the readout from the TI-83/4 when we use it to compute the one variable statistics on tabulated data.

Now at the end of the lecture on Friday, I asked you to think about the meaning of a certain conditional probability that at first may not have made sense to you, involving blocks in a box. 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 $P(2^{nd}R|1^{st}R\&K) = 2/4$. In fact, we can also see that $P(1^{st}R|2^{nd}R) = 2/4$ makes sense as well here even though at first you might think this cannot make sense. But, if we imagine the blocks replaced by playing cards, say three diamonds (red) and two clubs (blue), then you easily see that the same results could be obtained by putting the five cards in a stack and shuffling the stack and dealing one after another from the top of the stack. Then $P(1^{st}R|2^{nd}R)$ is simply asking the chance that the top card is a red card given that the card underneath it is red. This means that we can just as well analyze the blocks in the box problem by thinking of the blocks as stacked but we cannot see the stack. Our question is then to find the probability the top block is red given that the second block from the top is red. From this, we see that we can easily answer very complicated probability questions easily for drawing blocks from a box without replacement. Just imagine the blocks are in a stack which we cannot see and view the information as giving us information about specific locations in the stack. Thus if we want

$$P(4^{th}R|1^{st}B\&5^{th}R),$$

then the answer is obviously $2/3$. That is,

$$P(4^{th}R|1^{st}B\&5^{th}R) = 2/3$$

since we know 2 of the 5 locations in the stack so only three are unknown to us and only 2 red blocks are available for those remaining three locations.

The preceding problems involving blocks should not be thought of as being special. The method used here of thinking in terms of the blocks being stacked instead of drawn one after another is an example of a very general property of the Expectation. Namely, it is all about the state of our information, and in any situation we can where convenient totally re-imagine the set up as long as it has all the same information relationships. The more you practice thinking about such problems, the more creative you become in thinking up useful ways to look at things. As Albert Einstein once said, "imagination is more useful than information".

6. LECTURE FRIDAY 23 JANUARY 2009

We begin by reviewing our basic rules for Expectation.

NORMALIZATION RULE: If K implies that $X = c$, then [2.1]

$$(6.1) \quad E(X|K) = c.$$

POSITIVITY RULE: If K implies that $a \leq X \leq b$, then [1.3]

$$(6.2) \quad a \leq E(X|K) \leq b.$$

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 [1.4]

$$(6.3) \quad E(X + Y|K) = E(X|K) + E(Y|K).$$

HOMOGENEITY: If K implies that $Y = cX$, then [1.5]

$$(6.4) \quad E(Y|K) = E(cX|K) = cE(X|K).$$

MULTIPLICATION RULE:

For any unknown X and any statement N and any background statement K , we have [3.4]

$$(6.5) \quad E(XI_N|K) = E(X|N \& K)P(N|K).$$

We reviewed probability and expectation calculations for simple problems involving the model of equally likely outcomes or fair gambling games.

We practiced using the four basic rules of expectation to calculate expected values of binomial expressions. For example, suppose that we have $E(X) = 5$ and $E(Y) = 7$ and further suppose we are given the fact that $E(XY) = 50$. Notice that 50, the expected value of the product is NOT equal to the product of the expected values of X and Y as that is only 35. A look up at our rules will tell you there is no such rule and in fact such a multiplication rule only works in complete generality when every unknown is actually completely known. There are useful situations where such a rule works, but in general it does not. But any case, if we try to use the rules to figure out

$$E((X - 3)(Y + 4))$$

we begin by multiplying out the binomial expression $(X - 3)(Y + 4)$ so we have, using high school algebra,

$$(X - 3)(Y + 4) = XY + 4X - 3Y - 12$$

and therefore,

$$E((X - 3)(Y + 4)) = E(XY + 4X - 3Y - 12) = E(XY) + 4E(X) - 3E(Y) - 12,$$

so using the given information we find

$$E((X - 3)(Y + 4)) = E(XY) + 4E(X) - 3E(Y) - 12 = 50 + 4 * 5 - 3 * 7 - 12 = 37.$$

That is finally we arrive at the result

$$E((X - 3)(Y + 4)) = 37.$$

For practice try to work out

$$E((X - 2)(Y - 3))$$

and

$$E((X + 3)(Y + 4))$$

and finally,

$$E((X - 5)(Y - 7)).$$

You might notice that in the last case there is a lot of cancellation and in fact the answer is just $50 - 35 - 35 + 35 = 15$. It happens because the number subtracted in each factor is the actual expected value. This is a general fact that follows from the rules of expectation. Notice that $X - 5$ is the *Deviation* of X from its mean or expected value, or in other words, it is the ERROR when you guess X has the value 5. We can denote it simply as D_X . Easy problem. Calculate $E(D_X) = E(X - 5)$. Notice the result has to be 0. That is, if you think your guess should be 5, then when you go ahead and guess 5, then you should be also guessing that your error, D_X , is going to be zero. But, remember in the fair dice problem the guess for the number up should be 3.5 and that is not even a possible value for the number up on the dice, so the error is definitely not going to be zero. What is happening is that some of the errors are positive and some are negative, so overall the best guess is 0, again not even a possible value for the error. This is a general property which follows from the rules. In general, if we write $E(X) = \mu_X$, then $E(X - \mu_X) = \mu_X - \mu_X = 0$. This is because the positive errors are being exactly balanced out by the negative errors. In order to prevent this to get a better idea of our error when we guess, we should actually square the deviation before taking the expected value. Since the Positivity property guarantees that if all possible values are zero and at least one possible value is positive, then the expected value is positive. We define the *Variance* of X to accomplish this.

DEFINITION OF VARIANCE

$$(6.6) \quad \text{Var}(X|K) = E(D_X^2) = E((X - \mu_X)^2|K).$$

To make up for having squared things we take the square root and arrive at what we call the *Standard Deviation*, usually denoted σ_X .

DEFINITION OF STANDARD DEVIATION

$$(6.7) \quad \sigma_X = \sqrt{\text{Var}(X)}.$$

Thus, the standard deviation is always the same as the square root of the variance and the variance is always the same as the square of the standard deviation. If you know one you easily calculate the other with your calculator.

Let us return to the calculation of $E((X - 5)(Y - 7)) = E(D_X D_Y)$ where we noticed a lot of cancellations. In fact that always happens and in fact, we find generally

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

From this we see that $E(D_X D_Y)$ is telling us how far off we would be if we try to guess the product by just multiplying our guesses for the factors. But there is an even better reason for considering $E(D_X D_Y)$ when we have two variables X and Y to deal with. For instance, think of a population of fish where X is length and Y is weight. If we see a fish is longer than average, then it is a case of positive value for D_X , whereas if it is heavier than average it is a case of positive value for D_Y . Now, we know that usually longer fish weigh more even though it is not a hard and fast rule. Likewise, if one of the deviations is negative, the other will often

be negative (a fish shorter than average usually weighs less than average. Thus, the product $D_X D_Y$ will be positive in either of these case. The only time the product of deviations would be negative is for a fish that is longer than average but weighing less than average or for a fish that is shorter than average but weighs more than average. It is very reasonable then that we should have $E(D_X D_Y) > 0$. in this example. Here we say that length and weight are positively correlated and $E(D_X D_Y)$ is a preliminary measure of how well they are related. But there may be a large value for $E(D_X D_Y)$ just due to large deviations in the fish. That is if there are many fish far from average, the deviations will be very large making us think there is a better relationship between the two unknowns than there really is. To compensate for this, we divide by standard deviations. We begin by defining the *Covariance* of X and Y .

DEFINITION OF COVARIANCE

$$(6.8) \quad Cov(X, Y) = E(D_X D_Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X \mu_Y.$$

Notice that $Var(X) = Cov(X, X)$. That is if we think of the variance as some sort of squaring then covariance is the corresponding multiplication.

We then define the *correlation coefficient* of X and Y , generally denoted by $\rho_{X,Y}$, or simply ρ , by dividing by standard deviations.

DEFINITION OF THE CORRELATION COEFFICIENT

$$(6.9) \quad \rho = \rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

It is a mathematical fact that $\rho^2 \leq 1$ and therefore $-1 \leq \rho \leq 1$. I can make up examples by choosing the numbers $E(X), E(Y), \sigma_X, \sigma_Y, \rho$ any way I like as long as the number chosen for ρ is between -1 and 1. We can then go backwards and calculate $Cov(X, Y)$ by noticing that

$$(6.10) \quad Cov(X, Y) = \rho \sigma_X \sigma_Y.$$

From here we then easily calculate $E(XY)$ since

$$(6.11) \quad E(XY) = \mu_X \mu_Y + Cov(X, Y).$$

Try this example. Imagine all the land in Duckburg is divided into rectangular properties whose boundaries run east-west and north-south. Suppose that the average east-west boundary length is 100 feet and that the average north-south boundary length is 200 feet. Suppose that the standard deviation for east-west length is 40 feet and that the standard deviation for north-south length is 70 feet and that the correlation coefficient is $\rho = .8$. Calculate the expected value for the area of these rectangular properties.

Taking the case $X = Y$ in (6.11), we find that the variance of X can be calculate when we know the mean of X and the mean of the square of X , for

$$(6.12) \quad E(X^2) = \mu_X^2 + Var(X),$$

and therefore,

$$(6.13) \quad Var(X) = E(X^2) - \mu_X^2 = E(X^2) - E(X)^2.$$

7. LECTURE MONDAY 26 JANUARY 2009

Today we discuss properties of covariance and variance and go over the practice test. We also computed sample correlation coefficient, r using the TI calculator and made sure that the diagnostics are ON in the calculator so it gives the values for r and r^2 in the readout.

The main thing to keep in mind is that $Cov(X, Y)$ behaves like a multiplication, so FOIL works just as in high school algebra. Therefore, if U, W, X, Y are all variables, then

$$(7.1) \quad Cov(U + W, X + Y) = Cov(U, X) + Cov(U, Y) + Cov(W, X) + Cov(X, Y).$$

Also, $Cov(X, Y) = Cov(Y, X)$. Since $\sigma_X^2 = Var(X) = Cov(X, X)$, if we know $Cov(X, Y)$ and σ_X and σ_Y , then we can calculate $Var(X + Y)$ and take the square root to find the standard deviation for $X + Y$. therefore, when we apply FOIL here, we get

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

and therefore

$$(7.2) \quad Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y).$$

Remember that $Cov(X, Y) = E(D_X D_Y)$ where $D_X = X - \mu_X$ is the deviation of X from its expected value $\mu_X = E(X|K)$. We noticed some simple properties of deviations. In particular, as for a definite number, say the number 3, we have $\mu_3 = E(3) = 3$, and therefore $D_3 = 3 - 3 = 0$. Therefore, the deviation of any definite known number is 0. If c is any definite known number, then $D_c = 0$. This means that $Cov(c, X) = Cov(X, c) = 0$ and $Var(c) = 0$. Putting this together with [7.2] we see that $Var(X + c) = Var(X)$ and therefore adding a constant to a variable cannot change its standard deviation. Thus if salaries of Scrooge's employees average 80K dollars with a standard deviation of 4K dollars, and if Scrooge gives everyone a 10K dollar raise, then all deviations remain unchanged so the standard deviation remains unchanged but the average salary becomes 90K dollars. On the other hand, we observed that if salaries are all doubled, then the average salary doubles, and all the deviations from the mean also double and the standard deviation doubles. If everyone gets a 10% raise, then that is the same as multiplying all salaries by 1.1 and this multiplies the standard deviation by 1.1. Putting these two things together, we see that if Scrooge gives everyone a 20K dollar raise plus 10% of their original salary, then the new mean or average salary is 108K dollars whereas the new standard deviation is 4.4K dollars, the 20K raise everyone got had no effect on the standard deviation.

We reviewed [6.10] the equation $Cov(X, Y) = \rho\sigma_X\sigma_Y$ so that when you are given ρ, σ_X, σ_Y , then you can calculate the standard deviation of $X + Y$ because you can use the fact that variance is the square of the standard deviation to calculate the variance terms on the right hand side of [7.2], use [6.10] to calculate the covariance term on the right hand side, total up, and then take the square root of the result to get the required standard deviation.

8. **LECTURE** WEDNESDAY 28 JANUARY 2009

NO LECTURE. TEST 1 IN LECTURE MEETING.

9. **LECTURE** FRIDAY 30 JANUARY 2009

NO LECTURE.

10. **LECTURE** MONDAY 02 FEBRUARY 2009

We reviewed all the rules of expectation and probability so far, paying particular attention to the multiplication rule

$$(10.1) \quad E(XI_N|K) = E(X|N\&K)P(N|K).$$

We discussed Bayes' Rule and its use in probability computations and computations of expected value. We discussed **INDEPENDENCE** for events and statements. Refer to sections 6 and 7 of Chapter 4 in the textbook. For expected value, we assume we have statements A, B, C, \dots and we know exactly one of these is true, so their indicator unknowns must add up to one and thus the sum of their probabilities must be one.

$$(10.2) \quad 1 = I_A + I_B + I_C + \dots,$$

$$(10.3) \quad 1 = P(A) + P(B) + P(C) + \dots,$$

so multiplying through each term of [10.2] by the unknown X we find

$$(10.4) \quad X = XI_A + XI_B + XI_C + \dots,$$

and therefore we can apply the addition rule to find

$$(10.5) \quad E(X|K) = E(XI_A|K) + E(XI_B|K) + E(XI_C|K) + \dots,$$

and then applying the MULTIPLICATION RULE to each term we find

$$(10.6) \quad E(X|K) = E(X|A\&K)P(A|K) + E(X|B\&K)P(B|K) + E(X|C\&K)P(C|K) + \dots,$$

which means that if we know the expected value or optimal guess in each of the exclusive cases that are possible and if we know the probabilities of each of these cases, then we can simply add up all the products to find the expected value of X . If we take the case where $X = I_R$, where R is some statement or event, then as $E(I_R|L) = P(R|L)$ and as $I_R I_L = I_{R\&L}$ for any statements K, L , we see that from [10.6] we get the general Bayes' Rule for probability.

$$(10.7) \quad P(R|K) = P(R|A\&K)P(A|K) + P(R|B\&K)P(B|K) + P(R|C\&K)P(C|K) + \dots,$$

which allows us to calculate the probability of statement R when we know the chance it happens in each of a bunch of exclusive cases provided that we know the probability each case happens.

We discussed **UNCORRELATION** and **INDEPENDENCE** for statements and for unknowns. Recall that from [6.11] and [6.9]

$$(10.8) \quad \rho\sigma_X\sigma_Y = Cov(X, Y) = E(XY) - E(X)E(Y),$$

so when X and Y are uncorrelated, meaning $\rho = 0$, we can calculate the expected value of the product by simply multiplying the individual expected values. If I guess X has the value 10 and I guess Y has the value 5, then if I also know these unknowns are uncorrelated, I will

guess XY has the value 50. In general this does not work if there is correlation. For instance if I guess that the room is 80 feet wide and 50 feet deep, I would know that in general wider rooms are deeper, so there is positive correlation and therefore I would know that the product $(50)(80)=4000$ square feet would likely be an underestimate for the area of the room.

In case of statements, we say that they are uncorrelated if their indicators are uncorrelated. More generally, we say that the unknowns X, Y are **INDEPENDENT** provided that any statement we can possibly make about the value of X is uncorrelated with any statement we can possibly make about Y . For indicator unknowns it turns out that since their only possible values are either 0 or 1, there is little that can be said and in fact uncorrelation and independence are the same for statements. Thus all these are saying the same thing, namely that A and B are independent statements:

$$P((A\&B) = P(A)P(B),$$

$$P(A|B) = P(A),$$

$$P(B|A) = P(B),$$

$$Cov(I_A, I_B) = 0,$$

and if any one of these is true, then they are all true. Notice that saying $P(A|B) = P(A)$ is saying very directly that knowing B is irrelevant to our determination of the likelihood of A . Thus, if B is irrelevant to the determination of the likelihood of A , then symmetrically, A is irrelevant to the determination of the likelihood of B .

For general unknowns the situation is more complicated. It is possible to have unknowns which are not independent but are uncorrelated, in certain tricky situations. However it is the case that **if X and Y are independent, then they are uncorrelated**. This is useful, because independence is easy to recognize in applications. If we have a pair of dice and we toss them on the table, unless there are some magnets inside or springs connecting them somehow, the two dice are not going to have any real effect on each other. For instance if one is red and the other is blue and I tell you that the red one came up even, that is completely irrelevant to guessing whether or not the blue dice came up even. In physical setups, we can recognize independence easily. To guarantee independence we can also use one repeatable experiment which we do over and over. So, if instead of tossing a pair of dice I toss the same dice twice and tell you that on the first toss it came up even, you would generally think that is irrelevant to the question as to whether it will come up even on the second toss. When we see that unknowns are independent, then we know that they are uncorrelated and we can calculate the expected product by simply multiplying the individual expected values.

11. LECTURE WEDNESDAY 04 FEBRUARY 2009

We covered the addition and multiplication rules for counting as well as tree diagrams. Refer to section 4 of chapter 4.

12. LECTURE FRIDAY 06 FEBRUARY 2009

We covered examples of probability computations using counting and as well covered the hypergeometric distribution for counting successes when sampling (drawing) without replacement. We computed probabilities for all the different 5-card poker hands for the game of poker. For the success count X having the hypergeometric distribution (drawing without replacement) the formula for the probability of k successes in n trials where the population size is N and the population of successes has size R , then the probability $X = k$ is given by

$$(12.1) \quad P(X = k) = C(R, k)C(N - R, n - k)/C(N, n),$$

where $C(N, n)$ denotes the number of ways to choose n things from a set of N things.

13. LECTURE MONDAY 09 FEBRUARY 2009

We covered the binomial distribution for computing probabilities of various numbers of successes in independent trials. We counted the number of arrangements of objects which are not all distinguishable. For instance, the number of arrangements of the letters AABBBCCCC is $9!/(2!3!4!)$, whereas the number of arrangements of 9 distinguishable things would be $9!$. Suppose that we consider any repeatable experiment such as tossing a dice or flipping a coin, and a statement A about the outcome. If we are going to repeat the experiment n times, then we have the unknown X whose value is the number of times that A will happen over the course of the n trials. If we set $q = P(\text{not } A)$, then of course $p + q = 1$. The formula for the distribution of X giving the probability that X has the definite value k is

$$(13.1) \quad P(X = k|n, p) = C(n, k)p^k q^{n-k} = \text{binompdf}(n, p, k),$$

where $C(n, k)$ denotes the number of ways to chose k things from a set of n things. The last term of the equation is what you type in the TI-calculator to get the value of the probability. It is found in the calculators distribution menu which is the second function, "DISTR" of the variables button, denoted "VARS" which in turn is just to the right of the stat button, "STAT". To compute the probability that the number of successes is less than or equal to k , which in symbols is $P(X \leq k)$ use "binomcdf" in the distribution menu which follows "binompdf" in the menu list. For instance, to compute $P(X \leq 40|p = .4, n = 100)$ would take 41 computations using "binompdf" which would then all need to be added up, but "binomcdf" does all the computations and the additions for you all at once. Therefore, to compute $P(30 \leq X \leq 40)$ we must compute

$$P(30 \leq X \leq 40) = \text{binomcdf}(100, .4, 40) - \text{binomcdf}(100, .4, 29).$$

14. LECTURE WEDNESDAY 11 FEBRUARY 2009

Today we reviewed the binomial and hypergeometric distributions as well as methods of counting arrangements. We then discussed examples with the binomial distribution and learned about the Poisson distribution and worked some examples using it. We have now covered chapters 2,4, and 5 as well as a little bit of chapter 3. Test II on Wednesday 18 February in class will cover chapters 4 and 5. Be sure to check the practice test online.

15. LECTURE FRIDAY 13 FEBRUARY 2009

Today we went over the counting distributions-the hypergeometric (12.1), the binomial, and the poisson. We also began reviewing for TEST 2 on Wednesday, 2009-02-18.

16. LECTURE MONDAY 16 FEBRUARY 2009

Today we reviewed for TEST 2 on Wednesday 2009-02-18. We reviewed counting methods, computing the chance that in a room of 30 people at least two have the same birthday, and general methods to determine which counting distribution to use in problems. Remember, first if the sample size (what you examine to count successes) has a continuous measure, the success count is governed by the Poisson distribution. If the sample size consists of a certain whole number n of trials and half a trial makes no sense, then the distribution is binomial or hypergeometric. If successive trials are independent, then the distribution is binomial which is in your distribution menu in the calculator. If the successive trials are not independent such as in drawing without replacement (dealing cards from a deck or drawing blocks from a box without replacement) from a relatively small population with size N , then the distribution is hypergeometric (12.1) and the probability of a given number of successes must be computed directly using the methods we developed for computing probabilities of various poker hands. Whenever the population size is not specified, we assume that it is so large it does not matter whether we draw with or without replacement so we would use the binomial distribution. For instance, if I ask 100 different registered voters in a large city if they will vote in an upcoming election, and if it is known that in general 70 percent of the registered voters vote in elections, then the probability of that I find no more than 65 say they will vote in the upcoming election is *binomial*(100, .7, 65). Even though finding the first person will be a voter in the upcoming election means the next we ask is less likely to be, since the population size of the city is not specified, we assume it is so large that for all practical purposes the probability is still .7 that the next person will be a voter in the upcoming election.

17. LECTURE WEDNESDAY 18 FEBRUARY 2009

NO LECTURE. TEST 2 IN LECTURE MEETING.

18. LECTURE FRIDAY 27 FEBRUARY 2009

Today I announced that there will be an extra test on the second Wednesday in March, which is the eleventh of March. We will have then 5 tests so two test grades will be dropped.

We discussed continuous distributions in general as well as cumulative distribution functions and probability density curves for continuous random variables. Remember the chief distinction between discrete and continuous random variables is that for the discrete ones we count to observe the value whereas with continuous ones we measure to observe the value. This means that for continuous random variables, actual observation of a value entails dealing with the accuracy of measurement. We can never really measure anything with perfect accuracy. We saw with pictures that for a continuous unknown, $P(X = c) = 0$ is always true. That is the probability that X has value exactly c must be zero, no matter what we pick for c . On the other hand, when actual measurements are carried out, we must decide on the level of accuracy in advance. That is we decide how many decimal places we need for our accuracy. To say that X has the value 3.47 to two decimal place accuracy is really to say that the true value of X is somewhere between 3.465 and 3.475, which is a small interval of possible values and which can possibly have a positive probability. We discussed the normal distribution whose density curve is the bell curve. We discussed how to visualize the mean and standard deviation when looking at a bell curve. We used the calculator to calculate probabilities with the normal distribution. We discussed the fact that the normal distribution applies whenever the only information we have about the distribution is the true mean μ and the true standard deviation σ .

19. LECTURE MONDAY 2 MARCH 2009

We discussed the normal distribution and the use of the calculator to calculate centile scores as well as the way to handle expressions involving absolute value. Remember, that the absolute value of the number x is always denoted by $|x|$, and is defined by

$$(19.1) \quad |x| = \sqrt{x^2},$$

and that $|x - y|$ gives the DISTANCE between points x and y on the number line. Therefore an inequality such as

$$|x - c| \leq r$$

is the same as saying that

$$c - r \leq x \leq c + r.$$

Thus, if X is a normal variable, compute $P(|X - c| \leq r)$ with the TI-83or84 calculator we use

$$P(|X - c| \leq r) = \text{normalcdf}(c - r, c + r, \mu, \sigma).$$

Given a number c , to find the score x for which $P(X \leq x) = c$, when X is normal, we use the INVERSE PROCESS in the calculator which in the distribution menu is the `invNorm`. Thus, for normal X ,

$$P(X \leq x) = c$$

is exactly the same as

$$x = \text{invNorm}(c, \mu, \sigma).$$

We also discussed the UNIFORM DISTRIBUTION which is the distribution you would use if the only information you have is the minimum value and the maximum value of the unknown or variable. The picture of a uniform distribution is thus a flat horizontal straight line segment over the interval of possible values whose height is simply determined by the fact that the total area under any distribution curve must equal 1. Probabilities are easy to calculate with the uniform distribution, as

$$(19.2) \quad P(a \leq X \leq b) = \frac{b - a}{\text{max} - \text{min}}.$$

The mean of any distribution can be visualized as the balance point, so obviously the mean of a uniform distribution is the midpoint between the minimum and maximum possible values and therefore,

$$(19.3) \quad \mu = \frac{\text{min} + \text{max}}{2}.$$

To find the standard deviation requires a little calculus, and the result is simply

$$(19.4) \quad \sigma = \frac{\text{max} - \text{min}}{\sqrt{12}}.$$

If we use the fact that

$$2(\text{max} - \mu) = \text{max} - \text{min} = 2(\mu - \text{min}),$$

we can express the standard deviation more simply as

$$(19.5) \quad \frac{\text{max} - \mu}{\sqrt{3}} = \sigma = \frac{\mu - \text{min}}{\sqrt{3}}.$$

Now, $1/\sqrt{3}$ is approximately .5773502692, or roughly 58 percent. Therefore, the distance from the center to a point 58 percent of the way to the edge gives the approximate standard deviation

of a uniform distribution, and this means also that roughly 58 percent of any uniformly distributed population is within one standard deviation of the true mean. For comparison with the normal distribution, by what statisticians call the "rule of thumb" for the normal distribution, we have roughly 68 percent of any normal distribution is within one standard deviation of the true mean. On the other hand, again by the rule of thumb, only roughly 95 percent of any normal distribution is within two standard deviations of the true mean, whereas we see that 100 percent of any uniform distribution is within two standard deviations of the true mean. Knowing the min and the max therefore has its uses.

20. **LECTURE** WEDNESDAY 4 MARCH 2009

Today we reviewed all the distributions covered so far and their comparisons with each other. We also discussed Tchebeyechev's inequality which applies to any distribution.

The distributions you need to know are

NORMAL given μ and σ

UNIFORM given min and max , then $\mu = (max - min)/2$ and $\sigma = (\mu - min)/\sqrt{3}$

POISSON given μ , then $\sigma = \sqrt{\mu}$

BINOMIAL given n and p , then $\mu = np$ and $\sigma = \sqrt{\mu(1 - p)}$

HYPERGEOMETRIC given N and n and either R or p since $R = Np$, then $\mu = np = nR/N$
and

$$\sigma = \sqrt{\mu(1 - p)} \sqrt{\frac{N - n}{N - 1}}.$$

21. LECTURE FRIDAY 6 MARCH 2009

Today we begin **SAMPLING**. The setup is we have a variable X which can be repeatedly observed. We decide to take a sample of size n . This means we will make n observations of X in succession. In advance we have no idea what the values will turn out to be, so they become new unknowns. We use X_1 to denote the unknown value of the first observation, X_2 to denote the unknown value of the second observation, and so on, so finally, X_n denotes the value of the last observation for the sample.

We will be interested in the sample mean as it can be used to make an estimate of the true population mean, and how reliable this procedure is for finding the true mean is what our theory here will tell us.

To begin, we know we will need to add up all the observed values and divide by n to get the sample mean. That is, before we divide by n we must first have the total T_n of all the observed values, which is a new unknown called the **sample total unknown**. Thus we know

$$(21.1) \quad T_n = \sum X_k = X_1 + X_2 + \dots + X_n.$$

Now each observation has the same distribution as X itself. For instance, if $n = 4$ and we are going to roll a fair dice, then X_1 is the result which will come up on the first roll, X_2 is the number which will come up on the second roll, X_3 is the number which will come up on the third roll, and X_4 is the number which will come up on the fourth and last roll. If we ask for the probability that the third roll results in a 5 the answer is $1/6$ just like the probability that the second roll results in a five, which is the probability that X is 5. The probabilities do not change from roll to roll.

Since in general $X_1, X_2, X_3, \dots, X_n$ all have the same distribution as X , it follows that they all have the same mean as X and they all have the same standard deviation as X . We express this with the equations

$$(21.2) \quad E(X_1) = \mu_X, E(X_2) = \mu_X, \dots, E(X_n) = \mu_X,$$

and

$$(21.3) \quad \text{Var}(X_1) = \sigma_X^2, \text{Var}(X_2) = \sigma_X^2, \dots, \text{Var}(X_n) = \sigma_X^2.$$

Now, as pointed out above, when we compute a sample mean of some scores, we have to add up all the scores and divide by the number of scores. The process begins obviously with totalling all the scores. In our present situation, we can wonder as to what the total will turn out to be. We gave the total the symbol T_n from (21.1) which as stated above is now a new unknown.

What do we expect to get for the sample total? What should we guess for the value of T_n ? Well, since we expect to get the true mean μ_X for each observation before we actually take the sample, that is to say, in advance of actually taking the sample, that means that the expected sample total should be simply $n\mu$. For instance, if you expect to get $7/2$ for rolling a dice, then you expect to get a total of 7 when you roll the dice twice and you expect to get a total of 14 when you roll the dice 4 times and you expect to get a total of $35/2$ when you roll the dice 5 times, and so on. This is just a simple result of applying the SUM or ADDITION RULE for expectation. Specifically, here we have by the ADDITION RULE and equation (21.1),

$$(21.4) \quad E(T_n) = E(X_1) + E(X_2) + \dots + E(X_n) = \mu_X + \mu_X + \dots + \mu_X = n\mu_X.$$

or simply

$$(21.5) \quad E(T_n) = n\mu_X.$$

Now, the sample mean itself is just the total divided by the number n of observations. In advance of actually making the observations, the sample mean is also an unknown, and it is given the symbol \bar{X}_n , so we call it the **sample mean unknown**. We can therefore write the equation

$$(21.6) \quad \bar{X}_n = \frac{1}{n}T_n.$$

Well we know that $E(cY) = cE(Y)$ for any number c and any unknown Y by the HOMOGENEITY of expectation, so here this means that

$$(21.7) \quad E(\bar{X}_n) = E((1/n)T_n) = (1/n)E(T_n) = (1/n)n\mu_X = \mu_X,$$

or simply,

$$(21.8) \quad E(\bar{X}_n) = \mu_X.$$

In other words, every time we go to take a sample, in advance of actually getting the data, we always expect to get the true population mean as a result. For instance, if I know the dice is fair and I decide to roll the dice 58 times and average the results, in advance of actually doing this I would know the best guess as to what the average will turn out to be is simply $7/2$, just the same as for a single dice roll.

Now remember, we know it is impossible to get $7/2$ when you roll a dice once, and in general, we know we do not usually get what we expect. However, the standard deviation is in general the number that gives us some indication as to how close the result will actually be to our expected value. If I am about to weigh a fish which comes from a population with average weight 47.3 pounds, and the standard deviation is only 2 pounds, then I am much more likely to find the weight to be near 47.3 pounds than if in the situation of having standard deviation 23 pounds. When standard deviation is large, observed values can stray far from what is expected, whereas when the standard deviation is small, all observed values tend to be close the what is expected.

This means that equation (21.8) is not very useful unless we can find the standard deviation of \bar{X}_n . Here we must recall the way covariance and independence work by using equation (7.2). When the observations are all uncorrelated, that is X_1, X_2, \dots, X_n are all mutually uncorrelated, then the variance of the sum is simply the sum of the variances. We therefore can easily compute the variance of T_n in this situation. More specifically, if X_1, X_2, \dots, X_n are mutually independent we say we have an **INDEPENDENT RANDOM SAMPLE (IRS)**. Since independence implies uncorrelation, this means that assuming IRS, the variance of T_n is simply the sum of the variances of the individual observation unknowns, and they are all equal to σ_X^2 , according to (21.3). This means that

$$(21.9) \quad \text{Var}(T_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) = \sigma_X^2 + \sigma_X^2 + \dots + \sigma_X^2 = n\sigma_X^2,$$

or simply

$$(21.10) \quad \text{Var}(T_n) = n\sigma_X^2.$$

Now we can find the standard deviation of T_n by just taking the square roots of both sides of equation (21.10). Obviously, the result is simply

$$(21.11) \quad \sigma_{T_n} = \sqrt{n}\sigma_X.$$

Notice that the equation giving the expected sample total says we should expect n times what we expect for a single observation whereas the standard deviation is only multiplied by \sqrt{n} .

This means that we can use (21.6) to easily find the standard deviation of \bar{X}_n using the general fact that $\sigma_{cY} = |c|\sigma_Y$ for any unknown or variable Y . We calculate

$$\sigma_{\bar{X}_n} = \frac{1}{n}\sigma_{T_n} = \frac{1}{n}\sqrt{n}\sigma_X = \frac{\sigma_X}{\sqrt{n}},$$

so finally, ASSUMING IRS,

$$(21.12) \quad \sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}.$$

This means that even though the expected sample mean is the true mean, the standard deviation for the sample mean unknown is less than the standard deviation for the unknown. Thus equation (21.12) tells us that the larger the sample size n , the smaller the standard deviation. We can make the standard deviation for the sample mean unknown as small as we please by just making the sample size large enough, provided we are using independent random sampling. In practice, in a finite population, this means sampling WITH REPLACEMENT. This is fortunate, because in a finite population, you can only take a sample of size larger than the population by sampling with replacement.

If we take the case where X is the indicator of the event A , so $X = I_A$, then T_n simply tells the number of times A happened. If we regard A happening as being a success, then T_n is giving the success count. Remember, that $E(I_A) = P(A)$ is really the definition of probability, and when dealing with success counts, we usually denote the success rate by p . Thus we have now $\mu_X = E(I_A) = P(A) = p$. That is to say, when X is an indicator, then $\mu_X = p$ is simply the success rate, T_n is the success count, and now (21.7) says that $E(T_n) = np$. So this means that any time we count successes, we expect to get number of successes equal to the product of the number of trials multiplied by the success rate. This applies then to both the hypergeometric or binomial distributions. However, to find the standard deviation of the success count using our formula we need to assume IRS, which means the success count must be binomial, so we find that the standard deviation of the binomial random variable is $\sqrt{n}\sigma_{I_A}$. To calculate σ_{I_A} we begin by calculating the variance using the general fact (6.13) that

$$\text{Var}(Y) = E(Y^2) - E(Y)^2,$$

for any unknown Y . Since an indicator can only take the values 0 and 1, it is the case that

$$I_A^2 = I_A.$$

So,

$$\text{Var}(I_A) = p - p^2 = p(1 - p),$$

and therefore, for any **binomial** variable with n trials and success rate p , the mean is

$$\mu = np$$

and the standard deviation is

$$(21.13) \quad \sigma = \sqrt{np(1 - p)} = \sqrt{\mu(1 - p)}.$$

22. LECTURE MONDAY 9 MARCH 2009

Today we reviewed the formulas for the sampling theory that was covered on Friday, including formulas for the expected sample total and the expected sample mean. Remember for samples of size n , the expected sample total is $n\mu$ whereas the expected sample mean is exactly the true population or variable mean, μ . Review formulas (21.7) and (21.8) from the previous lecture. In order to compute the standard deviations for these new unknowns, we must assume something about the sampling method. In case of INDEPENDENT RANDOM SAMPLING (IRS), we assume that all the observations are conducted so as to be independent of each other. This means for a finite population, that we must sample WITH REPLACEMENT. The result is that the standard deviation of the sample total is that for the population multiplied by the square root of the sample size, n , whereas for the standard deviation of the sample mean, we DIVIDE by the square root of n . The formulas to look at are (21.11) and (21.12), from the previous lecture.

If we sample randomly but without replacement from a finite population of size N it is called a SIMPLE RANDOM SAMPLE (SRS). This means for a sample of size n we must insure that all possible samples of the same size n are equally likely to be the one we choose. For instance, if we are dealing 5 card poker hands from a standard deck of 52 cards, then the hand constitutes a SRS if all the different possible 5 card hands are equally likely. We do not mean to say that a hand with a pair is as likely as a hand with 4 of a kind, we mean that for instance the hand which has 4 kings and the ace of spades has the same probability as the hand that has the 2,3 7 of hearts, the king of diamonds and the ace of spades. All specific 5 card hands are equally likely. Thus when playing cards, a hand which is a fair deal is a simple random sample. Now, clearly the successive draws in a simple random sample are not independent of each other. For instance, if we draw all 52 cards from the deck there is no doubt as to what we will get. If the first 4 draws result in aces, I know I will not get an ace on the fifth draw. The way that this effects the standard deviation for the sample mean and sample total is to require a correction factor for SRS. That is to say that if we calculate the sample mean standard deviation or the sample total standard deviation under the assumption of IRS, and if the reality is the sample is a SRS, then the standard deviations need to be multiplied by the correction factor $c_{(SRS)}$ which is given by the formula

$$(22.1) \quad c_{(SRS)} = \sqrt{\frac{N-n}{N-1}}.$$

The distinction between the binomial and hypergeometric distributions is simply that the binomial is counting successes using IRS whereas the hypergeometric is counting successes using SRS. Both distributions have mean $\mu = np$ where n is the sample size or number of trials or number of things drawn with or without replacement. But, the binomial has standard deviation $\sqrt{\mu(1-p)}$ whereas the hypergeometric has standard deviation $c_{(SRS)}\sqrt{\mu(1-p)}$. Of course, in a population of size N if there are R successes in the whole population, then $p = R/N$ and therefore $\mu = np = nR/N$, using either IRS or SRS.

23. LECTURE WEDNESDAY 11 MARCH 2009

TEST 2-3 TODAY IN LECTURE MEETING.

24. LECTURE FRIDAY 13 MARCH 2009

Today we began discussing confidence intervals for true population mean μ . We begin with the idea of trying to estimate the true population mean of an unknown population using sampling. If we take a sample of size n and get the sample mean \bar{x} , then that is our initial guess as to the true population mean, but we know there is likely an error which we have reason to hope is not too big if our sample is large. The real question is to say something useful about how large the error might be.

We call the **MARGIN OF ERROR** = ME the bound on the error, so we can say that as $error = \mu - \bar{x}$, saying $error \leq \Delta$ is the same as saying the number Δ is the margin of error, $ME = \Delta$. Thus, if I say that I guess your weight is 120 pounds give or take 10 pounds, I am saying that your actual weight w which I do not really know is somewhere between 110 and 130. Notice I do NOT mean that your weight can only be 110 or 130. Thus, in every day common parlance, the give or take number is the margin of error. In mathematical terms, a direct translation of the phrase "give or take" is \pm , so we will (by what mathematicians sometimes refer to as "abuse of language") write $w = a \pm ME$ to mean that $a - ME \leq w \leq a + ME$.

Now, in dealing with an unknown population, say unknown X , if we cannot take a population census, then there is nothing that we can say with absolute certainty from a mere sample. We say we have confidence C in our margin of error provided that $P(error \leq ME) = C$. In this case, we denote this by writing ME_C for the margin of error. We call the number C here the **LEVEL OF CONFIDENCE**. Thus, we say that $\mu = \bar{x} \pm ME_C$ is the $100 * C$ percent **CONFIDENCE INTERVAL** for the true mean μ in case we have a sample whose sample mean is \bar{x} .

We first deal with the case of a normal population with population mean unknown but with known standard deviation, σ . You might think that it is not realistic to know the standard deviation when you do not know the mean for a population, but in certain industrial settings it is very reasonable because the standard deviation in a machine process is often due merely to the wobbles in the machine whereas the mean of the output itself is usually a result of particular settings of the machine. We have to live with the wobbles, and over a sufficiently long period of time we would say that for practical purposes we know the standard deviation caused by the wobbles in the machine. What might not be clear however is how the machine was set to make a specific batch that we have sitting on the shop floor. If this needs to be determined, then our method here will apply. In many cases, measurements require destroying the things sampled, so sample sizes need to be kept to a reasonable minimum on that consideration, but we will see that it is also advantageous to use large samples when possible.

If we are dealing with a sample of size n , then we know that \bar{X}_n , the sample mean random variable has standard deviation σ/\sqrt{n} . For instance, from the rule of thumb we know there is roughly a 95 percent chance that the value of \bar{X}_n falls within 2 standard deviations of the true mean $E(\bar{X}_n) = \mu_X = \mu$. This means that approximately we can write

$$ME_{.95} = 2 \frac{\sigma}{\sqrt{n}}.$$

If $n = 1600$ and $\sigma = 10$, then $ME_{.95} = 2 * 10/40 = .5$ so even though our standard deviation of the population is 10, using such a large sample guarantees that we have a 95 percent chance that our error is no more than .5 when we use a sample mean as an estimate of the true mean with $n = 1600$.

Notice that we have actually assumed much more than we needed because we really do not need X to be normal, we ONLY need \bar{X}_n to be normal. This is certainly true if X itself is normal, but by the **CENTRAL LIMIT THEOREM**, this is also very approximately true whenever n is very large (in fact exactly true in the limit as n becomes infinite). How large in practice turns out to be simply $n \geq 30$. Thus, we assume here that either X is normal or $n \geq 30$.

Now using the rule of thumb only gives us the approximate Margin of Error, and it limits us to considering only levels of confidence .68, .95, or .997. We need to find ME_C with accuracy and for any value of C which might be assigned. Since the rule of thumb only deals with the numbers of standard deviations, we can notice that the number 2 is really just an approximation of the standard normal score z_C which has the property that $P(-z_C \leq Z \leq z_C) = .95$. To find the actual value, notice that the area under the bell curve outside the region $-z_C \leq Z \leq z_C$ is only .05 with half of this, .025, on each side. Thus, $P(Z \leq -z_C) = .025$ and $P(z_C \leq Z) = .025$. This means that $.975 = P(Z \leq z_C) = \text{invNorm}(.975, 0, 1)$. We can therefore get the precise value that the rule of thumb should have given us for the middle 95 percent of a normal population. It is $\text{invNorm}(.975, 0, 1)$ which is approximately 1.960 to three decimal place accuracy. This is certainly close to 2 and for rough calculations that is often close enough. But for three decimal place accuracy in our calculations, we need to use the inverse normal distribution in the calculator to get the accurate z_C .

More generally, if we want to have confidence C in our ME then we need to choose the number of standard deviations z_C for \bar{X}_n so that

$$(24.1) \quad P(-z_C \leq Z \leq z_C) = C,$$

where Z denotes the standard normal variable. The amount of area outside is $1 - C$ with half on each side again, so the amount to the left of $-z_C$ is $(1 - C)/2$. We therefore have

$$P(Z \leq -z_C) = \frac{1 - C}{2}$$

and

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

Clearly this means on adding up these two areas that

$$P(Z \leq z_C) = C + \frac{1 - C}{2} = \frac{1 + C}{2}.$$

Using the inverse normal distribution we now have

$$z_C = \text{invNorm}\left(\frac{1 + C}{2}, 0, 1\right).$$

We can now say that more generally and accurately than the rule of thumb, that in any normal population there is probability C of being within z_C standard deviations of the mean. In particular, for our sample we can say that the margin of error is given by the formula

$$ME_C = z_C \frac{\sigma}{\sqrt{n}}.$$

25. LECTURE MONDAY 2009 MARCH 16

Today we reviewed for TEST 3 which will be given in lecture class on Wednesday. We observed that if (a, b) is the confidence interval for μ , then it must have been the case that

$$(25.1) \quad \bar{x} = \frac{a + b}{2}$$

and whenever we know a, b , and \bar{x} then we can find the Margin of Error ME as

$$(25.2) \quad ME = b - \mu.$$

We worked examples using stats as well as examples using actual data for both z -intervals and t -intervals.

Given a sample of size n the Margin of Error for the confidence interval with confidence level C is given by

$$(25.3) \quad ME_C = z_C \frac{\sigma}{\sqrt{n}},$$

when σ , the population standard deviation, is known and \bar{X} can be assumed to be normally distributed (which is the case as long as either X is normal or $n \geq 30$). Here z_C is the cutoff score for the standard normal which cuts off a lower tail of area $(1 + C)/2$. Thus,

$$z_C = \text{invNorm}\left(\frac{1 + C}{2}, 0, 1\right).$$

In case the population standard deviation is unknown ($\sigma = ?$), then the margin of error formula is

$$(25.4) \quad ME_C = t_C \frac{s}{\sqrt{n}},$$

when X is normal and t_C is the cutoff score for the t -distribution for $n - 1$ degrees of freedom which cuts off a lower tail of area, again, $(1 + C)/2$. Thus, here

$$t_C = \text{invt}\left(\frac{1 + C}{2}, n - 1\right).$$

Some of you do not have the *invt* in your TI calculator distribution list which is only mildly annoying. This is because the calculator gives the confidence interval for the z -interval or t -interval and if you need the margin of error you can find it from the formulas (25.1) and (25.2) using (25.4). Suppose that d is the number of degrees of freedom. Thus, if we want to know the score $\text{invt}(A, d) = t_d(A)$ which cuts off a lower tail of area A , meaning $P(t \leq t_d(A)) = A$, then the right tail must have area $1 - A$ so if we cut off this same area from the left end of the left tail we arrive at $C = 2A - 1$ as the area from $-t_C$ to $t_C = t_d(A)$, meaning $P(-t_C \leq t \leq t_C) = C$. Thus we have $t_d(A) = t_C$. This means that we can use the t -interval in the calculator with confidence $C = 2A - 1$. Report $\bar{x} = 0$. Since the degrees of freedom d for samples of size n is always $d = n - 1$, we want to report a sample of size $n = d + 1$. Now from equation (25.4) we see that if we report $s_x = \sqrt{n} = \sqrt{d + 1}$, then in the Margin of Error formula we see that s and \sqrt{n} cancel each other out and the formula gives $ME_C = t_C$. Since we reported $\bar{x} = 0$, the resulting confidence interval reported is simply $(-t_C, t_C)$. In this way we find $\text{invt}(A, d) = t_d(A) = t_C = ME_C$. Thus, by reporting the proper fictitious sample statistics in a special way, we fool the calculator into giving us the inverse t -distribution it is using to calculate t -intervals.

26. LECTURE MONDAY 30 MARCH 2009

Today we discussed the **CENTRAL LIMIT THEOREM (CLT)** and its applications to sampling and confidence intervals. The precise statement of the CLT follows.

Theorem 26.1. CENTRAL LIMIT THEOREM. *Suppose that*

$$X_1, X_2, X_3, \dots, X_n, \dots$$

is an infinite sequence of independent unknowns all having the same distribution, and in particular, the same mean μ and standard deviation σ . Let

$$T_n = X_1 + X_2 + \dots + X_n$$

and

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

be the average of $X_1, X_2, X_3, \dots, X_n$. Then the limit as n tends to infinity of T_n and \bar{X}_n are normal. More precisely, if Z_n is the standardization of \bar{X}_n , so

$$Z_n = \frac{\bar{X}_n - \mu}{(\sigma/\sqrt{n})},$$

then the limit as n tends to infinity of Z_n , is Z , a standard normal random variable.

In practical terms the experience of statisticians is that X_n and T_n are reasonably normal when $n \geq 30$.

The way to think of the sequence of unknowns in the CLT is to think of an infinite sequence of independent observations of the same unknown X . Thus we are sampling X with an independent random sample. For instance, if we have a loaded dice which we are going to toss over and over ad infinitum, or if we are going to over and over observe the weight of a fish taken from a large fish pond and we are going to throw each fish caught back into the pond after weighing it, then the sequence of observed values gives such an infinite sequence of unknowns. Before we start sampling, we do not know what the sequence of observations will turn out to be, so they are unknowns. But, since they are all observations of the same unknown X , they all must have the same distribution as X . For instance, in the case of tossing the loaded dice, supposing X is the number up on the dice when tossed, if $P(2 \leq X \leq 5) = .7$, then the probability $P(2 \leq X_8 \leq 5) = .7$ and $P(2 \leq X_{59} \leq 5) = .7$, and so on, ad infinitum. Since the mean and standard deviation of an unknown are completely determined by the distribution, it follows that

$$E(X_n) = \mu_X$$

for every n , and as well, the standard deviation of X_n is

$$\sigma_{X_n} = \sigma_X,$$

for every n .

In practice, in our class, because of the CLT, whenever $n \geq 30$, we will consider both T_n and \bar{X}_n to be NORMALLY DISTRIBUTED.

Notice that in applying the CLT to this case of sampling X , we know that

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

for every n , that

$$\sigma_{T_n} = \sqrt{n}\sigma_X$$

for every n , that

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

for every n , and finally,

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

for every n .

As a result, the standardization of X is Z_X given by

$$Z_X = \frac{X - \mu}{\sigma},$$

so Z_X has mean zero and standard deviation 1. Therefore the standardization of T_n is Z_{T_n} given by

$$Z_{T_n} = \frac{T_n - n\mu}{\sqrt{n}\sigma},$$

and the standardization of \bar{X} is $Z_{\bar{X}_n} = Z_n$, where Z_n is given by the formula in the CLT.

As an illustration of the CLT, we calculated the binomial distribution in a specific example with $n = 20$ and observed the graphical shape of the distribution pictured as a sequence of spikes on a horizontal axis. We noticed that if a curve was drawn through the tops of the spikes it had the shape of a bell curve. We then computed the mean and standard deviation for the binomial distribution and used these values to compute approximate probabilities for the binomial distribution using the normal distribution. The results using the normal distribution were good to almost three decimal place accuracy. In the case of political polling to make confidence intervals the sample sizes need to be in the thousands in order to obtain reasonably small margins of error, and consequently, the normal distribution is highly accurate for estimating proportions using sample sizes that large.

To estimate a probability or true proportion, keep in mind we are dealing with the example where $X = I_A$ is simply an indicator for event A which we generically refer to as success. Thus, recall that $p = P(A) = E(I_A) = \mu_{I_A} = \mu_X$ is called the *success rate*. Since $\sigma_X = \sqrt{p(1-p)}$ in this case, the success count for a given sample is simply an observed value of the total T_n . The sample proportion is denoted \hat{p}_n here, read "pee hat". Since the sample proportion is simply the number of successes divided by the number of trials, we see that $\hat{p}_n = (1/n)T_n = \bar{X}_n$. Thus, for large n , we expect that T_n and \hat{p}_n should be normal. That is to say, that for large samples as used in practical political polling, the sample proportion and sample totals are very accurately normally distributed, unless the true proportions are very extreme (here extreme means near zero or near one—more about this in the next lecture).

Notice that from our previous standardization formulas, that the standardization of \hat{p}_n is $Z_{\hat{p}_n}$ given by

$$Z_{\hat{p}_n} = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

The margin of error for the confidence interval with confidence level C for the estimation of true proportion p is then ME given by

$$ME = z_C \frac{\sigma_{I_A}}{\sqrt{n}} = z_C \frac{\sqrt{p(1-p)}}{\sqrt{n}} = z_C \sqrt{\frac{p(1-p)}{n}} = z_C \sigma_{\hat{p}_n}.$$

When the sample size n is in the thousands, \hat{p}_n is essentially normal to a high degree of accuracy, and the normal distribution can be used to compute the confidence interval margin of error. Your TI-83/4 calculator does this conveniently using the "one-prop z-interval" or "1-PropZInt" from the TEST menu in the stat menu section of the calculator.

27. LECTURE WEDNESDAY 1 APRIL 2009

Today we discussed the normal approximation to the binomial and the method of using the MARGIN OF ERROR formula to determine the sample size when the margin of error must be controlled.

In general, if T_n has the binomial distribution with n the sample size and $p = P(A)$ the success rate, then we are really dealing with sampling $X = I_A$ where A is an event. In this case we have

$$\mu_X = \mu_{I_A} = E(I_A) = P(A) = p,$$

so

$$\mu_{T_n} = np.$$

For the standard deviation of the event indicator $X = I_A$ we have

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

and therefore,

$$\sigma_{T_n} = \sqrt{n}\sigma_X = \sqrt{n}\sqrt{p(1-p)} = \sqrt{np(1-p)} = \sqrt{\mu(1-p)}.$$

Naturally then, if T_n is approximately normally distributed, then the normal distribution which makes the approximation must have mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. On the other hand, as the actual distribution is binomial-remember, we are just counting successes and T_n is the success count for n trials, we know that

$$0 \leq T_n \leq n.$$

Recall now, how we effected the normal approximation to the binomial. The binomial distribution is discrete-the values of T_n can only be whole numbers. We pictured the distribution of T_n as a sequence of spikes on the horizontal axis. The spike over k is the probability that T_n has the value k . We noticed that the tops of the spikes followed a bell shaped curve, so to turn the probabilities into areas, we thickened each spike to a block. The spike that sits over k becomes a block whose left edge is at $k - .5$ and whose right edge is at $k + .5$. We then noticed that as the base length of the block is exactly 1 unit, the area of the block is numerically equal to the height of the spike that it thickened. Thus if we wish to use the normal distribution to approximate the probability that T_n takes the value k , then we should calculate the area under the normal curve from $k - .5$ to $k + .5$ using $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. That is, with the calculator, we should find that

$$\text{binompdf}(n, p, k)$$

and

$$\text{normalcdf}(k - .5, k + .5, np, \sqrt{np(1-p)})$$

are approximately the same when the approximation actually works. One thing we noticed in our picture was the fact that even though the region under the bell curve from $k - .5$ to $k + .5$ was not the same shape as the block, it did look like areas were the same because the part of the block not under the curve appears to be about the same size as the part of the curve that was above the top of the block.

We can see if the approximation actually works by calculating the tail areas where the area should be negligible. Because

$$0 \leq T_n \leq n,$$

if this normal approximation method is working properly, then it should be the case that tails

$$\text{normalcdf}(-\infty, -.5, np, \sqrt{np(1-p)})$$

and

$$\text{normalcdf}(n + .5, \infty, np, \sqrt{np(1-p)})$$

are approximately zero. That is, if we are needing 3 decimal place accuracy, then these tails should each have area less than 0.0005 whereas if we want four decimal place accuracy, these tails should each have area less than 0.00005, and so on.

When the normal approximation is working, that is, assuming each of those tails has area approximately zero to the required number of decimal places, then to compute

$$P(k \leq T_n \leq m),$$

you would add up the areas of the blocks approximated by the corresponding area under the bell curve. This means that

$$P(k \leq T_n \leq m) = \text{binomcdf}(n, p, m) - \text{binomcdf}(n, p, k - 1)$$

and

$$\text{normalcdf}(k - .5, m + .5, np, \sqrt{np(1-p)})$$

should be approximately equal.

For instance, we observed in class that for $n = 30$ and $p = .5$, the tails have area zero with 4 decimal place accuracy, whereas with $n = 100$ and $p = .01$ or with $p = .99$, the tails had to much area for any approximation to work. Now, we know in general, that there is very little area beyond 3 standard deviations in any normal distribution. In fact, by the rule of thumb, we know that each tail has area only about 0.0015. This means that for two decimal place accuracy, we should require that our binomial distribution has the interval from $np - 3\sigma$ to $np + 3\sigma$ entirely contained in the interval from 0 to n . If we calculate the tail areas outside 4 standard deviations for the standard normal, we find each tail has area .00006337 which means that if we require 4 standard deviations to be between 0 and n , then we almost get 4 decimal place accuracy. The tail areas outside 5 standard deviations each have area .0000005742, which means that requiring 5 standard deviations either side of the mean are between 0 and n will almost give 6 decimal place accuracy.

In general, then we can get a high degree of accuracy in the normal approximation when we have a good number of standard deviations either side of the mean np inside the interval $(0, n)$. For k standard deviations, either way to be inside $(0, n)$, we want

$$0 \leq np - k\sqrt{np(1-p)}$$

and

$$np + k\sqrt{np(1-p)} \leq n.$$

Beginning with

$$0 \leq np - k\sqrt{np(1-p)},$$

this inequality is the same as

$$k\sqrt{np(1-p)} \leq np.$$

Squaring both sides we see the inequality is equivalent to

$$k^2 np(1-p) \leq (np)^2,$$

so cancelling np from each side this simplifies to

$$k^2(1-p) \leq np.$$

This last inequality says that k^2 multiplied by the failure rate should be no more than the expected number of successes. Now consider the other inequality,

$$np + k\sqrt{np(1-p)} \leq n.$$

It is equivalent to

$$k\sqrt{np(1-p)} \leq n - np = n(1-p).$$

When we square both sides of this inequality, we see it is equivalent to

$$k^2 p [n(1-p)] = k^2 np(1-p) \leq [n(1-p)]^2,$$

so cancelling $n(1-p)$ from both sides gives the equivalent inequality

$$k^2 p \leq n(1-p).$$

This inequality says that k^2 multiplied by the success rate should be no more than the expected number of failures.

Thus, in terms of standard deviations, we can say that we have k standard deviations either side of the mean within the interval $(0, n)$ provided that the expected number of successes is more than k^2 multiplied by the failure rate and k^2 multiplied by the success rate is no more than the expected number of failures. For instance, let us take the case of $k = 4$. This gives almost 4 decimal place accuracy. If we have $n = 30$ and $p = .4$, then we expect 12 successes. Also $k^2 = 16$ and the failure rate is .6, and 16 multiplied by .6 is only 9.6 which is less than 12, so on this side the approximation is good. For the other side, the expected number of failures is 18 and .4 multiplied by 16 is only 6.4, so we are good on this side as well. Consider the case where $n = 100$ and $p = .03$. Now, with k only 2 we have $k^2 = 4$ so k^2 multiplied by the failure rate is still almost 4 whereas the expected number of successes is now only 3. This means that the left tail of the normal distribution we would be trying to use to make the approximation would have way too much area even for two standard deviations either side of the mean to fit inside $(0, n)$. We would not even get 2 decimal place accuracy here.

In general, for the normal approximation to the binomial to be effective, we need the success rate to stay well above 0 and well below 1. A typical *minimal* criterion is that one should always have at least 5 successes expected and at least 5 failures expected.

When dealing with sample data as in political polling or estimating proportions or probabilities using experimental data, we usually do not know the true value of p , so we assume that we are taking samples large enough that our sample proportion \hat{p} is a good approximation to p . This means that if we are using the 1-propZInt in the calculator to compute a confidence interval for a proportion, then in our data we should have at least 5 successes and at least 5 failures for the method to be accurate. If not, you need a bigger sample.

In general for any unknown X when we want to use a sample mean to estimate μ_X we are often given a number E and are required to make the sample large enough that the margin of error ME is no more than E . We are also required to have the assigned level of confidence C in our interval estimate. That is, we need to control the margin of error and simultaneously maintain our level of confidence.

Recall that with confidence C , the margin of error, ME_C , for the case where we know the standard deviation σ_X is given by

$$ME_C = z_C \frac{\sigma_X}{\sqrt{n}}.$$

Keep in mind that $z_C = \text{invNorm}((1+C)/2, 0, 1)$ so z_C is determined by the level of confidence. To make $ME_C \leq E$, the only control we have is on the sample size. Thus the inequality

$$ME_C \leq E$$

is the same as

$$z_C \frac{\sigma_X}{\sqrt{n}} \leq E$$

which in turn is the same as

$$z_C \frac{\sigma_X}{E} \leq \sqrt{n}.$$

Squaring both sides then tells us that

$$(z_C \frac{\sigma_X}{E})^2 \leq n.$$

In other words, we can find the required sample size by solving the equation

$$z_C \frac{\sigma_X}{\sqrt{n}} = E$$

for n and then rounding up to the nearest whole number.

If we do not know what σ_X is, then we often do know a number B which is at least as big as σ_X . Such a number we call a *bound* for σ_X . In this case, we work conservatively by simply using B in place of σ_X as that will guarantee that $ME_C \leq E$. Thus, if we know that $\sigma_X \leq B$ and we want $ME_C \leq E$, then we choose the sample size

$$n \geq (\frac{z_C B}{E})^2.$$

We should first notice how this inequality causes the sample size to change when one of the factors on the right-hand side changes. Notice that if we double the allowed margin of error E , then the required sample size is only one fourth as large. If we double the bound B , then the sample size is multiplied by four. If we change the level of confidence so as to double z_C , then the required sample size is multiplied by four. More generally, if we change confidence level so as to cause z_C to be multiplied by a factor k , then the sample size is multiplied by k^2 and similarly for a change in B , whereas if we allow the error E to grow k times as big, then the required sample size is divided by k^2 .

Let us return now to the case of political polling. Here, $X = I_A$, so $\sigma_X = \sqrt{p(1-p)}$. But before taking the poll, we do not know what p is. However, as $0 \leq p \leq 1$, it must be that $\sigma_X^2 = p(1-p) \leq (1/2)(1/2)$, as is easily seen by looking at a graph of $y = x(1-x)$ for $0 \leq x \leq 1$. This means that

$$\sigma_{I_A} \leq \frac{1}{2},$$

for any event A . That is, in the polling situation, the bound $B = .5$.

Let's use this result to quickly arrive at some required sample sizes. If we begin with allowed error $E = .01$ and confidence level such that $z_C = 2$, then $z_C B = 1$, so $z_C B/E$ is just $1/E$. This means that n must be at least $(1/E)^2 = (100)^2 = 10000$. Remember that $z_C = 2$ corresponds to a level of confidence slightly better than 95 percent according to the rule of thumb. If we allow our margin of error to be at most $E = .02$ thus doubling it, we can cut our required sample size down by one fourth, so we only need a sample size of 2500, which is in the realm of what is manageable. If we want 99 percent confidence, we are increasing z_C from 2 up to 2.576 and this means multiplying by the factor $(2.576/2)$ which means the required sample size is multiplied by a factor of $(2.576/2)^2 = 1.658944$. For allowing only $E = .01$ this means the required sample size is 16590, since we have to round up to the nearest whole number. If we allow the margin of error to go up to $E = .03$ then we divide the required sample size by 9 and the result is that we only need $n = 1844$. Thus, for a political poll with 99 percent accuracy and margin of error at most 3 percentage points, we need a sample of size 1844. Thus, in T.V. polls, typically sample sizes are 2000.

Of course, when trying to predict election outcomes using exit poll data, the results are sometimes very close to a tie which means to tell who will win requires a very large sample.

If the election results are not close, then a relatively small sample suffices to give enough accuracy to determine a winner. We will deal with this problem next using the method of *HYPOTHESIS TESTING*. This method is better adapted to dealing with decisions. It is one thing to use sample data to estimate the true percentage of voters who vote a certain way. It is another thing to determine from sample data if a certain candidate will win an election.

28. LECTURE FRIDAY 3 APRIL 2009

Today we began the discussion of hypothesis testing with the example of testing rope for mountain climbing in order to determine if it is safe to use. Clearly, we do not want to find ourselves falling through the air attached to a safety rope and at the same time not knowing whether or not the rope will hold us when it comes under the tension caused by our fall. The laws of physics will enable us to calculate a breaking strength for the rope which will be required to hold us in case this dangerous circumstance should come about. In order to determine the breaking strength of the rope we will use, we first begin by noticing that a rope is made of a very large number of identical fibers which all have the same distribution in breaking strength. This means that the actual breaking strength of a piece of rope is the TOTAL of the breaking strengths of the identical fibers making up the rope so by the CENTRAL LIMIT THEOREM, it is very reasonable that rope breaking strength is normally distributed.

If we need the rope to have a 99.9999 percent chance of holding when subjected to a sudden 2000 pound tension force, then we need the mean rope breaking strength of the rope we use to have a large enough mean that the left tail of the distribution having tail area .0001 should have cutoff score at least 2000. If we assume say that the breaking strength has a standard deviation of 50 pounds, then as

$$\text{invNorm}(.0001, 0, 1) = -3.719016541,$$

we need the mean rope breaking strength to be about 3.72 standard deviations above 2000 pounds. With a standard deviation of 50 pounds, this means that having mean breaking strength of 2200 pounds or 4 standard deviations above 2000 will give us a reasonable margin of safety. We therefore begin with the reasonable choice that 2200 pound breaking strength is what we need when dealing with rope having a 50 pound standard deviation in breaking strength.

Next, we must obtain our rope and test some of it to make sure that its breaking strength is above the 2200 pound requirement. To this end, we would take a sample and after testing the rope, we compute the average breaking strength of the pieces of rope in the sample. Now, suppose we test 10 pieces of rope and find that the average breaking strength is 2216 pounds. Does this mean that we can be assured that the population mean breaking strength of the rope we use is over 2200 pounds? NO IT DOES NOT. It could have just been a "lucky" sample, since after all, the rope does vary somewhat in breaking strength. Clearly we need our sample mean to be above 2200 pounds, but how much more?

The relevant question here is: GIVEN THAT THE ROPE IS TOO WEAK ($\mu \leq 2200$), WHAT IS THE CHANCE OUR SAMPLE DATA WOULD LEAD US TO CONCLUDE THAT IT IS STRONG ENOUGH? For instance, if we decide to be satisfied that having sample mean at least 2216, if the true mean is ONLY 2200 pounds what is the chance the sample mean will turn out to be at least 2216? Suppose that the sample size is $n = 16$, so there were 16 pieces of rope actually subjected to breaking strength tests. The sample mean random variable has standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 50/4 = 12.5$ This means that the true mean could be only 2200 pounds in which case the sample mean 2216 is less than 1.4 standard deviations above the mean which is a common occurrence with a normal distribution. This is not good enough. We see here that we could be climbing with rope that is too weak if we accept this as being evidence that the rope is strong enough.

To make a precise probabilistic evaluation of how good our sample data is for proving the rope is strong enough, we calculate the chance we could have found the sample mean to be 2216 or higher (remember the probability that it is 2216 exactly equals zero). We evaluate our data by computing

$$P(\bar{X} \geq 2216 | \mu_X = 2200, \sigma_X = 50, n = 16, \bar{X} \text{ normal})$$

or

$$\text{normalcdf}(2216, \infty, 2200, 12.5) = .100272634,$$

and this means that we could have rope which has mean breaking strength no more than 2200 pounds (including possibly less than 2200) and there is still better than a 10 percent chance that our sample mean turns out to be 2216 or more.

Suppose we replace the 2216 with a larger number. Let us replace the 2216 with 2276. The same calculation now gives a probability of only .00000000128. Roughly one in a billion chance that we could find a sample mean that high if the population of rope we have is too weak. This is more like what we want.

In this type of setting, we call the conditional probability we just calculated the **SIGNIFICANCE** or **P-VALUE** of our data. Clearly we want it to be very small in this situation of mountain climbing. In this situation, if the P-value is very small, we say that the data is *highly significant* as evidence for what we are trying to prove. In this case, we have highly significant evidence that the rope is indeed strong enough to be safe. On the other hand, in the case of the lower sample mean where the P-value is over 10 percent, we would say that the sample data is insignificant to us. Notice that when the P-value is larger as a number we say the data is less significant. Big P-value means low significance whereas small P-value means high significance. Think of the P-value as the chance that you have been fooled by sample data when the actual population of rope is too weak. We will never be able to use the actual piece of rope we have tested, since the test damages the rope. We must take a batch of rope and cut off pieces to test and then infer that the piece we use for climbing is strong enough on the basis of our sample data. This means that there is always a chance that the sample data could accidentally fool us into thinking the rope is strong enough when it is not. Our aim here is to know what the probability of being fooled is so we can properly decide whether to use the rope for climbing.

Because the rope breaking strength is normal, in the TI-83/4 calculator's test menu, we would use the *z-test* which is at the top of the test menu list. We call this a **HYPOTHESIS TEST**. We have here two competing hypotheses, the hypothesis that the rope is too weak ($\mu \leq 2200$) and the hypothesis we are trying to prove which states that the rope is strong enough ($\mu > 2200$). We summarize this by writing-

$$H_0 : \mu \leq 2200$$

versus

$$H_1 = H_{ALT} : \mu > 2200.$$

We call H_0 the **NULL HYPOTHESIS** and we call $H_1 = H_{ALT}$ the **ALTERNATE HYPOTHESIS**. It is the alternate hypothesis that we are trying to prove. The negation of the alternate hypothesis is the null hypothesis.

Now for hypothesis testing to begin, always key in on the alternate hypothesis. That is what you are trying to prove with the data. So to see if (the data proves that) the rope is strong enough, we realize that the alternate hypothesis is $\mu > 2200$. Under the null hypothesis we will calculate the P-value of our evidence under the assumption of that $\mu = 2200$ since that is the number that gives the boundary between the null and alternate hypotheses. We therefore write

$$\mu_0 = 2200$$

here to indicate that we will use the null hypothesis to give us this value for the true population mean. Once that is done we have our hypothesis test in a general form

$$H_0 : \mu \leq \mu_0$$

versus

$$H_1 = H_{ALT} : \mu > \mu_0.$$

To run such a hypothesis test using the calculator, begin in this case by picking the z -test from the test menu. Then we simply report the value of μ_0 which is here 2200, the value of σ which is here 50, the value of \bar{x} which is here 2276, the value of n which is here 16, and then select the alternate hypothesis choice from the possibilities $\mu < \mu_0, \mu \neq \mu_0, \mu > \mu_0$ as the case may be (here $\mu > \mu_0$). Highlight your choice with the cursor and hit the enter button and the read out gives the standardized score of your data as $z = \dots$, the P-value as $p = \dots$, the value of \bar{x} that you entered as well as the value of n that you entered.

29. LECTURE MONDAY 6 APRIL 2009

Today we discussed the basics of hypothesis testing and compared a hypothesis test to a criminal trial. In fact, we can say that a criminal trial is a special case of a hypothesis test. In a criminal trial we have two competing hypotheses.

The CRIMINAL TRIAL NULL HYPOTHESIS = H_0 : **The accused is innocent.**

versus

The CRIMINAL TRIAL ALTERNATE HYPOTHESIS = H_1 : **The accused is guilty.**

It is up to the prosecutor to prove the alternate hypothesis here, the accused need not put up any defense. If the prosecutor has no evidence against the accused, the accused goes free-no defense argument is required.

This is a general feature of **ALL HYPOTHESIS TESTS: WHAT YOU ARE TRYING TO PROVE IS THE ALTERNATE HYPOTHESIS.**

Suppose we consider an example of a murder trial. Suppose that the prosecutor begins by [1] presenting a deadly weapon and [2] presenting expert testimony that the deadly weapon presented has the victim's blood on it, and [3] testimony by an investigator that the weapon was found in a car owned by the accused. Clearly this makes the accused look somewhat guilty, but certainly there might be explanations that a defender could give to counter this evidence.

Suppose next that the prosecutor [4] presents expert testimony and forensic evidence that the blood of the accused is on the murder weapon. This now looks a lot worse for the accused, but the defender might still have an explanation for that.

Suppose next the prosecutor [5] presents eye-witness testimony from a relative of the accused that the victim had a violent argument with the accused approximately one hour before the victim's time of death at which time the accused threatened to kill the victim.

Suppose next that the prosecutor [6] presents as evidence an item of clothing owned by the accused having the victim's blood on it.

The typical juror at this point is probably thinking the accused must be guilty.

Notice that each time the prosecutor presents another piece of evidence we are more inclined to think the accused is guilty. Unless the defender can counter some of this evidence, the accused would probably be convicted by a jury hearing this evidence.

Let us consider this evidence from the standpoint of probability theory. In a criminal trial, the accused enjoys the presumption of innocence. Thus, we must evaluate the evidence under the assumption that H_0 is true. This means we should consider the *conditional* probabilities

$$\begin{aligned} P([1] | H_0), \\ P([1] \& [2] | H_0), \\ P([1] - [3] | H_0), \\ P([1] - [4] | H_0), \\ P([1] - [5] | H_0), \\ P([1] - [6] | H_0). \end{aligned}$$

Obviously as we go down this list, the probabilities are getting smaller and smaller. As the evidence piles up, the conditional probability that evidence so contradictory of H_0 *under the assumption of H_0* begins to be very small. It is in this sense that the evidence must be evaluated.

There is a problem here with simply evaluating the probability of the evidence under the assumption of H_0 . Almost anything that happens in the real world has a very low probability. For instance, suppose that the prosecutor [7] presents evidence that the accused knows how to speak Russian and that in general only a very small percent of the population knows how to

speak Russian. The prosecutor next [8] points out that the accused has one green eye and one blue eye, a very very rare condition. The prosecutor [9] presents evidence that the accused is from a tiny village in the Himalayas and points out that the probability of being from such a place is very small. Finally the prosecutor [10] presents evidence that it is snowing in New Orleans. We might have

$$P([7] - [10] | H_0) < P([1] - [6] | H_0).$$

But evidence [7]-[10] is completely irrelevant unless the prosecutor has a way of making it relevant. But, if we only went on the basis of probability, the prosecutor could make a stronger case by only using the irrelevant evidence. Clearly that is not what we want.

What we have to do is consider all the possibilities for evidence that is as or more contradictory than the evidence we have. Notice that anything is as or more contradictory of H_0 than [7]-[10], and the probability of anything is 1. It is in this way that the irrelevant evidence is ruled out with probability. On the other hand, it is definitely not the case that anything is as or more contradictory of H_0 than [1]-[6]. We can thus say that if A denotes the set of all possible statements the prosecutor can make that are as or more contradictory of H_0 than the evidence [1]-[6] and if B is the set of all possible statements the prosecutor can make that are as or more contradictory of H_0 than [7]-[10], then clearly

$$P(A|H_0) < P(B|H_0).$$

It is $P(A|H_0)$ that we call the **SIGNIFICANCE OF THE EVIDENCE** or the **P-VALUE OF THE EVIDENCE**. In order for the prosecutor to convince the jury that the accused is guilty, it must be the case that the jury thinks that $P(A|H_0)$ is some very tiny number-zero for all practical purposes, which would mean "the accused is guilty beyond all reasonable doubt".

In a more typical situation involving statistics and probability directly, we can consider a box containing many many blocks. Suppose we consider the hypothesis test

H_0 : the box contains no red blocks

versus

H_{ALT} : the box contains at least one red block.

To work this hypothesis test, we need evidence or data. Suppose that we reach into the box, grab a block and see that it is red. Our evidence is in direct contradiction of the null hypothesis, H_0 . We see that $P(\text{get a red block from the box} | H_0) = 0$, which means that the significance or P-value of this data is numerically equal to zero. We say that the data is *highly significant* in this case. Notice that "high" significance means significance is low in numerical value, so the terminology can be confusing. In the case here, the evidence is as significant as it can get, since it is in direct contradiction of the null hypothesis.

Consider now a modification of these hypotheses to form the new hypothesis test

H_0 : the box contains at least 70 percent red blocks

versus

H_1 : the box contains less than 70 percent red blocks.

Now finding a red block or a block that is not red either way is virtually useless. Suppose that we take a sample of 1000 blocks from the box and find only 657 red blocks. If we assume H_0 then we expect at least 700 red blocks, so we see that our data is somewhat contradictory of the null hypothesis. How do we measure this? Well again we must consider all possible values for the number of red blocks in a sample of size 1000 that are as or more contradictory than what we found in our data. That is, if X is the number of red blocks in a sample of size 1000, then as or more contradictory than our data means $X \leq 657$. Let $P = P(\text{red})$ be the true probability of getting a red block from the box. Thus P is the true proportion of red blocks in the box. With these symbols, we have

H_0 : $P \geq .7$

versus

H_1 : $P < .7$,

and the condition of being as or more contradictory than our data

$$X \leq 657.$$

See how the last inequality whose probability gives the P-value is exactly "parallel" to the inequality in the alternate hypothesis. This is a general feature of hypothesis tests, and it allows the calculator to compute the probability for the P-value as soon as you choose the alternate hypothesis. To work this with the calculator, begin by going to the "1-prop Z Test". We set $p_0 := .7$ as the null hypothesis value of the true proportion, enter $x := 657$, enter $n := 1000$, choose the alternate hypothesis $p < p_0$, and hit the enter button. In the readout you need to be aware that the P-value is reported as p . The value reported as \hat{p} is the proportion in the sample data, here .657.

If we are interested in knowing if our poll data proves Sen. Snort will win the upcoming election, our alternate hypothesis is what we are trying to prove

$$H_1 : p > .5$$

because to win he needs more than 50 percent of those who vote to vote for him. Thus, in this case $p_0 = .5$

Our hypothesis test is therefore

$$H_0 : p \leq .5$$

versus

$$H_1 : p > .5.$$

30. LECTURE WEDNESDAY 8 APRIL 2009

Today I discussed the effect of irrelevant evidence in a criminal trial as a way to see that when we evaluate the significance of data in a hypothesis test, we must compute not just the probability of our evidence but in fact the probability of all evidence that is as or more contradictory of the null hypothesis than our evidence. For instance if you say you got up at 9am this morning, and we consider the probability that you got up at exactly 9am, the result is zero, since time is a continuous variable and the probability that it takes any definite exact value is zero. Anything that happens in the world is something of very low probability. It is only when we consider many similar possibilities that we can find a reasonably positive probability. If you awake almost every morning between 8:30am and 9:30am, then the probability that you awoke at some time between 8:55am and 9:05am is probably not so small. In the case of evidence in a criminal trial, if the jury members simply compute the probability of the presented evidence given the null hypothesis, then the prosecutor can be very successful by simply presenting a lot of evidence for irrelevant things that have low probability. If the prosecutor [1] presents DNA evidence that the blood of the accused is on the victim and DNA evidence that the blood of the victim is on the accused, then this is very strong evidence of very low probability. However, there are irrelevant facts that have even lower probability. For instance, the prosecutor could [2] present the entire DNA sequence of the victim and claim that the probability that someone had that sequence is far less than the probability of [1] given that the accused is innocent. However let us imagine a sample space S consisting of all possible true facts that the prosecutor could bring up. Then the evidence [1] is a very tiny subset of S . Consider now the set A of all statements which are as or more contradictory of H_0 than [1]. It is slightly bigger set of statements than [1]. Let B be the set of all statements which are as or more contradictory of the null hypothesis than [2]. Since [2] is irrelevant to H_0 , it follows that $B = S$, that is, any statement the prosecutor might make is as contradictory of H_0 as [2]. It follows that however we define probability here, we should have

$$P(A|H_0) \ll P(B|H_0) = P(S|H_0) = 1.$$

Thus, the method of evaluating the evidence by throwing in all possibilities as or more contradictory than our evidence leads to a P-value equal to 1 for irrelevant information. We can also see here that if C is the set of all possible statements that are as or more contradictory of H_0 than "[1]and[2]", then $C = A$. That is, when we calculate the P-value of evidence by this method, the irrelevant information has no effect on the resulting P-value.

In the case of the hypothesis test

$$H_0 : \mu \leq 1000$$

versus

$$H_1 : \mu > 1000,$$

if the sample mean is $\bar{x} = 1023$, then we can say the data definitely favors the alternate hypothesis as it is somewhat contradictory of the null hypothesis. However, we can also see here that if the sample mean had been 1025, it would be even more contradictory of the null hypothesis. To calculate the probability of all possible results which could have been the result of our sample method but as or more contradictory of H_0 than ours, we thus calculate

$$P - value = P(\bar{X} \geq 1023 | H_0).$$

Notice we are calculating the area of a tail of the distribution, so we call this a *one tail test* or a *one-sided test*.

On the other hand, if we are dealing with the hypothesis test

$$H_0 : \mu = 1000$$

versus

$$H_1 : \mu \neq 1000,$$

then there is no difference between being too high or too low as far as contradicting the null hypothesis is concerned. Thus, if we are dealing with a normal distribution, as it is symmetric about the mean, as we compute by assuming H_0 , we are looking at a distribution which is symmetric about 1000. This means that being 23 units below 1000 is just as contradictory of H_0 as being 23 units above 1000. That is, now the P-value is

$$\begin{aligned} P - \text{value} &= P(\bar{X} \geq 1023 \text{ or } \bar{X} \leq 977 | H_0) \\ &= P(\bar{X} \geq 1023 | H_0) + P(\bar{X} \leq 977 | H_0) \\ &= 2P(\bar{X} \geq 1023 | H_0). \end{aligned}$$

We therefore finally see that for disproving a null hypothesis of exact equality, we have

$$P - \text{value} = 2P(\bar{X} \geq 1023 | H_0).$$

Notice that the P-value is the area of two tails of the distribution in this case, so we call this a *two tail test* or a *two-sided test*. Any time we try to prove something not equal to a definite value, we have a two tail test. The P-value is thus double the resulting P-value for the one tail test where we try to prove the one-sided inequality favored by the data. Thus, if our sample mean had been $\bar{x} = 977$, then our P-value would have been calculated as twice the P-value of this data as evidence that $\mu < 1000$, since that inequality is favored by the sample mean of our data. Thus the P-value for this one-sided test is

$$P - \text{value} = P(\bar{X} \leq 977 | H_0)$$

and then the P-value for the two tail test is

$$P - \text{value} = 2P(\bar{X} \leq 977 | H_0).$$

In general, we should also notice that if we use the data $\bar{x} = 1023$ to try to prove $H_1 : \mu < 1000$, which is clearly not favored by our data in this case, then the P-value is given by

$$P - \text{value} = P(\bar{X} \leq 1023 | H_0) = 1 - P(\bar{X} \geq 1023 | H_0)$$

which shows that if we arrive at the P-value p using a one tail test, then reversing the one-sided inequality of the alternate hypothesis changes the P-value to $1 - p$. This means that with a symmetric distribution in particular, if the data favors the alternate hypothesis in a one tail test, then the P-value must be less than .5 whereas if it favors the null hypothesis the P-value will be over .5. A common mistake when using the calculator to run hypothesis tests is forgetting to chose the proper alternate hypothesis. As it is usually obvious which hypothesis the data favors, we can always easily tell whether the P-value should be bigger than .5 or less than .5 before we run a one tail test. If the P-value in a one tail test comes out bigger than .5 when the data favors the alternate hypothesis, we know we made a mistake. If you are using the calculator and have chosen the proper test in the test menu and have entered all the data correctly, then you have probably forgotten to chose the proper alternate hypothesis.

Suppose that we were dealing with a two tail test concerning the true mean of a continuous variable or unknown and the distribution of \bar{X} is not symmetric about the mean. For instance, in the discrete case, this is the situation when dealing with binomial distributions where the

success rate is not .5. If we have a sample mean of 1023, and we want to run the two tail test to try to prove say that the true mean is not 1000, then we do not know what value below 1000 is just as contradictory on the low side as 1023 is on the high side. But there is some number $b < 1000$ which is just as contradictory as is 1023 of the null hypothesis that $\mu = 1000$. But, what does it mean to say $\bar{X} = b$ is just as contradictory of $\mu = 1000$ as $\bar{X} = 1023$. It really just means that

$$P(\bar{X} \leq b|H_0) = P(\bar{X} \geq 1023|H_0).$$

Thus even in this non-symmetric case, we can still say that

$$P - value = 2P(\bar{X} \geq 1023|H_0).$$

If we have a situation where we are running a two tail test, that is a test to try to prove that $\mu \neq 1000$, then which ever side of the hypothetical mean 1000 the data favors, if we run the one-sided tests, the P-values will be p and $1 - p$. One of these is less than .5 and thus which ever that is, multiplying by 2 gives the P-value for the two tail test. To summarize, when dealing with continuous unknowns, if we run any one tail test and arrive at the number p as the P-value, then the P-value for the two tail test is just

$$P - value = 2[\min\{p, 1 - p\}],$$

where $\min\{p, 1 - p\}$ is the minimum of the two numbers p and $1 - p$.

If we are dealing with a discrete distribution, such as the binomial distribution which would be the case in political polling situations, to disprove true proportion = .3, if the sample result is 4 out of 20, then we see the sample proportion is $\hat{p} = .2$ which favors true proportion less than .3, so the P-value is

$$P - value = 2 * \text{binomcdf}(20, .3, 4) = .4750155578.$$

Thus, in the discrete case, we can say that we should compute the P-value for the one tail test with alternate hypothesis favored by the data, and then we simply double that to arrive at the P-value for the two tail test.

31. **LECTURE** FRIDAY 10 APRIL 2009

EASTER VACATION NO LECTURE MEETING.

32. **LECTURE** MONDAY 13 APRIL 2009

EASTER VACATION NO LECTURE MEETING.

33. **LECTURE** WEDNESDAY 15 APRIL 2009

Today we reviewed hypothesis testing and discussed the *TYPE I ERROR* and the *TYPE II ERROR*. Given the hypothesis test H_0 versus H_A , where H_0 denotes the null hypothesis and H_A denotes the alternate hypothesis, there are two different errors that are possible. Remember that in the hypothesis test, we look at the data to see if it is *statistically contradictory* of the null hypothesis, H_0 . We are attempting to use our data to disprove H_0 and prove H_A . If we draw the conclusion that we have proven H_A , which is to say we *reject* H_0 , and if in fact H_0 is actually true, then we have made an error called the **TYPE I ERROR**. The other error we could make, the **TYPE II ERROR**, is that actually H_A is true, but our data fails to convince us of that fact so that we regard our data as inconclusive. Notice that when we make the type I error we have actually drawn a conclusion which might then lead us to act in a dangerous fashion. In the case of the type II error, we are left in the dark as to which hypothesis is true even though actually H_A is true. The type II error does not usually lead to action so is less dangerous. For instance in the case of a criminal trial, remember the null hypothesis is that the accused is actually innocent whereas, the alternate hypothesis is that the accused is guilty. The type I error here is that the prosecutor proves the alternate hypothesis but the null hypothesis is actually true. This means an innocent person is convicted and penalized, possibly losing his life. The type II error here is that a guilty person goes free. Here, if the person is guilty, there is always the possibility that a subsequent trial on a slightly different charge results in guilt. For instance, O. J. Simpson was found innocent in his criminal trial for murder, but was subsequently found guilty in a civil lawsuit and penalized as a result. Thus, we generally think of the worst outcome of a criminal trial is to penalize an innocent person, and the criminal trial procedure is set up to first minimize that possibility.

In the case of mountain climbing with rope, we would set up our hypothesis test so that we use the rope only if our data definitely proves the rope's strength exceeds what is necessary to be safe. Thus we would take this hypothesis as our alternate hypothesis and this means that our null hypothesis should be H_0 : *the rope is not strong enough*. The type I error leads us to conclude that the rope strength definitely exceeds what is necessary so will be safe for climbing when in fact it is not safe for climbing. The type II error is where the rope actually is strong enough, but our data does not lead us to conclude that it is strong enough so we pass up using the rope for mountain climbing. In this situation, if we need rope for mountain climbing our action would be to continue our search, clearly not a dangerous outcome by comparison to the type I error.

Because of the danger of a type I error, we always set up a hypothesis test so that the most dangerous error is the type I error. In hypothesis testing we call the **conditional probability**:

$$P(\text{Type I Error}) = P(\text{reject } H_0 | H_0 \text{ true}),$$

the probability of the type I error. Notice the sloppiness in terminology here. The probability of the type I error is a conditional probability. We really would want to calculate the probability that "we reject H_0 and H_0 is true", but we usually do not have sufficient information to calculate this probability. We must settle for the conditional probability probability that we reject H_0 under the assumption (given) that H_0 is true.

If we work at the level of significance α , then

$$P(\text{Type I Error}) = \alpha,$$

so we see that hypothesis testing is designed to give control of the type I error probability. To see this more clearly, let us remember that we will reject H_0 if the P-value of our data does not exceed the level of significance, $P - value \leq \alpha$. If we are testing say

$$H_0 : \mu \leq 100$$

versus

$$H_A : \mu > 100,$$

then obviously, the larger our sample mean, the smaller the P-value of our data and the more likely the P-value will not exceed α . Clearly, there is a minimum or critical value \bar{x}_c with the property that its P-value is exactly α . Thus,

$$P(\bar{X} \geq \bar{x}_c | H_0) = \alpha.$$

We then see that if the sample mean of our data $\bar{x}_{data} \geq \bar{x}_c$, then the P-value of our data is less than or equal to α and we will reject the null hypothesis. That is, here rejecting the null hypothesis is the same as $\bar{x}_{data} \geq \bar{x}_c$. Remember, the P-value is always computed under the assumption that the null hypothesis is true. Thus, if $\bar{x} \geq \bar{x}_c$ and H_0 is true, we will make the type I error. This means

$$P(\text{Type I Error}) = P(\bar{X} \geq \bar{x}_c | H_0) = \alpha.$$

34. LECTURE FRIDAY 17 APRIL 2009

We are still discussing hypothesis testing. Remember that to calculate the P-value of our data we always calculate

$$P(\text{data as or more contradictory of } H_0 \text{ than ours} | H_0 \text{ true}).$$

For a hypothesis test concerning the true mean of the continuous variable X which is assumed normal,

$$H_0 : \mu \leq 100$$

versus

$$H_A : \mu > 100,$$

if \bar{x}_d denotes the sample mean of our data, then anyone else taking a sample and finding a sample mean $\bar{x} \geq \bar{x}_d$ has data as or more contradictory of H_0 than our data. Thus the P-value of our data is

$$P\text{-value} = P(\bar{X} \geq \bar{x}_d | \mu = 100).$$

Notice that the form of the inequality used to calculate the P-value exactly mimics that of the alternate hypothesis H_A . With the inequalities arranged so that the symbols are on the same side and the definite numbers on the same side (\bar{x}_d is a definite number determined by the data), then the the direction of both inequalities is the same. We can always obtain the inequality whose probability needs to be calculated for the P-value by simply mimicking the inequality in the alternate hypothesis.

As an example, suppose that we are counting the number of tadpoles per gallon of pond water. If we suspect that data will prove the number of tadpoles per gallon is less than 8, and if our sample is going to be the number of tadpoles we find in a sample of five gallons of pond water, then our hypothesis test should properly be formulated in terms of what we expect for the sample. If μ denotes the expected number of tadpoles in five gallons, then we are trying to prove the alternate hypothesis $H_A : \mu < 40$. Thus the null hypothesis is $H_0 : \mu \geq 40$. If we find only 35 tadpoles in our five gallon sample, then the P-value is $P(X \leq 35 | \mu = 40)$. Here, we use X to denote the number of tadpoles in a five gallon sample of pond water. This should reasonably obey the Poisson distribution with $\mu = 40$, or to remind us that this is the assumption resulting from the null hypothesis, we would better denote this as $\mu_0 = 40$. Again, notice how the inequality in the P-value calculation mimics the inequality in the alternate hypothesis. Thus, in this example, the P-value is simply (using the TI-83/84 calculator)

$$P\text{-value} = P(X \leq 35 | \mu_0 = 40) = \text{poissoncdf}(40, 35).$$

35. LECTURE MONDAY 20 APRIL 2009

We reviewed for TEST 4.

36. LECTURE WEDNESDAY 22 APRIL 2009

TEST 4 in lecture meeting.

ANNOUNCEMENT :

FINAL EXAM
JONES HALL ROOM 204
8AM-NOON FRIDAY 1 MAY 2009

Notice the exam time is NOT time for classes meeting MWF at our lecture time and the place for the exam will not be our usual classroom.

37. LECTURE FRIDAY 24 APRIL 2009**ANNOUNCEMENT:**

FINAL EXAM
JONES HALL ROOM 204
8AM-NOON FRIDAY 1 MAY 2009

Notice the exam time is NOT time for classes meeting MWF at our lecture time and the place for the exam will not be our usual classroom.

Today we begin reviewing for the **FINAL EXAM**. We reviewed basic probability rules, rules for conditional probability, and their use in calculating probabilities. We also discussed the normal distribution, the hypergeometric distribution and the binomial distribution, and calculations of probabilities and expectations.