Stat 110 Strategic Practice 3, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Continuing with Conditioning

1. Consider the Monty Hall problem, except that Monty enjoys opening Door 2 more than he enjoys opening Door 3, and if he has a choice between opening these two doors, he opens Door 2 with probability p, where $\frac{1}{2} \leq p \leq 1$.

   To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is Door 1. Monty Hall then opens a door to reveal a goat, and offers you the option of switching. Assume that Monty Hall knows which door has the car, will always open a goat door and offer the option of switching, and as above assume that if Monty Hall has a choice between opening Door 2 and Door 3, he chooses Door 2 with probability p (with $\frac{1}{2} \leq p \leq 1$).

   (a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of Doors 2,3 Monty opens).

   (b) Find the probability that the strategy of always switching succeeds, given that Monty opens Door 2.

   (c) Find the probability that the strategy of always switching succeeds, given that Monty opens Door 3.

2. For each statement below, either show that it is true or give a counterexample. Throughout, X, Y, Z are discrete random variables.

   (a) If X and Y are independent and Y and Z are independent, then X and Z are independent.

   (b) If X and Y are independent, then they are conditionally independent given Z.

   (c) If X and Y are conditionally independent given Z, then they are independent.

(d) If X and Y have the same distribution given Z, i.e., for all a and z, we have $P(X = a|Z = z) = P(Y = a|Z = z)$, then X and Y have the same distribution.

2 Simpson's Paradox

1. (a) Is it possible to have events A, B, E such that $P(A|E) < P(B|E)$ and $P(A|E^c) < P(B|E^c)$, yet $P(A) > P(B)$? That is, A is less likely under B given that E is true, and also given that E is false, yet A is more likely than B if given no information about E. Show this is impossible (with a short proof) or find a counterexample (with a "story" interpreting A, B, E).

   (b) Is it possible to have events A, B, E such that $P(A|B, E) < P(A|B^c, E)$ and $P(A|B, E^c) < P(A|B^c, E^c)$, yet $P(A|B) > P(A|B^c)$? That is, given that E is true, learning B is evidence against A, and similarly given that E is false; but given no information about E, learning that B is true is evidence in favor of A. Show this is impossible (with a short proof) or find a counterexample (with a "story" interpreting A, B, E).

2. Consider the following conversation from an episode of *The Simpsons*:

   > Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
   >
   > Homer: *Lisa, a guy who's got lots of ivory is less likely to hurt Stampy than a guy whose ivory supplies are low.*

   Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

   (a) Define clear notation for the various events of interest here.

   (b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

   (c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's

reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

3 Gambler's Ruin

1. A gambler repeatedly plays a game where in each round, he wins a dollar with probability 1/3 and loses a dollar with probability 2/3. His strategy is "quit when he is ahead by \$2," though some suspect he is a gambling addict anyway. Suppose that he starts with a million dollars. Show that the probability that he'll ever be ahead by \$2 is less than 1/4.

4 Bernoulli and Binomial

1. (a) In the World Series of baseball, two teams (call them A and B) play a sequence of games against each other, and the first team to win four games wins the series. Let p be the probability that A wins an individual game, and assume that the games are independent. What is the probability that team A wins the series?

   (b) Give a clear intuitive explanation of whether the answer to (a) depends on whether the teams always play 7 games (and whoever wins the majority wins the series), or the teams stop playing more games as soon as one team has won 4 games (as is actually the case in practice: once the match is decided, the two teams do not keep playing more games).

2. A sequence of n independent experiments is performed. Each experiment is a success with probability p and a failure with probability $q = 1 - p$. Show that conditional on the number of successes, all possibilities for the list of outcomes of the experiment are equally likely (of course, we only consider lists of outcomes where the number of successes is consistent with the information being conditioned on).

3. Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, independent of X.

   (a) Show that $X + Y \sim \text{Bin}(n + m, p)$, using a story proof.

   (b) Show that $X - Y$ is *not* Binomial.

   (c) Find $P(X = k|X + Y = j)$. How does this relate to the elk problem from HW 1?

Stat 110 Strategic Practice 3 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Continuing with Conditioning

1. Consider the Monty Hall problem, except that Monty enjoys opening Door 2 more than he enjoys opening Door 3, and if he has a choice between opening these two doors, he opens Door 2 with probability p, where $\frac{1}{2} \leq p \leq 1$.

   To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is Door 1. Monty Hall then opens a door to reveal a goat, and offers you the option of switching. Assume that Monty Hall knows which door has the car, will always open a goat door and offer the option of switching, and as above assume that if Monty Hall has a choice between opening Door 2 and Door 3, he chooses Door 2 with probability p (with $\frac{1}{2} \leq p \leq 1$).

   (a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of Doors 2,3 Monty opens).

   Let C_j be the event that the car is hidden behind door j and let W be the event that we win using the switching strategy. Using the law of total probability, we can find the unconditional probability of winning in the same way as in class:

$$\begin{aligned} P(W) &= P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) \\ &= 0 \cdot 1/3 + 1 \cdot 1/3 + 1 \cdot 1/3 = 2/3. \end{aligned}$$

   (b) Find the probability that the strategy of always switching succeeds, given that Monty opens Door 2.

   A tree method works well here (delete the paths which are no longer relevant after the conditioning, and reweight the remaining values by dividing by their sum), or we can use Bayes' rule and the law of total probability (as below).

Let D_i be the event that Monty opens Door i. Note that we are looking for $P(W|D_2)$, which is the same as $P(C_3|D_2)$ as we first choose Door 1 and then switch to Door 3. By Bayes' rule and the law of total probability,

$$\begin{aligned} P(C_3|D_2) &= \frac{P(D_2|C_3)P(C_3)}{P(D_2)} \\ &= \frac{P(D_2|C_3)P(C_3)}{P(D_2|C_1)P(C_1)+P(D_2|C_2)P(C_2)+P(D_2|C_3)P(C_3)} \\ &= \frac{1 \cdot 1/3}{p \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} \\ &= \frac{1}{1+p}. \end{aligned}$$

(c) Find the probability that the strategy of always switching succeeds, given that Monty opens Door 3.

The structure of the problem is the same as part (b) (except for the condition that $p \geq 1/2$, which was no needed above). Imagine repainting doors 2 and 3, reversing which is called which. By part (b) with $1-p$ in place of p, $P(C_2|D_3) = \frac{1}{1+(1-p)} = \frac{1}{2-p}$.

2. For each statement below, either show that it is true or give a counterexample. Throughout, X, Y, Z are discrete random variables.

(a) If X and Y are independent and Y and Z are independent, then X and Z are independent.

False: for a simple example, take $X = Z$.

(b) If X and Y are independent, then they are conditionally independent given Z.

False: this was discussed in class (the fire-popcorn example) in terms of events, for which we can let X, Y, Z be indicators.

(c) If X and Y are conditionally independent given Z, then they are independent.

False: this was discussed in class in terms of events (the chess opponent of unknown strength example); a coin with a random bias (as on HW 2) is another simple, useful example to keep in mind.

(d) If X and Y have the same distribution given Z, i.e., for all a and z, we have $P(X = a|Z = z) = P(Y = a|Z = z)$, then X and Y have the same distribution.

True: by the law of total probability, conditioning on Z gives

$$P(X = a) = \sum_z P(X = a|Z = z)P(Z = z).$$

Since X and Y have the same conditional distribution given Z, this becomes $\sum_z P(Y = a|Z = z)P(Z = z) = P(Y = a)$.

2 Simpson's Paradox

1. (a) Is it possible to have events A, B, E such that $P(A|E) < P(B|E)$ and $P(A|E^c) < P(B|E^c)$, yet $P(A) > P(B)$? That is, A is less likely under B given that E is true, and also given that E is false, yet A is more likely than B if given no information about E. Show this is impossible (with a short proof) or find a counterexample (with a "story" interpreting A, B, E).

   It is *not* possible, as seen using the law of total probability:

   $$P(A) = P(A|E)P(E)+P(A|E^c)P(E^c) < P(B|E)P(E)+P(B|E^c)P(E^c) = P(B).$$

   (b) Is it possible to have events A, B, E such that $P(A|B, E) < P(A|B^c, E)$ and $P(A|B, E^c) < P(A|B^c, E^c)$, yet $P(A|B) > P(A|B^c)$? That is, given that E is true, learning B is evidence against A, and similarly given that E is false; but given no information about E, learning that B is true is evidence in favor of A. Show this is impossible (with a short proof) or find a counterexample (with a "story" interpreting A, B, E).

   Yes, this is possible: this is the structure of Simpson's Paradox! For example, consider the Stampy problem above. Or, consider the two doctors example discussed in class: suppose that there are two doctors, Dr. Hibbert and Dr. Nick. Each performs two types of surgery, say heart transplants and bandaid removals. Let A be the event that a surgery is successful, let B be the event that Dr. Nick performs the surgery and B^c be the complement, that Dr. Hibbert performs the surgery. Let E be heart surgery and E^c be bandaid removal.

Is it possible that Dr. Hibbert is better than Dr. Nick at both heart transplants and bandaid removals, yet Dr. Nick has a higher success rate overall? Yes, this can happen if Dr. Nick performs mostly bandaid removals and Dr. Hibbert performs mostly heart transplants. That is, the better doctor may perform relatively more of the harder surgery, resulting in a lower success rate.

To make up specific numbers, suppose that Dr. Hibbert performed 90 heart transplants, with 70 successful, and 10 bandaid removals, with all 10 successful. Dr. Nick performed 10 heart transplants, with 2 successful, and 90 bandaid removals, with 81 successful. Formally, our probability space consists of randomly choosing one of the 200 surgeries, uniformly. Note that each individual surgery is more likely to be successful given that it's performed by Dr. Hibbert, but Dr. Hibbert's overall success rate (80%) is lower than Dr. Nick's (83%).

There are many real-life examples of Simpson's Paradox. For example, it is possible for one baseball player to have a higher batting average than another in each of two seasons, yet a lower batting average when the two seasons are aggregated. Simpson's Paradox illustrates the importance of controlling for additional variables that interfere with the analysis (known as *confounders*).

2. Consider the following conversation from an episode of *The Simpsons*:

> Lisa: *Dad, I think he's an ivory dealer! His boots are ivory, his hat is ivory, and I'm pretty sure that check is ivory.*
>
> Homer: *Lisa, a guy who's got lots of ivory is less likely to hurt Stampy than a guy whose ivory supplies are low.*

Here Homer and Lisa are debating the question of whether or not the man (named Blackheart) is likely to hurt Stampy the Elephant if they sell Stampy to him. They clearly disagree about how to use their observations about Blackheart to learn about the probability (conditional on the evidence) that Blackheart will hurt Stampy.

(a) Define clear notation for the various events of interest here.

Let H be the event that the man will hurt Stampy, let L be the event that a man has lots of ivory, and let D be the event that the man is an ivory dealer.

(b) Express Lisa's and Homer's arguments (Lisa's is partly implicit) as conditional probability statements in terms of your notation from (a).

Lisa observes that L is true. She suggests (reasonably) that this evidence makes D more likely, i.e., $P(D|L) > P(D)$. Implicitly, she suggests that this makes it likely that the man will hurt Stampy, i.e.,

$$P(H|L) > P(H|L^c).$$

Homer argues that

$$P(H|L) < P(H|L^c).$$

(c) Assume it is true that someone who has a lot of a commodity will have less desire to acquire more of the commodity. Explain what is wrong with Homer's reasoning that the evidence about Blackheart makes it less likely that he will harm Stampy.

Homer does not realize that observing that Blackheart has so much ivory makes it much more likely that Blackheart is an ivory dealer, which in turn makes it more likely that the man will hurt Stampy. (This is an example of Simpson's Paradox.) It may be true that, *controlling for whether or not Blackheart is a dealer*, having high ivory supplies makes it less likely that he will harm Stampy: $P(H|L, D) < P(H|L^c, D)$ and $P(H|L, D^c) < P(H|L^c, D^c)$. However, this does not imply that $P(H|L) < P(H|L^c)$.

3 Gambler's Ruin

1. A gambler repeatedly plays a game where in each round, he wins a dollar with probability 1/3 and loses a dollar with probability 2/3. His strategy is "quit when he is ahead by \$2," though some suspect he is a gambling addict anyway. Suppose that he starts with a million dollars. Show that the probability that he'll ever be ahead by \$2 is less than 1/4.

   This problem is a special case of the gambler's ruin. Let A_1 be the event that he is successful on the first play and let W be the event that he is ever ahead by \$2 before being ruined. Then by the law of total probability, we have

   $$P(W) = P(W|A_1)P(A_1) + P(W|A_1^c)P(A_1^c).$$

   Let a_i be the probability that the gambler achieves a profit of \$2 before being ruined, starting with a fortune of $\$i$. For our setup, $P(W) = a_i$, $P(W|A_1) = a_{i+1}$ and $P(W|A_1^c) = a_{i-1}$. Therefore,

   $$a_i = a_{i+1}/3 + 2a_{i-1}/3,$$

with boundary conditions $a_0 = 0$ and $a_{i+2} = 1$. We can then solve this difference equation for a_i (directly or using the result of the gambler's ruin problem):

$$a_i = \frac{2^i - 1}{2^{2+i} - 1}.$$

This is always less than 1/4 since $\frac{2^i-1}{2^{2+i}-1} < \frac{1}{4}$ is equivalent to $4(2^i-1) < 2^{2+i}-1$, which is equivalent to the true statement $2^{2+i} - 4 < 2^{2+i} - 1$.

4 Bernoulli and Binomial

1. (a) In the World Series of baseball, two teams (call them A and B) play a sequence of games against each other, and the first team to win four games wins the series. Let p be the probability that A wins an individual game, and assume that the games are independent. What is the probability that team A wins the series?

   Let $q = 1 - p$. First let us do a direct calculation:

$$\begin{aligned} P(\text{A wins}) &= P(\text{A winning in 4 games}) + P(\text{A winning in 5 games}) \\ &\quad + P(\text{A wins in 6 games}) + P(\text{A winning in 7 games}) \\ &= p^4 + \binom{4}{3}p^4q + \binom{5}{3}p^4q^2 + \binom{6}{3}p^4q^3. \end{aligned}$$

   To understand how these probabilities are calculated, note for example that

$$\begin{aligned} P(\text{A wins in 5}) &= P(\text{A wins 3 out of first 4}) \cdot P(\text{A wins 5th game}|\text{A wins 3 out of first 4}) \\ &= \binom{4}{3}p^3qp \end{aligned}$$

   (This value can also be found from the PMF of a distribution known as the *Negative Binomial*, which we will see later in the course.)

   An neater solution is to use the fact (explained in (b)) that we can assume that the teams play all 7 games no matter what. Then let X be the number of wins for team A, so that

$$X \sim \text{Binomial}(7, p).$$

   The probability that team A wins the series is

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$$

The PMF of the Bin(n, p) distribution is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

and therefore

$$P(X \geq 4) = \binom{7}{4} p^4 q^3 + \binom{7}{5} p^5 q^2 + \binom{7}{6} p^6 q + p^7,$$

which looks different from the above but is actually identical as a function of p (as can be verified by simplifying both expressions as polynomials in p).

(b) Give a clear intuitive explanation of whether the answer to (a) depends on whether the teams always play 7 games (and whoever wins the majority wins the series), or the teams stop playing more games as soon as one team has won 4 games (as is actually the case in practice: once the match is decided, the two teams do not keep playing more games).

The answer to (a) does not depend on whether the teams play all seven games no matter what. Imagine telling the players to continue playing the games even after the match has been decided, just for fun: the outcome of the match won't be affected by this, and this also means that the probability that A wins the match won't be affected by assuming that the teams always play 7 games!

2. A sequence of n independent experiments is performed. Each experiment is a success with probability p and a failure with probability $q = 1 - p$. Show that conditional on the number of successes, all possibilities for the list of outcomes of the experiment are equally likely (of course, we only consider lists of outcomes where the number of successes is consistent with the information being conditioned on).

   Let X_j be 1 if the jth experiment is a success and 0 otherwise, and let $X = X_1 + \cdots + X_n$ be the total number of successes. Then for any k and any

$a_1, \ldots, a_n \in \{0,1\}$ with $a_1 + \cdots + a_n = k$,

$$\begin{aligned} P(X_1 = a_1, \ldots, X_n = a_n | X = k) &= \frac{P(X_1 = a_1, \ldots, X_n = a_n, X = k)}{P(X = k)} \\ &= \frac{P(X_1 = a_1, \ldots, X_n = a_n)}{P(X = k)} \\ &= \frac{p^k q^{n-k}}{\binom{n}{k} p^k q^{n-k}} \\ &= \frac{1}{\binom{n}{k}}. \end{aligned}$$

This does not depend on $a_1, \ldots, a_n$. Thus, for n independent Bernoulli trials, given that there are exactly k successes, the $\binom{n}{k}$ possible sequences consisting of k successes and $n-k$ failures are equally likely. Interestingly, the conditional probability above also does not depend on p (this leads to the notion of a *sufficient statistic*, which is studied in Stat 111).

3. Let $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$, independent of X.

(a) Show that $X + Y \sim \text{Bin}(n+m,p)$, using a story proof.

Interpret X as the number of successes in n independent Bernoulli trials and Y as the number of successes in m more independent Bernoulli trials, where each trial has probability p of success. Then $X+Y$ is the number of successes in the $n+m$ trials, so $X + Y \sim \text{Bin}(n+m,p)$.

(b) Show that $X - Y$ is *not* Binomial.

A Binomial can't be negative, but $X - Y$ is negative with positive probability.

(c) Find $P(X = k | X + Y = j)$. How does this relate to the elk problem from HW 1?

By definition of conditional probability,

$$P(X = k | X + Y = j) = \frac{P(X = k, X + Y = j)}{P(X + Y = j)} = \frac{P(X = k)P(Y = j - k)}{P(X + Y = j)}$$

since the event $X = k$ is independent of the event $Y = j - k$. This becomes

$$\frac{\binom{n}{k} p^k (1-p)^{n-k} \binom{m}{j-k} p^{j-k} (1-p)^{m-(j-k)}}{\binom{m+n}{j} p^j (1-p)^{m+n-j}} = \binom{n}{k}\binom{m}{j-k} \Big/ \binom{n+m}{j}.$$

Note that the p disappeared! This is exactly the same distribution as in the elk problem (it is called the *Hypergeometric* distribution). To see why, imagine that there are n male elk and m female elk, each of which is tagged with the word "success" with probability p (independently). Suppose we then want to know how many of the male elk are tagged, given that a total of j elk have been tagged. For this, p is no longer relevant, and we can "capture" the male elk and count how many are tagged, analogously to the original elk problem.

Stat 110 Homework 3, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. (a) Consider the following 7-door version of the Monty Hall problem. There are 7 doors, behind one of which there is a car (which you want), and behind the rest of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door. Monty Hall then opens 3 goat doors, and offers you the option of switching to any of the remaining 3 doors.

Assume that Monty Hall knows which door has the car, will always open 3 goat doors and offer the option of switching, and that Monty chooses with equal probabilities from all his choices of which goat doors to open. Should you switch? What is your probability of success if you switch to one of the remaining 3 doors?

(b) Generalize the above to a Monty Hall problem where there are $n \geq 3$ doors, of which Monty opens m goat doors, with $1 \leq m \leq n-2$.

2. The *odds* of an event with probability p are defined to be $\frac{p}{1-p}$, e.g., an event with probability 3/4 is said to have odds of 3 to 1 in favor (or 1 to 3 against). We are interested in a hypothesis H (which we think of as a event), and we gather new data as evidence (expressed as an event D) to study the hypothesis. The *prior* probability of H is our probability for H being true before we gather the new data; the *posterior* probability of H is our probability for it after we gather the new data. The *likelihood ratio* is defined as $\frac{P(D|H)}{P(D|H^c)}$.

(a) Show that Bayes' rule can be expressed in terms of odds as follows: *the posterior odds of a hypothesis H are the prior odds of H times the likelihood ratio.*

(b) As in the example from class, suppose that a patient tests positive for a disease afflicting 1% of the population. For a patient who has the disease, there is a 95% chance of testing positive (in medical statistics, this is called the *sensitivity* of the test); for a patient who doesn't have the disease, there is a 95% chance of testing negative test (in medical statistics, this is called the *specificity* of the test).

The patient gets a second, independent test done (with the same sensitivity and specificity), and again tests positive. Use the odds form of Bayes' rule to find the probability that the patient has the disease, given the evidence, *in two ways*: in one step, conditioning on both test results simultaneously, and in two steps, first updating the probabilities based on the first test result, and then updating again based on the second test result.

3. Is it possible to have events A_1, A_2, B, C with $P(A_1|B) > P(A_1|C)$ and $P(A_2|B) > P(A_2|C)$, yet $P(A_1 \cup A_2|B) < P(A_1 \cup A_2|C)$? If so, find an example (with a "story"

interpreting the events, as well as giving specific numbers); otherwise, show that it is impossible for this phenomenon to happen.

4. Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the probability that Calvin wins the match (in terms of p), in two different ways:

(a) by conditioning, using the law of total probability.

(b) by interpreting the problem as a gambler's ruin problem.

5. A fair die is rolled repeatedly, and a running total is kept (which is, at each time, the total of all the rolls up until that time). Let p_n be the probability that the running total is ever *exactly* n (assume the die will always be rolled enough times so that the running total will eventually exceed n, but it may or may not ever equal n).

(a) Write down a recursive equation for p_n (relating p_n to earlier terms p_k in a simple way). Your equation should be true for all positive integers n, so give a definition of p_0 and p_k for $k < 0$ so that the recursive equation is true for small values of n.

(b) Find p_7.

(c) Give an intuitive explanation for the fact that $p_n \to 1/3.5 = 2/7$ as $n \to \infty$.

6. Players A and B take turns in answering trivia questions, starting with player A answering the first question. Each time A answers a question, she has probability p_1 of getting it right. Each time B plays, he has probability p_2 of getting it right.

(a) If A answers m questions, what is the PMF of the number of questions she gets right?

(b) If A answers m times and B answers n times, what is the PMF of the total number of questions they get right (you can leave your answer as a sum)? Describe exactly when/whether this is a Binomial distribution.

(c) Suppose that the first player to answer correctly wins the game (with no predetermined maximum number of questions that can be asked). Find the probability that A wins the game.

7. A message is sent over a noisy channel. The message is a sequence $x_1, x_2, \ldots, x_n$ of n bits ($x_i \in \{0, 1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error (a 0 becomes a 1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual

bit has an error ($0 < p < 1/2$). Let $y_1, y_2, \ldots, y_n$ be the received message (so $y_i = x_i$ if there is no error in that bit, but $y_i = 1 - x_i$ if there is an error there).

To help detect errors, the nth bit is reserved for a parity check: x_n is defined to be 0 if $x_1 + x_2 + \cdots + x_{n-1}$ is even, and 1 if $x_1 + x_2 + \cdots + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + \cdots + y_{n-1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors.

(a) For $n = 5, p = 0.1$, what is the probability that the received message has errors which go undetected?

(b) For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.

(c) Give a simplified expression, not involving a sum of a large number of terms, for the probability that the received message has errors which go undetected.

Hint for (c): Letting

$$a = \sum_{k \text{ even}, k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} \text{ and } b = \sum_{k \text{ odd}, k \geq 1} \binom{n}{k} p^k (1-p)^{n-k},$$

the binomial theorem makes it possible to find simple expressions for $a+b$ and $a-b$, which then makes it possible to obtain a and b.

Stat 110 Homework 3 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. (a) Consider the following 7-door version of the Monty Hall problem. There are 7 doors, behind one of which there is a car (which you want), and behind the rest of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door. Monty Hall then opens 3 goat doors, and offers you the option of switching to any of the remaining 3 doors.

Assume that Monty Hall knows which door has the car, will always open 3 goat doors and offer the option of switching, and that Monty chooses with equal probabilities from all his choices of which goat doors to open. Should you switch? What is your probability of success if you switch to one of the remaining 3 doors?

Assume the doors are labeled such that you choose Door 1 (to simplify notation), and suppose first that you follow the "stick to your original choice" strategy. Let S be the event of success in getting the car, and let C_j be the event that the car is behind Door j. Conditioning on which door has the car, we have

$$P(S) = P(S|C_1)P(C_1) + \cdots + P(S|C_7)P(C_7) = P(C_1) = \frac{1}{7}.$$

Let M_{ijk} be the event that Monty opens Doors i, j, k. Then

$$P(S) = \sum_{i,j,k} P(S|M_{ijk})P(M_{ijk})$$

(summed over all i, j, k with $2 \leq i < j < k \leq 7$.) By symmetry, this gives

$$P(S|M_{ijk}) = P(S) = \frac{1}{7}$$

for all i, j, k with $2 \leq i < j < k \leq 7$. Thus, the conditional probability that the car is behind 1 of the remaining 3 doors is $6/7$, which gives $2/7$ for each. So you should switch, thus making your probability of success $2/7$ rather than $1/7$.

(b) Generalize the above to a Monty Hall problem where there are $n \geq 3$ doors, of which Monty opens m goat doors, with $1 \leq m \leq n-2$.

By the same reasoning, the probability of success for "stick to your original choice" is $\frac{1}{n}$, both unconditionally and conditionally. Each of the $n-m-1$ remaining doors has conditional probability $\frac{n-1}{(n-m-1)n}$ of having the car. This value is greater than $\frac{1}{n}$, so you should switch, thus obtaining probability $\frac{n-1}{(n-m-1)n}$ of success (both conditionally and unconditionally).

2. The *odds* of an event with probability p are defined to be $\frac{p}{1-p}$, e.g., an event with probability 3/4 is said to have odds of 3 to 1 in favor (or 1 to 3 against). We are interested in a hypothesis H (which we think of as a event), and we gather new data as evidence (expressed as an event D) to study the hypothesis. The *prior* probability of H is our probability for H being true before we gather the new data; the *posterior* probability of H is our probability for it after we gather the new data. The *likelihood ratio* is defined as $\frac{P(D|H)}{P(D|H^c)}$.

(a) Show that Bayes' rule can be expressed in terms of odds as follows: *the posterior odds of a hypothesis H are the prior odds of H times the likelihood ratio.*

We want to show that

$$\frac{P(H|D)}{P(H^c|D)} = \frac{P(H)}{P(H^c)}\frac{P(D|H)}{P(D|H^c)}$$

By Bayes' rule, we have

$$P(H|D) = P(D|H)P(H)/P(D),$$
$$P(H^c|D) = P(D|H^c)P(H^c)/P(D).$$

Dividing the first of these by the second, we immediately obtain the desired equation, which is a useful alternative way to write Bayes' rule.

(b) As in the example from class, suppose that a patient tests positive for a disease afflicting 1% of the population. For a patient who has the disease, there is a 95% chance of testing positive (in medical statistics, this is called the *sensitivity* of the test); for a patient who doesn't have the disease, there is a 95% chance of testing negative test (in medical statistics, this is called the *specificity* of the test).

The patient gets a second, independent test done (with the same sensitivity and specificity), and again tests positive. Use the odds form of Bayes' rule to find the probability that the patient has the disease, given the evidence, *in two ways*: in one step, conditioning on both test results simultaneously, and in two steps, first updating the probabilities based on the first test result, and then updating again based on the second test result.

To go from odds back to probability, we divide odds by (1 plus odds), since $\frac{p/q}{1+p/q} = p$ for $q = 1 - p$. Let H be the event of having the disease. The prior odds are 99 to 1 against having the disease. The likelihood ratio based on one test result is $\frac{0.95}{0.05}$. Now we can immediately carry out either the one update or the two update method.

One update method: The likelihood ratio based on both test results is $\frac{0.95^2}{0.05^2}$ since the tests are independent. So the posterior odds of the patient having the disease are

$$\frac{1}{99} \cdot \frac{0.95^2}{0.05^2} = \frac{361}{99} \approx 3.646,$$

which corresponds to a probability of $361/(361+99) = 361/460 \approx 0.78$ (whereas based on one positive test, the probability was only 0.16 of having the disease).

Two updates method: After the first test, the posterior odds of the patient having the disease are

$$\frac{1}{99} \cdot \frac{0.95}{0.05} \approx 0.19,$$

which corresponds to a probability of $0.19/(1+0.19) \approx 0.16$ (agreeing with the result from class). These posterior odds become the new prior odds, and then updating based on the second test gives $\left(\frac{1}{99} \cdot \frac{0.95}{0.05}\right) \frac{0.95}{0.05}$, which is the same result as above.

3. Is it possible to have events A_1, A_2, B, C with $P(A_1|B) > P(A_1|C)$ and $P(A_2|B) > P(A_2|C)$, yet $P(A_1 \cup A_2|B) < P(A_1 \cup A_2|C)$? If so, find an example (with a "story" interpreting the events, as well as giving specific numbers); otherwise, show that it is impossible for this phenomenon to happen.

Yes, this is possible. First note that $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$, so it is *not* possible if A_1 and A_2 are disjoint, and that it is crucial to consider the intersection. So let's choose examples where $P(A_1 \cap A_2|B)$ is much larger than $P(A_1 \cap A_2|C)$, to offset the other inequalities.

Story 1: Consider two basketball players, one of whom is randomly chosen to shoot two free throws. The first player is very streaky, and always either makes both or misses both free throws, with probability 0.8 of making both (this is an extreme example chosen for simplicity, but we could also make it so the player has good days (on which there is a high chance of making both shots) and bad days (on which there is a high chance of missing both shots) without requiring *always* making both or missing both). The second player's free throws go in with probability 0.7, independently. Define the events as A_j: the jth free throw goes in; B: the free throw shooter is the first player; $C = B^c$. Then

$$P(A_1|B) = P(A_2|B) = P(A_1 \cap A_2|B) = P(A_1 \cup A_2|B) = 0.8,$$

$$P(A_1|C) = P(A_2|C) = 0.7, P(A_1 \cap A_2|C) = 0.49, P(A_1 \cup A_2|C) = 2 \cdot 0.7 - 0.49 = 0.91.$$

Story 2: Suppose that you can either take Good Class or Other Class, but not both. If you take Good Class, you'll attend lecture 70% of the time, and you will understand the material if and only if you attend lecture. If you take Other Class, you'll attend lecture 40% of the time and understand the material 40% of the time, but because the class is so poorly taught, the only way you understand the material

is by studying on your own and not attending lecture. Defining the events as A_1: attend lecture; A_2: understand material; B: take Good Class; C: take Other Class,

$$P(A_1|B) = P(A_2|B) = P(A_1 \cap A_2|B) = P(A_1 \cup A_2|B) = 0.7,$$

$$P(A_1|C) = P(A_2|C) = 0.4, P(A_1 \cap A_2|C) = 0, P(A_1 \cup A_2|C) = 2 \cdot 0.4 = 0.8.$$

4. Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the probability that Calvin wins the match (in terms of p), in two different ways:

(a) by conditioning, using the law of total probability.

Let C be the event that Calvin wins the match, $X \sim \text{Bin}(2, p)$ be how many of the first 2 games he wins, and $q = 1 - p$. Then

$$P(C) = P(C|X = 0)q^2 + P(C|X = 1)(2pq) + P(C|X = 2)p^2 = 2pqP(C) + p^2,$$

so $P(C) = \frac{p^2}{1-2pq}$. This can also be written as $\frac{p^2}{p^2+q^2}$, since $p + q = 1$.

Miracle check: Note that this should (and does) reduce to 1 for $p = 1$, 0 for $p = 0$, and $\frac{1}{2}$ for $p = \frac{1}{2}$. Also, it makes sense that the probability of Hobbes winning, which is $1 - P(C) = \frac{q^2}{p^2+q^2}$, can also be obtained by swapping p and q.

(b) by interpreting the problem as a gambler's ruin problem.

The problem can be thought of as a gambler's ruin where each player starts out with \$2. So the probability that Calvin wins the match is

$$\frac{1 - (q/p)^2}{1 - (q/p)^4} = \frac{(p^2 - q^2)/p^2}{(p^4 - q^4)/p^4} = \frac{(p^2 - q^2)/p^2}{(p^2 - q^2)(p^2 + q^2)/p^4} = \frac{p^2}{p^2 + q^2},$$

which agrees with the above.

5. A fair die is rolled repeatedly, and a running total is kept (which is, at each time, the total of all the rolls up until that time). Let p_n be the probability that the running total is ever *exactly* n (assume the die will always be rolled enough times so that the running total will eventually exceed n, but it may or may not ever equal n).

(a) Write down a recursive equation for p_n (relating p_n to earlier terms p_k in a simple way). Your equation should be true for all positive integers n, so give a definition of p_0 and p_k for $k < 0$ so that the recursive equation is true for small values of n.

We will find something to condition on to reduce the case of interest to earlier, simpler cases. This is achieved by the useful strategy of *first step anaysis*. Let p_n be the probability that the running total is ever *exactly* n. Note that if, for example, the first throw is a 3, then the probability of reaching n exactly is p_{n-3} since starting from that point, we need to get a total of $n-3$ exactly. So

$$p_n = \frac{1}{6}(p_{n-1} + p_{n-2} + p_{n-3} + p_{n-4} + p_{n-5} + p_{n-6}),$$

where we define $p_0 = 1$ (which makes sense anyway since the running total is 0 before the first toss) and $p_k = 0$ for $k < 0$.

(b) Find p_7.

Using the recursive equation in (a), we have

$$p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{6}(1 + \frac{1}{6}), \quad p_3 = \frac{1}{6}(1 + \frac{1}{6})^2,$$
$$p_4 = \frac{1}{6}(1 + \frac{1}{6})^3, \quad p_5 = \frac{1}{6}(1 + \frac{1}{6})^4, \quad p_6 = \frac{1}{6}(1 + \frac{1}{6})^5.$$

Hence,

$$p_7 = \frac{1}{6}(p_1 + p_2 + p_3 + p_4 + p_5 + p_6) = \frac{1}{6}\left((1 + \frac{1}{6})^6 - 1\right) \approx 0.2536.$$

(c) Give an intuitive explanation for the fact that $p_n \to 1/3.5 = 2/7$ as $n \to \infty$.

An intuitive explanation is as follows. The average number thrown by the die is (total of dots)/6, which is $21/6 = 7/2$, so that every throw adds on an average of $7/2$. We can therefore expect to land on 2 out of every 7 numbers, and the probability of landing on any particular number is $2/7$. This result can be proved as follows (a proof was *not* required):

$$\begin{aligned}
&p_{n+1} + 2p_{n+2} + 3p_{n+3} + 4p_{n+4} + 5p_{n+5} + 6p_{n+6} \\
=&p_{n+1} + 2p_{n+2} + 3p_{n+3} + 4p_{n+4} + 5p_{n+5} \\
&+ p_n + p_{n+1} + p_{n+2} + p_{n+3} + p_{n+4} + p_{n+5} \\
=&p_n + 2p_{n+1} + 3p_{n+2} + 4p_{n+3} + 5p_{n+4} + 6p_{n+5} \\
=&\cdots \\
=&p_{-5} + 2p_{-4} + 3p_{-3} + 4p_{-2} + 5p_{-1} + 6p_0 = 6.
\end{aligned}$$

Taking the limit of the lefthand side as n goes to ∞, we have

$$(1+2+3+4+5+6)\lim_{n\to\infty} p_n = 6,$$

so $\lim_{n\to\infty} p_n = 2/7$.

6. Players A and B take turns in answering trivia questions, starting with player A answering the first question. Each time A answers a question, she has probability p_1 of getting it right. Each time B plays, he has probability p_2 of getting it right.

(a) If A answers m questions, what is the PMF of the number of questions she gets right?

The r.v. is $\text{Bin}(m, p_1)$, so the PMF is $\binom{m}{k} p_1^k (1-p_1)^{m-k}$ for $k \in \{0, 1, \ldots, m\}$.

(b) If A answers m times and B answers n times, what is the PMF of the total number of questions they get right (you can leave your answer as a sum)? Describe exactly when/whether this is a Binomial distribution.

Let T be the total number of questions they get right. To get a total of k questions right, it must be that A got 0 and B got k, or A got 1 and B got $k-1$, etc. These are disjoint events so the PMF is

$$P(T=k) = \sum_{j=0}^{k} \binom{m}{j} p_1^j (1-p_1)^{m-j} \binom{n}{k-j} p_2^{k-j} (1-p_2)^{n-(k-j)}$$

for $k \in \{0, 1, \ldots, m+n\}$, with the usual convention that $\binom{n}{k}$ is 0 for $k > n$.

This is the $\text{Bin}(m+n, p)$ distribution if $p_1 = p_2 = p$, as shown in class (using the story for the Binomial, or using Vandermonde's identity). For $p_1 \neq p_2$, it's not a Binomial distribution, since the trials have different probabilities of success; having some trials with one probability of success and other trials with another probability of success isn't equivalent to having trials with some "effective" probability of success.

(c) Suppose that the first player to answer correctly wins the game (with no predetermined maximum number of questions that can be asked). Find the probability that A wins the game.

Let $r = P(A \text{ wins})$. Conditioning on the results of the first question for each player, we have

$$r = p_1 + (1-p_1)p_2 \cdot 0 + (1-p_1)(1-p_2)r,$$

which gives $r = \frac{p_1}{1-(1-p_1)(1-p_2)} = \frac{p_1}{p_1+p_2-p_1p_2}$.

7. A message is sent over a noisy channel. The message is a sequence $x_1, x_2, \ldots, x_n$ of n bits ($x_i \in \{0,1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error (a 0 becomes a 1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual bit has an error ($0 < p < 1/2$). Let $y_1, y_2, \ldots, y_n$ be the received message (so $y_i = x_i$ if there is no error in that bit, but $y_i = 1 - x_i$ if there is an error there).

To help detect errors, the nth bit is reserved for a parity check: x_n is defined to be 0 if $x_1 + x_2 + \cdots + x_{n-1}$ is even, and 1 if $x_1 + x_2 + \cdots + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + \cdots + y_{n-1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors.

(a) For $n = 5, p = 0.1$, what is the probability that the received message has errors which go undetected?

Note that $\sum_{i=1}^{n} x_i$ is even. If the number of errors is even (and nonzero), the errors will go undetected; otherwise, $\sum_{i=1}^{n} y_i$ will be odd, so the errors will be detected.

The number of errors is $\text{Bin}(n, p)$, so the probability of undetected errors when $n = 5, p = 0.1$ is

$$\binom{5}{2} p^2 (1-p)^3 + \binom{5}{4} p^4 (1-p) \approx 0.073.$$

(b) For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.

By the same reasoning as in (a), the probability of undetected errors is

$$\sum_{k \text{ even},\, k \geq 2} \binom{n}{k} p^k (1-p)^{n-k}.$$

(c) Give a simplified expression, not involving a sum of a large number of terms, for the probability that the received message has errors which go undetected.

Hint for (c): Letting

$$a = \sum_{k \text{ even},\, k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} \text{ and } b = \sum_{k \text{ odd},\, k \geq 1} \binom{n}{k} p^k (1-p)^{n-k},$$

the binomial theorem makes it possible to find simple expressions for $a+b$ and $a-b$, which then makes it possible to obtain a and b.

Let a, b be as in the hint. Then

$$a + b = \sum_{k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} = 1,$$

$$a - b = \sum_{k \geq 0} \binom{n}{k} (-p)^k (1-p)^{n-k} = (1-2p)^n.$$

Solving for a and b gives

$$a = \frac{1 + (1-2p)^n}{2} \text{ and } b = \frac{1 - (1-2p)^n}{2}.$$

$$\sum_{k \text{ even}, \, k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1 + (1-2p)^n}{2}.$$

Subtracting off the possibility of no errors, we have

$$\sum_{k \text{ even}, \, k \geq 2} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1 + (1-2p)^n}{2} - (1-p)^n.$$

Miracle check: note that letting $n = 5, p = 0.1$ here gives 0.073, which agrees with (a); letting $p = 0$ gives 0, as it should; and letting $p = 1$ gives 0 for n odd and 1 for n even, which again makes sense.