Mathematical Modeling of Infectious Diseases, UCSF

Lecture 2

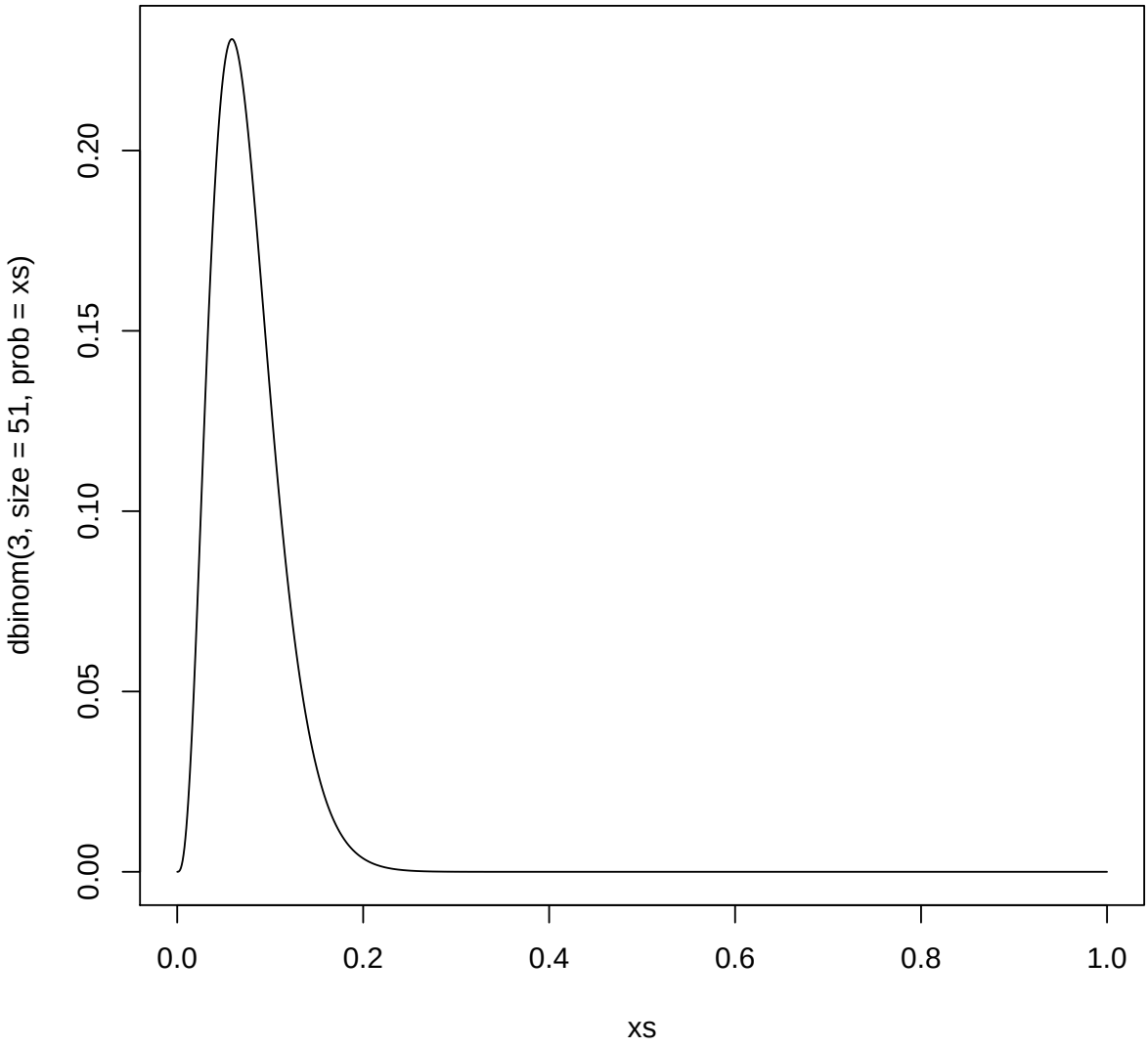
Version 1.1—modified 15 Jan 2015, time 1400.

Likelihood

Before we go much further, let's look at likelihood. This is the probability of the data given various parameters. If you pick bad parameters, the data is much less likely than if you pick good parameters. For example, we know that the chance of getting 3 infections out of 51 injuries is

$$P(X = 3) = \binom{51}{3} p^3 (1-p)^4 8.$$

```
xs <- seq(0,1,by=1/1024)
plot(xs,dbinom(3,size=51,prob=xs),type="l")
```



Let's find out for what value of p this likelihood is the largest. The trick will be to find where the hilltop is flat. It turns out to be much easier to work with the log of the likelihood; since the log is monotone ($x > y$ implies $\log x > \log y$), we can just find where that is maximized:

$$\ell = \log \binom{51}{3} + 3 \log p + 48 \log(1-p).$$

$$\frac{d\ell}{dp} = \frac{3}{p} - \frac{48}{1-p} = 0$$

$$\frac{3}{p} - \frac{48}{1-p}.$$

In fact this works for any $0 < x < N$, so why not write it in general:

$$\frac{x}{p} - \frac{N-x}{1-p}.$$

Solving this gives $p = x/N$ when $0 < x < N$; when $x = 0$ or $x = N$ this doesn't work, but you get a solution on the boundary anyway. If $x = 0$ the most likely value for p is 0, etc.

Quick review

Remember that the expected value of a function of a random variable is

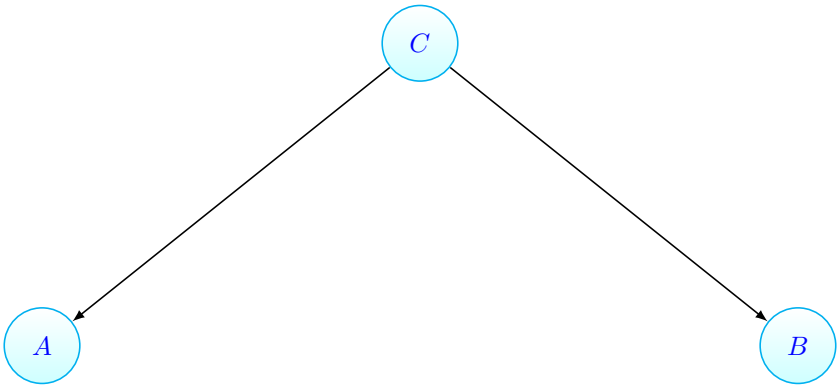
$$E[g(X)] = \sum_x g(x)P(X = x).$$

Exercise 1. Consider a Bernoulli random variable X with success probability p. Suppose that we consider the following function: for every value X takes, let the transformed value be minus the log of the probability, basically the loglikelihood (with a minus sign, since the log of a probability is negative). Write the expected value of this transformed variable. ∎

More on conditional independence

Last time we reviewed the definition of conditional independence. Events A and B are conditionally independent given C if $P(AB|C) = P(A|C)P(B|C)$. We showed that if this is true, $P(A|BC) = P(A|C)$ (knowing B doesn't help once we already know C), and $P(B|AC) = P(B|C)$ (A doesn't help with B, once you know C). An example you see in the literature sometimes: it could be that crime (A) and ice cream sales (B) are conditionally independent given temperature (C) in some city.

When you have a lot of random variables in some analysis, sometimes conditional independence relationships of this sort are available through reasoning, and can be used to simplify the problem. The idea is to represent known dependencies graphically, and assume conditional independence unless specified otherwise. For example, we might write:



This graph is meant to indicate that while we know C influences A and C influences B, A and B are conditionally independent given C, since we have not indicated any relationship.

Now, we can always write

$$P(ABC) = P(A|BC)P(BC) = P(A|BC)P(B|C)P(C).$$

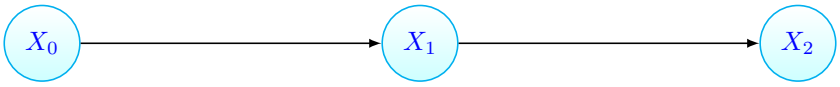
This is just plain true, all the time. But now that we assume that A and B are conditionally independent given C, we can just turn $P(A|BC) = P(A|C)$. So we get

$$P(ABC) = P(C)P(A|C)P(B|C).$$

We are representing independence relationships between random variables using a graphical method in which each random variable is a node in a *directed acyclic graph*. To talk about this, we need some formalism. A *directed graph* is a set of *nodes*, and a set of *edges* which are ordered pairs of nodes. We represent an edge by an arrow from one node to another; if there is an arrow from node A to node B, we say A is a *parent* of B and that B is a *child* of A. We say there is a *directed path* (but usually we just say *path*) from node A to node E (say) if either A is a parent of node E, or there is a node B such that A is the parent of B and there is a path from B to E. (This is a typical example of a recursive definition.) If there is a path from A to E, then A is an *ancestor* of E and E is a *descendant* of A. If we order the nodes of a directed graph so that descendants always follow ancestors, the ordering is called *topological* and we are said to have *topologically sorted* the nodes according to the graph.

The idea will be that any node is assumed to be conditionally independent of all non-descendants given its parents. In general we will have nodes and their descendants related, but otherwise, we assume that once we have conditioned on parents we have taken care of all the dependencies there are; if we haven't included it explicitly, it is not there. A graphical representation of a joint distribution in this form is called a *Bayes network*. Bayes networks are a very large topic that we can barely scratch the surface of.

Let's do another one.



Here, we mean to say that X_0 directly influences X_1, and X_1 directly influences X_2. But equally importantly, we mean to say that X_0 and X_2 are conditionally independent given X_1. We never drew any arrow from X_0 to X_2.

If we topologically sort the vertices, we would write X_0, X_1, and X_2, in that order. Let's go through in reverse, writing down the joint distribution as a product:

$$P(X_0, X_1, X_2) = P(X_2|X_0, X_1)P(X_0, X_1)$$

That is just the definition of conditional probability.

Then, the next variable in line is X_1, so let's apply the definition again:

$$P(X_0, X_1, X_2) = P(X_2|X_0, X_1)P(X_1|X_0)P(X_0)$$

This is just plain always true; it's just the definitions. And it would have been true if we had used any order of the variables too. But using the order we did allows us to apply the conditional independence assumptions in a very nice way—because none of the descendants of a variable ever appear after a vertical bar. We have assumed that we are conditionally independent of all non-descendants given the parents, so we just drop the nondescendants from after the bar. For the example we are working on, we will have $P(X_2|X_0, X_1) = P(X_2|X_1)$. So now

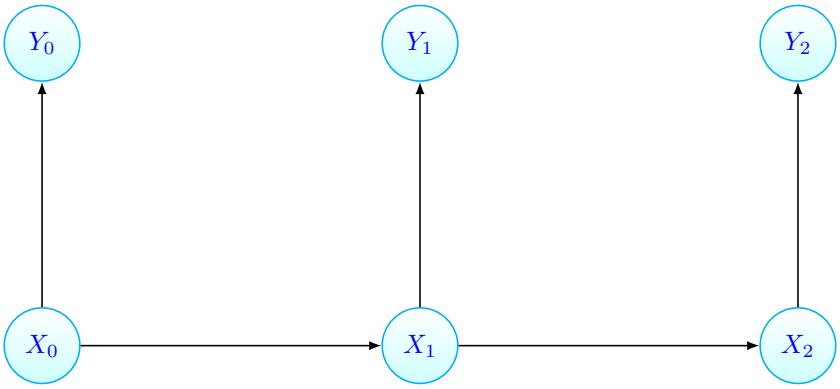
$$P(X_0, X_1, X_2) = P(X_0)P(X_1|X_0)P(X_2|X_1).$$

A set of variables whose graph can be represented $X_0 \to X_1 \to X_2 \to \cdots \to X_n$ is called a *Markov chain*, and the reasoning we just used gives us

$$P(X_0, X_1, X_2, \ldots, X_n) = P(X_0)P(X_1|X_0)P(X_2|X_1)\cdots P(X_n|X_{n-1}).$$

We will usually be thinking of the subscript as representing time. For instance, X_0 might be the prevalence of trachoma in a village at the beginning of a study, X_1 the prevalence at time 1, and X_2 at time 2. This equation will continue to be used even when we (soon) go over to consider the X_i's to be vectors.

Let's do another one.



Here, imagine that X_i is still the prevalence of trachoma at time i, but now Y_i is an imperfect measurement or observation, or any other variable that is influenced by X_i. Let's topologically sort the variables: $X_0, Y_0, X_1, Y_1, X_2, Y_2$. Applying the chain rule in the same way:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_0, Y_0, X_1, Y_1, X_2)P(X_0, Y_0, X_1, Y_1, X_2)$$

Now, apply conditional independence:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_0, Y_0, X_1, Y_1, X_2)$$

Let's take the next one in the topological order:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_0, Y_0, X_1, Y_1)P(X_0, Y_0, X_1, Y_1)$$

The only parent of X_2 is X_1, so applying conditional independence again:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(X_0, Y_0, X_1, Y_1)$$

Another round:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(Y_1|X_0, Y_0, X_1)P(X_0, Y_0, X_1)$$

The only parent of Y_1 is X_1:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(Y_1|X_1)P(X_0, Y_0, X_1)$$

Continuing:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(Y_1|X_1)P(X_1|X_0, Y_0)P(X_0, Y_0)$$

The only parent of X_1 is X_0:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(Y_1|X_1)P(X_1|X_0)P(X_0, Y_0)$$

Finally:

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(Y_2|X_2)P(X_2|X_1)P(Y_1|X_1)P(X_1|X_0)P(Y_0|X_0)P(Y_0)$$

We can rearrange this to give

$$P(X_0, Y_0, X_1, Y_1, X_2, Y_2) = P(X_0)P(X_1|X_0)P(X_2|X_1)P(Y_0|X_0)P(Y_1|X_1)P(Y_2|X_2) \quad (1)$$

The first part of this gives the distribution of the values X, and then given that, the latter part tells the distribution of the Y's given the X's. Models of this form are known as *hidden Markov* models, and they are used all over the place in applications, including medicine and public health. We will see this equation again later in the semester.

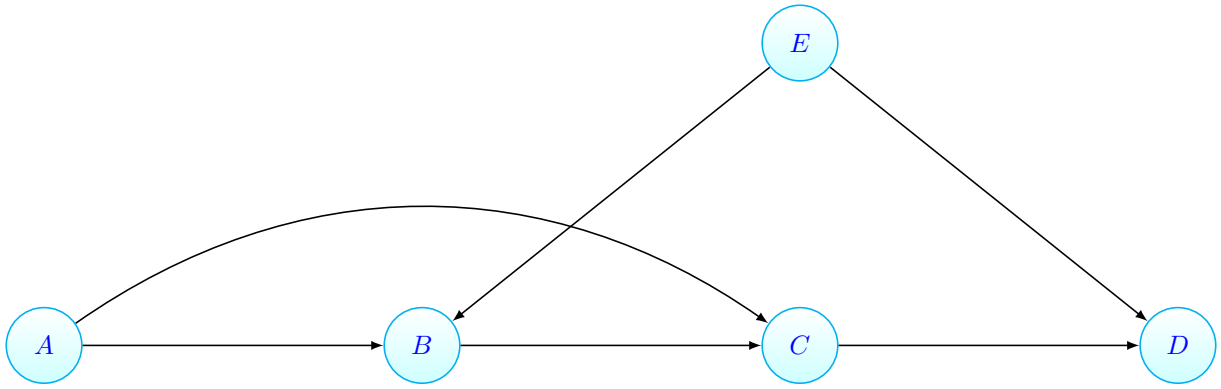
Exercise 2. The topological sort is not unique. Write down a different topologically sorted ordering of the variables, and apply the same chain rule. Did it make any difference? ∎

The general chain rule for a Bayes net over variables X_1, X_2, …, X_n is

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{parents}(X_i)).$$

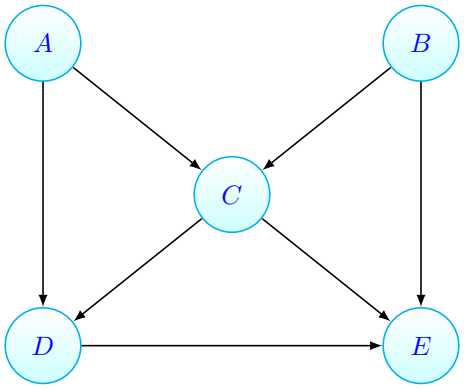
The general proof is given in various books, but is essentially just the general version of the steps we've gone through.

Exercise 3.



The graph above is called a *Verma graph*. Use the chain rule for Bayes nets to write the full joint distribution in terms of conditional probabilities. ∎

Exercise 4.



Use the chain rule for Bayes nets to write the full joint distribution in terms of conditional probabilities. ∎

A simple Markov model

Let's now look at a simple Markov chain. We return to the needle reuse example from last time (not the needlestick injury example; that is completely different.) To review, we supposed a needle is being—improperly!—reused between patients, contrary to acceptable policy. The prevalence of infection in the patient pool is p and the needle is used on a random patient. A needle is first rinsed, decontaminating it with probability b. Then, if it is used on an infected patient, the needle becomes contaminated. We found that the chance the needle is contaminated after one round. If the needle was not contaminated, then it will be contaminated with probability p. If it was contaminated, it will be contaminated with probability $1 - b + bp$.

Let X_τ be an indicator variable (1 if yes, 0 if no) for whether the needle is contaminated at time τ. We write

$$P(X_{\tau+1} = 1) = P(X_\tau = 0)P(X_{\tau+1} = 1|X_\tau = 0) + P(X_\tau = 1)P(X_{\tau+1} = 1|X_\tau = 1).$$

This is just the law of total probability. But we worked out those conditional probabilities already. So

$$P(X_{\tau+1} = 1) = P(X_\tau = 0)p + P(X_\tau = 1)(1 - b + bp)$$

or

$$P(X_{\tau+1} = 1) = (1 - P(X_\tau = 1))p + P(X_\tau = 1)(1 - b + bp).$$

This is easier to read if we just write (for example) $y_\tau = P(X_\tau = 1)$:

$$y_{\tau+1} = (1 - y_\tau)p + y_\tau(1 - b + bp) = p - py_\tau + y_\tau - by_\tau + bpy_\tau = p + (1 - p)(1 - b)y_\tau \tag{2}$$

Worked example using a few concepts we haven't used before. Assume $y_0 = 0$. Find a formula for y_τ. *Solution.* For simplicity let $k = (1-p)(1-b)$, so that $y_{\tau+1} = p+ky_\tau$. Equation **??** guarantees $y_1 = p+ky_0 = p$.

Let's plug in again: $y_2 = p + ky_1 = p + kp = p(1+k)$. Again: $y_3 = p + ky_2 = p + kp(1+k) = p(1+k+k^2)$. We conjecture $y_\tau = p(1 + k + \cdots + k^{\tau-1})$. We know this is true for $\tau = 3$ and below by direct example just given. If this is true, then $y_{\tau+1} = p + ky_\tau = p + kp(1 + k + \cdots + k^{\tau-1})$. This is $p(1 + k + k^2 + \cdots + kk^{\tau-1}) = p(1 + k + \cdots + k^\tau)$. It must therefore be true for all τ by mathematical induction. Also it turns out that $1 + k + k^2 + \cdots + k^{\tau-1} = \frac{1-k^\tau}{1-k}$. So our solution is

$$y_\tau = \frac{p(1-k^\tau)}{1-k} = \frac{p(1-(1-p)^\tau(1-b)^\tau}{1-(1-p)(1-b)}.$$

If $y_\tau = p + (1-p)(1-b)y_\tau$, then the value $y_\tau = y_{\tau+1}$ and the values don't change any more. Call this value $\bar{y}$.

Exercise 5. Show $\bar{y} = p/(1-(1-p)(1-b))$. If $0 < r < 1$, what is $\lim_{\tau\to\infty} r^\tau$? (Just the answer intuitively, not a formal proof.) Use your formula from the previous problem and show that as $\tau \to \infty$, $y_\tau \to \bar{y}$. This is our first example of something quite important: the convergence of a Markov chain to a stationary distribution. ∎

Exercise 6. Consider the value of $\bar{y}$ for these extreme special cases: $p = 0$, $p = 1$, $b = 0$, $b = 1$. Does the formula give interpretable results? What if $p = b = 1$? ∎

Worked exercise. We showed that for the two state Markov chain, the equilibrium value is $\bar{y} = p/(1-(1-p)(1-b))$. In one step, the probability of going from uncontaminated to contaminated (if you are uncontaminated) is p; the probability of going from contaminated back to uncontaminated (given the needle is contaminated) is $1-(1-b+bp) = 1-1+b-bp = b(1-p)$. Show that at equilibrium, the chance the needle is uncontaminated times the chance it moves to contaminated equals the chance the needle is contaminated times the chance it moves back to uncontaminated.

Solution.

$$(1-\bar{y})p = \bar{y}b(1-p)$$

Substituting,

$$p\frac{1-(1-p)(1-b)-p}{1-(1-p)(1-b)} = \frac{pb(1-p)}{1-(1-b)(1-p)},$$

$$p\frac{1-(1-p-b+bp)-p}{1-(1-p)(1-b)} = \frac{b(1-p)p}{1-(1-b)(1-p)},$$

$$p(1-1+p+b-bp-p) = b(1-p)p,$$

$$p(b-bp) = b(1-p)p,$$

and

$$pb(1-p) = b(1-p)p.$$

This is an identity and the proof is done. This is an extremely important property called *detailed balance*.

Conditional Expectation

The conditional expectation of Y given a particular value of $X = x$ is

$$E[Y|X = x] = \sum_y yP(Y = y|X = x).$$

Exercise 7. Suppose that in a certain region, the prevalence Q of infection is either 10%, 20%, or 40% in a district, and that the probabilities are 40%, 50%, and 10%, respectively. (In other words, forty percent of the districts have a prevelance of 10%, fifty percent have a prevalence of 20%, and ten percent have a prevalence of 40%.) In each district, suppose you sample N children and estimate the number X with the infection. Assume simple random sampling; given the prevalence, the number of infected children in the sample is going to be assumed binomial (the districts are much larger than the sample size).

Prevalence π	$P(Q = \pi)$	$E[X\|Q = \pi]$
0.1	0.4	
0.2	0.5	
0.4	0.1	

The quantity $E[X|Q]$ is a random variable taking what values with what probabilities? Find the expected value of the random variable $E[X|Q]$. ∎

We will show something that gets called "Adam's Theorem", supposedly because it's been around forever:

$$E[E[Y|X]] = E[Y].$$

So

$$E[E[Y|X]] = E[\sum_y yP(Y = y|X = x)] = \sum_x \sum_y yP(Y = y|X = x)P(X = x)$$

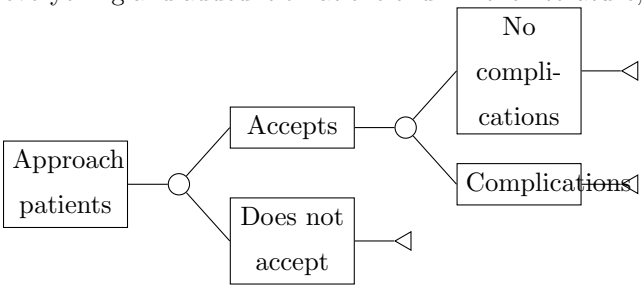
$$E[E[Y|X]] = \sum_x \sum_y yP(Y = y, X = x) = \sum_y y \sum_x P(Y = y, X = x) = \sum_y yP(Y = y) = E[Y].$$

Here is a different example. Suppose we do the following: reach out to people known to have latent tuberculosis infection, offering them therapy. We have a variable A indicating whether or not the person accepted the therapy. We have a second variable C for whether or not the person had complications from therapy.

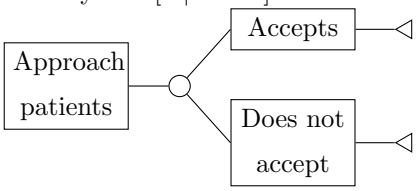
We will assume that each outcome is associated with certain costs and probabilities, as shown in this table

Acceptance $A = a$	Complications $C = c$	Cost
0	0	\$2
0	1	Impossible
1	0	\$2 + \$200
1	1	\$2 + \$200 + \$500

We assume that everyone we reach out too costs us 2 per person in administrative overhead or postage. Assume that the therapy itself costs $200, and that the cost of complications (if they occur) is $500. Of course, we didn't have to carry the $2 outreach cost through the whole computation; we could have done everything and added it on at the end. In the literature, such analyses are often represented as *decision trees*:



It's much simpler usually to work with conditional probabilities in the tree form. For instance, we may know that 40% will refuse the therapy, and we may know that of people who accept it, 10% will develop complications. (Note: these are illustrative numbers only.) We might decide to compute the conditional expectation of the cost Y given acceptance of therapy: $E[Y|A=1] = 0.9 \times 202 + 0.1 \times 702 = 252$. And we can compute the conditional expectation of the cost given refusal, which formally is $E[Y|A=0] = 1 \times 2 = 2$. That tree above would be replaced (conceptually) by a simpler one:



We would then compute the expected cost of the policy by computing $0.4 \times 2 + 0.6 \times 252 = 152$. This process is sometimes called "folding back the tree" and it's not hard to see that it is nothing more than an application of the law of iterated expectation/"Adam's Theorem". In cost-effectiveness analysis, software usually automates it.

Optional exercise. Suppose that an intracameral injection of a certain antibiotic costs C on average. Suppose the probability of endophthalmitis when it is not used is p, and when it is not is $(1-e)p$, $0 \leq e \leq 1$. Here e is the efficacy. Suppose the cost of endophthalmitis is X. What is the expected cost break-even point for the efficacy of the antibiotic? (When does the expected cost of the therapy equal the expected cost of no therapy?) ∎

Conditional variance

The conditional variance $var(Y|X)$ is defined as the expected squared difference from the conditional mean. Let me have a some toy random variables. For instance, X is a binary factor, and say Y is some measurement like the cost.

X	Y	$P(X=x, Y=y)$
0	10	0.16
0	100	0.04
1	10	0.72
1	110	0.08

We can look just at the conditional distribution of Y given $X = 0$.

x	Y	$P(Y=y\|X=0)$
0	10	$0.16/0.2 = 0.8$
0	100	$0.04/0.2 = 0.2$

So with X frozen at 0, Y is a random variable with the given distribution. Its expectation value is 10 times the chance of 10 plus 100 times the chance of 100: $10 \times 0.8 + 100 \times 0.2 = 28$. We can compute its second moment too: the sum of each squared value times the chance of the value you're squaring: $10^2 \times 0.8 + 100^2 \times 0.2$ which is 2080. The variance here, which will be the *conditional variance* of Y given $X = 0$ is the conditional second moment minus the conditional expectation, or 2080-28^2, or 1296. We can also write this as

$$\text{var}(Y|X=x) = E[(Y - E[Y|X=x])^2|X=x]$$

$$\text{var}(Y|X=x) = E[Y^2 - 2YE[Y|X=x] + (E[Y|X=x])^2|X=x]$$

$$\text{var}(Y|X=x) = E[Y^2|X=x] - 2E[YE[Y|X=x]|X=x] + E[(E[Y|X=x])^2|X=x]$$

$$\text{var}(Y|X=x) = E[Y^2|X=x] - 2E[YE[Y|X=x]|X=x] + E[(E[Y|X=x])^2|X=x]$$

Now: when $X = x$, $E[Y|X=x]$ is *just some number*; it's a constant, so outside the expectation it goes. For the same reason, the squared conditional expectation goes outside too (but then nothing is left under the expectation sign; the expectation of a constant is just the constant.)

$$\text{var}(Y|X=x) = E[Y^2|X=x] - 2E[Y|X=x]E[Y|X=x] + (E[Y|X=x])^2$$

Then combine the last two terms:

$$\text{var}(Y|X=x) = E[Y^2|X=x] - (E[Y|X=x])^2$$

Exercise 8. Show the conditional variance of Y given $X = 1$ is 900 in our example. ∎

We can write

$$\text{var}(Y|X) = E[Y^2|X] - (E[Y|X])^2.$$

Here, we are now talking about a relation between random variables.

So we can think of $E[Y|X]$ as a *transform of* X, and in the same way, the conditional variance of Y given X as another transform of X. Whenever X takes one of its values, as it does with whatever probability it is supposed to, we take instead a smoothed version of Y—one smoothed over that value of X.

x	$P(X=x)$	$E[Y\|X=x]$	$\text{var}(Y\|X=x)$	$(E[Y\|X=x])^2$
0	0.2	28	1296	784
1	0.8	20	900	400

Now: if the conditional expectation of Y given X is just a random variable (a transform of X, of X, of X, of X!), what is its expectation? We already know $E[E[Y|X]] = E[Y]$; this is 28*0.2 + 20*0.8 = 21.6. Let's compute its variance. We can take the fourth column of the table and compute the second moment: $784 \times 0.2 + 400 \times 0.8 = 476.8$. So the variance *of the conditional expectation of* Y *given* X is the second moment minus the square of the mean, which comes out to be 10.24 here.

And while we're at it, the conditional variance of Y given X is a random variable, and it's also a transform of X. What is its expectation? The sum of the values times the chance of each. Thus, $E[\text{var}(Y|X)]$ is $1296 \times 0.2 + 900 \times 0.8 = 979.2$.

The variance of Y is now going to be the second moment (1456) minus the square of *its* mean: 466.56, for 989.44. Now, lets add up the expectation of the conditional variance (979.2) to the variance of the conditional expectation: 10.24, and get 989.44.

The variance of Y is the variance of the conditional expectation of Y given X, plus the expectation of the conditional variance of Y given X.

Now, let's prove this in general. We had

$$\text{var}(Y|X) = E[Y^2|X] - (E[Y|X])^2$$

This is a relationship between random variables. Let's take the expectation of both sides:

$$E[\text{var}(Y|X)] = E[E[Y^2|X]] - E[(E[Y|X])^2]$$

Use Adam's Law:

$$E[\text{var}(Y|X)] = E[Y^2] - E[(E[Y|X])^2]$$

What about

$$\text{var}(E[Y|X])?$$

That's the variance of a random variable, specifically $E[Y|X]$. The variance is the second moment minus the square of the mean.

$$\text{var}(E[Y|X]) = E[(E[Y|X])^2] - (E[E[Y|X]])^2$$

On the right, the expectations are all under the squared sign, and we know $E[E[Y|X]] = E[Y]$.

$$\text{var}(E[Y|X]) = E[(E[Y|X])^2] - (E[Y])^2$$

So if we add

$$E[\text{var}(Y|X)] + \text{var}(E[Y|X]) = E[Y^2] - E[(E[Y|X])^2] + E[(E[Y|X])^2] - (E[Y])^2$$

The middle terms cancel.

$$E[\text{var}(Y|X)] + \text{var}(E[Y|X]) = E[Y^2] - (E[Y])^2$$

On the right—for Y, the second moment minus the square of the mean.

$$\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}(E[Y|X])$$

The first part of this is the part "UNexplained" by X; the second term is the part "explained" by X. We can think of this formula as partitioning the variance of Y into these two components. Notice this has nothing to do with the normal distribution or any kind of parameterization at all.

Consider the following optional worked example. Suppose that in some country the prevalence of infection is π. For each district, however, suppose the prevalence has mean π but variance σ^2. Suppose we pick M districts and N people per district, with N much less than the district population. Assume the districts are all the same size. Let X_i be the number of positive people in our sample from district i, $i = 1, \ldots, M$. Let $\hat{p}$ be $\frac{\sum_{X_i}}{NM}$. Assume districts are independently sampled and people are independently sampled within the district. Find the mean and variance of $\hat{p}$.

Solution.

$$E[\hat{p}] = E[\frac{1}{MN}\sum_i^M X_i] = \frac{1}{MN}\sum_i^M E[X_i]$$

Then:

$$E[X_i] = E[E[X_i|W_i]]$$

where W_i is the unobserved true prevalence in the village. Once we have the true prevalence, the number of positives in our sample is binomial with W_i as the probability and N as the number of trials. So $E[X_i|W_i] = NW_i$.

$$E[X_i] = E[E[X_i|W_i]] = E[NW_i] = NE[W_i] = N\pi.$$

So

$$E[\hat{p}] = \frac{1}{MN}\sum_i^M N\pi = \frac{MN}{MN}\pi = \pi.$$

For the variance,

$$\text{var}[\hat{p}] = \text{var}(\frac{1}{MN}\sum_{i}^{M} X_i) = \frac{1}{M^2N^2}\sum_{i}^{M}\text{var}(X_i) = \frac{1}{MN^2}\text{var}(X_i)$$

Then

$$\text{var}(X_i) = \text{var}(E[X_i|W_i]) + E[\text{var}(X_i|W_i)].$$

We know $E[X_i|W_i] = NW_i$ from before. And given W_i and the binomial, we have $\text{var}(X_i|W_i) = NW_i(1-W_i)$.

$$\text{var}(X_i) = \text{var}(NW_i) + E[NW_i(1-W_i)].$$

$$\text{var}(X_i) = N^2\text{var}(W_i) + NE[W_i(1-W_i)] = N^2\sigma^2 + NE[W_i] - NE[W_i^2]$$

$$\text{var}(X_i) = N^2\sigma^2 + N\pi - N(\sigma^2 + \pi^2)$$

$$\text{var}(X_i) = N(N-1)\sigma^2 + N\pi(1-\pi)$$

Finally,

$$\text{var}[\hat{p}] = \frac{N(N-1)\sigma^2 + N\pi(1-\pi)}{MN^2} = \frac{\pi(1-\pi) + (N-1)\sigma^2}{MN}.$$

If this were just a straight binomial sample, with no extra between village variability, then this variance would just be $\pi(1-\pi)/MN$. The extra term in the numerator, $(N-1)\sigma^2$, is called the *variance inflation factor*. Notice the village size N actually hurts you in the numerator even though it helps you in the denominator. Getting more villages is pure benefit, since making M bigger increases the denominator at no cost to your numerator.

Exercise 9. How long did this assignment take you? ∎