Modeling Probabilities of All-or-Nothing Events

Often we're looking at all-or-nothing outcomes: life or death, admitted or rejected from Harvard med school, and so on (obviously in ascending order of importance). We represent each outcome as a 0 or 1. The best we can do is make a model that predicts the *probability* of getting a 1 (e.g., life) for any given set of conditions.

There are a variety of different models used for predicting these dichotomous outcomes, including logistic regression and probit. There's no snap rule to choose between them. Sometimes understanding the logic of the model will tell you- we'll see how that can work for one simple model. More often, it's just a matter of seeing what works best. We'll focus on logistic regression just because it's the most widely used.

There will be one change of strategies that affects *all* these dichotomous predictions, *regardless of the particular model chosen*. Instead of adjusting model parameters to minimize some squared prediction error, we typically adjust them to *maximize the probability of getting our observed results, given our model.*

How should we pick the best-fit model?

The criterion for the best fit really depends on what you want the model to do.

Say you're an insurance company trying to set the price for a group policy for a college. You need to predict the probability of people getting an expensive ($1M) disease. Your model (e.g. logistic regression) gives you a probability p_i for each college employee of them costing you $1M. Those p_i's depend on things like the age and sex of the employee. You set parameters in the model for the effects of age and sex on p_i by fitting data from similar colleges. You add up those probabilities to estimate how much you expect to pay out. For each employee, whose modeled probability is p_i, the error in your estimate is just $(1-p_i)$*$1M or (p_i)*$1M, depending on if they get the disease or not. You can do a good job just by minimizing the sum of those errors squared in your old data. So you wouldn't need any new methods.

Still, the model you get with those least-squares parameters might be far from the best model for other purposes. Say that one person per 100,000 for whom the model says $p=0.0000000001$ gets the disease. That's not too big a deal for estimating total costs. It's a very big deal if you want to use your model to make predictions for a much different population, where people like that one may not be so rare. Let's think of a more extreme case. Say you had a model that gave $p_i = 0$ or $p_i = 1$ for some people. Then a single case of disease for a $p_i = 0$ case or of no disease for a $p_i = 1$ case tells you that it's impossible that your model is right. A single case of disease where your model gives $p_i = 0.0000000001$ doesn't quite say the model logically has to be wrong, but it does say that if it were right you'd be extremely unlikely to see the results you saw.

Likelihood Methods

If we want a model that's "right" in the sense that it does a good job of predicting the probability of rare events as well as common ones, then we need a different way to evaluate what fits best. What we do is to find the parameters that make the observed data as probable as possible to be what you'd see. For this type of "maximum likelihood" fit, the difference between saying a sick person had p_i =0.0000000001 and saying they had p_i =0.1 is a big deal, but the difference between saying they had p_i =0.4 and p_i =0.6 isn't a big deal. For simple least squares, those differences are equally important. Likelihood measures how much you should be surprised by the data, if you think your model is true. Least squares measures what sort of payoff errors you might expect for another population similar to the one you already observed.

For example, if a model predicts p_i = 0.5 of costing $1M for 4 people, and they all get sick, the company is out $2M more than they expected. Even if the model is right over the long run for the population, that type of error isn't very unlikely. If just one person for whom the model gave p_i =0 got sick, the company would only be out $1M more than they expected. However, that one person would prove that the model was wrong. You can see why you want different criteria depending on what use you're planning for your model.

This maximum likelihood technique is really used all over the place in more advanced statistics, so it's good to get a taste of it here. The idea is this. Any model gives probabilities for getting your results. This probability is called the *likelihood* of the model. We search around (via a computer) among all the parameters for our type of model to find the ones that would have made our results most probable, i.e. the model with the biggest likelihood. We can start looking at simple examples where there are just lists of different hypotheses to evaluate, not a continuum of different parameters for a family of hypotheses.

Simple Likelihood Examples

You've got 3 boxes, *A* , *B,* and *C*, filled with money. *A* has 2000 $1 bills. *B* has 1000 $1 bills and 1000 $10 bills. *C* has 2000 $10 bills. You don't know which box is which.
You get to draw 1 bill from a box of your choice. Your chances of drawing from any of the boxes are equal: 1/3. So your chances of drawing any particular bill are equal, 1/6000.

You're in luck, you draw a $10.

What's the probability that you drew from each of *A* , *B,* and *C*?

The bill you drew was one of 1000 $10 bills from *B* or one of 2000 10 bills from *C*. The possibilities have narrowed from 6000 bills to 3000.

So any of those 3000 $10 bills were equally probable. None were in *A*, 1000 were in *B*, and 2000 were in *C*. So the probability that you drew from *A* is zero,
the probability that you drew from *B* is 1000/3000= 1/3, and the probability that you drew from *C* is 2000/3000= 2/3.

You're on a roll, so you draw again from the same box. Luck again, another $10!

Now what's the probability that you drew from *A* , *B,* or *C*?
Any pair of $10 bills from the same box are equally likely to be the ones you drew.
Still no chance it was from *A*.
There are 500*499/2 or about 500*250 pairs of bills you might have drawn from *B*.
There are 1000*999/2 or about 500*1000 pairs of bills you might have drawn from *C*.

So now the result you got is ~4 times more probable if you've been drawing from *C* than if you've been drawing from *B.* The probability you've been drawing from *B* is now about:

1/10 1/5 ¼ 1/3 ½ 2/3 ?

Simple Likelihood Examples

You can think of what you just did as evaluating 3 competing hypotheses in terms of how probable they were to generate the data. The 3 hypotheses were that you were drawing from *A* , *B,* or *C*. Before you had any data, each hypothesis had probability 1/3.

Then you drew a $10 bill. You evaluated the probability that would happen for each of the 3 hypotheses, and we call that the *likelihood* of the hypothesis L(A)=0, L(B)=1/2, L(C)=1. (Weird notation, but don't blame us!)

After you got the data (and the $10 bill!) you had new probabilities for the 3 hypotheses. For each hypothesis you multiplied your old probabilities by the probability of getting your result. Then you renormalized to make those probabilities still add up to one.

Hypothesis	A	B	C	
Initial probability, P_0	1/3	1/3	1/3	
Probability of $10, L	0	½	1	
P_0*L	0	1/6	1/3	(These sum to ½.)
New probability P_1=P_0*L/sum	0	1/3	2/3.	

What did you do when you drew the second bill? Let's say the likelihoods for $10 were still given by 0, ½, 1, which is very close since we had so many bills to start with.

Hypothesis	A	B	C	
Starting P_1, from above	0	1/3	2/3	
Probability of $10, L	0	½	1	
P_1*L	0	1/6	2/3	(These sum to 5/6.)
New probability P_2=P_1*L/sum	0	1/5	4/5.	

You could see that you can continue this re-evaluation of your hypotheses as you get more data. What would the next P's be if your next draw from that box were (alas) a $1?

Simple Likelihood Exercises

Now let's say that each box only has 2 bills in it. (*A* has 2 $1's, *B* has 1 each, and C has 2 $10's).

Say you've drawn 1 $10 bill. What are the probabilities of *A* , *B,* and *C*? It's just like the case with lots of bills in the same proportions:

Hypothesis	A	B	C	
Initial probability, P_0	1/3	1/3	1/3	
Probability of $10, L	0	½	1	
P_0*L	0	1/6	1/3	(These sum to ½.)
New probability P_1=P_0*L/sum	0	1/3	2/3.	

Now you draw a second $10 bill from the same box.
What are the probabilities of *A* , *B,* and *C*?

Likelihoods with Several Possible Outcomes

So far we've just looked at likelihoods for processes with all-or-nothing outcomes, such as life or death, $1 or $10. Likelihood methods can be used just as well to pick hypotheses when there are more possible outcomes.

Let's say those boxes could have $1, $10, or $20 bills in them. Let's say we've got 3 boxes with lots of bills again: A has 10% $20's, 20% $10's, and 70% $1's. B has 20% $20's, 30% $10's, and 50% $1's. C has 30% $20's, 30% $10's, and 40% $1's.

Now let's say we make 5 draws: $20, $1, $1, $10, $1.
Let's figure out how likely we were to get that result from each of A, B, and C. Then we can figure out the probability that each hypothesis (A, B, and C) was correct.

Hypothesis	A	B	C	
Initial probability, P_0	1/3	1/3	1/3	
Likelihood, $20	0.1	0.2	0.3	
Likelihood, $1	0.7	0.5	0.4	
Likelihood, $1	0.7	0.5	0.4	
Likelihood, $1	0.7	0.5	0.4	
Likelihood, $10	0.2	0.3	0.3	
	(events independent so L= product)			
L	0.0069	0.0075	0.0058	(sum to 0.0202)
$P=P_0*L/sum$	0.343	0.372	0.285	

We'd need to draw many more bills to have much confidence in knowing which box we were drawing from.

Exactly the same procedure could be extended to include any number of categorical outcomes. So long as the results are independent, this likelihood procedure is very simple and straightforward for anything (a computer) with the patience to do the calculations.

Likelihoods, Odds, and Odds Ratios

Notice that in each problem when we wanted to calculate the probability of each hypothesis we only needed the *ratios* of their probabilities, because we know that the *probabilities have to add up to one.* Let's explore this in a simple two-hypothesis case.
***When we have just two possibilities, we call the ratios of their probabilities the Odds*.**
What we've been doing is that for each new piece of data we took our old A/B odds and multiplied it by some A/B Odds Ratio based on the new data.
Take the last example, but with only boxes A and B. At the end we just need the Odds.

We've got 2 boxes with lots of bills again: A has 10% \$20's, 20% \$10's, and 70% \$1's. B has 20% \$20's, 30% \$10's, and 50% \$1's. Now let's say we make 5 draws: \$20, \$1, \$1, \$10, \$1. We can figure out the probability that each hypothesis (A or B) was correct.

Hypothesis	A	B	A/B	
Initial probability, P_0	1/2	1/2	initial Odds	1
Likelihood, \$20	0.1	0.2	odds ratio	1/2
Likelihood, \$1	0.7	0.5	odds ratio	7/5
Likelihood, \$1	0.7	0.5	odds ratio	7/5
Likelihood, \$1	0.7	0.5	odds ratio	7/5
Likelihood, \$10	0.2	0.3	odds ratio	2/3
L	0.0069	0.0075	final odds	343/375 (product)
$P=P_0*L/\text{sum}$	0.478	0.522		

If we were just keeping track of Odds, how would we get back p= P(A) at the end?

Try that for this example.

We often like to use *additive* models, so instead of multiplying the OR's, we take their ln's and add them, then antilog at the end.

Ln(initial A/B Odds)	ln(1)=	0
Ln(A/B odds ratio \$20)	ln(1/2)=	-0.693
Ln(A/B odds ratio \$1)	ln(7/5)=	+0.337
Ln(A/B odds ratio \$1)	ln(7/5)=	+0.337
Ln(A/B odds ratio \$1)	ln(7/5)=	+0.337
Ln(A/B odds ratio \$10)	ln(2/3)=	-0.406
Ln(final Odds)	ln(343/375)=-0.09	(sum of column)
Final Odds = $e^{\ln(\text{final odds})}$ =	343/375	

In real life, this procedure is used to evaluate more important hypotheses, e.g. whether a cancer is a lethal one. Many somewhat independent pieces of diagnostic evidence each contribute some OR. The final ln(odds) can be obtained from an additive model of the ln(OR) for each diagnostic trait. That's what we were doing in Logistic Regression.

Likelihood with an Almost Continuous Parameter

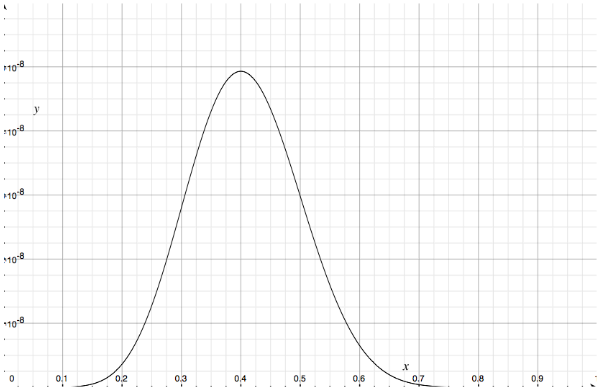
Say we had 101 boxes, each with huge numbers of \$1 and \$10 bills.
One has 0% \$10's, the next has 1% \$10's, etc. on up to 100% \$10's.
So these boxes have p(\$10)= 0, 0.1, 0.02,…, 1.00.

You draw say 25 bills, 10 \$10's and 15 \$1's in some order, e.g. \$1, \$10, \$10, \$10, \$1, .. We call the likelihoods of these 25 independent outcomes L_1, L_2. L_3, …, L_{25}.
L_j is p when you draw \$10 and (1-p) when you draw \$1.
The likelihood of our particular result for box "p" is then
$(1-p)*p*p*p*(1-p)\ldots = p^{10}{}_{*}(1-p)^{15}$

Hypothesis ($p=\%_*100$)	0,	0.01	0.02	…………..1.00
P_0	1/101	1/101	1/101	……………..1/101
Likelihood of \$10	0	0.01	0.02	……………1.00
Likelihood of \$1	0	0.01	0.02	……………1.00
$L_{1*}L_{2*}L_3\ldots L_{25}$=	(Each result is $p^{10}{}_{*}(1-p)^{15}$ for the hypothesis p.)			

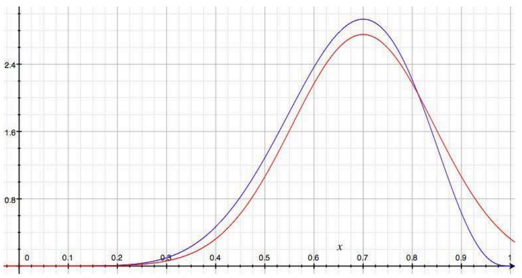
You can plot that familiar binomial function, $p^{10}{}_{*}(1-p)^{15}$ as a function of p to see how the likelihood varies with p. No surprise, the highest L is for p=0.40= 10/25.

(Why didn't we bother with the factor 25!/(10!15!) that comes from counting all the different patterns of getting the 10 \$10's and 15 \$1's? That factor doesn't depend on our hypothesis so it doesn't change the shape of L(p))



Likelihoods and Continuous Parameters

Here's a really simple example. You already know the answer but now we'll do it by likelihoods. Say we observed 10 people and 7 survived. We want to model survival by a single parameter p, the probability of survival. For some p, how probable is it that 7 would survive? It's just given by our old binomial expression:
$L(p)= (10!/3!*7!)p^7(1-p)^3$. That peaks at p=0.7, so that's the maximum likelihood estimate for p. The shape of L(p) is a bit different from the normal curve. For example, the normal curve always spills out into the impossible region where p>1, unlike L(p). That's a hint that if we develop methods to use likelihoods to estimate confidence intervals, they won't give exactly the same results as the normal approximation.



Null Hypothesis Testing via Likelihood

Sometimes we wish to use likelihood methods to do the same things we did with our least-squares fitting methods- find a p-value for a null hypothesis. Remember what we did with least squares. We said that the probability that our estimate of a parameter was off from the true value fell off as the squared error of the estimate got larger. Our χ^2 statistic was a properly normalized way to keep track of how far off our results were from the null prediction. We could then look up on a χ^2 table to see how likely it would be to get χ^2 that big or bigger as our results if the null were true. That was then called the p-value.

Now in general the way the likelihood L of the null falls off as the results become more different from the null prediction does not follow our old normal curve. However, for large samples our old central limit theorem kicks in again, and the form of how L falls off away from its maximum (L_M) takes on the old universal normal shape. So then we can use our old χ^2 tables to figure out how badly the observed results fit with the null hypothesis. We need a statistic, based on likelihood calculations, that plays the same role as χ^2.

For our old χ^2 statistic we had a probability density ρ that fell from its maximum value ρ_{Max} as. Taking the natural log of both sides gives:

Thus, brushing past some philosophical issues about the relation between ρ and L, if we want a statistic that plays the same role as χ^2 we want to use

.

How would this deviance, G, get used? Say that we have a logistic regression problem where we can't use least-squares fitting. Our computers are fast enough, however, to calculate some likelihoods. We find out L_{Null}/L_{Max}. What was the probability of getting that low or lower a value for L_{Null}/L_{Max} if the null were true? We calculate the deviance,

$G= -2*\ln(L_{Null}/L_{Max})$, and plug it into a standard χ^2 table with the right number of degrees of freedom to get a p-value.

On General Likelihood Methods
Notice that whenever we have discrete outcomes, we can calculate a probability that we would have gotten our particular set of outcomes from any model we're given for how the probabilities depend on the conditions. So if we wanted, we could have used this same likelihood method to see how well different models fit any discrete-outcome process. We could have even done that for problems that we treated using least-squared error methods. Usually, we would have gotten very nearly the same estimates and about the same confidence intervals. With a little tweak, thinking of probabilities of results *near* the exact one found, we can use these methods for continuous-outcome processes too.

The same ideas work for all sort of different boxes (or even continuous families of hypotheses, with one or more parameters) and all sorts of different data. In most practical examples the likelihoods are harder to calculate so you use computer programs.

Any model gives a set of probabilities, one for each category of result. The categories could be draw one and draw two from the same big box, where the models are the different fractions of bills in each box. Or the models could be two different diagnoses for a patient, and the results are categorical facts about the patient, e.g. gender, ethnicity, whether or not they have high blood pressure,… From those probabilities, we can calculate the probability that we would have drawn our particular set of results from a box model for that category. That is the *likelihood* assigned to that model. We find the model parameters with the highest likelihood for the results we saw.

Sometimes we can assume that each result is independent of each other result. When that's true, the probability, given some model, that we would get our particular set results is just the *product* of the probabilities for each category of result. That's like for draws with replacement. It's a good approximation for draws from the boxes with lots of bills. In real-life problems it's often a decent starting point.

When the likelihoods of the models for the data set are given by the *products* of the model likelihoods taken from each individual data point, it's often more convenient to work with sums rather than products.

The log of a product is the sum of the logs of the factors: $\ln(ab) = \ln(a) + \ln(b)$.
So often people convert that product of likelihoods into the sum of the natural *logs* of the likelihoods. Then at the end you can antilog the sum to get back the likelihood. That would be a fairly pointless detour for small simple data sets and discrete hypotheses like the ones above. For bigger data sets, where there are lots of points with a variety of different types of data, it can be more convenient. Perhaps more importantly, though outside the scope of this course, when you search for the parameters that maximize the likelihood, you're finding the parameters for which the (calculus) derivative of the likelihood with respect to any changes in the parameters is zero. (A maximum is a spot where the slope is zero.) It's very convenient to take the derivative of a sum: the slope of the sum is just the sum of the slopes. It's less convenient to take the derivative of a product. So you will frequently see discussions framed in terms of the natural logs of the likelihoods.

Likelihood Exercises with a Continuous Parameter

Say you've moved to a new small town and have no idea what political views your new neighbors have. Let's simplify and pretend that each one is either a Democrat or Republican. You talk with three people you run into at random, and each is a Republican. You want to make some sort of informed guess about what fraction P of the town is Democratic. Our old methods would say to start with the obvious estimate P=0/3, the result you got in your sample. That's also the P that gives the highest probability (1.0 !) of getting your result, no D's in the sample. The question is what sort of error bars to set on that estimate.

Our chi-sq type statistic is clearly not the way to go. The calculated SE has a factor of sqrt(p(1-p)), which comes out zero for our estimated p=0. Yet we know that it's quite possible that there are many D's in town but none happened to be in our sample. You don't have any underlying normal distribution to excuse using "t" either. Likelihood methods work for this common type of problem, in which your samples are small.

Let's make a table of the probabilities of different outcomes {0,1,2,3} for various different population P's.

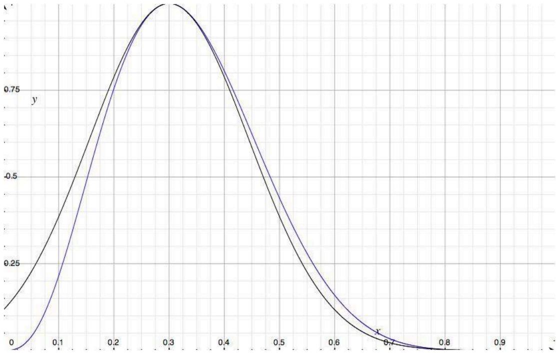
P(D)	**D=0**	**D=1**	**D=2**	**D=3**
0.0	**1**	**0**	**0**	**0**
0.1	**0.729**	**0.243**	**0.027**	**0.001**
0.2	**0.512**			
0.3	**0.343**			
0.4	**0.216**			
0.5	**1/8=0.125**	**3/8=0.375**	**3/8=0.375**	**1/8=0.375**
0.6	**0.0.64**			
0.7	**0.027**			
0.8	**0.008**			
0.9	**0.001**	**0.027**	**0.243**	**0.729**
1.0	**0**	**0**	**0**	**1**

You can fill in the rest, it's just our old binomial probabilities.

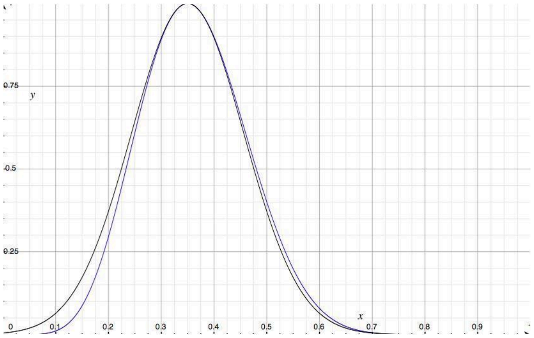
We could calculate the probability of any result if we knew P. That's just drawing tickets from a known box. But here our situation is that *we know what tickets we have but we don't know what's in the box*. All we can do here is ask "how probable is it to see this D=0 result if P had some actual value?" Knowing D but not knowing P means we know what *column* we're in here but not what *row*.

P(D)	**D=0**	**D=1**	**D=2**	**D=3**
0.0	**1**	**0**	**0**	**0**
0.1	**0.729**	**0.243**	**0.027**	**0.001**
0.2	**0.512**			
0.3	**0.343**			
0.4	**0.216**			
0.5	**1/8=0.125**	**3/8=0.375**	**3/8=0.375**	**1/8=0.375**
0.6	**0.064**			
0.7	**0.027**			
0.8	**0.008**			
0.9	**0.001**	**0.027**	**0.243**	**0.729**
1.0	**0**	**0**	**0**	**1**

That D=0 column fits best with P=0, in that P=0 is the value which was most likely to give D=0. However, even P=0.6 could give D=0 with probability 0.064, so we can hardly exclude P=0.6 with confidence. Our confidence interval is completely asymmetrical, since our estimate was P=0, and that's as low as you can go.



Now let's say we asked 10 people and got 3 D's, so P=0.30 is our estimate. We can try to calculate error bars around that estimate using SE= $(0.3*0.7/10)^{1/2}$=0.1449. We can then use the normal curve to estimate our confidence that P is within some range around that. Or we can calculate how the probability of that result depends on P, and use that to set our confidence limits. This curve gives you a feel for the difference between the likelihood curve and the normal curve. The curve shifted a little to the right is the likelihood curve for getting 3/10 as a function of P. The other curve is the normal curve with our calculated SE. They are pretty dramatically different for P<0.1, The likelihood curve indicates that such low P are not very compatible with the data but the normal approximation doesn't rule them out. The normal approximation always gives symmetrical confidence intervals, even when that doesn't make any sense. It doesn't even rule out P<0, although P<0 makes no sense.



We can go through the same exercise for 7 D's out of 20 people. You get a milder version of the same sort of difference between the likelihood calculation and the normal approximation, as shown here.

Toward Bayesian Statistics

Bayes' Rule.

This procedure you've been using is just like what you did in Stat 100 to evaluate the hypotheses that someone had or didn't have a dread disease after testing positive for it in a screening test where some results are false positives. You figured out how many results like the one you saw would be true positives and how many would be false positives. You divided the number of true positives by the total number of positives to get the new probability that the screened person was sick. These are examples of applying Bayes' Rule.

We so far just looked at cases where all the initial P's were the same, so we didn't have to think about them and can discuss just the likelihoods. The likelihoods are the things that standard computer models know how to calculate.

We can use these same techniques in more interesting cases. Say that you have 4 boxes, with two versions of the 50/50 mixed box, B_1 and $\underline{B_2}$. We still have just 3 hypotheses, *A, B,* and *C,* because you have no reason to care which B box you might have drawn from.
Now let's say you drew a $10 bill. What are the probabilities that you drew it from a box of type *A, B,* or *C*? What were the starting probabilities?

Hypothesis	A	B	C	
Initial probability, P_0	1/4	1/2	1/4	
Likelihood, $10	0	½	1	
P_0*L	0	1/4	1/4	(These sum to ½.)
New probability P_1=P_0*L/sum	0	1/2	1/2.	

This case, where the hypotheses have different initial probabilities, is a simple lead-in to **Bayesian** inference, a flexible and increasingly widely used alternative approach to statistical inference. If you go on in statistics you will learn much more about it.

Remembering **a more important example from Stat 100**:
Say you have a test for HIV that is 95% accurate-
5% of infected people test negative and 5% of uninfected people test positive.
In a screening a population in which 1% of the people are positive, someone tests positive.

a) If you were to take the two hypotheses for that person (infected, not-infected) what are their likelihoods given that data?
b) Which is the "maximum likelihood" hypothesis?
c) What are the odds that they are actually infected?

Notice that when you have information ahead of time the maximum likelihood hypothesis does not have to be the most reasonable guess about reality.

Simple Bayes Example

Remember the old clinical example: You screen for HIV using an old test that is only 95% accurate. Someone with the virus tests positive 95% of the time and negative 5% of the time. Someone without the virus tests negative 95% of the time and positive 5% of the time.
Now let's say that 1% of the males aged 20-30 in Chicago are in fact positive. You screen some randomly chosen male in that age bracket and get a positive result. What's the probability he has HIV?
Remember (unlike most MDs, who get this very wrong) that the probability turns out low. That's easy to see intuitively by imaging a population of 10,000 people and seeing how many positive results are from infected people (95) and how many are from uninfected people (495). So P(true pos|pos test)= 95/(95+495)=0.1638. Let's calculate P using the routine technique we just developed.

Hypothesis	Neg	Pos	Neg/Pos	
Initial probability, P_0	0.99	0.01	initial Odds	99
Likelihood, test pos	0.05	0.95	odds ratio	1/19
L	0.0495	0.0095	final odds	99/19 (product)
$P=P_0$*L/sum	0.8362	0.1638.		

Nothing new there.

Real-Life Bayes examples

But in real life, nothing is usually that clean. You almost always have more or less information than used in that simple calculation.

E.g. Say that the guy screened is actually 31 years old. You don't have any tabulated screening results on people over 30 years old. So you have *less* information than in our example.
Discussion exercise: what probability that this person has HIV?
Do you say 95% because the test is the only info? Seriously?
Do you say "I have no idea because I have no population data on the group?" Seriously?
What would a rational conclusion be? Is it purely objective?

E.g. Say that you notice numerous needle marks in the arms of the guy screened. So you have *more* information than in our simple example.
Discussion exercise: what probability that this person has HIV?
Do you say 16% because that's what the formal calculation gives? Seriously?
Do you say "I have no idea since I have no population data on this narrower group?" Seriously?
What would a rational conclusion be? Is it purely objective?

Bayes estimation of a continuous parameter

(The following example is a slightly fictionalized composite of two real medical examples.)

Say you want to know whether flu vaccine works well in people over 65.
There's an obvious null hypothesis:

- It works just as well in the old as it does in the general population. (say 70% effective in a typical year)

And there's another obvious null used for medical treatments:

- It doesn't work. (0% effective)

How would you pick which null to use?

Why should you have to pick either of these dull-witted, implausible nulls?

What you really know ahead of time is that the vaccine probably works a bit but not as well as in younger people. You could approximate this *prior knowledge* by saying there's a *prior probability distribution* for the effectiveness, E, that's uniform from 0% to 70%. Maybe you should pick a little different prior, peaked a bit in the middle and extending out a bit to negative E and to E > 70%, but let's not quibble.

Now let's say that there's a randomized clinical trial with the vaccine given to 100 elderly and not to another 10,000. Say that 50% of the control got the flu and 30% of the vaccinated group got the flu. From that it looks like E was ~40%, i.e. (50-30)/50, but of course with statistical uncertainty. With our big control group, we know the background rate well. So the expected number of flus in the treatment group is 50*(1-E) . Now we can just calculate the probability of getting 30 flus when 50*(1-E) are expected. But that's just the binomial probability we've calculated in problems above.

So we just multiply that binomial probability of getting 30 out of 100 when we expect 50*(1-E) by our prior uniform probability density. We take the resulting product, now a function of E as E goes from 0% to 100%, and divide by its integral. That's the Bayesian posterior probability of E. It peaks at E=40% and has a width of around 14%.