Method 3) (The hard way) ("extending the conversation")

[Feature 51] we know \( P(\theta = 1 \mid B) \),

\[
P(\gamma_i = 1 \mid \theta = 1, B) \quad \text{and} \quad P(\gamma_i = 0 \mid \theta = 0, B)
\]

or put bluntly: information is lost.

we want \( P(\theta = 1 \mid \gamma_i = 1, B) = \frac{P(\gamma_i = 1 \mid B) \cdot P(\gamma_i = 1 \mid \theta = 1, B)}{P(\gamma_i = 1 \mid B)} \)

Both of the numerator probabilities are known, but what about the annoying denominator \( \sqrt{?} \)?

Can we

How get \( P(\gamma_i = 1 \mid B) \) from \( P(\theta = 1 \mid B) \), \( P(\gamma_i = 1 \mid \theta = 1, B) \) and \( P(\gamma_i = 0 \mid \theta = 0, B) \)?

Notice that \( P(\gamma_i = 1 \mid B) \) is hard, but \( P(\gamma_i = 1 \mid \theta = 1) \) and \( P(\gamma_i = 0 \mid \theta = 0) \) are easy.

In other words, we don't know how the data (the blood test) will come out, but we do know probabilistically how the data will come out if we knew the truth \( \{ P(\gamma_i = 1 \mid \theta = 1), P(\gamma_i = 0 \mid \theta = 0) \} \). So let's

extend the conversation by bringing \( \theta \) into
The picture. Since the only two possibilities for the truth are \((\theta = 1)\) and \((\theta = 0)\), those two propositions form a partition: a collection of mutually exclusive possibilities that is exhaustive of all the possibilities. Thus

\[ P(\gamma = 1 | B) = P(\gamma = 1, \theta = 0 | B) \]

\[ + P(\gamma = 1, \theta = 1 | B) \]

This is progress (\(\theta\) is now in the conversation), but more work is needed.

Step 1: \(\theta\) is on the wrong side of the conditioning bar in \(P(\gamma = 1, \theta = 0 | B)\) to be useful to us, so let's force it to move to the other side: as before,

\[ P(\gamma = 1, \theta = 0 | B) = \]  

doesn't seem more useful, \(P(\gamma = 1 | \theta = 0, B) \) 

so let's try \(P(\gamma = 1, \theta = 0 | B) = P(\gamma = 1 | \theta = 0, B) \).
\[ P(\gamma_1 = 1, \theta_0 \mid B) = P(\gamma_1 = 1 \mid \theta_0 = 0, B) \cdot \frac{P(\theta_0 = 0, B)}{P(B)} \]

So it works, and the answer is

\[ P(\gamma_1 = 1, \theta_0 \mid B) = P(\gamma_1 = 1 \mid \theta_0 = 0, B) \cdot \frac{P(\theta_0 = 0 \mid B)}{P(B)} \]

Thus

\[ P(\gamma_1 = 1 \mid B) = P(\theta_0 = 0 \mid B) \cdot P(\gamma_1 = 1 \mid \theta_0 = 0, B) \]

\[ + P(\theta_0 = 1 \mid B) \cdot P(\gamma_1 = 1 \mid \theta_0 = 1, B) \]

This is a special case of the Law of Total Probability:

\[ \text{(for a partition of all possibilities)} \]

\[ P(A) = \sum_{i=1}^{k} P(B_i) P(A \mid B_i) \]

If \( \{B_1, \ldots, B_k\} \) form a partition of \( \{\text{all possibilities}\} \),

\[ P(\gamma_1 = 1) = \sum_{i=1}^{k} P(\theta_0 = i) P(\gamma_1 = 1 \mid \theta_0 = i, B) \]

we've just computed the annoying denominator by partitioning over the truth.
\[ P(\gamma, 1 \mid B) = P(\theta = 0 \mid B) \cdot P(\gamma, 1 \mid \theta = 0, B) \\
+ P(\theta = 1 \mid B) \cdot P(\gamma, 1 \mid \theta = 1, B) \]

\[ = (0.99) \left[ 1 - P(\gamma, 0 \mid \theta = 0, B) \right] \\
+ (0.01) \cdot (0.999) \]

\[ = \frac{99}{100} \cdot \frac{6}{1000} + \frac{1}{100} \cdot \frac{999}{10000} = \frac{1593}{100000} \]

\[ P(\theta = 1 \mid \gamma, 1, B) = \frac{\frac{1}{100} \cdot \frac{999}{100000}}{\frac{1593}{100000}} = \frac{999}{1593} = 0.63 \]

Extending the conversation to include the unknown \( \theta \), which in applications of Bayesian learning amounts to partitioning over the truth, is a powerful technique that will come up a number of times in what follows. Bayes's Theorem in odds form is also highly useful.
The Big Picture, \( P = (Q, C) \rightarrow (\theta, D, B) \) revisited

1. I can make the identification of \( Q \) and \( C \) from \( P \) unique by adopting the convention that if \( Y \), given \( I \), have a different \( Q \) and/or \( C \) in mind, you're working on a different problem than I am.

2. In general (unfortunately) the mapping from \( (Q, C) \) to \( (\theta, D, B) \) is not necessarily unique.

In cor.
(If this will typically be true), for example, \( \theta \) and \( D \) are uniquely specified, but different reasonable choices of \( B \) are possible: to obtain \( P(\theta = 1 | B) \) from the medical literature, it was necessary to specify \( P = \{ all u.s. adults similar to Bob in all relevant ways \}; \) I chose \( B_1 = \) (male), \( B_2 = \) (age 28), \( B_3 = \) (gay), \( B_4 = \) (mostly safe sex), but it would be reasonable to also consider \( B_5 = \) (multiple partners) and \( B_6 = \)
Last tested @ 11 months ago if data were available on those variables as well. The next step in the model-building was to go from to In C54 $p(\theta | B) = \sum_p(\theta = 1 | B)$ and $p(D | \theta, B) = \begin{cases} p(\gamma_i = 1 | \theta = 1, B) & 1 \\ 0 & 0 \end{cases}$ This is the prior information called the sampling distribution ("sample") because it specifies how the data $D = \{\gamma_i\}$ is likely to come out if $\theta$ were known; in C52 we'll see how the sampling distribution is converted (function) into the likelihood information (distribution) $l(\theta | D, B)$. So the general paradigm is $P = (\theta, C) \rightarrow (\theta, D, B) \rightarrow \mathcal{M}_{\theta} \triangleq \{ p(\theta | B), p(D | \theta, B) \}$

Here $\mathcal{M}_{\theta}$ stands for Inference (drawing conclusions about the unknown $\theta$) and Prediction (estimating...
You need to go beyond science (inference, prediction) to make a choice (decision-making). The paradigm becomes "we'll concentrate on inference, prediction in this course:

\[ P = (\theta, C) \rightarrow (\theta, D, B) \rightarrow \{ \pi(\theta \mid B), p(D \mid \theta, B) \}\]

Statistical data science encompasses 4 main activities: 1. Description (data curation) (graphical and numerical) of existing datasets 2. Inference (drawing conclusions about the underlying scientific process that gave rise to the data); 3. Prediction (drawing inferences about new data \( D^* \)); and 4. Decision (choosing the best action [because you have to make a choice] even though you have uncertainty about relevant quantities \( \theta \)). Good data science
Almost always begins with a (possibly extensive) graphical & numerical descriptive exploration of the data set D, with a particular focus on missing data; more on this toward the end of the class.

In CS4 we had uncertainty about how to specify the prior information \( p(\theta | B) \); this will often be true in real-world applications. In CS4 we didn't have any uncertainty about the sampling distribution (because the sensitivity & specificity of the blood test were "known"); in more complicated problems, even there, the 0.999± sensitivity & typically 0.994± specificity were estimated from data.
also have uncertainty about $p(D|\theta B)$.

Thus in general you will have 2 levels of uncertainty: you're uncertain about ($\theta$ but you're also uncertain about $\lambda$.

(How to specify your uncertainty about $\theta$ through $B$, $p(\theta|B)$, and $p(D|\theta B)$; i.e., the mapping from $(\theta, \lambda, B) \rightarrow M_p = \{p(\theta|B), p(D|\theta B)\}$ is also typically not unique.

Level 2 uncertainty is called (naturally enough) model uncertainty; it has been systematically studied in detail since the 1990s. [I will offer some advice on how to cope with model uncertainty in this course (Draper (1995)].
Advice: A simple approach to assessing

The magnitude of model uncertainty is

A sensitivity analysis which varies the aspect

of the modeling about which you care the most about. Are the models robust to

difference in results that you

care the most about. Let $\alpha$ = prevalence

Specificity

Symbolically

The false positive rate

is $FP_R = \frac{\alpha (1-\beta)}{(1-\beta)/(1-\beta) + \alpha (1-\beta)}$. Now we

The false negative rate

is $FN_R = \frac{\alpha (1-\beta)}{(1-\beta)/(1-\beta) + \alpha (1-\beta)}$. Now we
can (eg.) hold $\beta$ at 0.999 and $\delta$ at 0.994 and vary $\alpha = \Pr(\theta = 1 | \beta)$ from (say) 0.005 to 0.02 (a factor of 2 lower & higher than the previous values of 0.01).

$(\beta = 0.999, \delta = 0.994)$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>FPR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>0.544</td>
<td>0.00000506</td>
</tr>
<tr>
<td>0.01</td>
<td>0.373</td>
<td>0.0000182</td>
</tr>
<tr>
<td>0.02</td>
<td>0.227</td>
<td>0.0000205</td>
</tr>
</tbody>
</table>

cutting the prevalence in half increases FPR by $\left| \frac{0.544 - 0.373}{0.373} \right| = 46\%$; doubling the prevalence decreases FPR by $\left| \frac{0.227 - 0.373}{0.373} \right| = 39\%$. (i.e., the false positive rate is quite sensitive to prevalence (prior information))

Cutting $\alpha$ in half almost exactly cuts the FNR in half, and doubling $\alpha$ almost exactly doubles FNR, so the false negative rate is also quite sensitive to prevalence (although its value remains extremely low).