method 3) (the hard way) ("extending the conversation") CS1 we know $P(\theta = 1 \mid \mathcal{B})$,

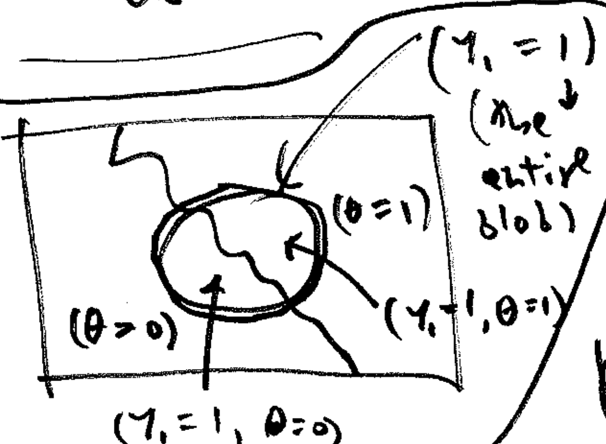
$P(y_1 = 1 \mid \theta_* = 1, \cancel{\mathcal{B}})$ and $P(y_1 = 0 \mid \theta_* = 0, \cancel{\mathcal{B}})$;
↖ or past blood test information in $\mathcal{B}$ ↗

we want $P(\theta = 1 \mid y_1 = 1, \mathcal{B}) = \dfrac{P(\theta = 1 \mid \mathcal{B})\, P(y_1 = 1 \mid \theta = 1, \mathcal{B})}{P(y_1 = 1 \mid \mathcal{B})}$

Both of the numerator probabilities are known; but what about the annoying denominator?

Q:[13] How can we get $P(y_1 = 1 \mid \mathcal{B})$ from $P(\theta = 1 \mid \mathcal{B})$, $P(y_1 = 1 \mid \theta = 1, \cancel{\mathcal{B}})$ and $P(y_1 = 0 \mid \theta = 0, \cancel{\mathcal{B}})$?

A:[13] Notice that $P(y_1 = 1 \mid \mathcal{B})$ is hard, but $P(y_1 = 1 \mid \theta = 1)$ and $P(y_1 = 0 \mid \theta = 0)$ are easy; in other words, we don't know how the data (the blood test) will come out but we do know how the data will come out probabilistically if we knew the truth $\{P(y_1 = 1 \mid \theta = 1), P(y_1 = 0 \mid \theta = 0)\}$. So let's extend the conversation (D.V. Lindley: 1923–2013) by bringing θ into the picture. Since the only two possibilities **Step 1** 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



$$P(y_1 = 1 \mid \mathcal{B}) = P(y_1 = 1 \overset{\text{and}}{,} \theta = 0 \mid \mathcal{B}) + P(y_1 = 1, \theta = 1 \mid \mathcal{B})$$

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

Step 2 θ is on the wrong side of the conditioning bar in $P(y_1 = 1, \theta = 0 \mid \mathcal{B})$ to be useful to us, so let's force it to move to the other side: as before, $P(y_1 = 1, \theta = 0 \mid \mathcal{B}) = P(y_1 = 1 \mid \theta = 0, \mathcal{B}) \pm \boxed{?}$

(1) doesn't go anywhere useful, so let's try $P(y_1 = 1, \theta = 0 \mid \mathcal{B}) = P(y_1 = 1 \mid \theta = 0, \mathcal{B}) \cdot \boxed{?}$

$$P(Y_1=1, \theta=0 \mid \mathcal{B}) = P(Y_1=1 \mid \theta=0, \mathcal{B}) \cdot \boxed{?}$$

(definition of conditional probability)

$$\frac{P(Y_1=1, \theta=0, \mathcal{B})}{P(\mathcal{B})} = \frac{P(Y_1=1, \theta=0, \mathcal{B})}{P(\theta=0, \mathcal{B})} \cdot \boxed{\frac{P(\theta=0, \mathcal{B})}{P(\mathcal{B})}}$$

So it works, and the answer is

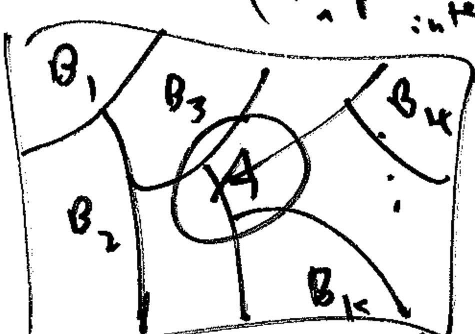
$$P(Y_1=1, \theta=0 \mid \mathcal{B}) = P(Y_1=1 \mid \theta=0, \mathcal{B}) \cdot \boxed{P(\theta=0 \mid \mathcal{B})}$$

Thus

$$\boxed{\begin{aligned} P(Y_1=1 \mid \mathcal{B}) = {} & P(\theta=0 \mid \mathcal{B}) \cdot P(Y_1=1 \mid \theta=0, \mathcal{B}) \\ & + P(\theta=1 \mid \mathcal{B}) \cdot P(Y_1=1 \mid \theta=1, \mathcal{B}) \end{aligned}}$$

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

(k a positive integer)



If $\{B_1, \ldots, B_k\}$ form a partition of {all possibilities}, then

$$P(A) = \sum_{i=1}^{k} P(B_i)\, P(A \mid B_i)$$

($P(B_i) > 0$ by definition of a partition)

we've just computed the annoying denominator by partitioning over the truth.

$$P(Y_1 = 1 \mid \mathcal{B}) = P(\theta = 0 \mid \mathcal{B}) \cdot P(Y_1 = 1 \mid \theta = 0, \mathcal{B}) + P(\theta = 1 \mid \mathcal{B}) \cdot P(Y_1 = 1 \mid \theta = 1, \mathcal{B})$$

$$= \underbrace{(0.99)}_{\text{prevalence}} \Big[1 - \underbrace{\boxed{P(Y_1 = 0 \mid \theta = 0, \mathcal{B})}}_{\text{specificity}} \Big] + \underbrace{(0.01)}_{\text{prevalence}} \cdot \underbrace{(0.999)}_{\text{sensitivity}}$$

$$= \frac{99}{100} \cdot \frac{6}{1000} + \frac{1}{100} \cdot \frac{999}{1000} = \frac{1593}{100000}$$ and

finally $$P(\theta = 1 \mid Y_1 = 1, \mathcal{B}) = \frac{\frac{1}{100} \cdot \frac{999}{100000}}{\frac{1593}{100000}} = \frac{999}{1593} \doteq 0.63$$ as before

Extending the conversation to include the unknown θ, 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, Revisited | $\mathbb{P} \underset{①}{=} (\mathbb{Q}, \mathbb{C}) \underset{②}{\to} (\theta, \mathcal{D}, \mathcal{B})$

① I can make the identification of $\mathbb{Q}$ and $\mathbb{C}$ from $\mathbb{P}$ unique by adopting the convention that if You, given $\mathbb{P}$, have a different $\mathbb{Q}$ and/or $\mathbb{C}$ in mind, You're working on a different problem than I am. ② In general (unfortunately) the mapping from $(\mathbb{Q}, \mathbb{C})$ to $(\theta, \mathcal{D}, \mathcal{B})$ is not necessarily unique. In CS1, for example, θ and $\mathcal{D}$ are uniquely specified ~~(this will typically be true)~~, but different reasonable choices of $\mathcal{B}$ are possible: to obtain $p(\theta = 1 \mid \mathcal{B})$ from the medical literature, it was necessary to specify $\mathcal{P}_{Bob} = \{$all U.S. adults similar to Bob in all relevant ways$\}$; I chose $\mathcal{B}_1 =$ (male), $\mathcal{B}_2 =$ (age 28), $\mathcal{B}_3 =$ (gay), $\mathcal{B}_4 =$ (mostly safe sex), but it would be reasonable to also consider, e.g., $\mathcal{B}_5 =$ (multiple partners) and $\mathcal{B}_6 =$

(last tested $\ominus$ 11 months ago) if population data were available on these variables as well. The next step in the model-building was to go from $(\theta, D, \mathcal{B})$ to $\{ p(\theta \mid \mathcal{B}), p(D \mid \theta \mathcal{B}) \}$.

In CS1 $p(\theta \mid \mathcal{B}) = \begin{cases} p(\theta = 1 \mid \mathcal{B}) \\ p(\theta = 0 \mid \mathcal{B}) \end{cases}$ this is the *prior* *information*

and $p(D \mid \theta \mathcal{B}) = \left\{ \begin{array}{cc} p(y_1 = 1 & \mid \theta = 1, \mathcal{B}) \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{array} \right\}$ ("sample")

this is called the *sampling* *distribution* because it specifies how the data $D = (y_1)$ is likely to come out if θ were known; in CS2 we'll see how the sampling distribution is converted into the likelihood *information* (function) (distribution) $\ell(\theta \mid D \mathcal{B})$. So the general paradigm so far is

$$\mathbb{P} = (\mathbb{Q}, \mathbb{C}) \to (\theta, D, \mathcal{B}) \to \mathbb{M}_{IP} \triangleq \{ p(\theta \mid \mathcal{B}), p(D \mid \theta \mathcal{B}) \}$$

Here IP stands for *Inference* (drawing conclusions about the unknown θ) and *Prediction* (estimating new (not yet observed) data values D^*). / If

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

$$\mathbb{P} = (Q, C) \to (\theta, D, \mathcal{B}) \to \mathcal{M}_{IPD} = \{ p(\theta \mid \mathcal{B}),\ p(D \mid \theta\,\mathcal{B}),\ (\mathcal{A} \mid \mathcal{B}),\ U(a, \theta \mid \mathcal{B}) \}$$

$(\mathcal{A} \mid C)$

(data wrangling)

Statistical data science encompasses 4 main activities: ① Description (graphical and numerical) (data curation) of existing data sets D; ② inference (drawing conclusions about the underlying scientific process (θ) 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 (θ)).

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 CS2 we had uncertainty about how to specify the prior information $p(\theta \mid \mathcal{B})$; this will often be true in real-world applications. In CS1 we didn't have any uncertainty about the sampling distribution $p(D \mid \theta\, \mathcal{B})$ (because the sensitivity & specificity of the blood test were "known"); in more complicated problems, even there, the 0.999 ± ? sensitivity & 0.994 ± ? specificity were estimated from data. You will typically also have uncertainty about $p(D \mid \theta \mathcal{B})$.

Thus in general You will have 2 levels of uncertainty: You're uncertain about (level 1) θ, but You're also uncertain about (how to specify Your uncertainty about (level 2) θ, through $\mathcal{B}$, $p(\theta \mid \mathcal{B})$, and $p(D \mid \theta \mathcal{B})$); i.e., the mapping from $(\theta, D, \mathcal{B})$ to $M_{IP} = \{ p(\theta \mid \mathcal{B}), p(D \mid \theta \mathcal{B}) \}$ is also typically not unique.

$\boxed{\text{Level 2 uncertainty is called (naturally enough) model uncertainty; it has been systematically studied in detail since the 1990s (Draper (1995)).}}$

I will offer some ~~in this course~~ advice on how to cope with model uncertainty.

Advice: A simple approach to assessing the magnitude of model uncertainty is <u>sensitivity analysis</u>: Vary the aspect of the modeling about which You're uncertain across a plausible range & see how much difference it makes to the results that You care the most about.

CS1: Let α = prevalence, β = sensitivity, γ = specificity

truth

blood test	HIV +	HIV −	
+	$\alpha\beta$	$(1-\alpha)(1-\gamma)$	$\alpha\beta + (1-\alpha)(1-\gamma)$
−	$\alpha(1-\beta)$	$(1-\alpha)\gamma$	$\alpha(1-\beta) + (1-\alpha)\gamma$
	α	$1-\alpha$	1

Symbolically the false positive rate is

$$FPR = \frac{(1-\alpha)(1-\gamma)}{\alpha\beta + (1-\alpha)(1-\gamma)}$$

and the false negative rate is

$$FNR = \frac{\alpha(1-\beta)}{\alpha(1-\beta) + \gamma(1-\alpha)}.$$

Now we

can (e.g.) hold β at 0.999 and γ at 0.994 and vary $\alpha = P(\theta = 1 \mid \mathcal{B})$ from (say) 0.005 to 0.02 (a factor of 2 lower & higher than the previous value of 0.01): 39

($\beta = 0.999$, $\gamma = 0.994$)		
α	FPR	FNR
0.005	0.544	0.00000506
0.01	0.373	0.0000102
0.02	0.227	0.0000205

cutting the prevalence in half increases FPR by $\left| \frac{.544 - .373}{.373} \right| \cdot 100\% \doteq 46\%$; doubling the prevalence decreases FPR by $\left| \frac{.227 - .373}{.373} \right| \cdot 100\% \doteq 39\%$. (15 Jun 19) (i.e., the false positive rate is quite sensitive to prevalence (prior information)).

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