this time:	Bayes's theorem
next time:	random variables

STAT 131
16 Apr 20

for A any event (true/false proposition), ① let $P(A) = p_A$; then

the odds ratio in favor of A being true is defined to be $o_A = \frac{P(A)}{P(\text{not } A)} = \frac{p_A}{1 - p_A}$

odds ratio in terms of probability

You can readily solve backwards to get $p_A = \frac{o_A}{1 + o_A}$ = probability in terms of odds ratio

p_A	o_A
0.01	$\frac{1}{99}$
0.1	$\frac{1}{9}$
0.5	$\frac{1}{1}$
0.9	$\frac{9}{1}$
0.99	99/1

odds ratios are used by people gambling on things like horse races, but they're also useful in medical research.

③

$$P(C \mid E) \stackrel{?}{=} P(E \mid C)$$

$$\frac{P(C \text{ and } E)}{P(E)} \qquad \frac{P(E \text{ and } C)}{P(C)}$$

$P(E) > 0$

$P(C) > 0$

$$P(C \mid E) = \frac{P(C \text{ and } E)}{P(E)}$$

①

$$\boxed{P(C \text{ and } E)} = \underline{P(E)\,P(C \mid E)}$$

$$P(E \mid C) = \frac{P(E \text{ and } C)}{P(C)}$$

②

$$\boxed{P(E \text{ and } C)} = \underline{P(C)\,P(E \mid C)}$$

$$P(E)\,P(C \mid E) = P(C)\,P(E \mid C)$$

$$P(C \mid E) = \frac{P(C)\,P(E \mid C)}{P(E)}$$

Bayes's theorem for single events (T/F propositions)

Case study: screening for COVID-19

truth:	C — really is COVID-19 positive	not C — really not
what collex test says:	+ — test thinks COVID-19 positive	− — negative

truth	test says
not C	−
not C	+
⋮	⋮

1,000,000

	truth: C	not C
test says +		
−		

1000000

have to know 3 things: $P(C) = 0.01$

1. prevalence of COVID-19 in U.S. pop.: case 1: $\frac{64{,}166}{328{,}200{,}000} \doteq 0.2\%$

   case 2: 1.0%

2. $P(+ \mid C) = 93.8\%$ = sensitivity of Cellex test

3. $P(- \mid \text{not } C)$ = specificity of test = 95.6%

culprit

method I

sens.

false positive

case 2

truth

Cellex test says	C	not C	
+	9,380	43,560	52,940
−	620	946,440	947,060
	10,000	990,000	1,000,000

false negative

$P(C) = 1\%$

$P(\text{not } C \mid -) = \frac{946{,}440}{947{,}060} \doteq .9993$

$P(C \mid +) = \frac{9{,}380}{52{,}940} = 0.177 \doteq 18\%$ (!)

antibody concentration →

cellex says ⊖ | $\boxed{c}$ | cellex says ⊕

case 2

cellex test says	truth: c	not c	
⊕	1,876	43,912	45,788
⊖	124	954,088	954,212
	2000	99,800	1,000,000

$P(c) = 0.002$

$$P(c \mid +) = \frac{1{,}876}{45{,}788} = 0.04\,(!)$$

$$P(\text{not } c \mid -) = \frac{954{,}088}{954{,}212} = .99987$$

case	$P(c \mid +)$	$P(\text{not } c \mid -)$	prevalence
1	4%	99.99%	0.2%
2	18%	99.93%	1.0%

Cellex test almost perfect when ⊖

but it's cheap when it says (+)

truth

cellex test says	c	not c
(+)	true positive	false positive (bad)
(−)	false negative (bad)	true negative

person told (+) but really (−) (bad)

person told (−) but really (+) | much worse

decision theory

action space $\mathcal{A}$

$$\mathcal{A} = \{ a_1, \; a_2, \; a_3 \}$$

a_3: cellex; if (−) conclude not c; if (+), gold standard & believe

a_1: give cellex (quicker, cheaper) to everybody & pretend it's right

a_2: give gold standard (slower, more costly) to everybody & pretend it's right know:

$P(C)$

$P(+ \mid C)$ = sens.

$P(- \mid \text{not } C)$ = spec.

want

$\boxed{P(C \mid +)}$

$P(\text{not } C \mid -)$

$P(A \mid B) \stackrel{?}{=} P(B \mid A)$ no

$P(\text{rain} \mid \text{clouds overhead})$ = low to in between

$P(\text{clouds overhead} \mid \text{rain})$ = high

prevalence sensitivity

$$P(C \mid +) = \frac{P(C)\,P(+ \mid C)}{\boxed{P(+)}} \qquad (1)$$

(annoying) normalizing constant

method II

$$p(\text{not } C \mid +) = \frac{p(\text{not } C)\, p(+ \mid \text{not } C)}{p(+)} \quad (2)$$

$$\frac{(1)}{(2)} = \frac{p(C \mid +)}{p(\text{not } C \mid +)} = \left[\frac{p(C)}{p(\text{not } C)}\right]\left[\frac{p(+ \mid C)}{p(+ \mid \text{not } C)}\right]$$

(likelihood ratio) → Bayes factor

(posterior odds ratio in favor of C given data) = (prior odds ratio in favor of C) (Bayes factor)

posterior information — prior information — (likelihood information)

$$\boxed{p(C \mid +)} = \frac{\boxed{p(C)}\ \boxed{p(+ \mid C)}}{p(+)}$$

data — unknown — (normalizing constant)

(before data) (a priori)

(after data) (a posteriori)

$$\begin{pmatrix}\text{posterior}\\ \text{info}\end{pmatrix} = \frac{\begin{pmatrix}\text{prior}\\ \text{info}\end{pmatrix}\cdot\begin{pmatrix}\text{likelihood}\\ \text{info}\end{pmatrix}}{\begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix}}$$

$$p(\underline{\text{unknown}} \mid \text{data}) = \frac{p(\underline{\text{unknown}})\cdot \underline{p(\text{data} \mid \text{unknown})}}{\begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix}}$$

prevalance (case 2) sensitivity

$$\frac{P(C \mid +)}{P(\text{not } C \mid +)} = \left(\frac{.01}{.99}\right)\left(\frac{0.938}{1-.956}\right)$$

10

1 − specificity

$$\boxed{P(+ \mid \text{not } C) = 1 - P(- \mid \text{not } C)}$$

$$\left(\frac{1}{99}\right)\left(\frac{469}{22}\right)$$