CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 2

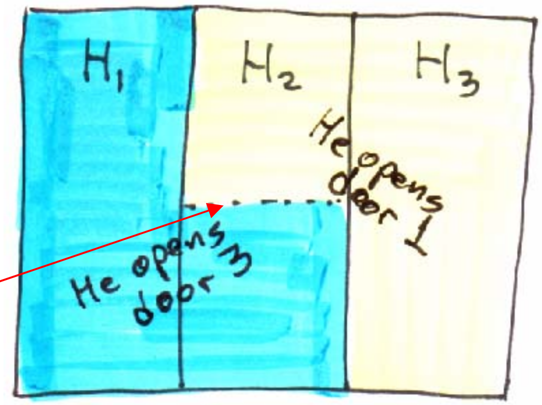
Example: The Monty Hall or Let's Make a Deal Problem

- Three doors
- Car (prize) behind one door
- You pick a door, but don't open it yet
- Monty then opens one of the other doors, always revealing no car (he knows where it is)
- You now get to switch doors if you want
- Should you?
- Most people reason: Two remaining doors were equiprobable before, and nothing has changed. So doesn't matter whether you switch or not.
- Marilyn vos Savant ("highest IQ person in the world") famously thought otherwise (Parade magazine, 1990)
- No one seems to know or care what Monty Hall thought!
  - he is alive at age 89
  - his daughter is Joanna Gleason, who starred in Sondheim's "Into the Woods"

$H_i =$ car behind door i, $i = 1, 2, 3$
Wlog, you pick door 2 (relabeling).
Wlog, Monty opens door 3 (relabeling).
$P(H_i|O3) \propto P(O3|H_i)P(H_i)$

"Without loss of generality…"

$$P(H_1|O3) \propto 1 \cdot \tfrac{1}{3} = \tfrac{2}{6}$$
$$P(H_2|O3) \propto \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6}$$
$$P(H_3|O3) \propto 0 \cdot \tfrac{1}{3} = 0$$



ignorance of Monty's preference between 1 and 3, so take 1/2

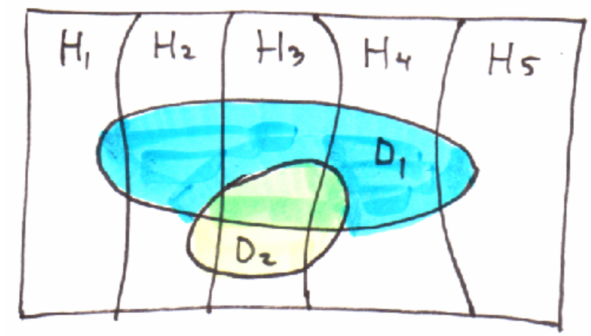
So you should always switch: doubles your chances!

Exegesis on Monty Hall

- Very important example! Master it.
- $P(H_i) = \frac{1}{3}$ is the "prior probability" or "prior"
- $P(H_i|O3)$ is the "posterior probability" or "posterior"
- $P(O3|H_i)$ is the "evidence factor" or "evidence"
- Bayes says posterior $\propto$ evidence $\times$ prior

Commutivity/Associativity of Evidence

$P(H_i|D_1D_2)$ desired



We see D_1:
$P(H_i|D_1) \propto P(D_1|H_i)P(H_i)$

Then, we see D_2:
$P(H_i|D_1D_2) \propto P(D_2|H_iD_1)P(H_i|D_1)$ ← this is now a prior!

But,
$= P(D_2|H_iD_1)P(D_1|H_i)P(H_i)$
$= P(D_1D_2|H_i)P(H_i)$

this being symmetrical shows that we would get the same answer regardless of the order of seeing the data

All priors $P(H_i)$ are actually $P(H_i|D)$, conditioned on previously seen data! Often write this as $P(H_i|I)$. ← background information

Bayes Law is a "calculus of inference", better (and certainly more self-consistent) than folk wisdom.

Example: Hempel's Paradox

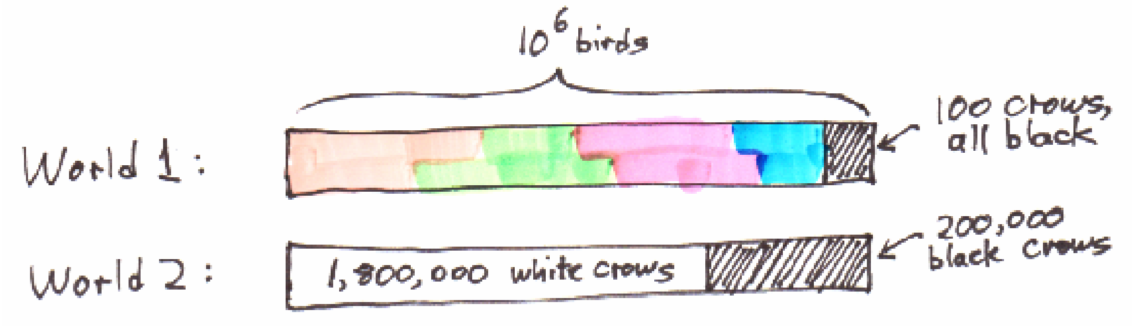
Folk wisdom: A case of a hypothesis adds support to that hypothesis.

Example: "All crows are black" is supported by each new observation of a black crow.

All crows are black $\Leftrightarrow$ All non-black things are non-crows

But this is supported by the observation of a white shoe.

So, the observation of a white shoe is thus evidence that all crows are black!



I.J. Good: “The White Shoe is a Red Herring” (1966)

We observe one bird, and it is a black crow.
a) Which world are we in?
b) Are all crows black?

Important concept, Bayes odds ratio:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}$$

$$= \frac{0.0001\,P(H_1)}{0.1\,P(H_2)} = 0.001\frac{P(H_1)}{P(H_2)}$$

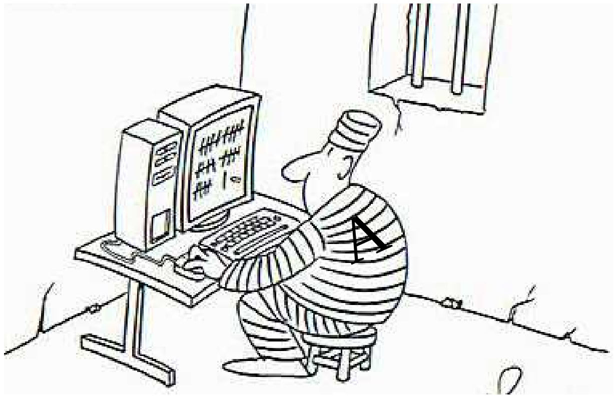
So the observation strongly supports H2 and the existence of white crows.

Hempel’s folk wisdom premise is not true.

Data supports the hypotheses in which it is more likely compared with other hypotheses. (This is Bayes!)

We must have <u>some</u> kind of background information about the universe of hypotheses, otherwise data has no meaning at all.

Our next topic is Bayesian Estimation of Parameters. We'll ease into it with an example that looks a lot like the Monte Hall Problem:



The Jailer's Tip:

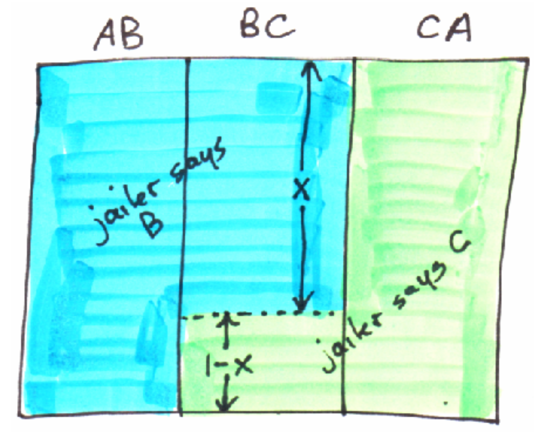
- Of 3 prisoners (A,B,C), 2 will be released tomorrow.
- A, who thinks he has a 2/3 chance of being released, asks jailer for name of one of the lucky – but not himself.
- Jailer says, truthfully, "B".
- "Darn," thinks A, "now my chances are only ½, C or me".

Is this like Monty Hall? Did the data ("B") change the probabilities?

Further, suppose (unlike Monty Hall) the jailer is not indifferent about responding "B" versus "C". Does that change your answer to the previous question?

$$P(S_B|BC) = x, \quad (0 \le x \le 1)$$

"says B"



$$P(A|S_B) = P(AB|S_B) + \overset{0}{\cancel{P(AC|S_B)}}$$

$$= \frac{\overset{1}{\cancel{P(S_B|AB)}}\,\overset{1/3}{\cancel{P(AB)}}}{P(S_B|AB)P(AB) + \underset{x}{\cancel{P(S_B|BC)}}P(BC) + P(S_B|CA)P(CA)}$$

$$= \frac{\frac{1}{3}}{1 \cdot \frac{1}{3} + x \cdot \frac{1}{3} + 0} = \frac{1}{1+x}$$

So if A knows the value x, he can calculate his chances.

If x=1/2 (like Monty Hall), his chances are 2/3, same as before; so (unlike Monty Hall) he got no new information.

If $x\neq$1/2, he does get new info – his chances change.

But what if he doesn't know x at all?

“Marginalization” (this is important!)

- When a model has unknown, or uninteresting, parameters we “integrate them out” …
- …multiplying by any knowledge of their distribution
  - At worst, just a prior informed by background information
  - At best, a narrower distribution based on data
- This is not any new assumption about the world
  - it’s just the Law of de-Anding

(e.g., Jailer’s Tip):

law of de-Anding:
x’s are EME!

$$
\begin{aligned}
P(A|S_B I) &= \int_x P(A|S_B x I)\, p(x|I)\, dx \\
&= \int_x \frac{1}{1+x}\, p(x|I)\, dx
\end{aligned}
$$

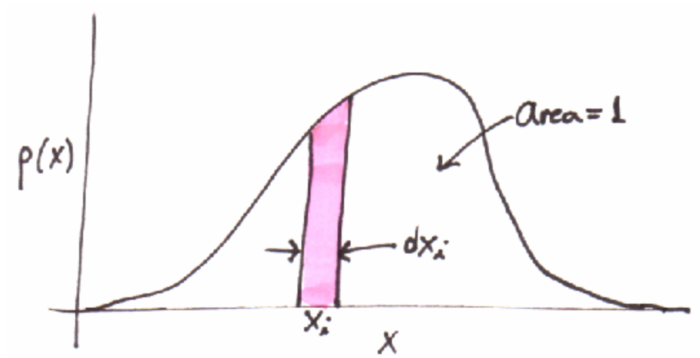
(repeating previous equation:)

$$P(A|S_B I) = \int_x P(A|S_B x I)\, p(x|I)\, dx$$
$$= \int_x \frac{1}{1+x}\, p(x|I)\, dx$$

first time we've seen a *continuous* probability distribution, but we'll skip the obvious repetition of all the previous laws

$$p(x) \equiv p(x|I)$$

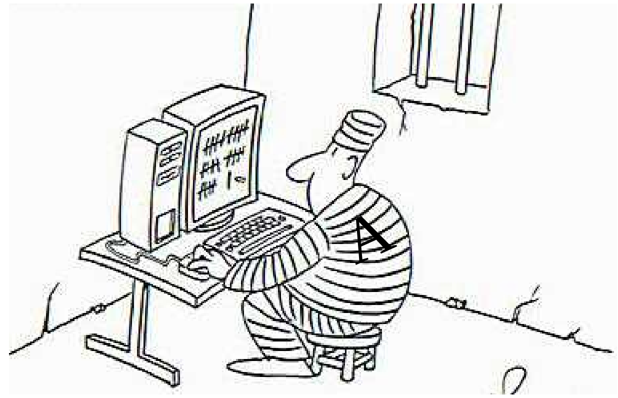
(Notice that p(x) is a probability of a probability! That is fairly common in Bayesian inference.)



$$\sum_i P_i = 1 \;\leftrightarrow\; \sum_i p(x_i) dx_i = 1 \;\leftrightarrow\; \int_x p(x)\, dx = 1$$

(repeating previous equation:)

$$P(A|S_B I) = \int_x P(A|S_B x I)\, p(x|I)\, dx$$
$$= \int_x \frac{1}{1+x}\, p(x|I)\, dx$$



What should Prisoner A take for *p(x)* ?
Maybe the “uniform prior”?

$$p(x) = 1, \quad (0 \le x \le 1)$$
$$P(A|S_B I) = \int_0^1 \frac{1}{1+x} dx = \ln 2 = 0.693$$

Not the same as the “massed prior at x=1/2”

“Dirac delta function”

$$p(x) = \delta(x - \tfrac{1}{2}), \quad (0 \le x \le 1)$$
$$P(A|S_B I) = \frac{1}{1+1/2} = 2/3$$

substitute value and
remove integral

Review where we are:

$$P(A|S_B I) = \int_x P(A|S_B x I)\, p(x|I)\, dx$$
$$= \int_x \frac{1}{1+x}\, p(x|I)\, dx$$

We are trying to estimate a parameter

$$x = P(S_B|BC), \quad (0 \le x \le 1)$$

The form of our estimate is a (Bayesian) probability distribution (of the parameter, itself here just happening to be a probability)

This is a sterile exercise if it is just a debate about priors. What we need is data! Data might be a previous history of choices by the jailer in identical circumstances.

BCBCCBCCCBBCBCBCCCCBBCBCCCBCBCBBCCB

$$N = 35, \quad N_B = 15, \quad N_C = 20$$

(What's wrong with: x=15/35=0.43? Hold on…)

We hypothesize (might later try to check) that these are i.i.d. "Bernoulli trials" and therefore informative about *x*

"independent and identically distributed"

As good Bayesians, we now need $P(\text{data}|x)$

$$P(\text{data}|x) \quad \left\{ \begin{array}{l} \text{means different things in frequentist vs. Bayesian contexts,} \\ \text{so this is a good time to understand the differences (we'll use} \\ \text{both ideas as appropriate)} \end{array} \right.$$

Frequentist considers the universe of what might have been, imagining repeated trials, even if they weren't actually tried, and needs <u>no prior</u>:

since i.i.d. only the $\mathcal{N}$'s can matter (a so-called "sufficient statistic").

$$P(\text{data}|x) = \underbrace{\binom{N}{N_B}}_{\text{no. of equivalent arrangements}} \overbrace{x^{N_B}(1-x)^{N_C}}^{\text{prob. of exact sequence seen}} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Bayesian considers only the <u>exact</u> data seen, <u>and has a prior</u>:

$$P(x|\text{data}) \propto x^{N_B}(1-x)^{N_C}\, p(x|I)$$

but we might first suppose that the prior it is **uniform**

No binomial coefficient, both conceptually and also since independent of x and absorbed in the proportionality. Use only the data you see, not "equivalent arrangements" that you didn't see. This issue is one we'll return to, not always entirely sympathetically to Bayesians (e.g., goodness-of-fit).