STSCI 4780:
Inference in discrete spaces

Tom Loredo, CCAPS & SDS, Cornell University

Recap: Bayesian inference in one slide

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem — Accounting for new info (learning)

$$p(\text{Hypothesis} \mid \text{Data}) \quad \propto \quad p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis $\propto$ ability of hypothesis to *predict* the data

Law of total probability — Decomposing a hypothesis

$$p(\text{Hypothese}\underline{\textbf{s}} \mid \text{Data}) \quad = \quad \sum p(\text{Hypothesi}\underline{\textbf{s}} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Contextual/prior/background information

Bayes's theorem moves the data and hypothesis propositions wrt the solidus:

$$P(H_i|D_{\text{obs}}, \mathcal{C}) = P(H_i|\mathcal{C}) \frac{P(D_{\text{obs}}|H_i, \mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})}$$

It lets us account for new facts by *augmenting the premises*

"Context" or "prior information" or "background information" = information that is **always** a premise (for the current calculation)

Notation: $P(\cdot|\cdot, \mathcal{C})$ or $P(\cdot|\cdot, I)$ or $P(\cdot|\cdot, M)$ or ...

The context can be a notational nuisance! "Skilling conditional":

$$P(H_i|D_{\text{obs}}) = P(H_i) \frac{P(D_{\text{obs}}|H_i)}{P(D_{\text{obs}})} \qquad || \mathcal{C}$$

Often just drop $\mathcal{C}$—but be careful if it gets altered!

Essential contextual information

We can only be uncertain about a proposition, A, if there are alternatives (at least $\overline{A}$!); what they are will bear on our uncertainty. *We must explicitly specify relevant alternatives.*

Hypothesis space: The set of alternative hypotheses of interest (and auxiliary hypotheses needed to predict the data, e.g., for LTP)

Data/sample space: The set of possible data we may have predicted before learning of the observed data

Predictive model: Information specifying the likelihood function (e.g., the conditional predictive dist'n/sampling dist'n)—the connection between data and hypotheses

Other prior information: Any further information available or necessary to assume to make the problem *well posed*

Bayesian literature often uses **model** to refer to *all* of the contextual information used to study a particular dataset and predictive model

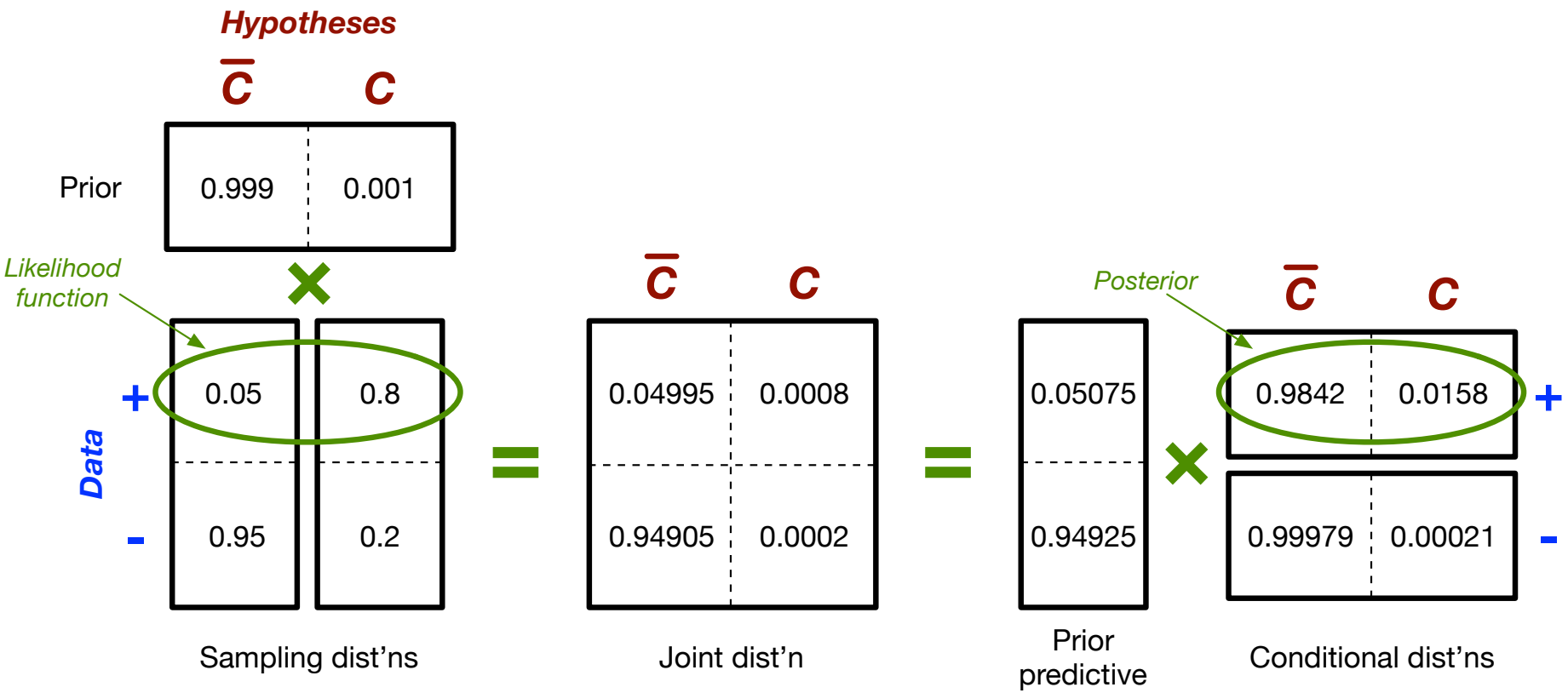
Recap: Binary hypotheses and data

Post-test probabilities after positive test result

Hypothesis H_i	Prior $\pi_i \equiv p(H_i)$	Likelihood $\mathcal{L}_i \equiv p(+\|H_i)$	Joint $\pi_i \times \mathcal{L}_i$	Posterior $p(H_i\|+)$
$\overline{C}$	0.999	0.05	0.04995	0.9842
C	0.001	0.8	0.0008	0.0158
Sums:	1.0	**NA**	0.05075 $= p(+)$	1.0

Bayes's theorem in terms of the *joint distribution*:

$$P(H_i|\mathcal{C}) \times P(D_{\text{obs}}|H_i, \mathcal{C}) = P(H_i, D_{\text{obs}}|\mathcal{C}) = P(H_i|D_{\text{obs}}, \mathcal{C}) \times P(D_{\text{obs}}|\mathcal{C})$$



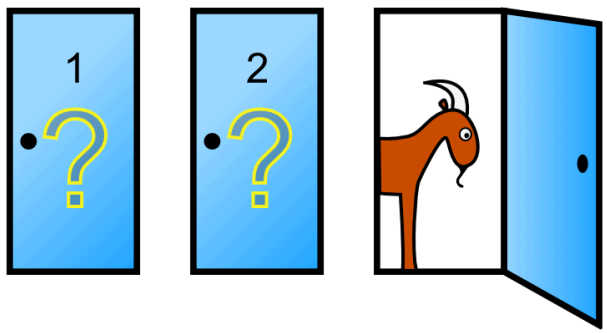
Bayesian conditioning

- Build a Bayesian probability model: A (joint) probability distribution over possible pre-data states of the world labeled by (H, D) pairs
- Focus on the slice with $D = D_{\text{obs}}$
- Normalize that slice via BT:

$$P(H_i|D_{\text{obs}}) = \frac{P(H_i, D_{\text{obs}})}{P(D_{\text{obs}})} = \frac{P(H_i, D_{\text{obs}})}{\sum_i P(H_i, D_{\text{obs}})} = \frac{P(H_i)P(D_{\text{obs}}|H_i)}{\sum_i P(H_i)P(D_{\text{obs}}|H_i)}$$

Monty Hall problem: 3 hypotheses

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then asks you, "Do you want to switch to door No. 2?" Is it to your advantage to switch?



Wikipedia

Important rules:

- The prize door is set randomly (with equal probability)
- The host will only reveal a goat
- When both unchosen doors hide a goat, the host picks one at random (with equal probability)

Setup

$\mathcal{C}$ = Rules of the game, player's door choice, and...

Hypothesis space: H_i = prize behind door i

Data/sample space: D_j = host opens door j; $D_{\text{obs}} = D_3$

$\Rightarrow$ Compute $P(H_i|D_3, \mathcal{C})$

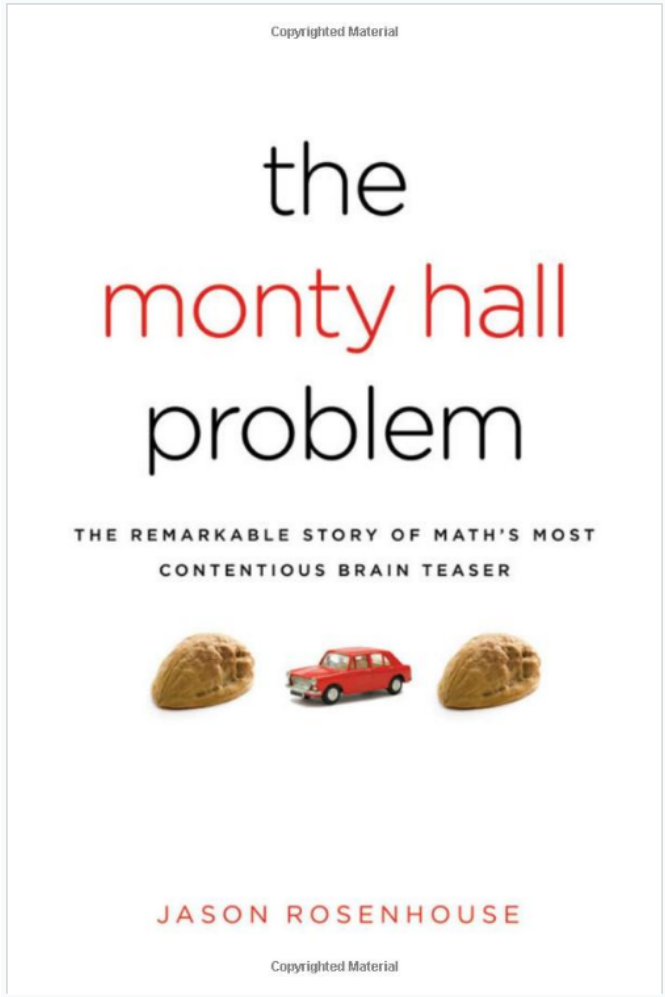
Controversy

Parade columnist Marilyn vos Savant (IQ $>$ 200!) posed the problem in her column and argued that one should switch.

She received $\approx$10,000 letters, the great majority disagreeing with her. The strongest opposition came from mathematicians and scientists.

A common line of intuitive reasoning: The doors have equal probability for hiding the prize. The host has ruled out one of them. The remaining doors now each have probabilty 1/2 of having the prize, so there is no advantage to switching.

See U. Iowa statistician Luke Tierney's summary: "Behind Monty Hall's Doors: Puzzle, Debate and Answer?" (NYT 1991)



The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser, by Jason Rosenhouse (2009)

From Tierney's essay:

> *Persi Diaconis, a former professional magician who is now a Harvard* University professor specializing in probability and statistics, said there was no disgrace in getting this one wrong.*
>
> *"I can't remember what my first reaction to it was," he said, "because I've known about it for so many years. I'm one of the many people who have written papers about it. But I do know that my first reaction has been wrong time after time on similar problems. Our brains are just not wired to do probability problems very well, so I'm not surprised there were mistakes."*

*Diaconis subsequently moved to Cornell, and is now at Stanford.

Whiteboard work!

Is coin flipping biased?

$\mathcal{C}$ specifies existence of two outcomes, S and F, in each of N cases or trials; for each case or trial, the probability for S is α; for F it is $(1-\alpha)$ (*independent, identically distributed* (IID) sampling)

Consider three hypotheses:

- $H_0 : \alpha = \alpha_0 = 1/2$ with $\pi_0 \equiv P(H_0|\mathcal{C}) = 0.9$
- $H_1 : \alpha = \alpha_1 = 0.4$ with $\pi_1 \equiv P(H_1|\mathcal{C}) = 0.05$ (tails bias)
- $H_2 : \alpha = \alpha_2 = 0.6$ with $\pi_2 \equiv P(H_2|\mathcal{C}) = 0.05$ (heads bias)

$D_{\text{obs}} =$ Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood (Bernoulli process)

$$\begin{aligned} P(D|H_i, \mathcal{C}) &= P(\text{failure}|H_i, \mathcal{C}) \times P(\text{failure}|H_i, \mathcal{C}) \times \cdots \\ &= \alpha_i^n (1-\alpha_i)^{N-n} \end{aligned}$$

Note that the result depends only on n, not the details of D: n is a *sufficient statistic*.

Prior Predictive

$$P(D|\mathcal{C}) = \sum_{i=0}^{2} \pi_i \, \alpha_i^n (1-\alpha_i)^{N-n} \equiv Z(n, N)$$

Posterior

$$P(H_i|D, \mathcal{C}) = \frac{\pi_i \, \alpha_i^n (1-\alpha_i)^{N-n}}{Z(n, N)}$$

For $D = D_{\text{obs}}$, plug in $N = 12$, $n = 8$

Coin flips with count data

Suppose datum is now $D' = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, \mathcal{C})$?

Likelihood

Let $\mathcal{S}$ = a sequence of flips with n heads.

$$p(n|\alpha, \mathcal{C}) = \sum_{\mathcal{S}} p(S|\alpha, \mathcal{C})\; p(n|\mathcal{S}, \alpha, \mathcal{C})$$

$\alpha^n\,(1-\alpha)^{N-n}$

$[\![\ \#\ \text{successes} = n]\!]$

$$= \alpha^n (1-\alpha)^{N-n} C_{n,N}$$

$C_{n,N} = \#$ of sequences of length N with n heads.

$$\rightarrow p(n|\alpha, \mathcal{C}) = \frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

The *binomial distribution* for n given α, N.

Prior Predictive

$$\begin{aligned} P(D'|\mathcal{C}) &= \sum_{i=0}^{2} \pi_i \frac{N!}{n!(N-n)!} \alpha_i^n (1-\alpha_i)^{N-n} \\ &= \frac{N!}{n!(N-n)!} \sum_{i=0}^{2} \pi_i \, \alpha_i^n (1-\alpha_i)^{N-n} \\ &= \frac{N!}{n!(N-n)!} \times Z(n,N) \end{aligned}$$

Posterior

$$P(H_i|D',\mathcal{C}) = \frac{\pi_i \, \alpha_i^n (1-\alpha_i)^{N-n}}{Z(n,N)}$$

Same result as when data specified the actual sequence

This is an example of the *likelihood principle* (more in Lec05):

All that matters in $P(D_{obs}|H_i,\mathcal{C})$ is the relative variation as H_i changes

Computation

We'll compute hypothesis table elements (likelihoods, joint probabilities, posterior probabilities) numerically in a Jupyter notebook $\Rightarrow$

$$
\begin{aligned}
P(H_0|D_{\text{obs}}, \mathcal{C}) &= 0.895 \\
P(H_1|D_{\text{obs}}, \mathcal{C}) &= 0.0173 \\
P(H_2|D_{\text{obs}}, \mathcal{C}) &= 0.0876
\end{aligned}
$$