Announcements

- Assignment01 (the Git/GitHub README exercise in the Lab01 slides) will be due *next Monday, 7 Feb*.
- As noted in Lab01, you may work on the assignment *in your own GitHub account* if you wish; you'll be able to copy or transfer your work to the `CU-BDA-2022` org later.
- Office hours: Tom on Tuesdays, 4–5:30pm (starting today); Georgia on Wednesdays, 2:30–4pm; Zoom links forthcoming
- Watch for a Qualtrics survey for your GitHub info (and other computing info) this evening.
- Lab02 will provide a brief introduction to Python and Jupyter notebooks; there will not be a new assignment until Lab03.

STSCI 4780:
Key theorems

Tom Loredo, CCAPS & SDS, Cornell University

Recap: Probability theory as logic

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

$$\begin{aligned} P &= 0 \;\rightarrow\; \text{Argument is \textit{invalid}; premises imply } \overline{H} \\ &= 1 \;\rightarrow\; \text{Argument is \textit{valid}} \\ &\in (0,1) \;\rightarrow\; \text{Degree of implication/deducibility} \end{aligned}$$

Mathematical model for induction

$$\text{‘AND’ (product rule):} \quad \begin{aligned} P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P})\, P(B|A \wedge \mathcal{P}) \\ &= P(B|\mathcal{P})\, P(A|B \wedge \mathcal{P}) \end{aligned}$$

$$\text{‘OR’ (sum rule):} \quad \begin{aligned} P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\ &\quad - P(A \wedge B|\mathcal{P}) \end{aligned}$$

$$\text{‘NOT’:} \quad P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation*.

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful*, none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities*, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Harold Jeffreys (1931)

If we like there is no harm in saying that a probability expresses a degree of reasonable belief. . . . 'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. *Essentially the notion can only be described by reference to instances where it is used.* It is intended to express *a kind of relation between data and consequence* that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

More On Interpretation

Physics uses words drawn from ordinary language—mass, weight, momentum, force, temperature, heat, etc.—but their technical meaning is more abstract than their colloquial meaning. We can map between the colloquial and abstract meanings associated with specific values by using specific instances as "calibrators."

A Thermal Analogy

Intuitive notion	*Quantification*	*Calibration*
Hot, cold	Temperature, T	Cold as ice = 273K Boiling hot = 373K
uncertainty	Probability, P	Certainty = 0, 1 $p = 1/36$: plausible as "snake's eyes" $p = 1/1024$: plausible as 10 heads

Arguments Relating Hypotheses, Data, and Models

We seek to appraise scientific hypotheses in light of observed data and modeling assumptions.

Consider the data and modeling assumptions to be the premises of an argument with each of various hypotheses, H_i, as conclusions: $H_i|D_{\text{obs}}, \mathcal{C}$. ($\mathcal{C} =$ "context," or background information—specifies hypothesis space $\{H_i\}$, connection b/t data and hypotheses, etc.)

$P(H_i|D_{\text{obs}}, \mathcal{C})$ measures the degree to which $(D_{\text{obs}}, \mathcal{C})$ supports asserting H_i. It provides an ordering and weighting among arguments for various H_i that share common premises.

Probability theory tells us how to analyze and appraise the argument, i.e., how to calculate $P(H_i|D_{\text{obs}}, \mathcal{C})$ from simpler, hopefully more accessible probabilities.

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises (sometimes "I")

'AND' (product rule):
$$\begin{aligned} P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\text{obs}}|H_i, \mathcal{C}) \\ &= P(D_{\text{obs}}|\mathcal{C})\, P(H_i|D_{\text{obs}}, \mathcal{C}) \end{aligned}$$

'OR' (sum rule):
$$\begin{aligned} P(H_1 \vee H_2|\mathcal{C}) &= P(H_1|\mathcal{C}) + P(H_2|\mathcal{C}) \\ &\quad - P(H_1, H_2|\mathcal{C}) \end{aligned}$$

'NOT':
$$P(\overline{H_i}|\mathcal{C}) = 1 - P(H_i|\mathcal{C})$$

(note that $\mathcal{C}$ is the part of the premises that is always given)

Three Important Theorems

Bayes's Theorem (BT)

Consider the *joint probability* for a hypothesis and the observed data, $P(H_i, D_{\rm obs}|\mathcal{C})$, using the product rule:

$$\begin{aligned} P(H_i, D_{\rm obs}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\rm obs}|H_i, \mathcal{C}) \\ &= P(D_{\rm obs}|\mathcal{C})\, P(H_i|D_{\rm obs}, \mathcal{C}) \end{aligned}$$

Solve for the *posterior probability* for H_i (adds a premise!):

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

Theorem holds for any propositions, but for hypotheses & data the factors have names:

posterior $\propto$ *prior* $\times$ *likelihood*

(all "for H_i")

norm. const. $P(D_{\rm obs}|\mathcal{C}) =$ *prior predictive* for $D_{\rm obs}$

Aside: likelihood vs. probability

Data influence inferences via the ability of rival hypotheses to predict the actually observed data, quantified by $P(D_{\rm obs}|H_i, \mathcal{C})$.

Consider the function of D and H_i,

$$p(D|H_i, \mathcal{C}).$$

This is a probability for choices of D, but not for choices of H_i (wrong side of the solidus!), although it does depend on H_i.

For a particular choice of H_i, the resulting function of D specifies the *sampling distribution* or (more descriptively!) the *conditional predictive distribution* for data.

The H_i dependence when we fix attention on the *observed* data is the *likelihood function for H_i* (*not* "for the data"):

$$\mathcal{L}(H_i) \equiv p(D_{\rm obs}|H_i, \mathcal{C})$$

The likelihood for H_i is not a probability for H_i, but the posterior probability for H_i is proportional to it.

Aside: Fisher on "likelihood"

"If we need a word to characterise this relative property of different values of p [the parameter], I suggest that we may speak without confusion of the likelihood of one value of p being thrice the likelihood of another, bearing always in mind that *likelihood is not here used loosely as a synonym of probability*, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample." (Fisher 1922)

Alas, this does not match the colloquial use of "likelihood," and thus is prone to confusion. E.g., from the OED:

likelihood:

the state or fact of something's being likely; probability

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ("suite;" $\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A,\mathcal{C})P(A|\mathcal{C}) = P(A|\mathcal{C}) \\ &= \sum_i P(B_i|\mathcal{C})P(A|B_i,\mathcal{C})\end{aligned}$$

If we do not see how to get $P(A|\mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{C})P(A|B_i,\mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Joseph Blitzstein (Harvard statistician) on LTP (paraphrased):

> In most areas of math, when you're stuck, saying, "I wish I knew this or that" doesn't help you. In probability theory, saying "I wish I knew this" suggests what to condition on; then you condition on it, compute *as if* you knew it, and then average over those possibilities.
>
> *I didn't name the law of total probability, but if I had, I would have just called it* **wishful thinking**.
>
> — YouTube lecture on conditional probability (15:48)

LTP example 1: With context $\mathcal{C}$, take $A = D_{\rm obs}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\rm obs}|\mathcal{C}) &= \sum_i P(D_{\rm obs}, H_i|\mathcal{C}) \\ &= \sum_i P(H_i|\mathcal{C}) P(D_{\rm obs}|H_i, \mathcal{C}) \end{aligned}$$

prior predictive for $D_{\rm obs}$ = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

LTP example 2: Take $\mathcal{C}$ to specify fair roll of a die, $A =$ "An even number comes up," $B_i =$ "face i comes up" ($i = 1$ to 6)

$$\begin{aligned} P(A|\mathcal{C}) &= \sum_{i=1}^{6} P(A, B_i|\mathcal{C}) \\ &= \sum_{i=1}^{6} P(B_i|\mathcal{C}) P(A|B_i, \mathcal{C}) \\ &= \frac{1}{6} \times (0 + 1 + 0 + 1 + 0 + 1) = \frac{1}{2} \end{aligned}$$

Aside: Fisher on "likelihood" (cont'd)

"If we need a word to characterise this relative property of different values of p [the parameter], I suggest that we may speak without confusion of the likelihood of one value of p being thrice the likelihood of another, bearing always in mind that *likelihood is not here used loosely as a synonym of probability*, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample." (Fisher 1922)

"Likelihood also *differs from probability* in that it is a differential element, and is *incapable of being integrated*: it is assigned to a particular point of the range of variation, not to a particular element [interval]." (Fisher 1922)

"...the integration with respect to m [p earlier] is illegitimate and has no definite meaning...." (Fisher 1912)

Normalization theorem

For *exclusive, exhaustive* H_i (a "suite" of hypotheses),

$$\sum_i P(H_i|\cdots) = 1$$

Principles of inference

View BT and LTP as **fundamental principles of inference**:

- **BT:** How uncertainties may change when *accounting for new factual information*—extending the premises, i.e., *learning*
- **LTP:** How uncertainty for a compound proposition should *account for possibilities*—all the ways that proposition could be true

Well-Posed Problems

The rules express desired probabilities in terms of other probabilities.

To get a numerical value *out*, at some point we have to put numerical values *in*.

Direct probabilities are probabilities with numerical values determined directly by premises (via modeling assumptions, symmetry arguments, previous calculations, desperate presumption...). Probabilites dreived from them are *indirect*.

An inference problem is *well posed* only if all the needed probabilities are assignable based on the specified premises (context, data...).

We may need to add new assumptions as we see what needs to be assigned. We may not be entirely comfortable with what we need to assume! (Remember Euclid's fifth postulate!)

We should explore how results depend on uncomfortable assumptions (*robustness*).

Tabular/diagrammatic Bayesian inference

Simplest case: *Binary classification*

- 2 hypotheses: $\{C, \overline{C}\}$
- 2 possible data values: $\{-, +\}$

Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(\overline{C})$?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
  - Test is 80% accurate if you have the condition:
    $P(+|C, \mathcal{C}) = 0.8$ ("sensitivity")
  - Test is 95% accurate if you are healthy:
    $P(-|\overline{C}, \mathcal{C}) = 0.95$ ("specificity," $1 - p(\text{false } +)$)

Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women

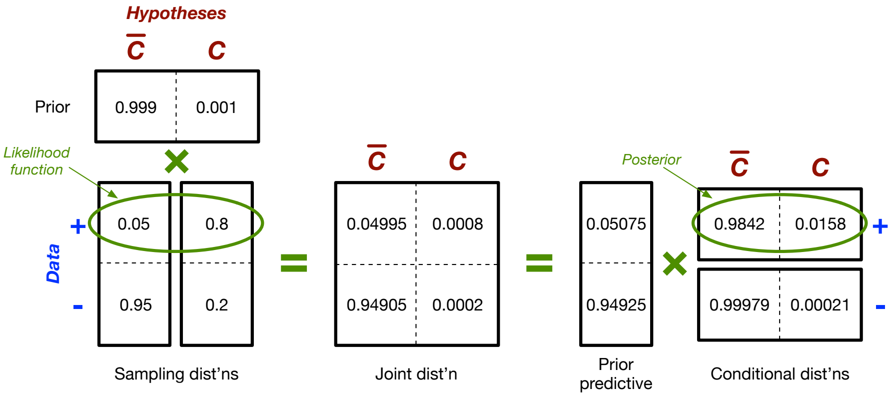
Tabular calculation

Hypothesis H_i	Prior $\pi_i \equiv p(H_i)$	Likelihood $\mathcal{L}_i \equiv p(+\vert H_i)$	Joint $\pi_i \times \mathcal{L}_i$	Posterior $p(H_i\vert +)$
$\overline{C}$	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

Inference as manipulation of the joint distribution

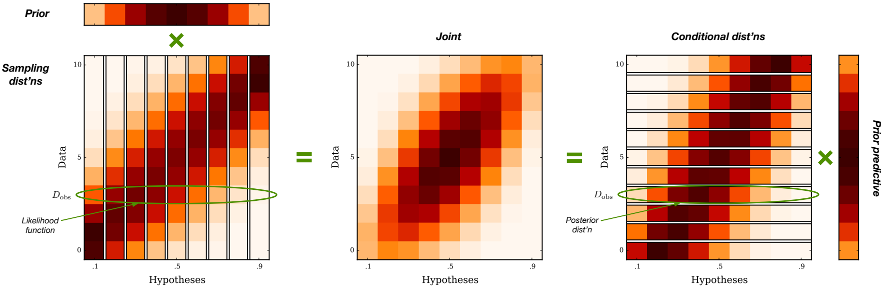
Bayes's theorem in terms of the *joint distribution*:

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

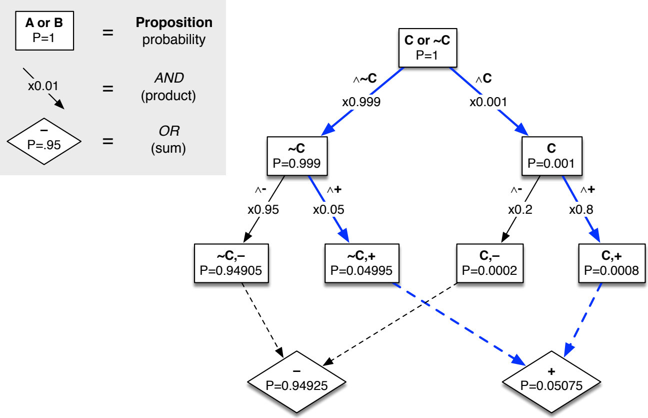


Larger discrete case—coin flipping:

- 9 hypotheses: For $\alpha \equiv P(\text{heads on a flip})$, consider $\alpha = 0.1, \ldots, 0.9$
- 11 possible data values: # of heads in 10 flips



Case diagram—probabilities



$$P(H_1 \vee H_2|\mathcal{C})$$

$$P(H_i|\mathcal{C})$$

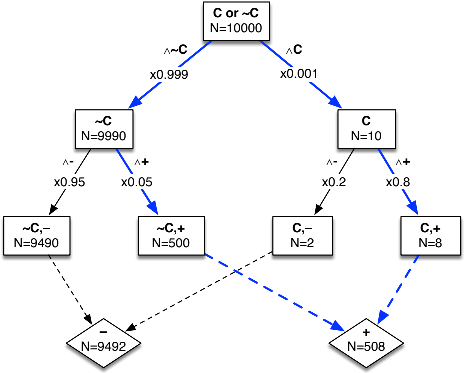
$$P(H_i, D_j|\mathcal{C}) = P(H_i|\mathcal{C})\, P(D_j|H_i, \mathcal{C})$$

$$P(D_j|\mathcal{C}) = \sum_i P(H_i, D_j|\mathcal{C})$$

$$P(C|+, \mathcal{C}) = \frac{0.0008}{0.05075} \approx 0.016$$

Case diagram—counts

Create a large collection of imaginary cases with ratios of counts chosen to approximate the probabilities.



$$P(C|+,\mathcal{C}) = \frac{8}{508} \approx 0.016$$

Of the 508 cases with positive test results, only 8 have the condition. The prevalence is so low that when there is a positive result, it's more likely to have been a mistake than accurate, even for a sensitive test.