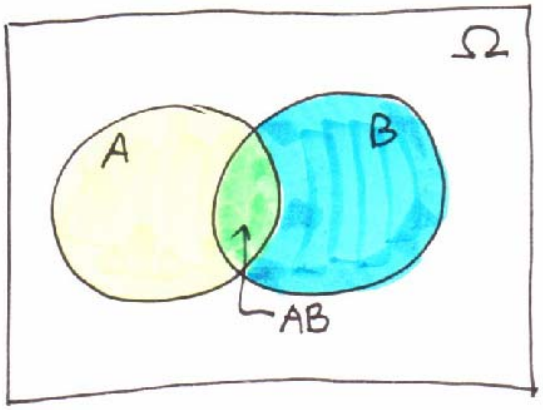
4th IMPRS Astronomy Summer School

Drawing Astrophysical Inferences from Data Sets

William H. Press
The University of Texas at Austin

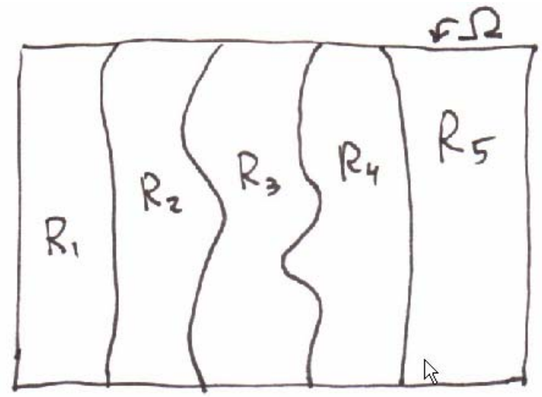
Lecture 1

Additivity or “Law of Or-ing”



$$P(A \cup B) = P(A) + P(B) - P(AB)$$

“Law of Exhaustion” for EME



$$\sum_i P(R_i) = 1$$

Multiplicative Rule or “Law of And-ing”

“given”

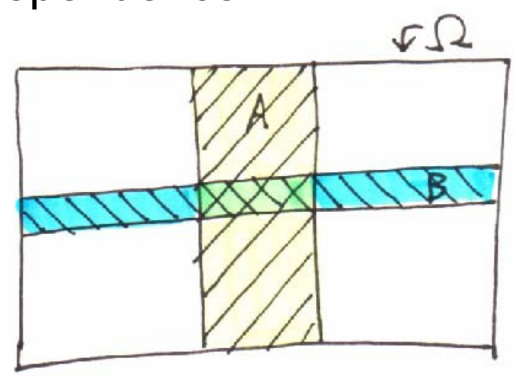
$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

$$P(B|A) = \frac{P(AB)}{P(A)}$$

“conditional probability”

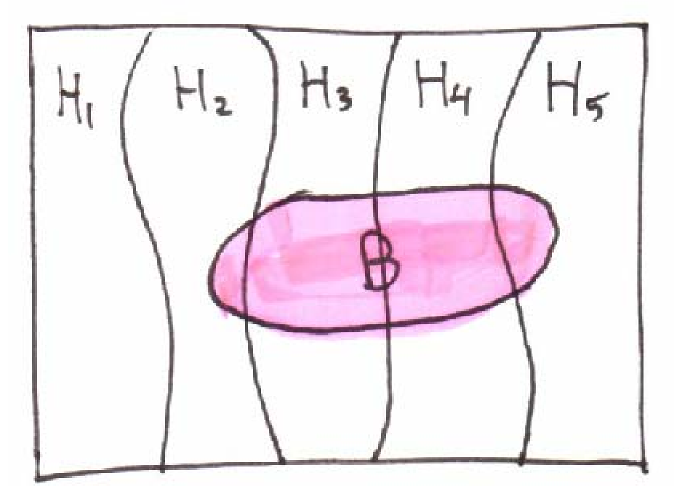
“renormalize the outcome space”

Independence:



Events A and B are independent if
$P(A|B) = P(A)$
so $P(AB) = P(B)P(A|B) = P(A)P(B)$

Law of Total Probability or “Law of de-Anding”



H’s are exhaustive and mutually exclusive (EME)

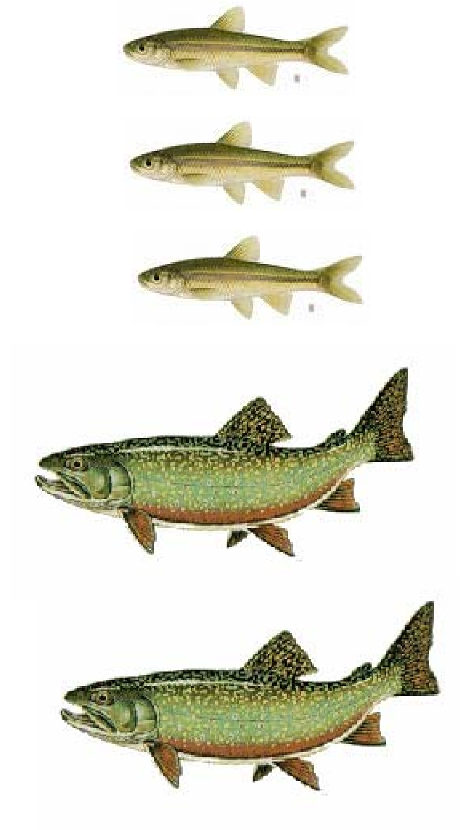
$$P(B) = P(BH_1) + P(BH_2) + \ldots = \sum_i P(BH_i)$$

$$P(B) = \sum_i P(B|H_i)P(H_i)$$

“How to put Humpty-Dumpty back together again.”

Example: A barrel has 3 minnows and 2 trout, with equal probability of being caught. Minnows must be thrown back. Trout we keep.

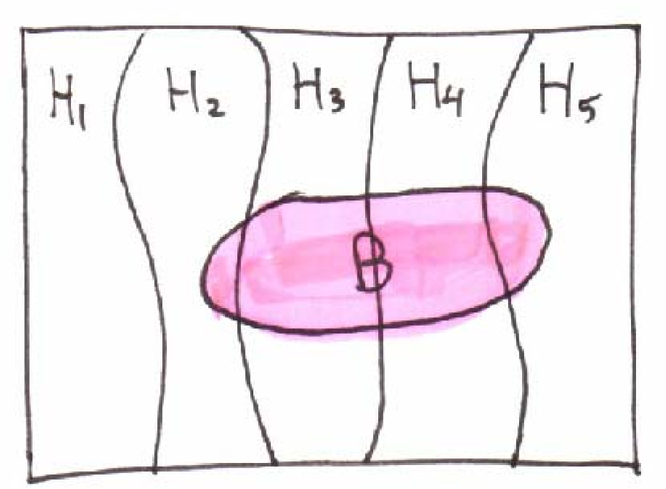
What is the probability that the <u>2nd</u> fish caught is a trout?



$$\begin{aligned}
H_1 &\equiv \text{1st caught is minnow, leaving } 3+2 \\
H_2 &\equiv \text{1st caught is trout, leaving } 3+1 \\
B &\equiv \text{2nd caught is a trout}
\end{aligned}$$

$$\begin{aligned}
P(B) &= P(B|H_1)P(H_1) + P(B|H_2)P(H_2) \\
&= \tfrac{2}{5}\cdot\tfrac{3}{5} + \tfrac{1}{4}\cdot\tfrac{2}{5} = 0.34
\end{aligned}$$

Bayes Theorem



Thomas Bayes
1702 - 1761

(same picture as before)

Law of And-ing

$$P(H_i|B) = \frac{P(H_i B)}{P(B)}$$

$$= \frac{P(B|H_i)P(H_i)}{\sum_j P(B|H_j)P(H_j)}$$

Law of de-Anding

We usually write this as

$$P(H_i|B) \propto P(B|H_i)P(H_i)$$

this means, "compute the normalization by using the completeness of the H_i's"

- As a theorem relating probabilities, Bayes is unassailable
- But we will also use it in inference, where the H's are hypotheses, while B is the data
  - "what is the probability of an hypothesis, given the data?"
  - some (defined as frequentists) consider this dodgy
  - others (Bayesians like us) consider this fantastically powerful and useful
  - in real life, the war between Bayesians and frequentists is long since over, and most statisticians adopt a mixture of techniques appropriate to the problem.
- Note that you generally have to know a complete set of EME hypotheses to use Bayes for inference
  - perhaps its principal weakness

Example: Trolls Under the Bridge

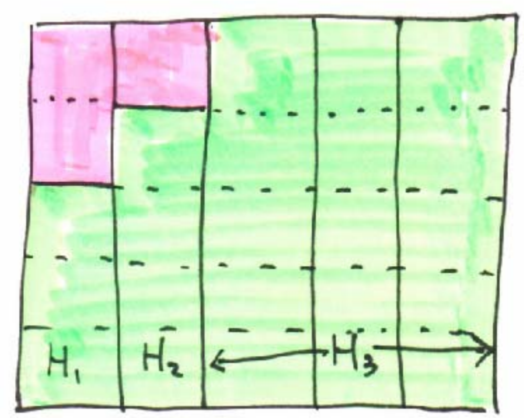

Trolls are bad. Gnomes are benign.
Every bridge has 5 creatures under it:

20% have TTGGG (H_1)
20% have TGGGG (H_2)
60% have GGGGG (benign) (H_3)

Before crossing a bridge, a knight captures one of the 5 creatures at random. It is a troll. "I now have an 80% chance of crossing safely," he reasons, "since only the case

20% had TTGGG (H1) → now have TGGG

is still a threat."

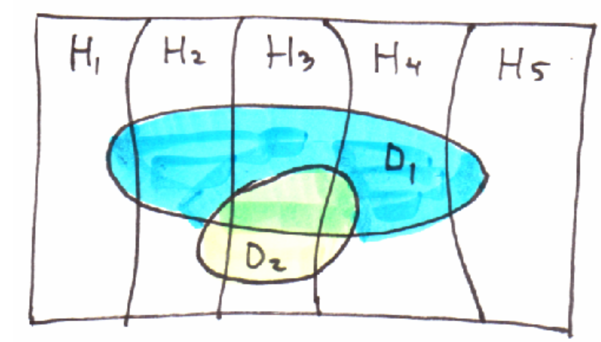


$$P(H_i|T) \propto P(T|H_i)P(H_i)$$

so, $$P(H_1|T) = \frac{\frac{2}{5}\cdot\frac{1}{5}}{\frac{2}{5}\cdot\frac{1}{5}+\frac{1}{5}\cdot\frac{1}{5}+0\cdot\frac{3}{5}} = \frac{2}{3}$$

The knight's chance of crossing safely is actually only 33.3%
Before he captured a troll ("saw the data") it was 60%.
Capturing a troll actually made things worse! [well…discuss]
(80% was never the right answer!)
Data changes probabilities!
Probabilities after assimilating data are called posterior probabilities.

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!

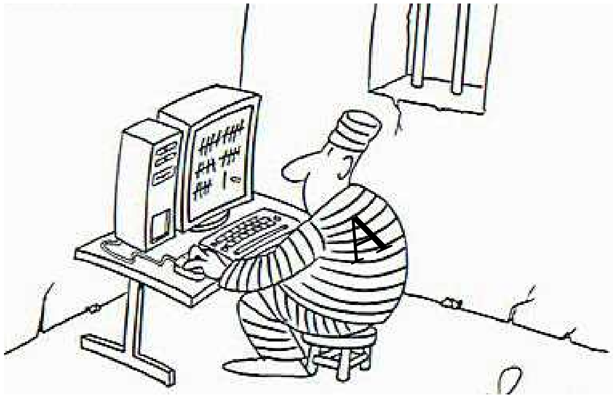
But,
$$= P(D_2|H_iD_1)P(D_1|H_i)P(H_i)$$
$$= P(D_1D_2|H_i)P(H_i)$$

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

Our next topic is Bayesian Estimation of Parameters. We'll ease into it with…

The Jailer's Tip:



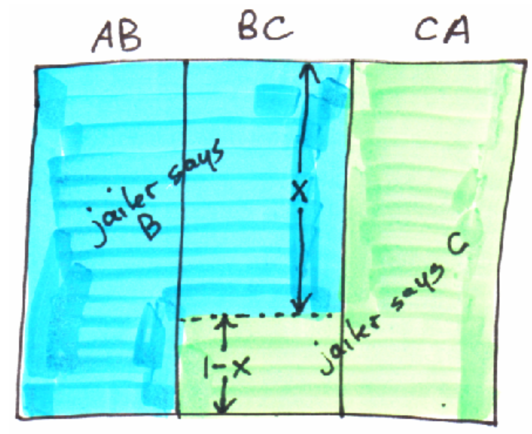
- Of 3 prisoners (A,B,C), 2 will be released tomorrow.
- A, who thinks he has a 2/3 chance of being released, asks jailer for name of one of the lucky – but not himself.
- Jailer says, truthfully, "B".
- "Darn," thinks A, "now my chances are only ½, C or me".

Here, did the data ("B") change the probabilities?

Further, suppose the jailer is not indifferent about responding “B” versus “C”.

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

“says B”



$$
\begin{aligned}
P(A|S_B) &= P(AB|S_B) + \cancelto{0}{P(AC|S_B)} \\
&= \frac{\cancelto{1}{P(S_B|AB)}\,\cancelto{1/3}{P(AB)}}{P(S_B|AB)P(AB) + \cancelto{x}{P(S_B|BC)}P(BC) + P(S_B|CA)P(CA)} \\
&= \frac{\frac{1}{3}}{1 \cdot \frac{1}{3} + x \cdot \frac{1}{3} + 0} = \frac{1}{1+x}
\end{aligned}
$$

So if A knows the value x, he can calculate his chances.
If x=1/2, his chances are 2/3, same as before; so he got no new information.
If x≠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

law of de-Anding

(e.g., Jailer’s Tip):

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

We are trying to estimate a parameter

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

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

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

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"

We now need $P(\text{data}|x)$

$P(\text{data}|x)$ { is the (forward) statistical model in both frequentist vs. Bayesian contexts. But it means something slightly different in each of the two.

A forward statistical model *assumes* that all parameters, assignments, etc., are known, and gives the probability of the observed data set. It is almost always the starting point for a well-posed analysis. If you can't write down a forward statistical model, then you probably don't understand your own experiment or observation!

the frequentist considers the universe of what might have been, imagining repeated trials, even if they weren't actually tried:

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

$$P(\text{data}|x) = \binom{N}{N_B} \overbrace{x^{N_\mathrm{B}}\,(1-x)^{N_\mathrm{C}}}^{\text{prob. of exact sequence seen}} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

no. of equivalent arrangements

the Bayesian considers only the <u>exact</u> data seen:

prior is still with us

$$P(x|\text{data}) \propto x^{N_\mathrm{B}}\,(1-x)^{N_\mathrm{C}}\,p(x|I)$$

No binomial coefficient, 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).

To get a normalized probability, we must integrate the denominator:

```
In[7]:= num = x^nb (1 - x) ^ (nn - nb)
```

Out[7]= $(1 - x)^{-nb+nn}\, x^{nb}$

we'll assume a uniform prior

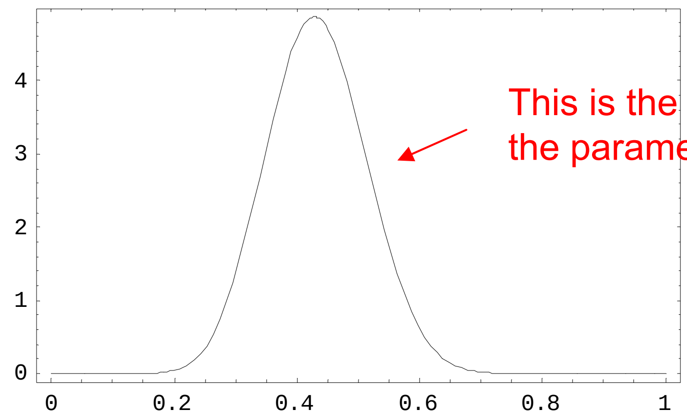
```
In[8]:= denom = Integrate[num, {x, 0, 1},
          GenerateConditions → False]
```

Out[8]= $\dfrac{\text{Gamma}[1 + nb]\ \text{Gamma}[1 - nb + nn]}{\text{Gamma}[2 + nn]}$

```
In[9]:= p[x_] = num / denom
```

Out[9]= $\dfrac{(1 - x)^{-nb+nn}\, x^{nb}\ \text{Gamma}[2 + nn]}{\text{Gamma}[1 + nb]\ \text{Gamma}[1 - nb + nn]}$

```
In[12]:= Plot[p[x] /. {nn → 35, nb → 15}, {x, 0, 1},
          PlotRange → All, Frame → True]
```



This is the Bayesian estimate of the parameter x, namely $p(x)$

Out[12]= - Graphics -

Properties of our Bayesian estimate of x:

derivative has this simple factor

```
In[20]:= Simplify[D[p[x], x]]
```

Out[20]= $-\frac{(1-x)^{-1-nb+nn}\, x^{-1+nb}\, (-nb+nn\, x)\, \text{Gamma}[2+nn]}{\text{Gamma}[1+nb]\, \text{Gamma}[1-nb+nn]}$

```
In[21]:= Solve[Simplify[D[p[x], x]] == 0, x]
```

Out[21]= $\left\{\left\{x \to \frac{nb}{nn}\right\}\right\}$

"maximum likelihood" answer is to estimate x as exactly the fraction seen

```
In[23]:= mean = Integrate[x p[x], {x, 0, 1},
           GenerateConditions → False]
```

mean is the 1st moment

Out[23]= $\frac{1+nb}{2+nn}$

```
In[27]:= sigma =
          Sqrt[FullSimplify[
            Integrate[x^2 p[x], {x, 0, 1},
              GenerateConditions → False] - mean^2]]
```

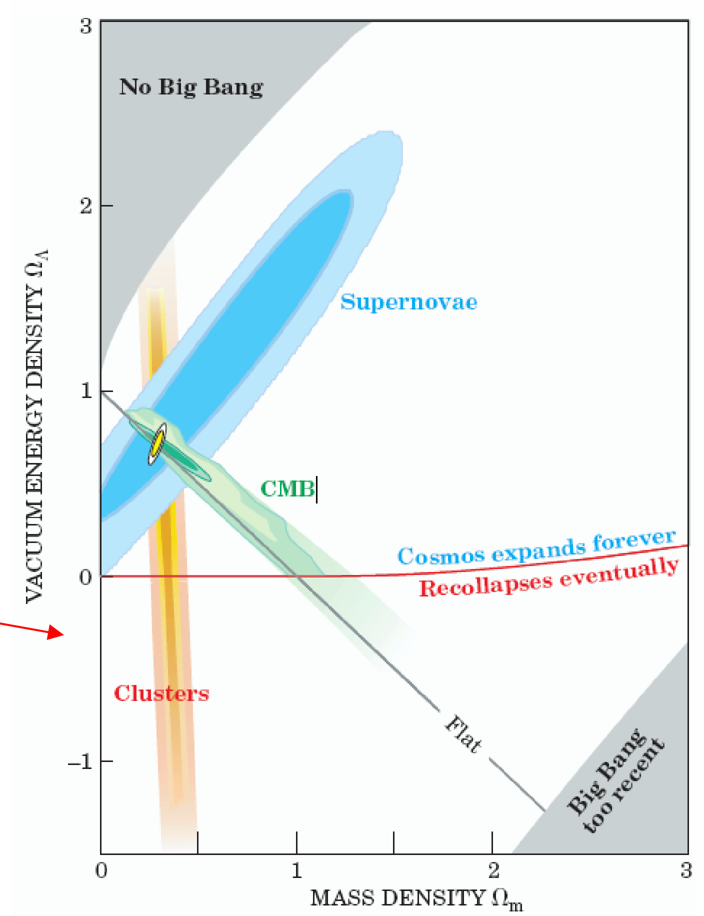
standard error involves the 2nd moment, as shown

Out[27]= $\sqrt{\frac{(1+nb)\,(1-nb+nn)}{(2+nn)^2\,(3+nn)}}$

This shows how *p(x)* gets narrower as the amount of data increases.

The basic paradigm of Bayesian parameter estimation :

- Construct a statistical model for the probability of the observed data as a function of all parameters
  - treat dependency in the data correctly
- Assign prior distributions to the parameters
  - jointly or independently as appropriate
  - use the results of previous data if available
- Use Bayes law to get the (multivariate) posterior distribution of the parameters
- Marginalize as desired to get the distributions of single (or a manageable few multivariate) parameters



Cosmological models are typically fit to many parameters. Marginalization yields the distribution of parameters of interest, here two, shown as contours.