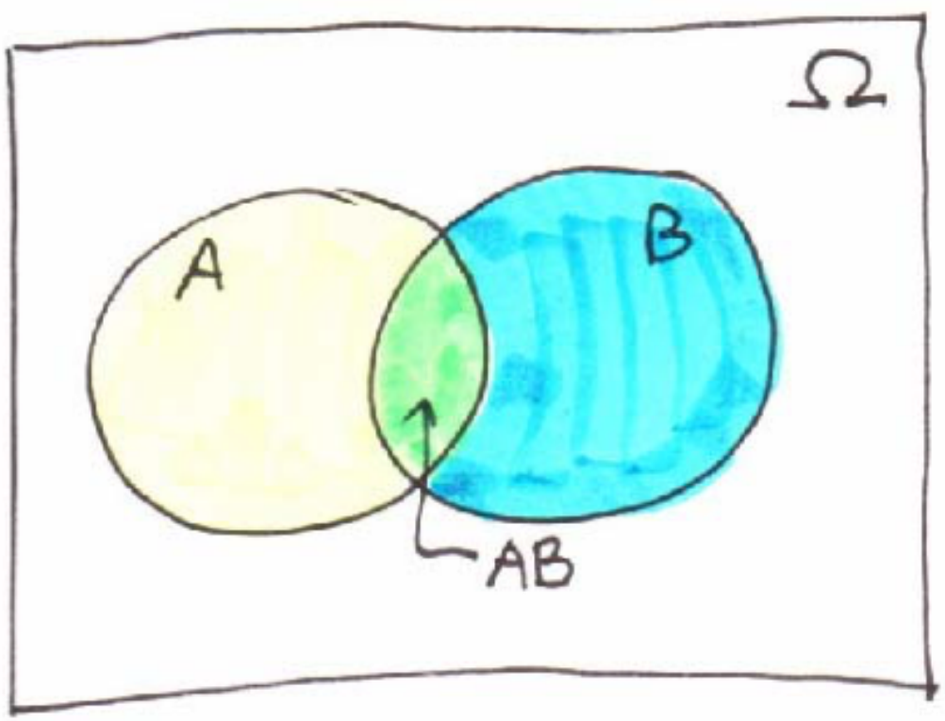
Lecture 4: Bayesian parameter estimates II.

Kolmogorov probability concept revisited

Additivity or “Law of Or-ing”



Venn diagrams at web site of Probability, Mathematical Statistics, Stochastic Processes:

http://www.math.uah.edu/stat/

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

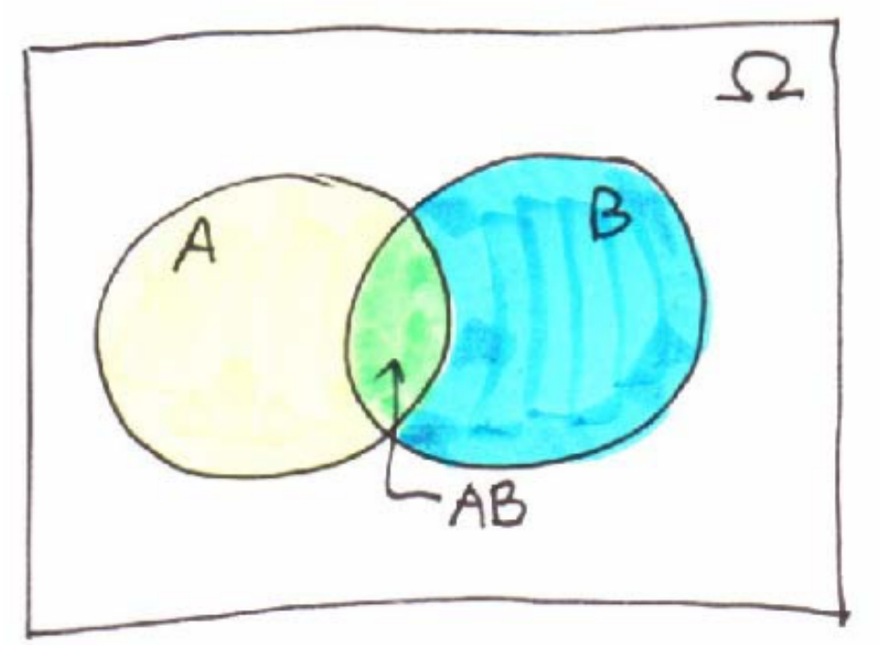
A or B

A and B

$P(A \cap B)$

Kolmogorov probability concept revisited

Multiplicative Rule or "Law of And-ing"



(same picture as before)

"given"

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

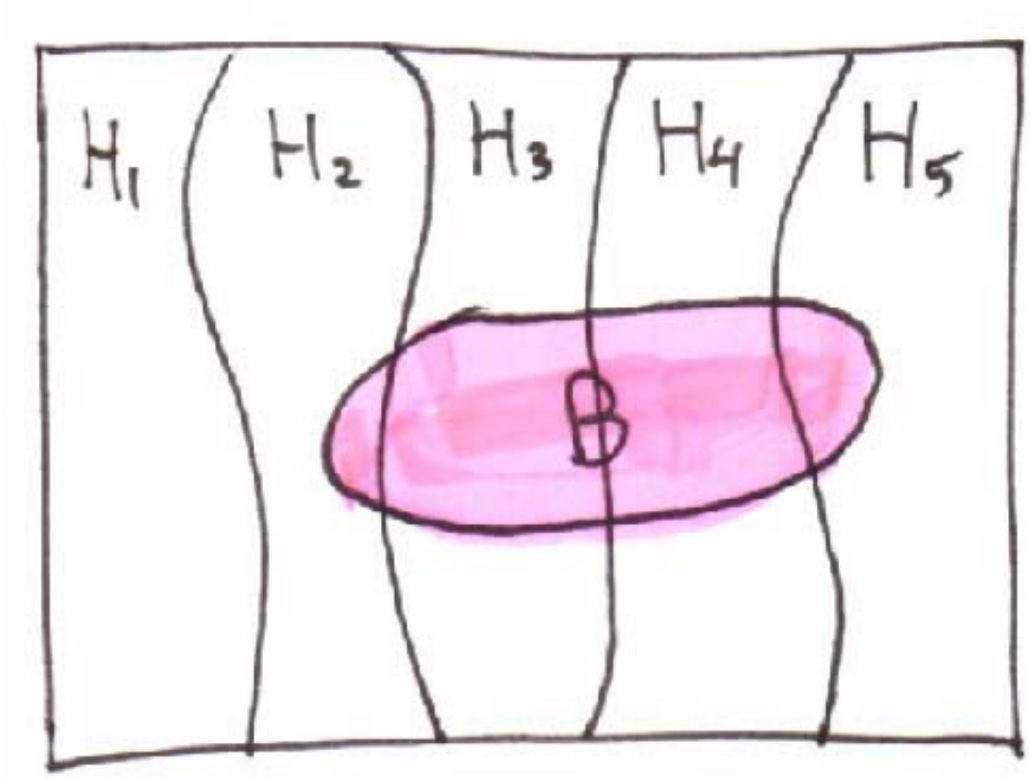
$$P(B|A) = \frac{P(AB)}{P(A)}$$

"conditional probability"

"renormalize the outcome space"

Kolmogorov probability concept revisited

Law of Total Probability or "Law of de-Anding"



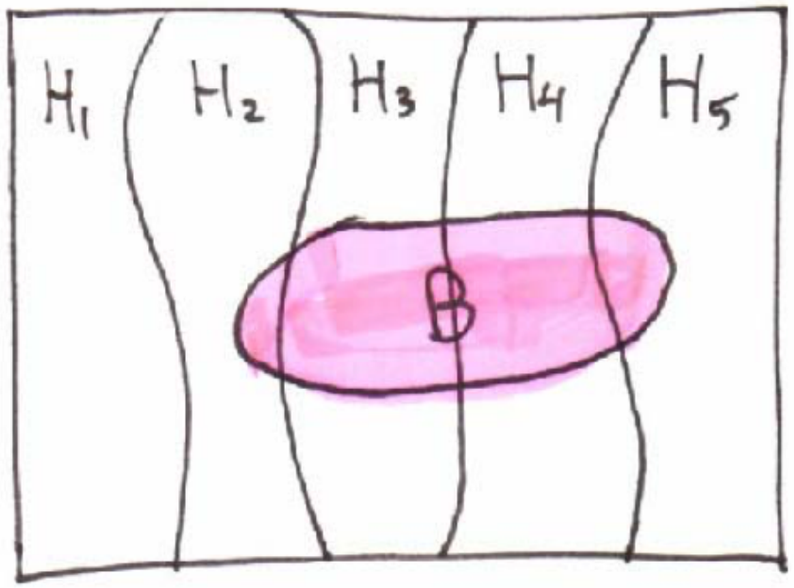
H's are exhaustive and mutually exclusive (EME)

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

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

Bayes' theorem revisited

Bayes Theorem



Thomas Bayes
1702 - 1761

(same picture as before)

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

Law of And-ing

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

Law of de-Anding

We usually write this as

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

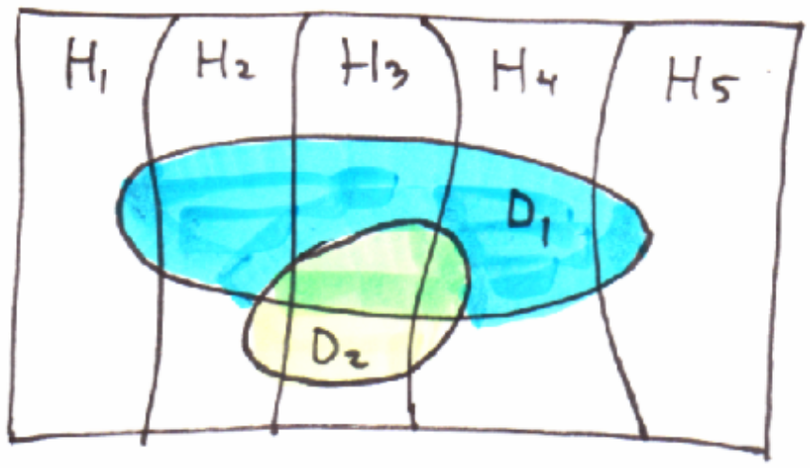
this means, "compute the normalization by using the completeness of the H_i's"

commutativity/associativity revisited

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

how to calculate Bayesian probabilities:

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, <u>always revealing no car</u> (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.

how to calculate Bayesian probabilities:

H_i = car behind door i, $i = 1, 2, 3$

Wlog, you pick door 2 (relabeling).

Wlog, Monty opens door 3 (relabeling).

$P(H_i|O3) \propto P(O3|H_i)P(H_i)$

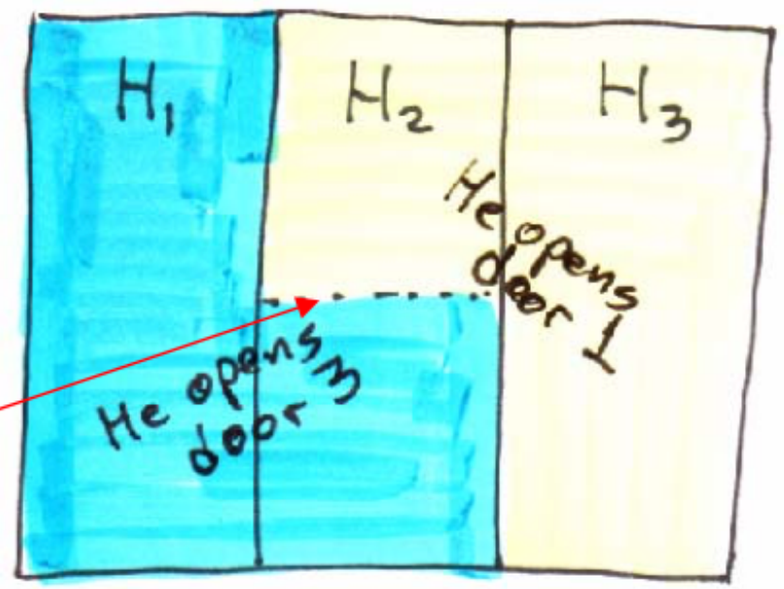
"Without loss of generality "

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

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

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

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



So you should always switch: doubles your chances!

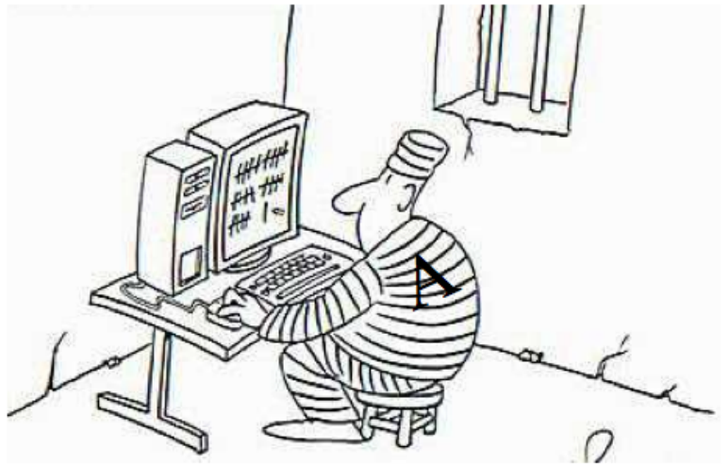
how to calculate Bayesian probabilities:

Monty Hall and the Reverend Bayes

- ★ 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

Bayesian parameter estimates:

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".

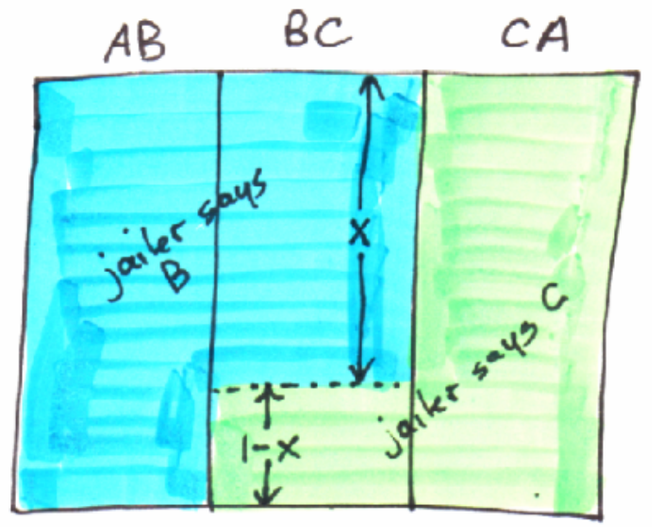
Is this like Monty Hall? Did the data ("B") change the probabilities?

Bayesian parameter estimates:

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"



$$P(A|S_B) = P(AB|S_B) + P(AC|S_B)$$

0

$$= \frac{P(S_B|AB)P(AB)}{P(S_B|AB)P(AB) + P(S_B|BC)P(BC) + P(S_B|CA)P(CA)}$$

1 1/3 x

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

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*≠1/2, he does get new info – his chances change.

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

Bayesian parameter estimates:

"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!

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

Bayesian parameter estimates:

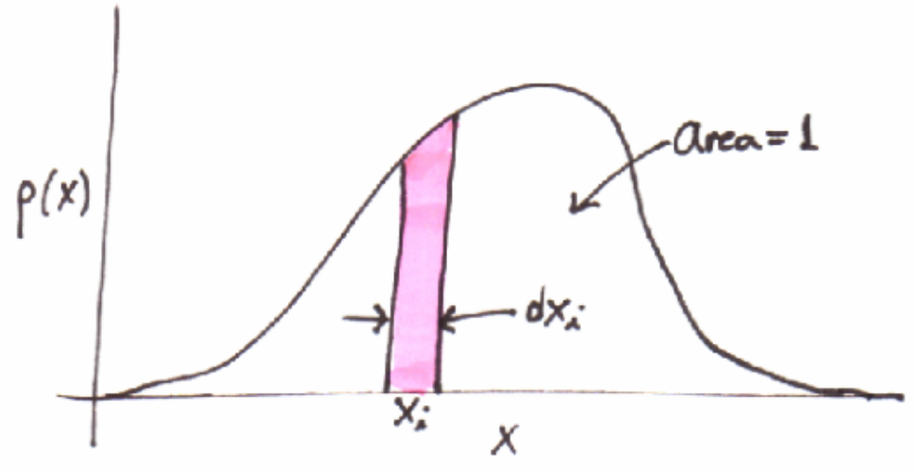
(repeating previous equation:)

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

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.)

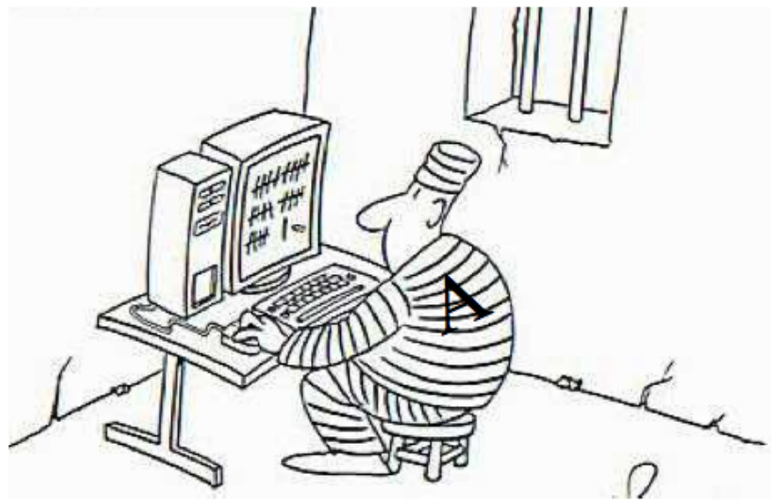


$$\sum_i P_i = 1 \;\leftrightarrow\; \sum_i p(x_i) dx_i = 1 \;\leftrightarrow\; \int_x p(x)\, dx = 1$$

Bayesian parameter estimates:

(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

Bayesian parameter estimates:

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 $P(\text{data}|x)$

Bayesian parameter estimates:

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

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

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

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

no. of equivalent arrangements

Bayesian considers only the exact data seen, and has a prior:

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

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).

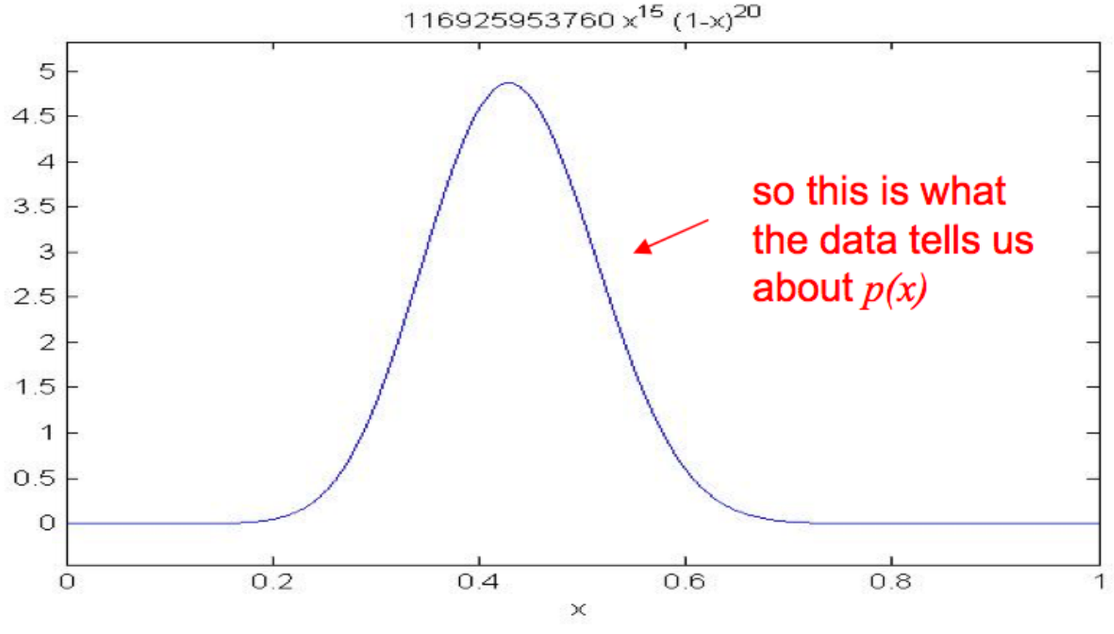
Bayesian parameter estimates:

Bayes numerator and denominator are:

$$P(x|\text{data}) = x^{N_B}(1-x)^{N-N_B} \times 1$$

$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B}(1-x)^{N-N_B}dx = \frac{\Gamma(N_B+1)\Gamma(N-N_B+1)}{\Gamma(N+2)}$$

Plot of numerator over denominator for N=35, N_B = 15:



Matlab code:

```
%%
syms nn nb x
num = x^nb*(1-x)^(nn-nb);
denom = gamma(nn-nb+1)*gamma(nb+1)/gamma(nn+2);
p = num/denom;
p=x^nb*(1-x)^(nn-nb)/gamma(nn- nb+1)/gamma(nb+1)*gamma(nn+2);
figure(30)
ezplot(subs(p,[nn,nb],[35,15]),[0,1]);

%%
```

Bayesian parameter estimates:

Find the mean, standard error, and mode of our estimate for x

$$P(x|\text{data}) = x^{N_B}(1-x)^{N-N_B}$$

$$\frac{dP(x|\text{data})}{dx} = 0 \;\Rightarrow\; x = \frac{N_B}{N}$$

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

$$\langle x \rangle = \int_0^1 x P(x|\text{data}) dx = \frac{N_B + 1}{N + 2}$$

mean is the 1st moment
notice it's different from ML!

variance involves the 2nd moment,

$$\text{Var}(x) = \langle x^2 \rangle - \langle x \rangle^2 = \int_0^1 x^2 P(x|\text{data}) dx - \langle x \rangle^2 = \frac{(N_B+1)(N-N_B+1)}{(N+2)^2(N+3)}$$

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

Bernoulli distribution:

(Let's leave behind the metaphor of the Jailer and Prisoner A.)

What we are illustrating is called Bernoulli trials:

- two possible outcomes
- i.i.d. events
- a single parameter x (the probability of one outcome)
- a sufficient statistic is the pair of numbers N and N_B

Jacob and Johann Bernoulli

$$P(\text{data}|x) = x^{N_B}(1-x)^{N-N_B}$$ (in the Bayesian sense)

$$P(x|\text{data}) \propto x^{N_B}(1-x)^{N-N_B} \times P(x|I)$$

for uniform prior, the Bayes denominator is, as we've seen, easy to calculate:

$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B}(1-x)^{N-N_B}dx = \frac{\Gamma(N_B+1)\Gamma(N-N_B+1)}{\Gamma(N+2)}$$

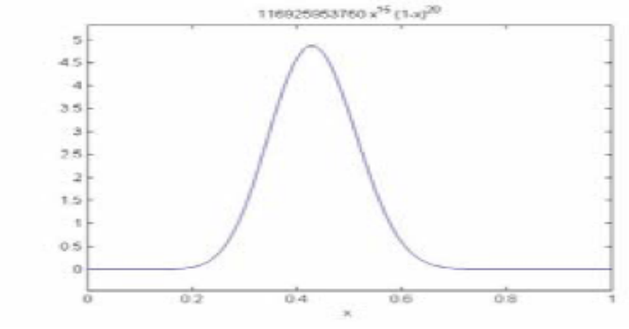
Bernoulli distribution:

Are there any other mathematical forms for the prior that would still leave the Bayes denominator easy to calculate?

Yes! try

$$P(x|I) \propto x^{\beta}(1-x)^{\alpha}$$

Choose α and β to make any desired center and width.



$$P(x|\text{data}) = x^{N_B}(1-x)^{N-N_B} \times x^{\beta}(1-x)^{\alpha}$$

$$\int_0^1 P(x|\text{data}) = \int_0^1 x^{N_B+\beta}(1-x)^{N-N_B+\alpha}dx$$

$$= \frac{\Gamma(N_B+\beta+1)\Gamma(N-N_B+\alpha+1)}{\Gamma(N+\alpha+\beta+2)}$$

Priors that preserve the analytic form of p(x) are called “conjugate priors”. There is nothing special about them except mathematical convenience.

If you start with a conjugate prior, you’ll also be able to assimilate new data trivially, just by changing the parameters of your estimate. This is because every posterior is in the right analytic form to be the new prior!