Probability Review and Statistical Estimation

Instructor: Alan Ritter

Many slides from Tom Mitchell

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment

- Examples

  A = True if a randomly drawn person from our class is female

  A = The hometown of a randomly drawn person from our class

  A = True if two randomly drawn persons from our class have same birthday

- Define P(A) as "the fraction of possible worlds in which A is true" or "the fraction of times A holds, in repeated runs of the random experiment"
  - the set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S

$$A: S \rightarrow \{0,1\}$$

A little formalism

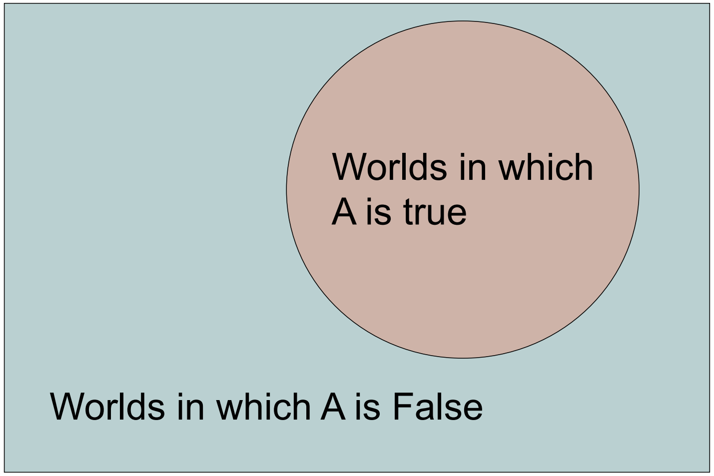
More formally, we have

- a <u>sample space</u> S (e.g., set of students in our class)
  - aka the set of possible worlds
- a <u>random variable</u> is a function defined over the sample space
  - Gender: S → { m, f }
  - Height: S → Reals
- an <u>event</u> is a subset of S
  - e.g., the subset of S for which Gender=f
  - e.g., the subset of S for which (Gender=m) AND (eyeColor=blue)
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

Visualizing A

Sample space of all possible worlds

Its area is 1



P(A) = Area of reddish oval

The Axioms of Probability

- $0 <= P(A) <= 1$
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

[di Finetti 1931]:

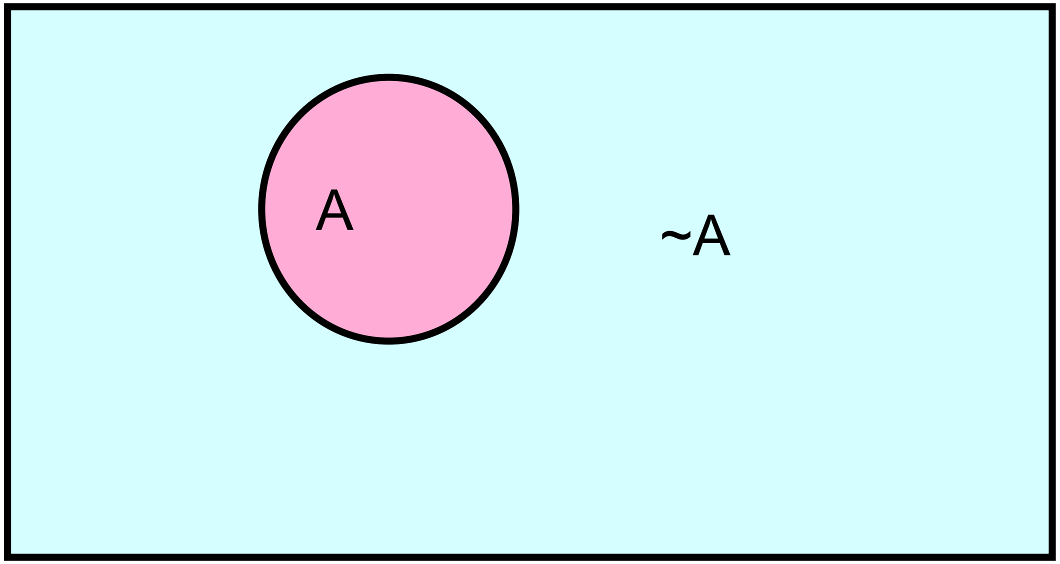
when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

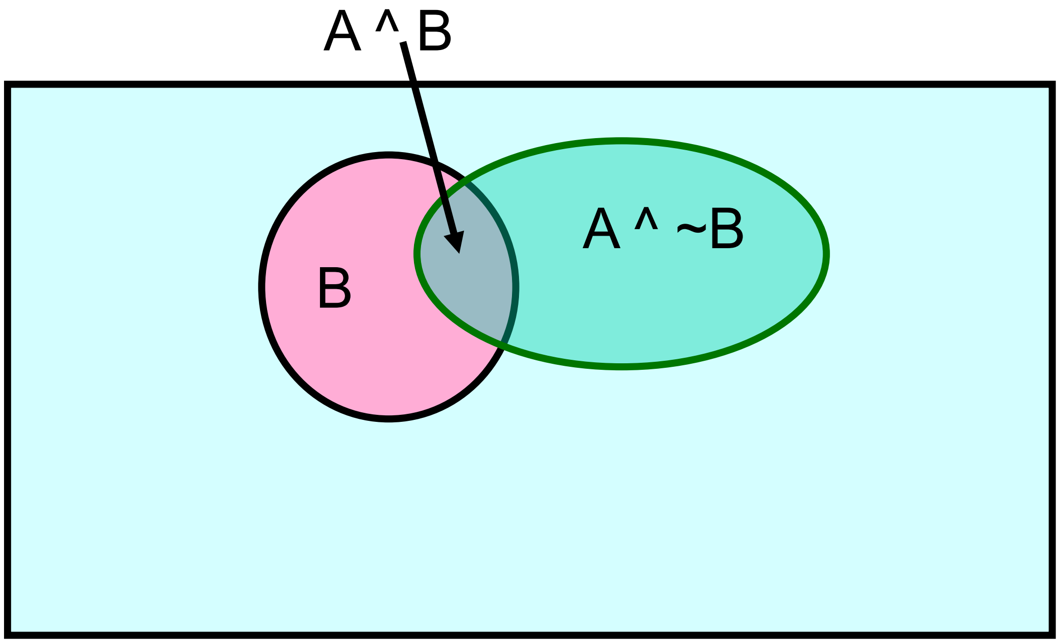
Elementary Probability in Pictures

- P(~A) + P(A) = 1



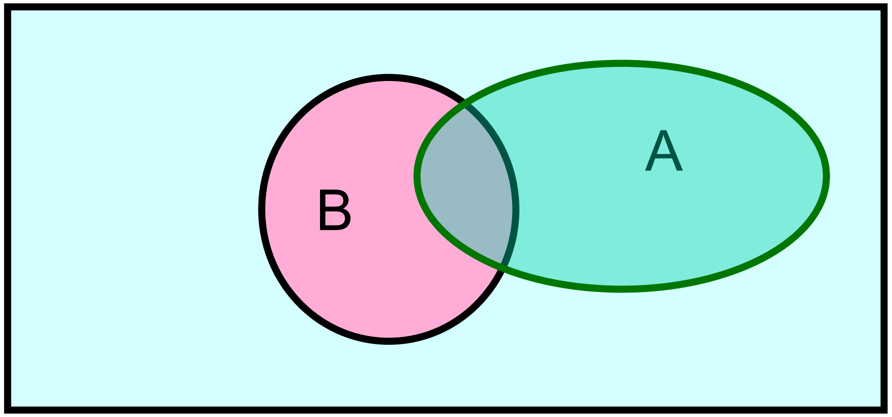
Elementary Probability in Pictures

- $P(A) = P(A \wedge B) + P(A \wedge \sim B)$



Definition of Conditional Probability

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



Definition of Conditional Probability

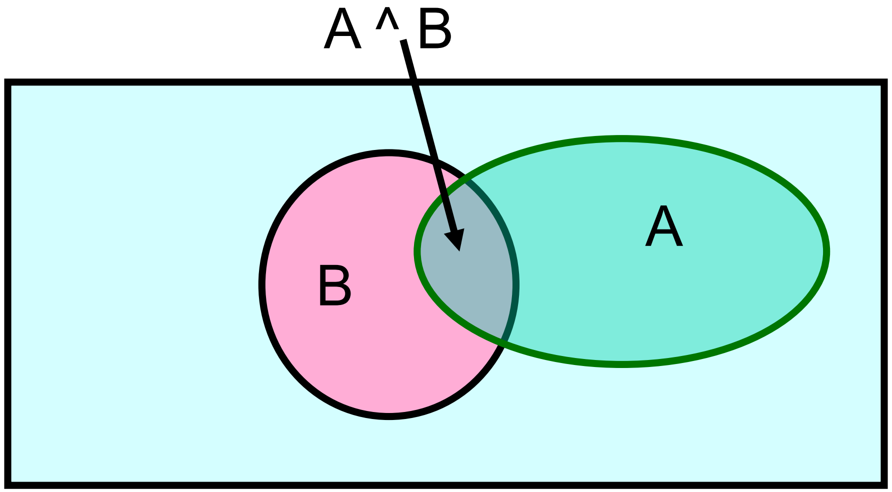
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

Bayes Rule

- let's write 2 expressions for P(A ^ B)



$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B \mid A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

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

A = you have the flu,   B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

what does all this have to do with function approximation?

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	**B**	**C**
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

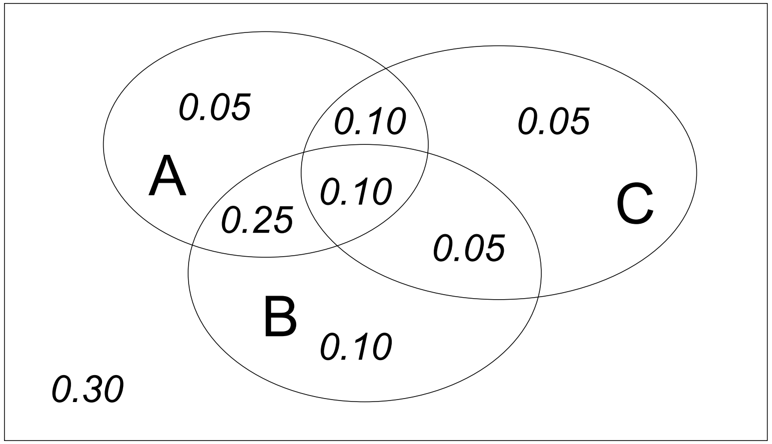
A	**B**	**C**	**Prob**
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of **any** logical expression involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

[A. Moore]

Using the Joint

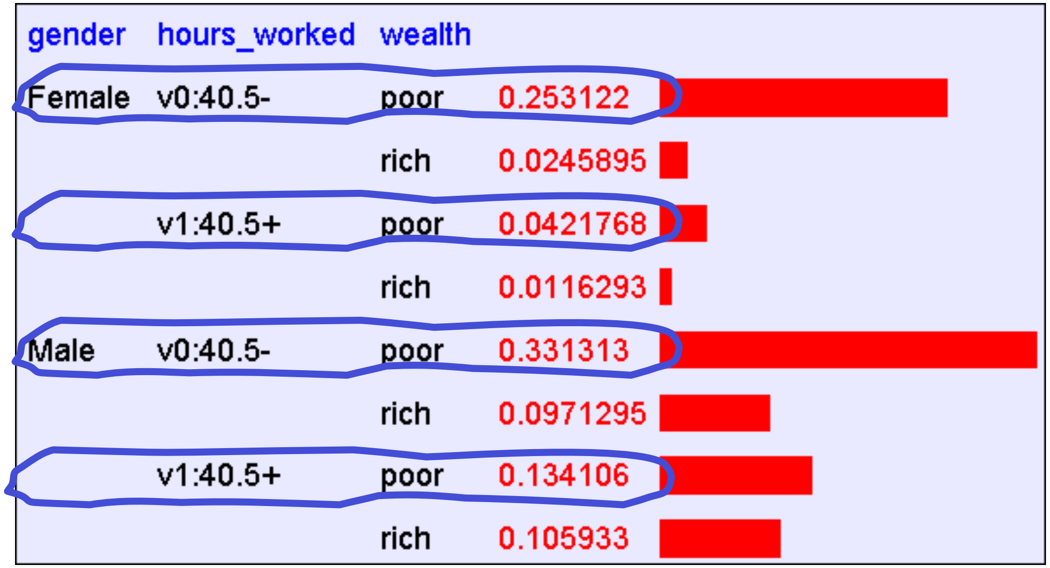
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

[A. Moore]

Using the Joint



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

[A. Moore]

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

[A. Moore]

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

[A. Moore]

sounds like the solution to learning $F: X \rightarrow Y$, or $P(Y \mid X)$.

Are we done?

sounds like the solution to learning F: X → Y, or P(Y | X).

Main problem: learning P(Y|X) can require more data than we have

consider learning Joint Dist. with 100 attributes

# of rows in this table?

# of people on earth?

fraction of rows with 0 training examples?

What to do?

1. Be smart about how we estimate probabilities from sparse data
   - maximum likelihood estimates
   - maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   - Bayes networks, graphical models

Bayesian Learning

A Game

- I choose a set of numbers
  - Prime Numbers
  - Numbers between 1 and 10

A Game

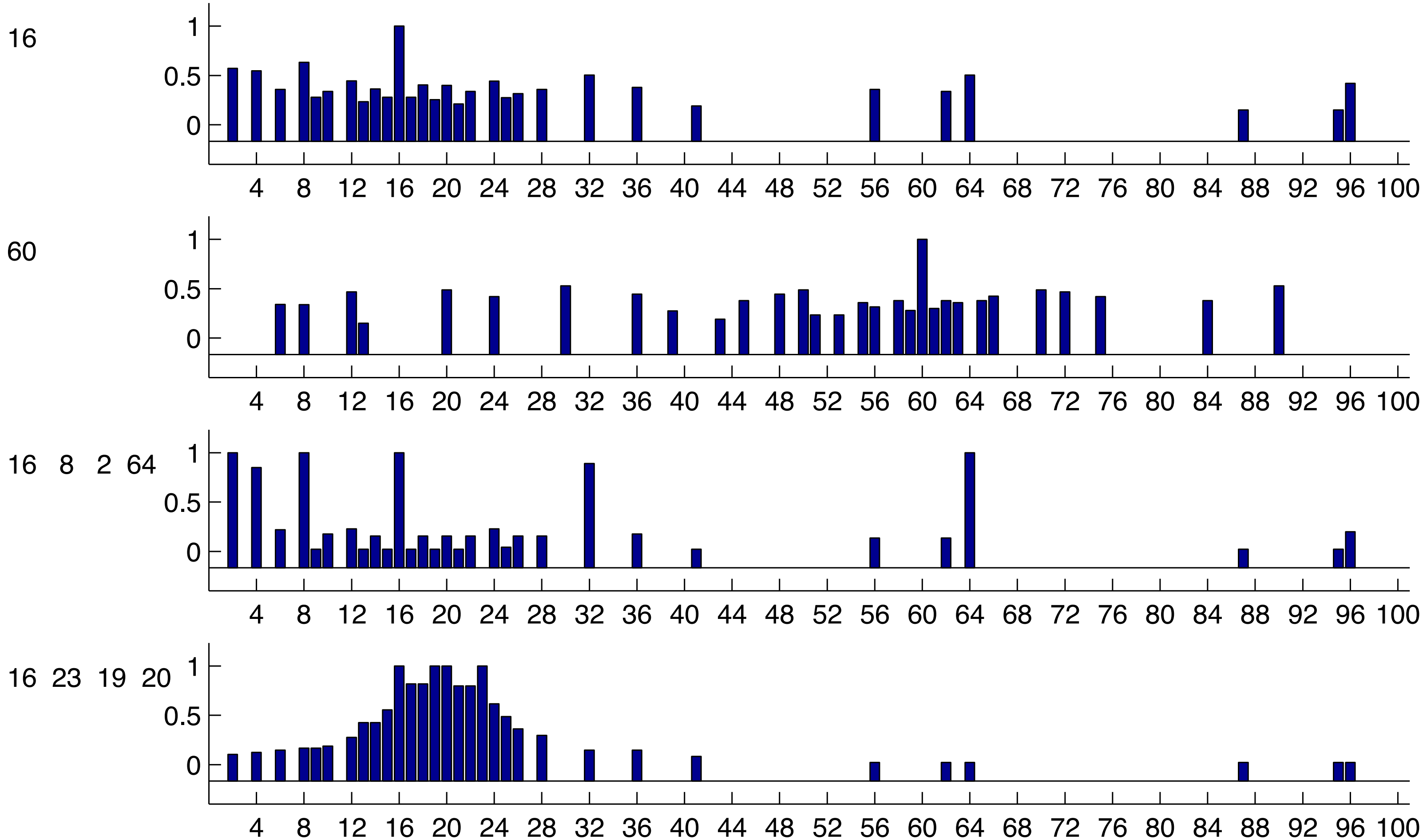
- I choose a set of numbers
  - Prime Numbers
  - Numbers between 1 and 10
- I give you a number of randomly chosen positive examples drawn from the set

A Game

- I choose a set of numbers
  - Prime Numbers
  - Numbers between 1 and 10
- I give you a number of randomly chosen positive examples drawn from the set
- Goal: predict whether a new number is in the set.

Real Data

Examples



Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

Powers of 2?

Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

  Powers of 2?

  Even numbers?

Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

Powers of 2?

Even numbers?

Powers of 2 except 32?

Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

$h_1 =$ Powers of 2?

$h_2 =$ Even numbers?

$h_3 =$ Powers of 2 except 32?

$$P(h_i | 16, 8, 2, 64) = ?$$

Assume we see: 16, 8, 2, 64

- Q: what is a reasonable hypothesis?

$h_1 =$ Powers of 2?

$h_2 =$ Even numbers?

$h_3 =$ Powers of 2 except 32?

$$P(h_i | 16, 8, 2, 64) = ?$$

Assume we see: 16, 8, 2, 64

- What is the probability of the data?

$$P(D|h) = P(16, 8, 2, 64|h)$$

Assume we see: 16, 8, 2, 64

- What is the probability of the data?

$$P(D|h) = P(16, 8, 2, 64|h)$$

$$= P(16|h) \cdot P(8|h) \cdot P(2|h) \cdot P(64|h)$$

Assume we see: 16, 8, 2, 64

- What is the probability of the data?

$$P(D|h) = P(16, 8, 2, 64|h)$$

$$= P(16|h) \cdot P(8|h) \cdot P(2|h) \cdot P(64|h)$$

$$= \left[\frac{1}{|h|}\right]^4$$

Prior Over Hypotheses?

$$P(h)$$

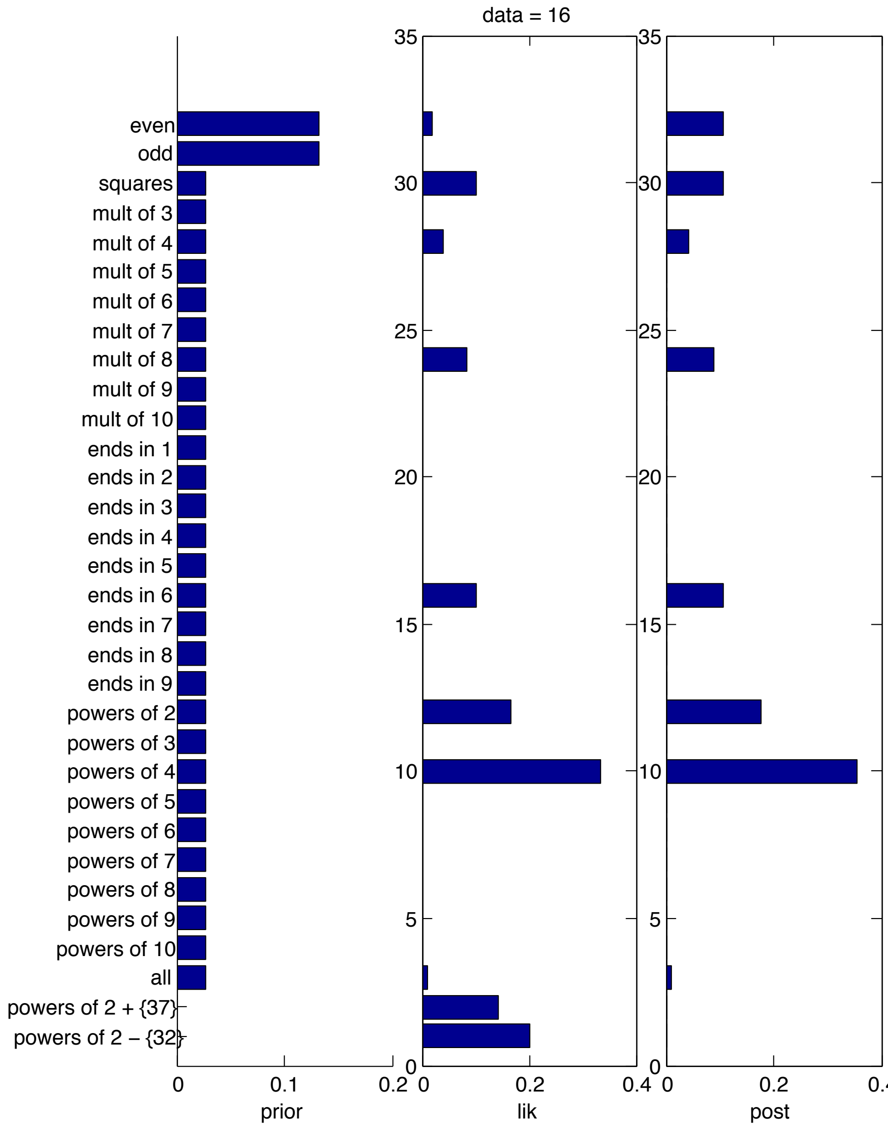
- Let’s assume some hypotheses are more likely than others
  - How likely is “powers of 2 except 32”?

Posterior Over Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{\sum_{h' \in \mathcal{H}} P(D,h')}$$

$$= \frac{P(h)\mathbb{1}(D \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} P(h')\mathbb{1}(D \in h')/|h'|^N}$$

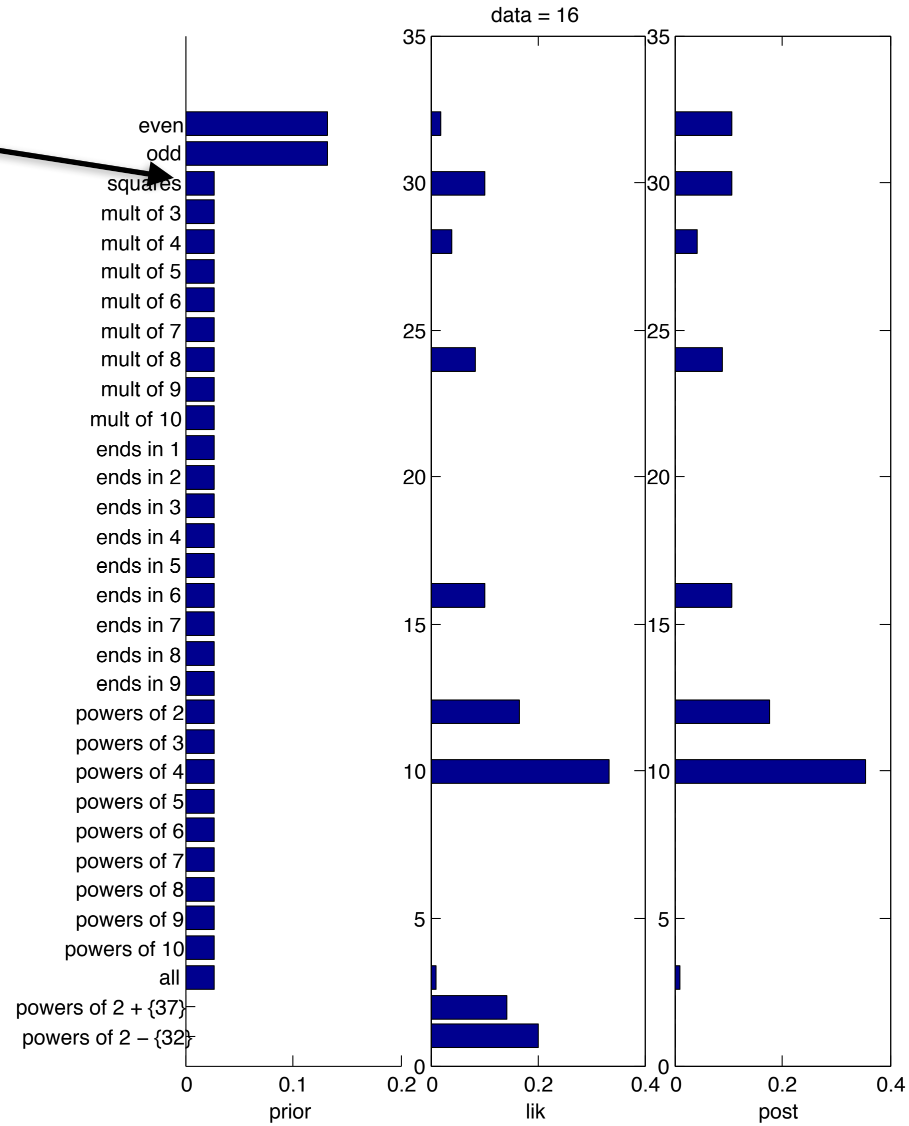
Data = {16}



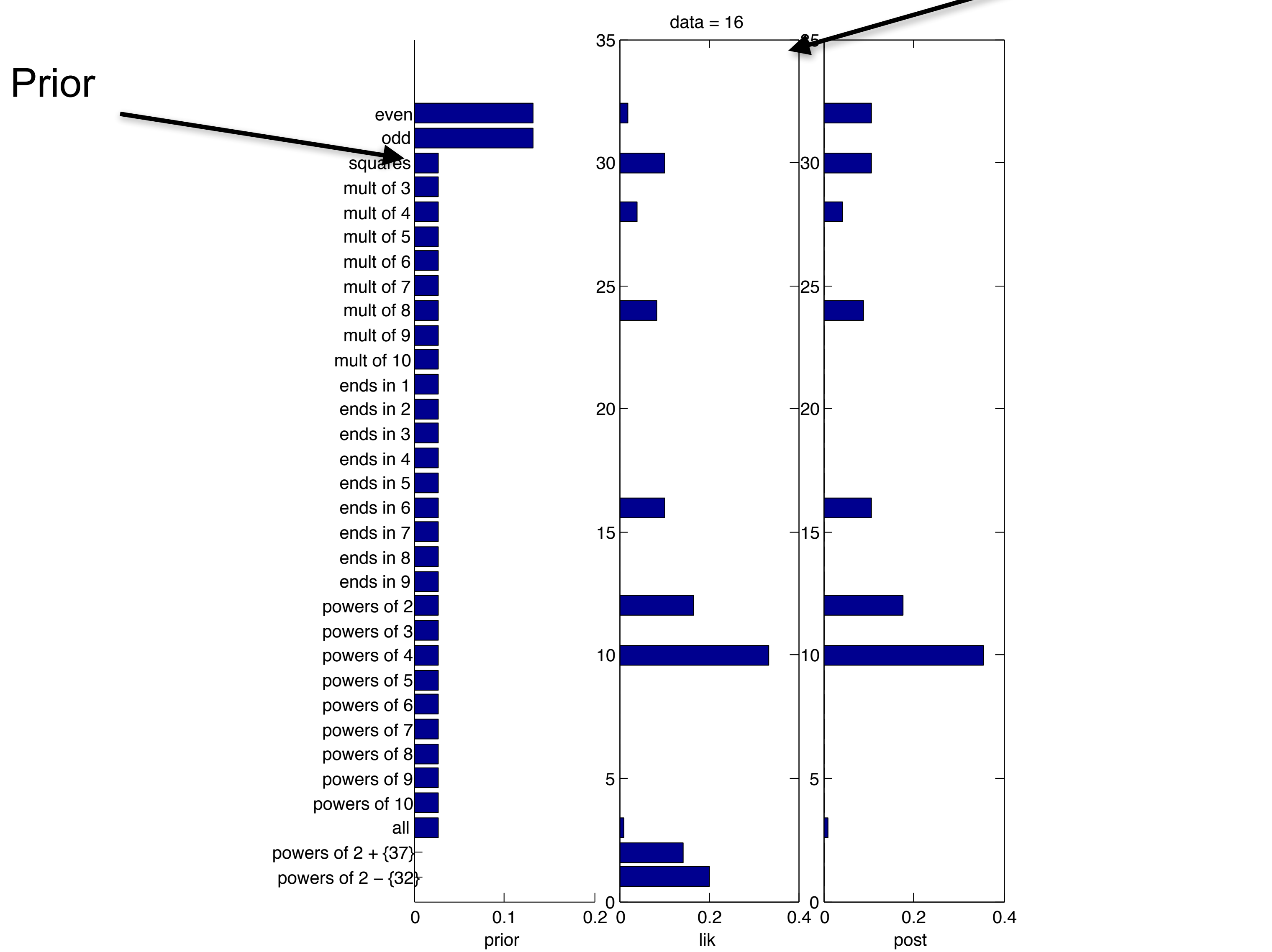
data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 − {32}
0
0.1
0.2
prior
35
30
25
20
15
10
5
0
0.2
0.4
lik
post

Data = {16}

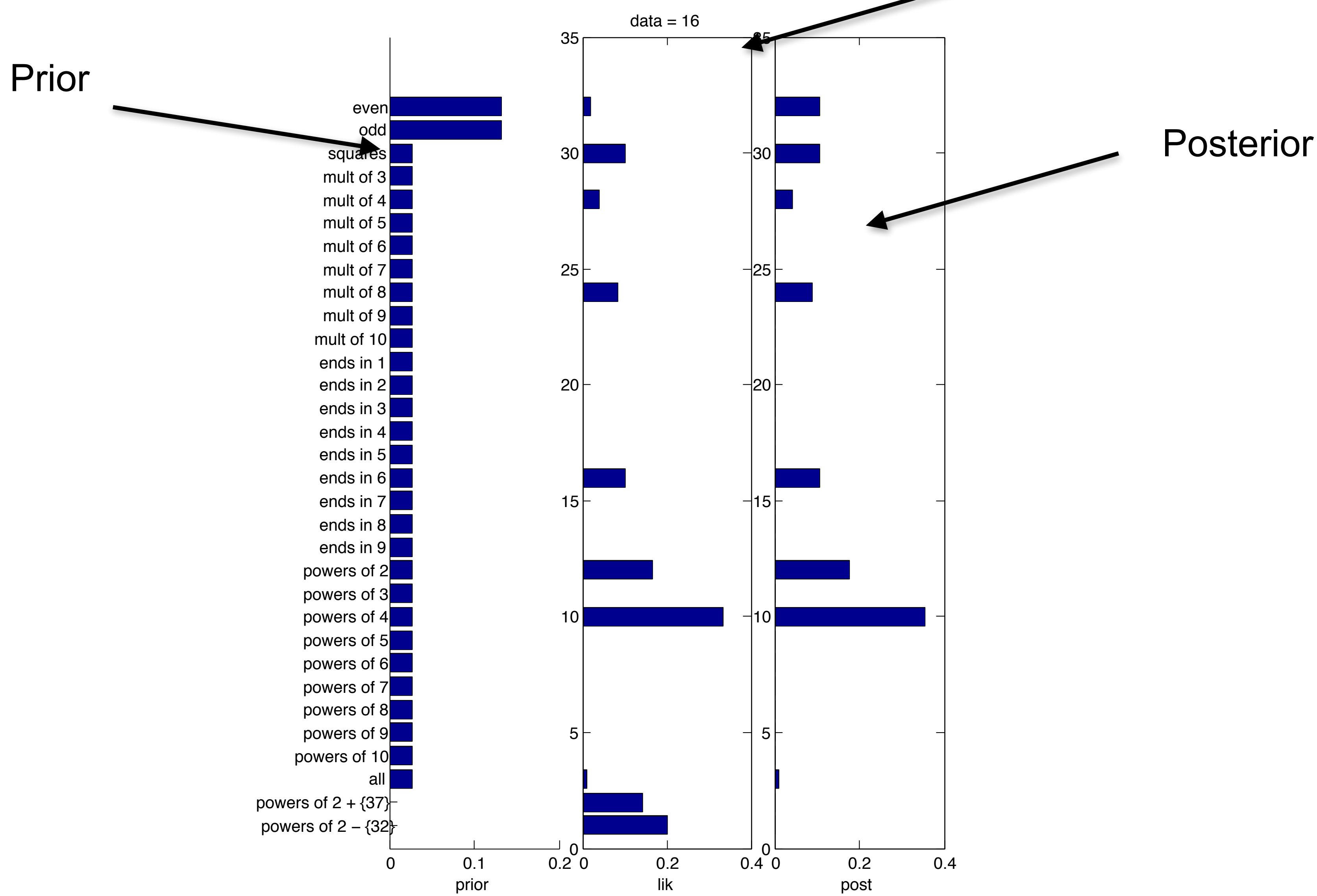
Prior



Data = {16}

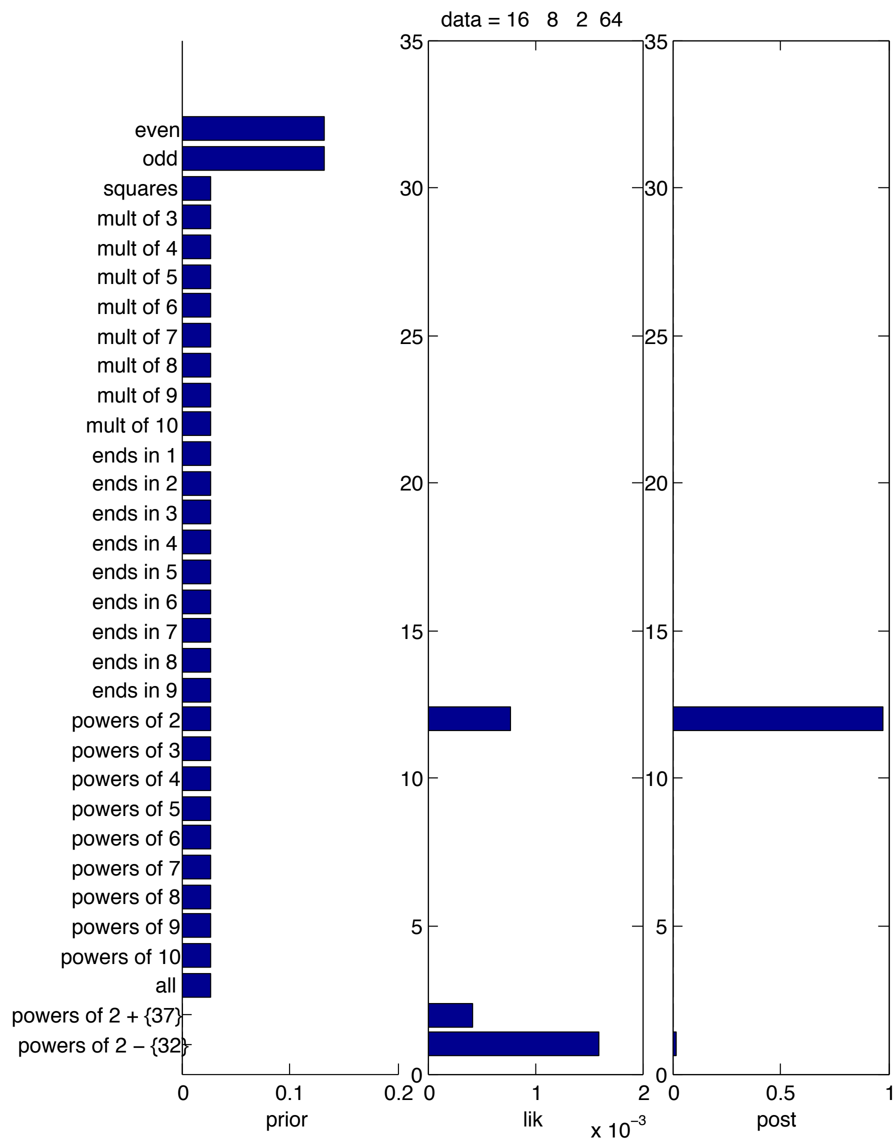

Prior
Likelihood
data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 – {32}
prior
lik
post

Data = {16}



Prior
Likelihood
Posterior
data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 – {32}
0 0.1 0.2
prior
0 0.2 0.4
lik
0 0.2 0.4
post
35 30 25 20 15 10 5 0

Data = {16,8,2,64}



data = 16 8 2 64
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 − {32}
0 0.1 0.2
prior
0 1 2
lik
x 10−3
0 0.5 1
post
0 5 10 15 20 25 30 35

MAP Hypothesis

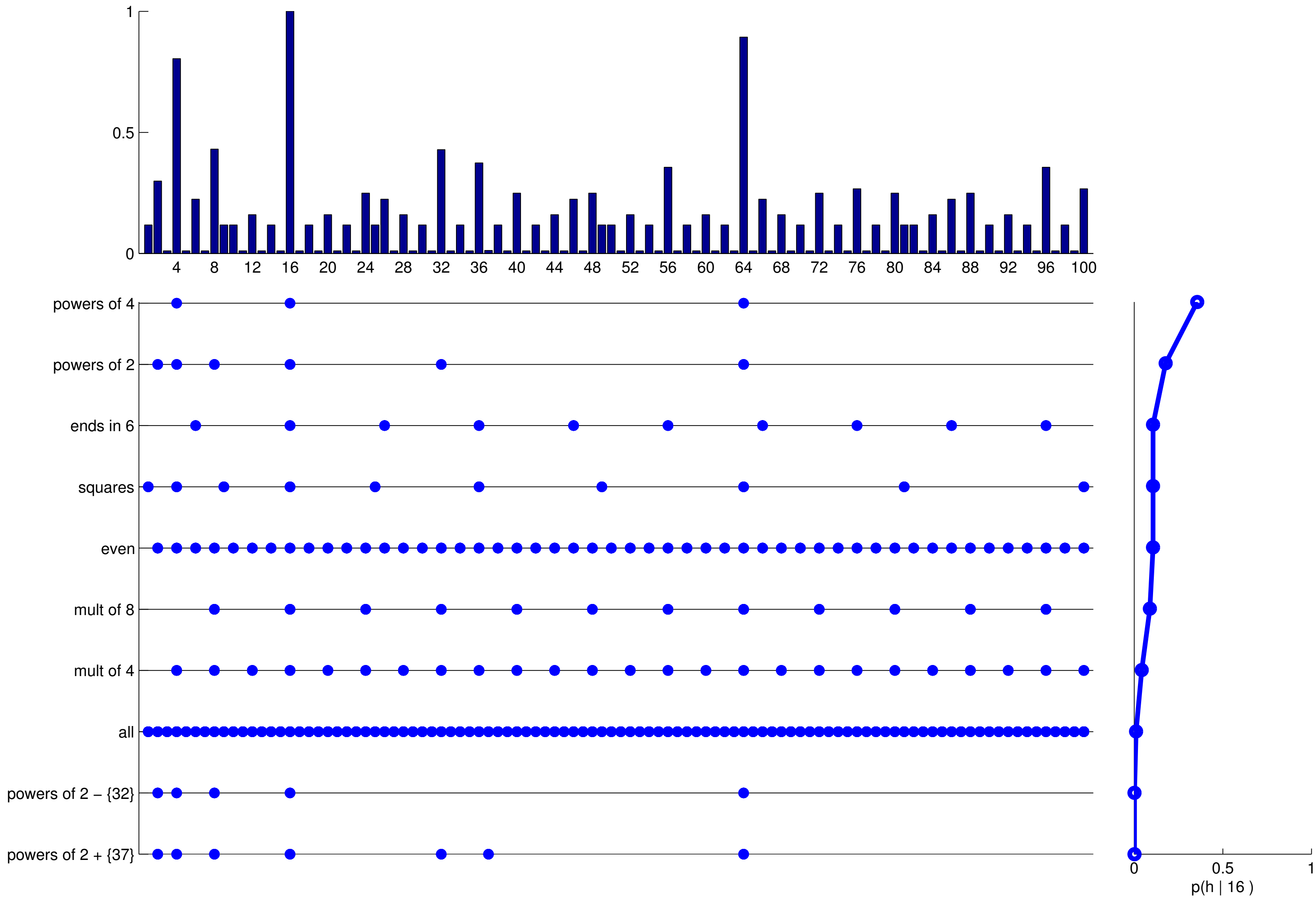
$$h^{MAP} = \arg\max_h \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg\max_h P(D|h)P(h)$$

Posterior Predictive Distribution

$$P(x \in C|D) = \sum_h P(x \in C|h) \; P(h|D)$$

Bayes Model Averaging



Parameter Estimation

- How to estimate parameters from data?

Parameter Estimation

- How to estimate parameters from data?

Maximum Likelihood Principle:

Choose the parameters that maximize the probability of the observed data!

Maximum Likelihood Estimation Recipe

1. Use the log-likelihood
2. Differentiate with respect to the parameters
3. *Equate to zero and solve

*Often requires numerical approximation (no closed form solution)

An Example

- Let's start with the simplest possible case
  - Single observed variable
  - Flipping a bent coin

An Example

- Let's start with the simplest possible case
  - Single observed variable
  - Flipping a bent coin

- We Observe:
  - Sequence of heads or tails
  - HTTTTTHTHT
- Goal:
  - Estimate the probability that the next flip comes up heads

Assumptions

- Fixed parameter θ_H
  - Probability that a flip comes up heads
- Each flip is independent
  - Doesn’t affect the outcome of other flips
- (IID) Independent and Identically Distributed

Example

- Let’s assume we observe the sequence:
  - HTTTTTHTHT
- What is the **best** value of θ_H ?
  - Probability of heads

Example

- Let's assume we observe the sequence:
  - HTTTTTHTHT
- What is the **best** value of θ_H ?
  - Probability of heads
- Intuition: should be 0.3 (3 out of 10)
- Question: how do we justify this?

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

$$P(x_1 = H, x_2 = T, \ldots, x_m = T; \theta_H)$$

$$= P(x_1 = H; \theta_H) P(x_2 = T; \theta_H), \ldots P(x_m = T; \theta_H)$$

$$= \theta_H \times (1 - \theta_H), \times \ldots \times \theta_H$$

$$= \theta_H^3 \times (1 - \theta_H)^7$$

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

$$P(x_1 = H, x_2 = T, \ldots, x_m = T; \theta_H)$$

$$= P(x_1 = H; \theta_H)P(x_2 = T; \theta_H), \ldots P(x_m = T; \theta_H)$$

$$= \theta_H \times (1 - \theta_H), \times \ldots \times \theta_H$$

$$= \theta_H^3 \times (1 - \theta_H)^7$$

This is the Likelihood Function

Maximum Likelihood Principle

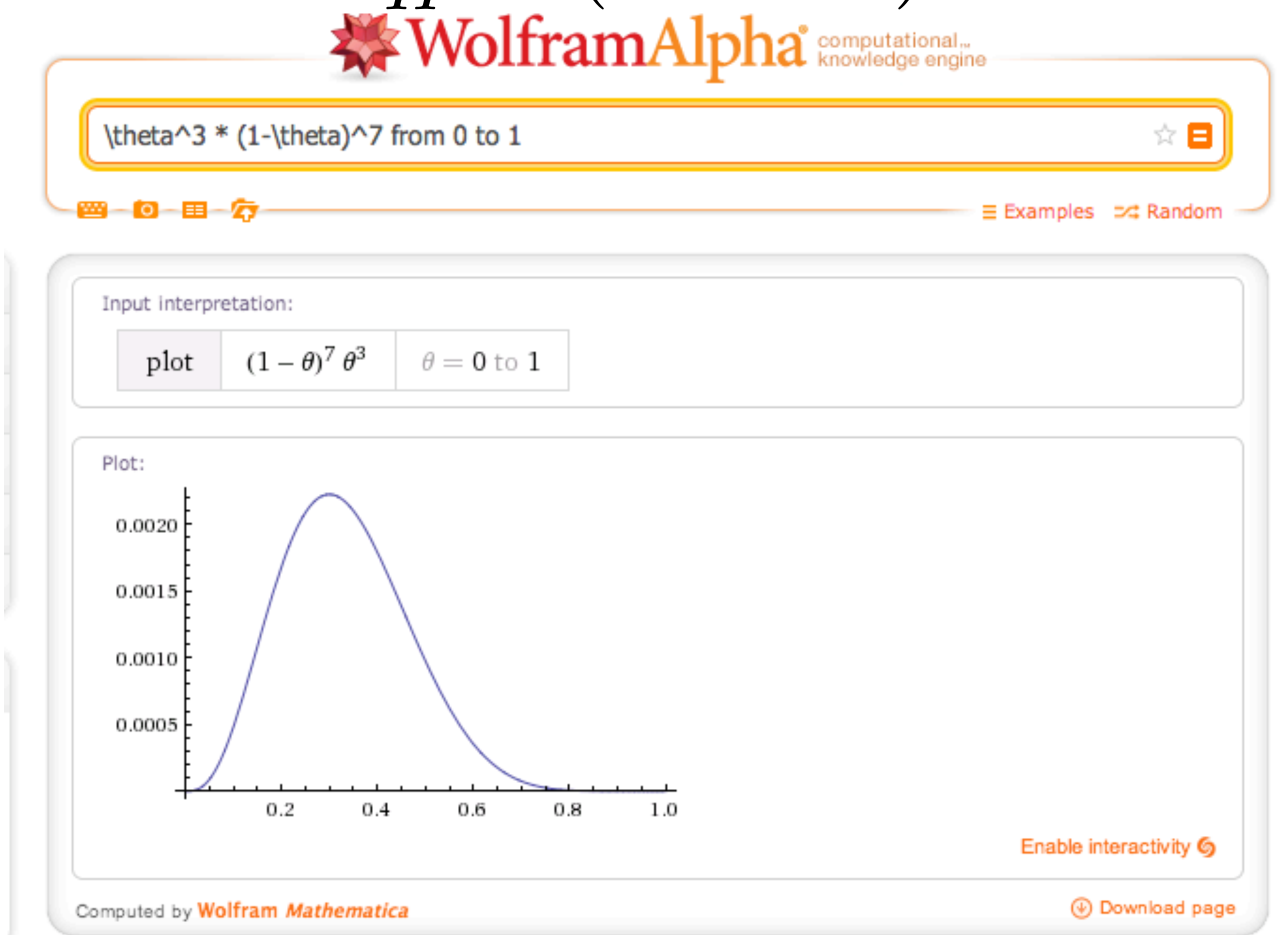
- Probability of “HTTTTTHTHT” as a function of θ_H

$$\theta_H^3 \times (1 - \theta_H)^7$$

Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

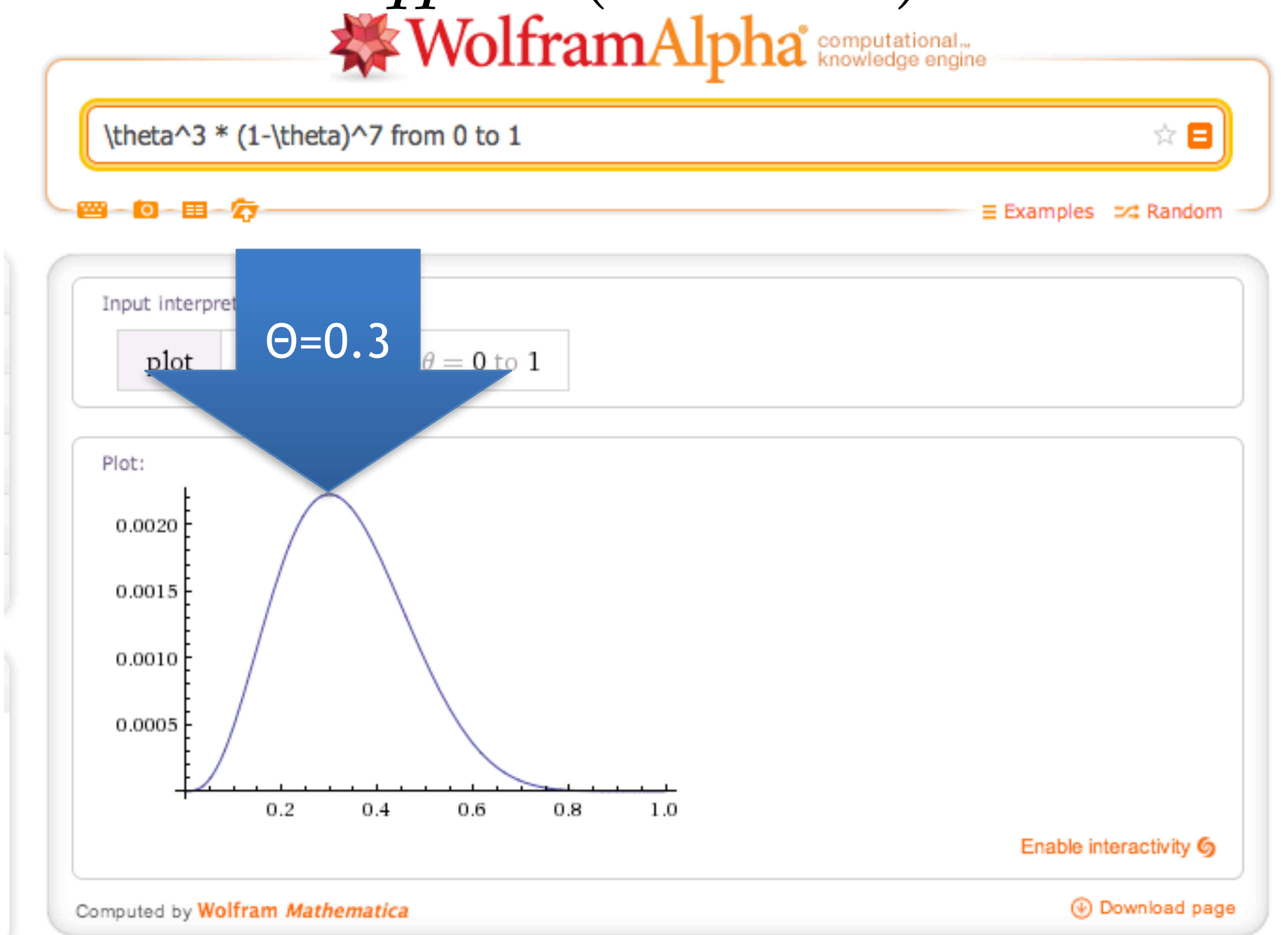
$$\theta_H^3 \times (1 - \theta_H)^7$$



Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

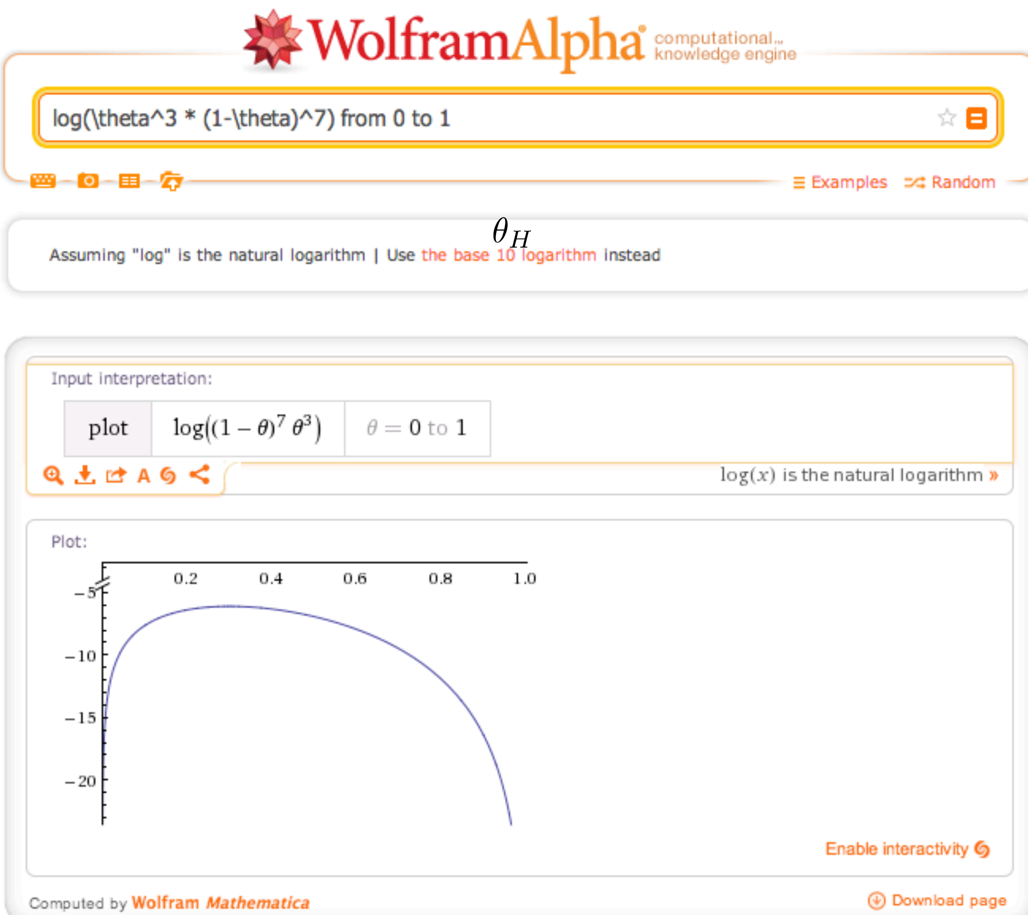
$$\theta_H^3 \times (1 - \theta_H)^7$$



Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

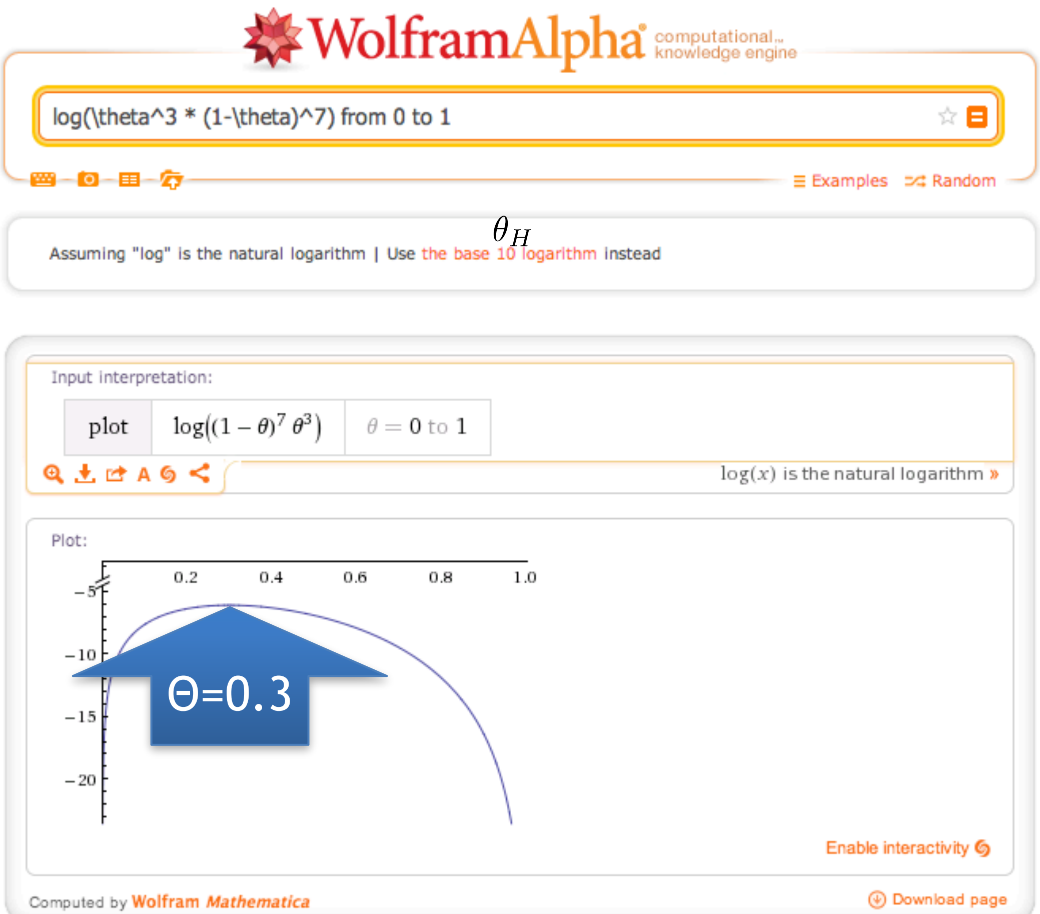
$$\log(\theta_H^3 \times (1-\theta_H)^7)$$



Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

$$\log(\theta_H^3 \times (1 - \theta_H)^7)$$



Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \log(\theta_H^{\#H}(1-\theta_H)^{\#T}) = 0$$

$$\frac{\partial}{\partial \theta_H} \log(\theta_H^{\#H}) + \log((1-\theta_H)^{\#T}) = 0$$

Log Identities

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1-\theta_H) = 0$$

Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1 - \theta_H) = 0$$

$$\frac{\#H}{\theta_H} - \frac{\#T}{1 - \theta_H} = 0$$

$$\hat{\theta} = \frac{\#H}{\#H + \#T}$$

Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1-\theta_H) = 0$$

$$\frac{\#H}{\theta_H} - \frac{\#T}{1-\theta_H} = 0$$

$$\vdots$$

$$\hat{\theta} = \frac{\#H}{\#H + \#T}$$

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.
  - No way to incorporate our prior knowledge!

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.
  - No way to incorporate our prior knowledge!

Q: how to deal with this problem?

Bayesian Parameter Estimation

- Let's just treat θ_H like any other variable
- Put a prior on it!
  - Encode our prior knowledge about possible values of θ_H using a probability distribution
- Now consider two probability distributions:

$$P(x_i|\theta_H) = \begin{cases} \theta_H, & \text{if } x_i = H \\ 1 - \theta_H, & \text{otherwise} \end{cases}$$

$$P(\theta_H) = ?$$

Posterior Over θ_H

$$P(\theta | x_1 = H, x_2 = T, \ldots, x_m = T)$$

Posterior Over θ_H

$$P(\theta | x_1 = H, x_2 = T, \ldots, x_m = T)$$

$$= \frac{P(x_1 = H, x_2 = T, \ldots, x_m = T | \theta) P(\theta)}{P(x_1 = H, x_2 = T, \ldots, x_m = T)}$$

Posterior Over θ_H

$$P(\theta|x_1 = H, x_2 = T, \ldots, x_m = T)$$

$$= \frac{P(x_1 = H, x_2 = T, \ldots, x_m = T|\theta)P(\theta)}{P(x_1 = H, x_2 = T, \ldots, x_m = T)}$$

$$= \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

Hyper-Parameters

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

Hyper-Parameters

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

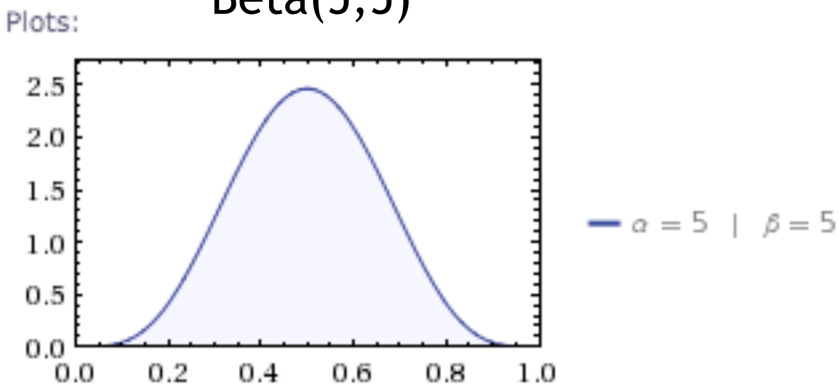
Hyper-Parameters

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

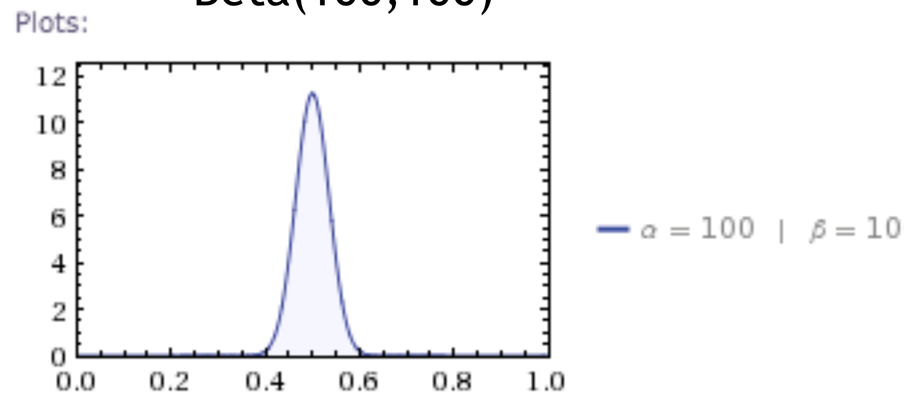
Gamma is a continuous generalization of the Factorial Function

Beta Distribution

Beta(5,5)

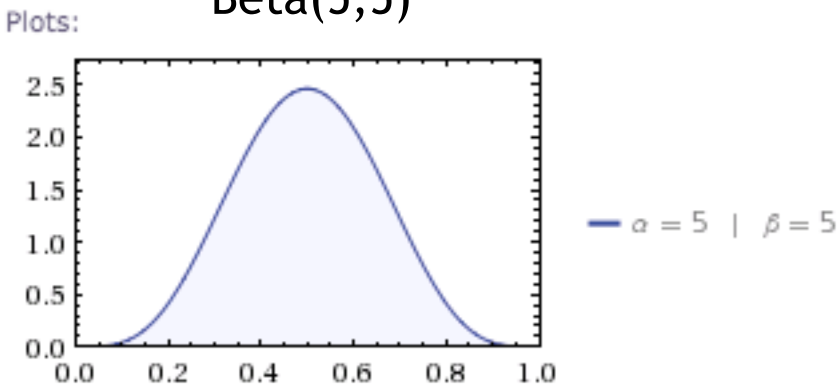


Beta(100,100)

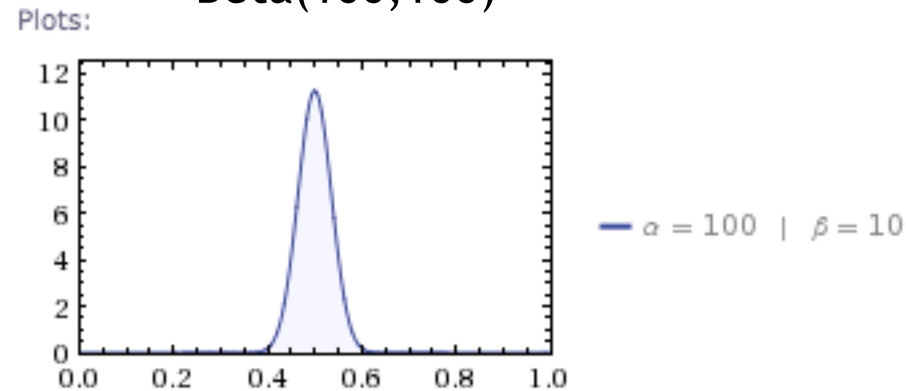


Beta Distribution

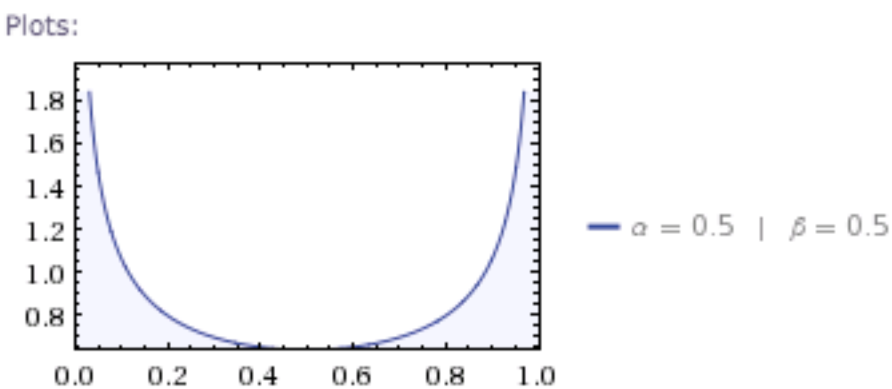
Beta(5,5)



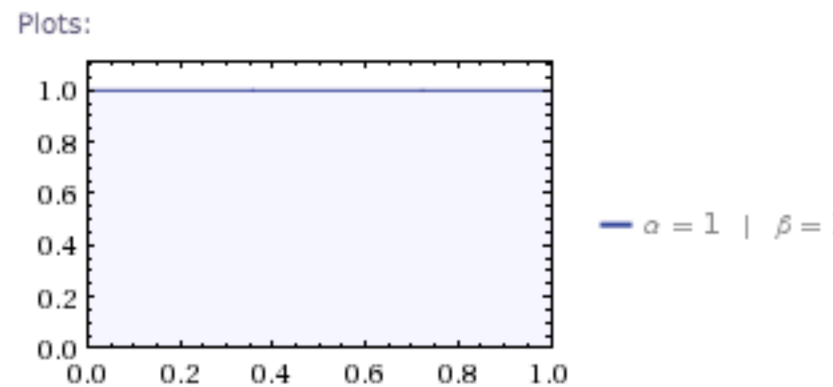
Beta(100,100)



Beta(0.5,0.5)



Beta(1,1)



MAP Estimate

$$\theta^{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \frac{\#H + \alpha - 1}{\#T + \#H + \alpha + \beta - 2}$$

MAP Estimate

$$\theta^{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \frac{\#H + \alpha - 1}{\#T + \#H + \alpha + \beta - 2}$$

-Add-N smoothing
-Pseudo-counts

Marginal Probability over single Toss

$$P(x_1 = H|\alpha, \beta)$$

$$= \int P(x_1 = H|\theta_H)P(\theta_H|\alpha, \beta)d\theta_H$$

$$= \int \theta P(\theta_H|\alpha, \beta)d\theta_H$$

Marginal Probability over single Toss

$$P(x_1 = H|\alpha, \beta)$$

$$= \int P(x_1 = H|\theta_H)P(\theta_H|\alpha, \beta)d\theta_H$$

$$= \int \theta P(\theta_H|\alpha, \beta)d\theta_H$$

$$\vdots$$

$$= \frac{\alpha}{\alpha + \beta}$$

Marginal Probability over single Toss

$$P(x_1 = H|\alpha, \beta)$$

$$= \int P(x_1 = H|\theta_H)P(\theta_H|\alpha, \beta)d\theta_H$$

$$= \int \theta P(\theta_H|\alpha, \beta)d\theta_H$$

$$\vdots$$

$$= \frac{\alpha}{\alpha + \beta}$$

Beta prior indicates α imaginary heads and β imaginary tails

More than one toss

$$P(\theta_H|x_1,\ldots,x_m) \propto P(x_1,\ldots,x_m|\theta)P(\theta|\alpha,\beta)$$

$$\propto \theta_H^{\#H}(1-\theta_H)^{\#T}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

$$= \theta_H^{\#H+\alpha-1}(1-\theta_H)^{\#T+\beta-1}$$

More than one toss

$$P(\theta_H | x_1, \ldots, x_m) \propto P(x_1, \ldots, x_m | \theta) P(\theta | \alpha, \beta)$$

$$\propto \theta_H^{\#H} (1 - \theta_H)^{\#T} \theta_H^{\alpha - 1} (1 - \theta_H)^{\beta - 1}$$

$$= \theta_H^{\#H + \alpha - 1} (1 - \theta_H)^{\#T + \beta - 1}$$

$$= Beta(\#H + \alpha, \#T + \beta)$$

More than one toss

$$P(\theta_H|x_1,\ldots,x_m) \propto P(x_1,\ldots,x_m|\theta)P(\theta|\alpha,\beta)$$

$$\propto \theta_H^{\#H}(1-\theta_H)^{\#T}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

$$= \theta_H^{\#H+\alpha-1}(1-\theta_H)^{\#T+\beta-1}$$

$$= Beta(\#H+\alpha, \#T+\beta)$$

- If the prior is Beta, so is posterior!

More than one toss

$$P(\theta_H | x_1, \ldots, x_m) \propto P(x_1, \ldots, x_m | \theta) P(\theta | \alpha, \beta)$$

$$\propto \theta_H^{\#H} (1 - \theta_H)^{\#T} \theta_H^{\alpha - 1} (1 - \theta_H)^{\beta - 1}$$

$$= \theta_H^{\#H + \alpha - 1} (1 - \theta_H)^{\#T + \beta - 1}$$

$$= Beta(\#H + \alpha, \#T + \beta)$$

- If the prior is Beta, so is posterior!
- Beta is **conjugate** to the Bernoulli likelihood

Main Takeaways

- Joint distribution encodes everything
  - But, hopeless except with very small # of variables
- Using Bayes Rule for Parameter Estimation
  - Model parameters are just like any other variables
  - Add-N smoothing