Artificial Intelligence
15-381

Mar 22, 2007

Probability and Uncertainty 2:
Probabilistic Reasoning

Review of concepts from last lecture

Making rational decisions when faced with uncertainty:

- *Probability*

  the precise representation of knowledge and uncertainty

- *Probability theory*

  how to optimally update your knowledge based on new information

- *Decision theory: probability theory + utility theory*

  how to use this information to achieve maximum expected utility

Basic concepts

- random variables
- probability distributions (discrete) and probability densities (continuous)
- rules of probability
- expectation and the computation of 1st and 2nd moments
- joint and multivariate probability distributions and densities
- covariance and principal components

Simple example: medical test results

- Test report for rare disease is positive, 90% accurate
- What's the probability that you have the disease?
- What if the test is repeated?

- This is the simplest example of reasoning by combining sources of information.

How do we model the problem?

- Which is the correct description of "Test is 90% accurate" ?

$$\begin{aligned} P(T = \text{true}) &= 0.9 \\ P(T = \text{true}|D = \text{true}) &= 0.9 \\ P(D = \text{true}|T = \text{true}) &= 0.9 \end{aligned}$$

- What do we want to know?

$$\begin{aligned} &P(T = \text{true}) \\ &P(T = \text{true}|D = \text{true}) \\ &P(D = \text{true}|T = \text{true}) \end{aligned}$$

- More compact notation:

$$\begin{aligned} P(T = \text{true}|D = \text{true}) &\rightarrow P(T|D) \\ P(T = \text{false}|D = \text{false}) &\rightarrow P(\bar{T}|\bar{D}) \end{aligned}$$

Evaluating the posterior probability through Bayesian inference

- We want P(D|T) = "The probability of the having the disease given a positive test"
- Use Bayes rule to relate it to what we know: P(T|D)

$$\underbrace{P(D|T)}_{\textit{posterior}} = \frac{\overbrace{P(T|D)}^{\textit{likelihood}}\overbrace{P(D)}^{\textit{prior}}}{\underbrace{P(T)}_{\textit{normalizing constant}}}$$

- What's the prior P(D)?
- Disease is rare, so let's assume

$$P(D) = 0.001$$

- What about P(T)?
- What's the interpretation of that?

Evaluating the normalizing constant

$$\underbrace{P(D|T)}_{\textit{posterior}} = \frac{\overbrace{P(T|D)}^{\textit{likelihood}}\overbrace{P(D)}^{\textit{prior}}}{\underbrace{P(T)}_{\textit{normalizing constant}}}$$

- P(T) is the marginal probability of P(T,D) = P(T|D) P(D)
- So, compute with summation

$$P(T) = \sum_{\text{all values of D}} P(T|D)P(D)$$

- For true or false propositions:

$$P(T) = P(T|D)P(D) + P(T|\bar{D})P(\bar{D})$$

What are these?

Refining our model of the test

- We also have to consider the negative case to incorporate all information:

$$\begin{aligned} P(T|D) &= 0.9 \\ P(T|\bar{D}) &= ? \end{aligned}$$

- What should it be?

- What about $P(\bar{D})$?

Plugging in the numbers

- Our complete expression is

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

- Plugging in the numbers we get:

$$P(D|T) = \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.1 \times 0.999} = 0.0089$$

- Does this make intuitive sense?

Same problem different situation

- Suppose we have a test to determine if you won the lottery.
- It's 90% accurate.
- What is P($ = true | T = true) then?

Playing around with the numbers

$$P(D|T) \quad = \quad \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

- What if the test were 100% reliable?

$$P(D|T) = \frac{1.0 \times 0.001}{1.0 \times 0.001 + 0.0 \times 0.999} = 1.0$$

- What if the test was the same, but disease wasn't so rare?

$$P(D|T) = \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.1 \times 0.999} = 0.5$$

Repeating the test

- We can relax, P(D|T) = 0.0089, right?
- Just to be sure the doctor recommends repeating the test.
- How do we represent this?

$$P(D|T_1, T_2)$$

- Again, we apply Bayes' rule

$$P(D|T_1, T_2) = \frac{P(T_1, T_2|D)P(D)}{P(T_1, T_2)}$$

- How do we model $P(T_1,T_2|D)$?

Modeling repeated tests

$$P(D|T_1, T_2) = \frac{P(T_1, T_2|D)P(D)}{P(T_1, T_2)}$$

- Easiest is to assume the tests are *independent*.

$$P(T_1, T_2|D) = P(T1|D)P(T_2|D)$$

- This also implies:

$$P(T_1, T_2) = P(T1)P(T_2)$$

- Plugging these in, we have

$$P(D|T_1, T_2) = \frac{P(T_1|D)P(T_2|D)P(D)}{P(T_1)P(T_2)}$$

Evaluating the normalizing constant again

- Expanding as before we have

$$P(D|T_1, T_2) = \frac{P(T_1|D)P(T_2|D)P(D)}{\sum_{D=\{t,f\}} P(T_1|D)P(T_2|D)P(D)}$$

- Plugging in the numbers gives us

$$P(D|T) = \frac{0.9 \times 0.9 \times 0.001}{0.9 \times 0.9 \times 0.001 + 0.1 \times 0.1 \times 0.999} = 0.075$$

- Another way to think about this:
  - What's the chance of 1 false positive from the test?
  - What's the chance of 2 false positives?

- The chance of 2 false positives is still 10x more likely than the a prior probability of having the disease.

Simpler: Combining information the Bayesian way

- Let's look at the equation again:

$$P(D|T_1, T_2) = \frac{P(T_1|D)P(T_2|D)P(D)}{P(T_1)P(T_2)}$$

- If we rearrange slightly:

$$P(D|T_1, T_2) = \frac{P(T_2|D)P(T_1|D)P(D)}{P(T_2)P(T_1)}$$

We've seen this before!

- It's the posterior for the first test, which we just computed

$$P(D|T_1) = \frac{P(T_1|D)P(D)}{P(T_1)}$$

The old posterior is the new prior

- We can just plugin the value of the old posterior
- It plays exactly the same role as our old prior

$$P(D|T_1,T_2) = \frac{P(T_2|D)P(T_1|D)P(D)}{P(T_2)P(T_1)}$$

$$P(D|T_1,T2) = \frac{P(T_2|D) \times 0.0089}{P(T_2)}$$

This is how Bayesian reasoning combines old information with new information to update our belief states.

- Plugging in the numbers gives the same answer:

$$P(D|T) \quad = \quad \frac{P(T|D)P'(D)}{P(T|D)P'(D) + P(T|\bar{D})P'(\bar{D})}$$

$$P(D|T) = \frac{0.9 \times 0.0089}{0.9 \times 0.0089 + 0.1 \times 0.9911} = 0.075$$

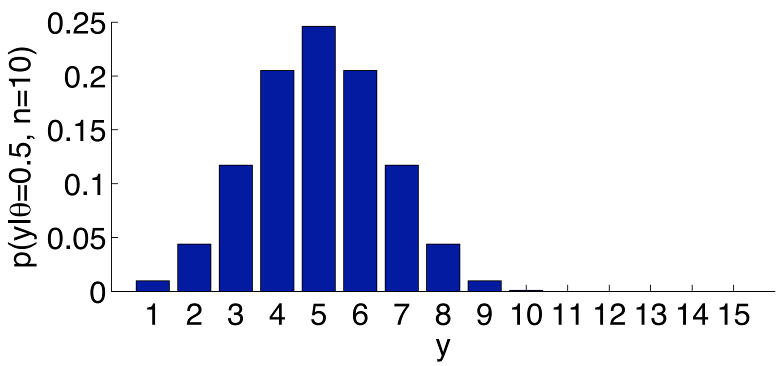
Bayesian inference for distributions

- The simplest case is true or false propositions
- The basic computations are the same for distributions

An example with distributions: coin flipping

- In Bernoulli trials, each sample is either 1 (e.g. heads) with probability θ, or 0 (tails) with probability 1 − θ.
- The binomial distribution specifies the probability of the total # of heads, y, out of n trials:

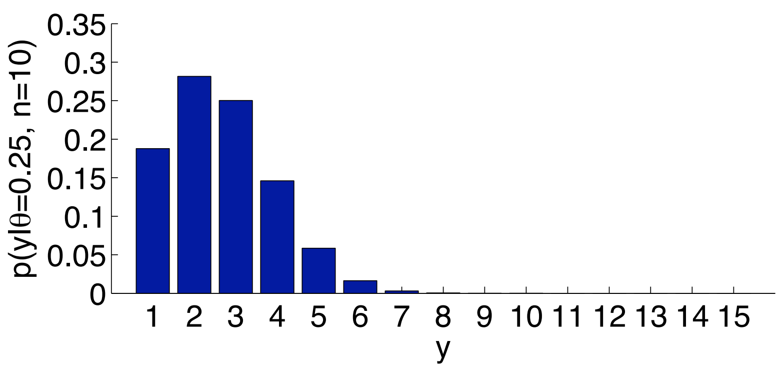
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



The binomial distribution

- In Bernoulli trials, each sample is either 1 (e.g. heads) with probability θ, or 0 (tails) with probability 1 − θ.
- The binomial distribution specifies the probability of the total # of heads, y, out of n trials:

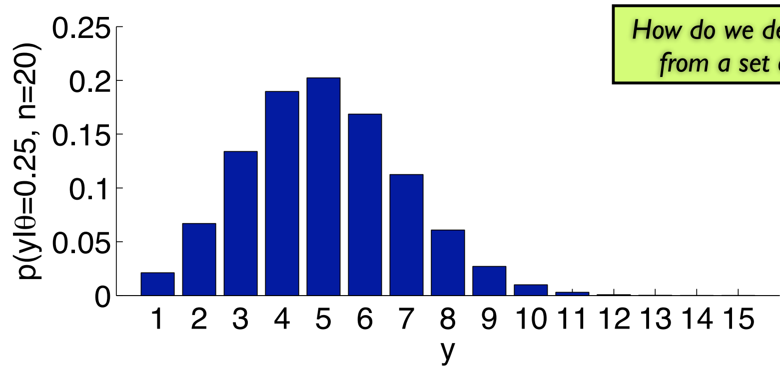
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



The binomial distribution

- In Bernoulli trials, each sample is either 1 (e.g. heads) with probability θ, or 0 (tails) with probability 1 − θ.
- The binomial distribution specifies the probability of the total # of heads, y, out of n trials:

$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



How do we determine θ from a set of trials?

Applying Bayes' rule

- Given n trials with k heads, what do we know about θ?
- We can apply Bayes' rule to see how our knowledge changes as we acquire new observations:

$$\underbrace{p(\theta|y,n)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta,n)}^{\text{likelihood}}\overbrace{p(\theta|n)}^{\text{prior}}}{\underbrace{p(y|n)}_{\text{normalizing constant}}} \quad p(y|n) = \int p(y|\theta,n)p(\theta|n)d\theta$$

- We know the likelihood, what about the prior?
- Uniform on [0, 1] is a reasonable assumption, i.e. "we don't know anything".
- What is the form of the posterior?
- In this case, the posterior is just proportional to the likelihood:

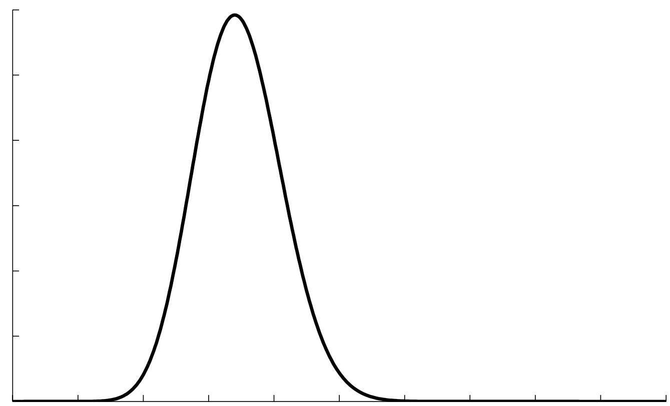
$$p(\theta|y,n) \propto \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Updating our knowledge with new information

- Now we can evaluate the poster just by plugging in different values of y and n.

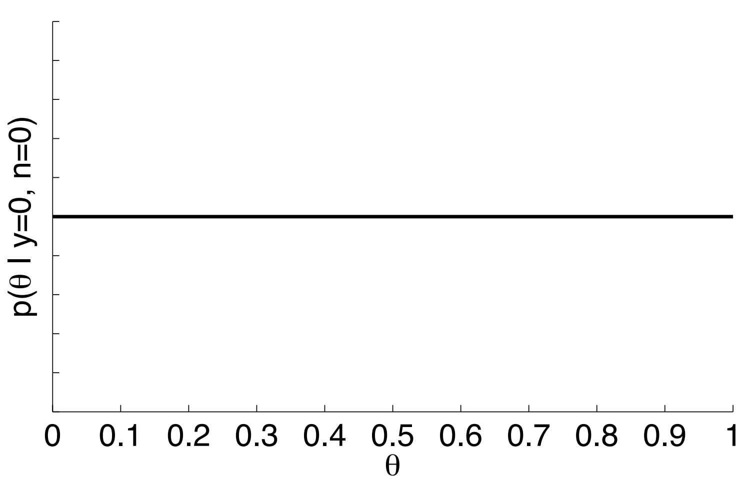
$$p(\theta|y,n) \propto \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Check: What goes on the axes?



Evaluating the posterior

- What do we know initially, before observing any trials?

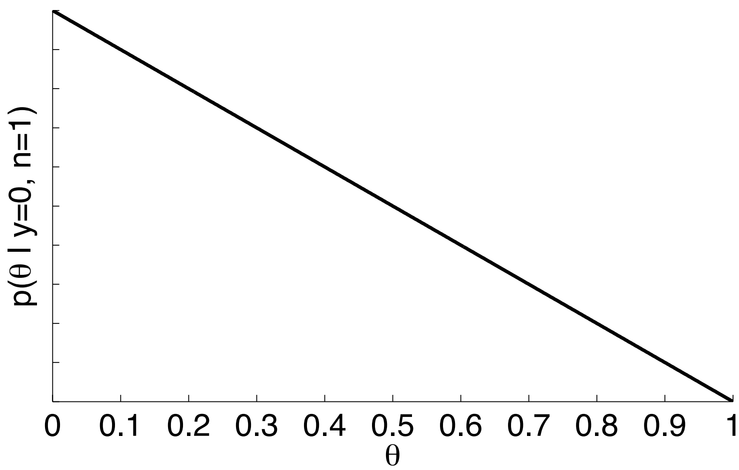


Coin tossing

- What is our belief about θ after observing one "tail" ? *How would you bet?*

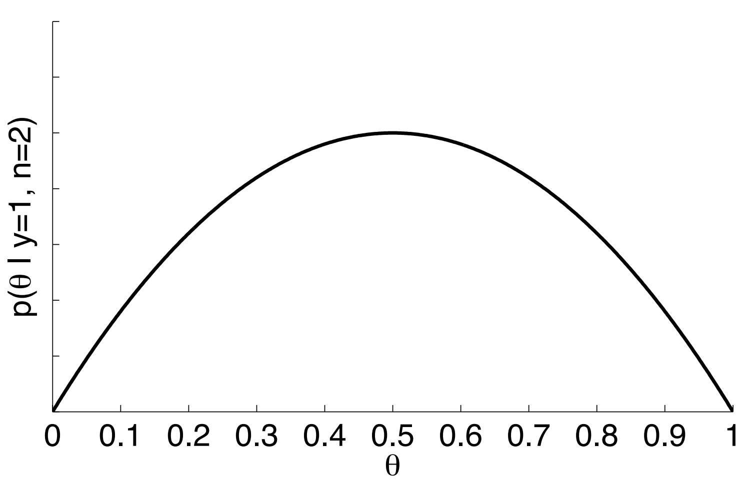
Is the p(θ >0.5) less or greater than 0.5?

What about p(θ >0.3)?



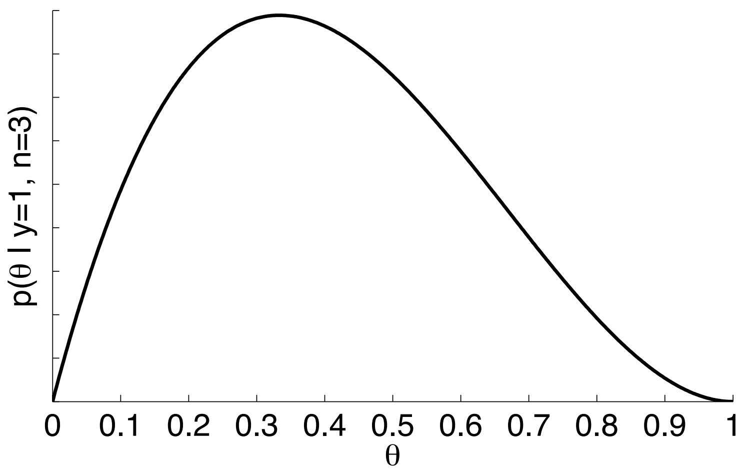
Coin tossing

- Now after two trials we observe 1 head and 1 tail.



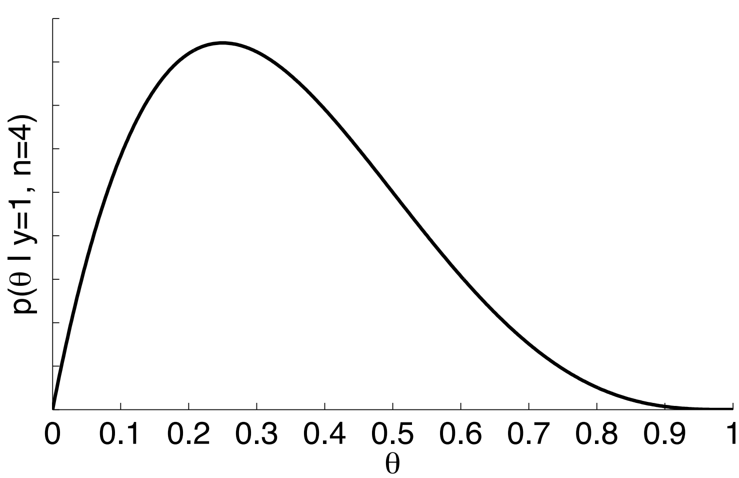
Coin tossing

- 3 trials: 1 head and 2 tails.



Coin tossing

- 4 trials: 1 head and 3 tails.



Coin tossing

- 5 trials: 1 head and 4 tails.

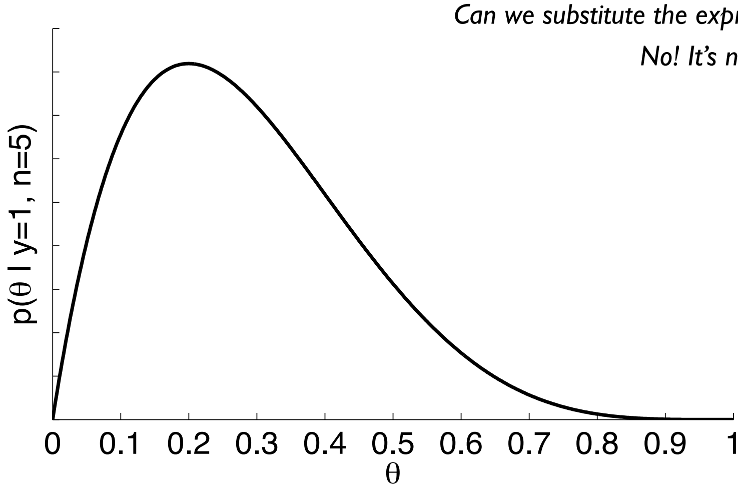
Do we have good evidence that this coin is biased?

How would you quantify this statement?

$$p(\theta > 0.5) = \int_{0.5}^{1.0} p(\theta|y,n)d\theta$$

Can we substitute the expression above?

No! It's not normalized.



Evaluating the normalizing constant

- To get proper probability density functions, we need to evaluate p(y|n):

$$p(\theta|y,n) = \frac{p(y|\theta,n)p(\theta|n)}{p(y|n)}$$

- Bayes in his original paper in 1763 showed that:

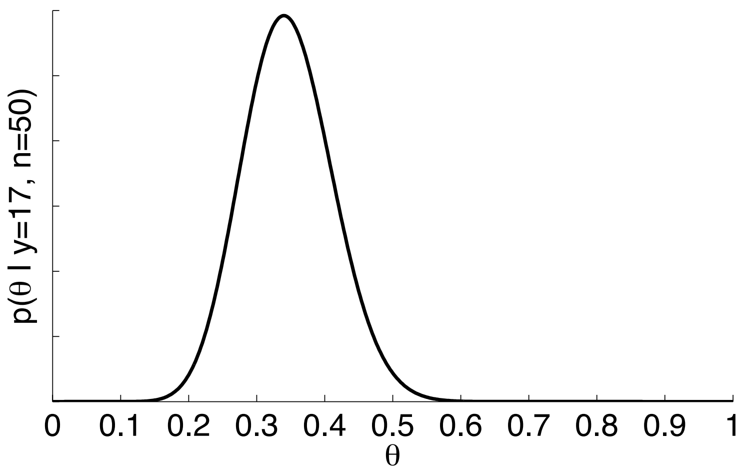
$$\begin{aligned} p(y|n) &= \int_0^1 p(y|\theta,n)p(\theta|n)d\theta \\ &= \frac{1}{n+1} \end{aligned}$$

$$\Rightarrow p(\theta|y,n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}(n+1)$$

More coin tossing

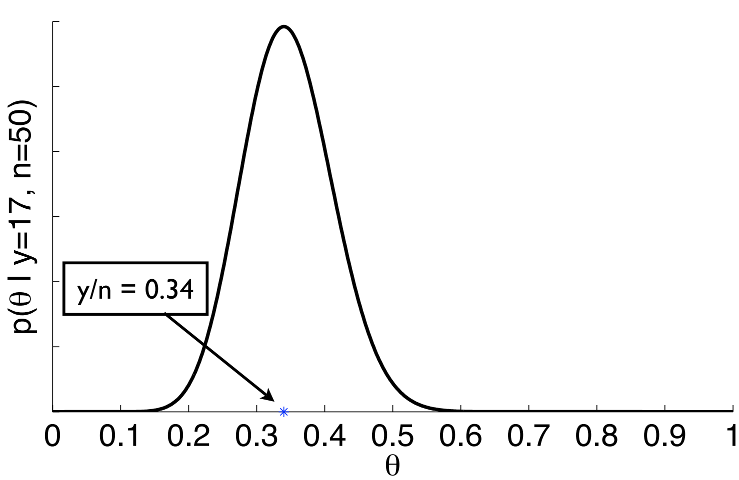
- After 50 trials: 17 heads and 33 tails.
- There are many possibilities.

What's a good estimate of θ?



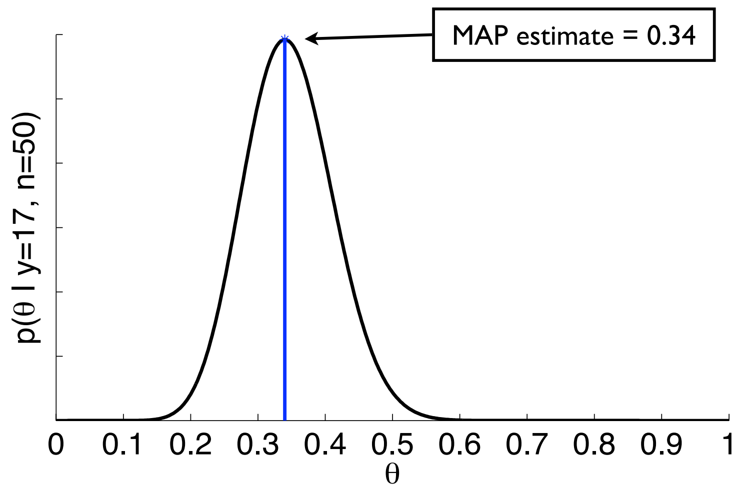
A ratio estimate

- Intuitive estimate: just take ratio θ = 17/50 = 0.34



The maximum a posteriori (MAP) estimate

- This just picks the location of maximum value of the posterior
- In this case, maximum is also at $\theta = 0.34$.

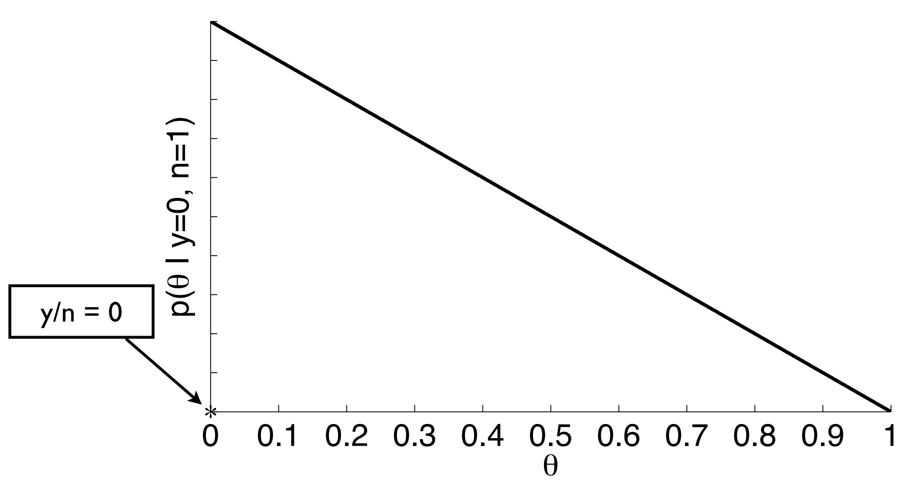


A different case

- What about after just one trial: 0 heads and 1 tail?
- MAP and ratio estimate would say 0.

Does this make sense?

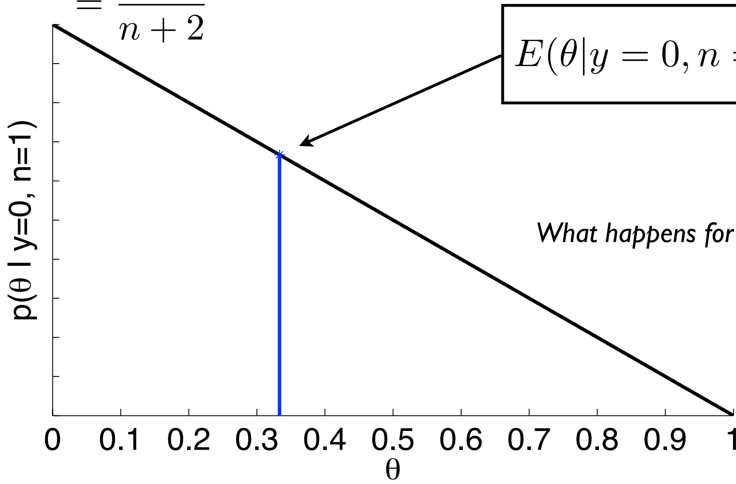
- What would a better estimate be?



The expected value estimate

- We defined the expected value of a pdf in the previous lecture:

$$E(\theta|y,n) = \int_0^1 \theta p(\theta|y,n)d\theta$$
$$= \frac{y+1}{n+2}$$



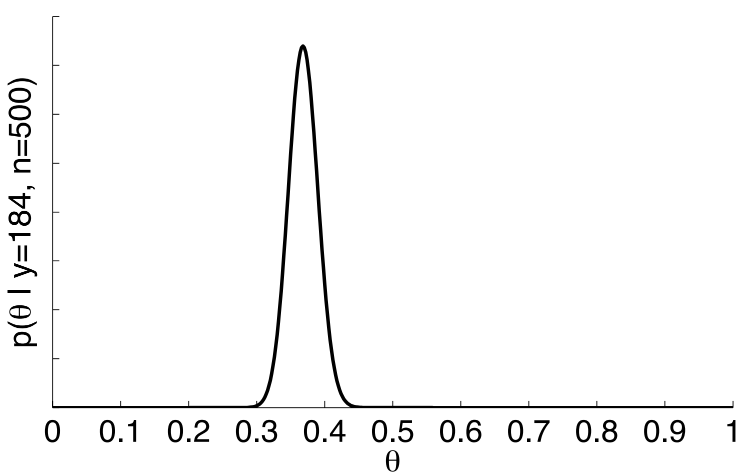
$$\boxed{E(\theta|y=0, n=1) = \frac{1}{3}}$$

What happens for zero trials?

Much more coin tossing

- After 500 trials: 184 heads and 316 tails.

What's your guess of θ?



Much more coin tossing

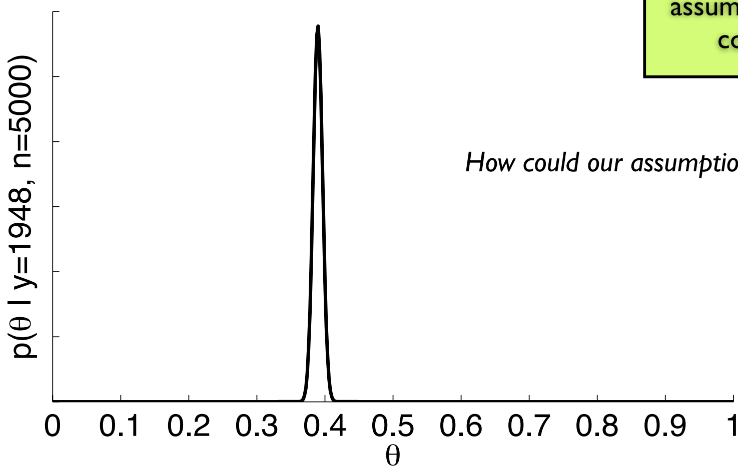
- After 5000 trials: 1948 heads and 3052 tails.
- Posterior contains true estimate.

True value is 0.4.

Is this always the case?

NO! Only if the assumptions are correct.

How could our assumptions be wrong?

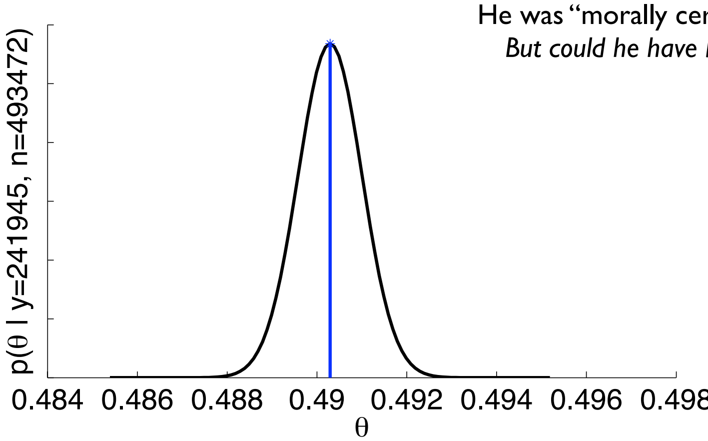


Laplace's example: proportion female births

- A total of 241,945 girls and 251,527 boys were born in Paris from 1745-1770.
- Laplace was able to evaluate the following

$$p(\theta > 0.5) = \int_{0.5}^{1.0} p(\theta|y,n)d\theta \approx 1.15 \times 10^{-42}$$

He was "morally certain" θ < 0.5.
But could he have been wrong?



Laplace and the mass of Saturn

- Laplace used "Bayesian" inference to estimate the mass of Saturn and other planets. For Saturn he said:

  It is a bet of 11000 to 1 that the error in this result is not within 1/100th of its value

Mass of Saturn as a fraction of the mass of the Sun	
Laplace (1815)	NASA (2004)
3512	3499.1

(3512 - 3499.1) / 3499.1 = 0.0037

Laplace is still wining.

Applying Bayes' rule with an informative prior

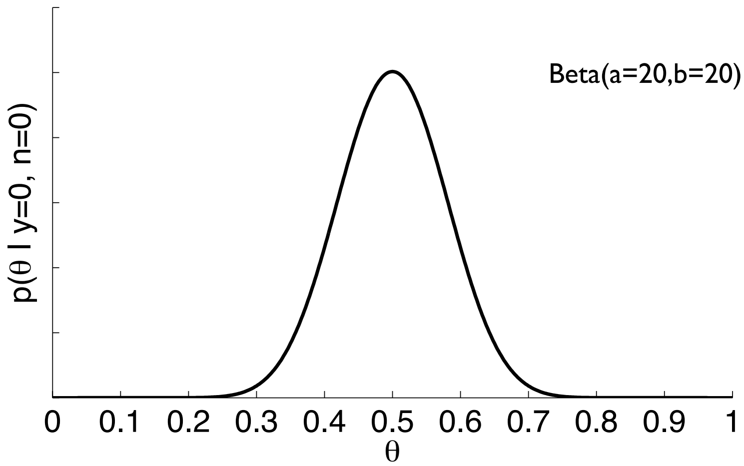
- What if we already know something about θ?
- We can still apply Bayes' rule to see how our knowledge changes as we acquire new observations:

$$p(\theta|y,n) = \frac{p(y|\theta,n)p(\theta|n)}{p(y|n)}$$

- But now the prior becomes important.
- Assume we know biased coins are never below 0.3 or above 0.7.
- To describe this we can use a beta distribution for the prior.

A beta prior

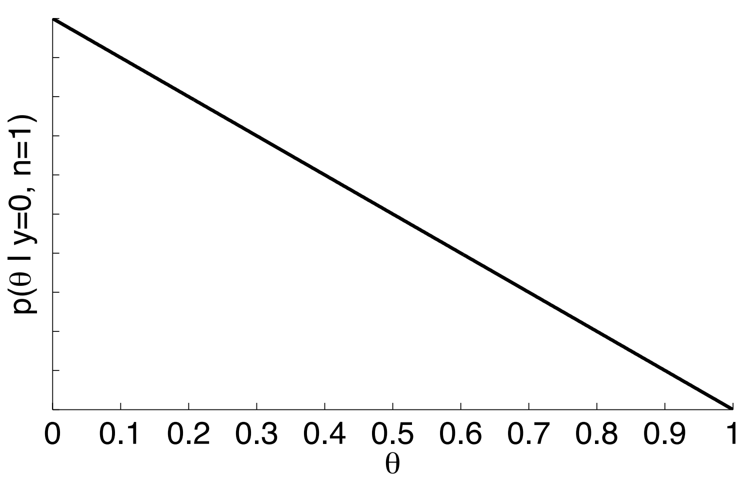
- In this case, before observing any trials our prior is not uniform:



Coin tossing revisited

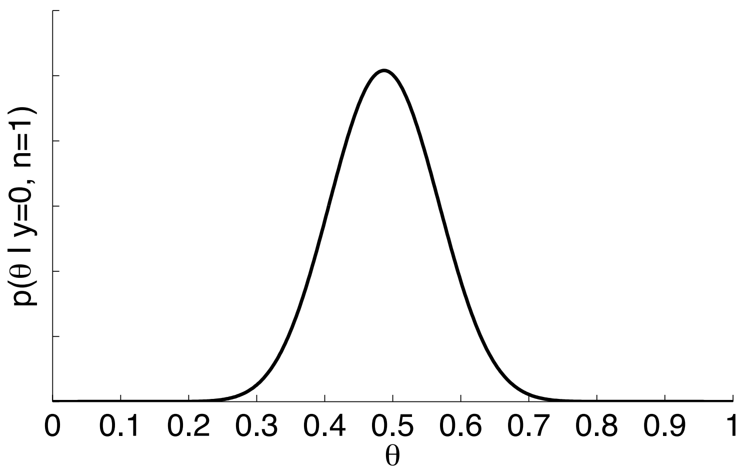
- What is our belief about θ after observing one "tail" ?
- With a uniform prior it was:

What will it look like with our prior?



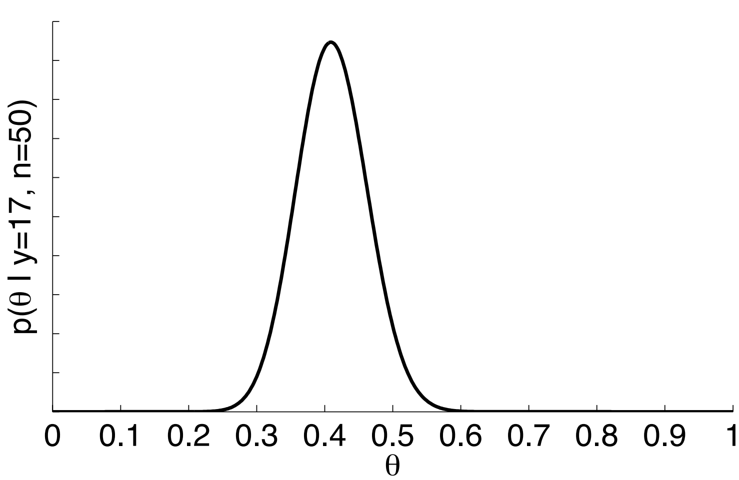
Coin tossing with prior knowledge

- Our belief about θ after observing one "tail" hardly changes.



Coin tossing

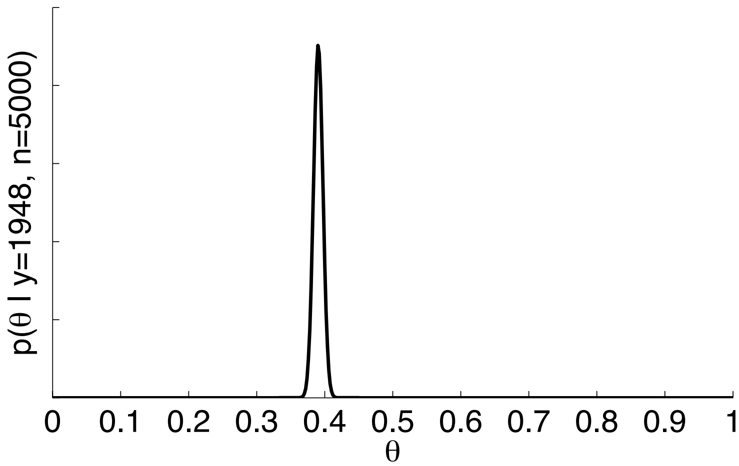
- After 50 trials, it's much like before.



Coin tossing

- After 5,000 trials, it's virtually identical to the uniform prior.

What did we gain?



Next time

- multivariate inference
- introduction to more sophisticated models
- belief networks