EECS 391
Intro to AI

Reasoning with Continuous Variables

L17 Tue Nov 6

How do you model a world?

How to you reason about it?

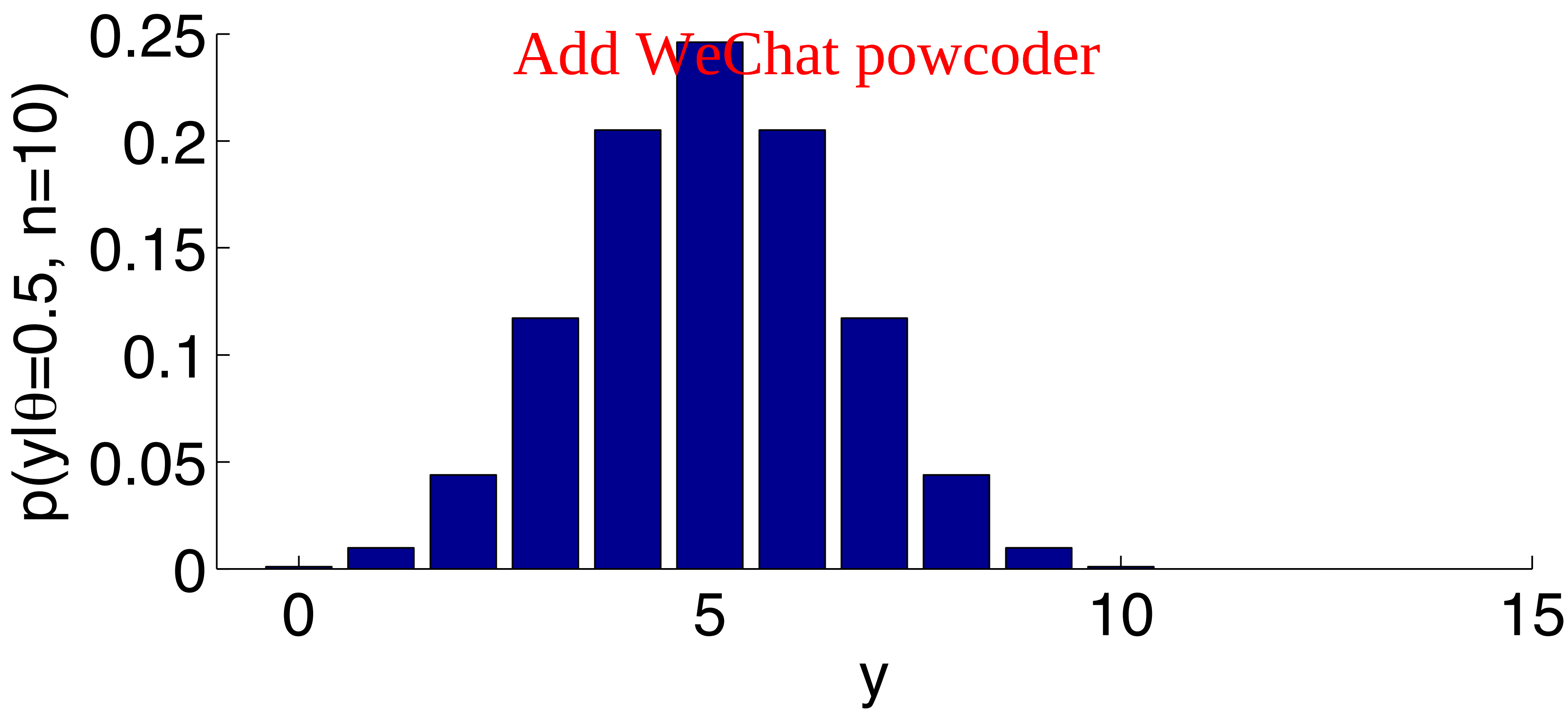
Bayesian inference for continuous variables

- The simplest case is true or false propositions
- Can easily extend to categorical variables
- The probability calculus is the same for continuous variables

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 $1 - \theta$.
- The binomial distribution specifies the probability of the total # of heads, y, out of n trials:

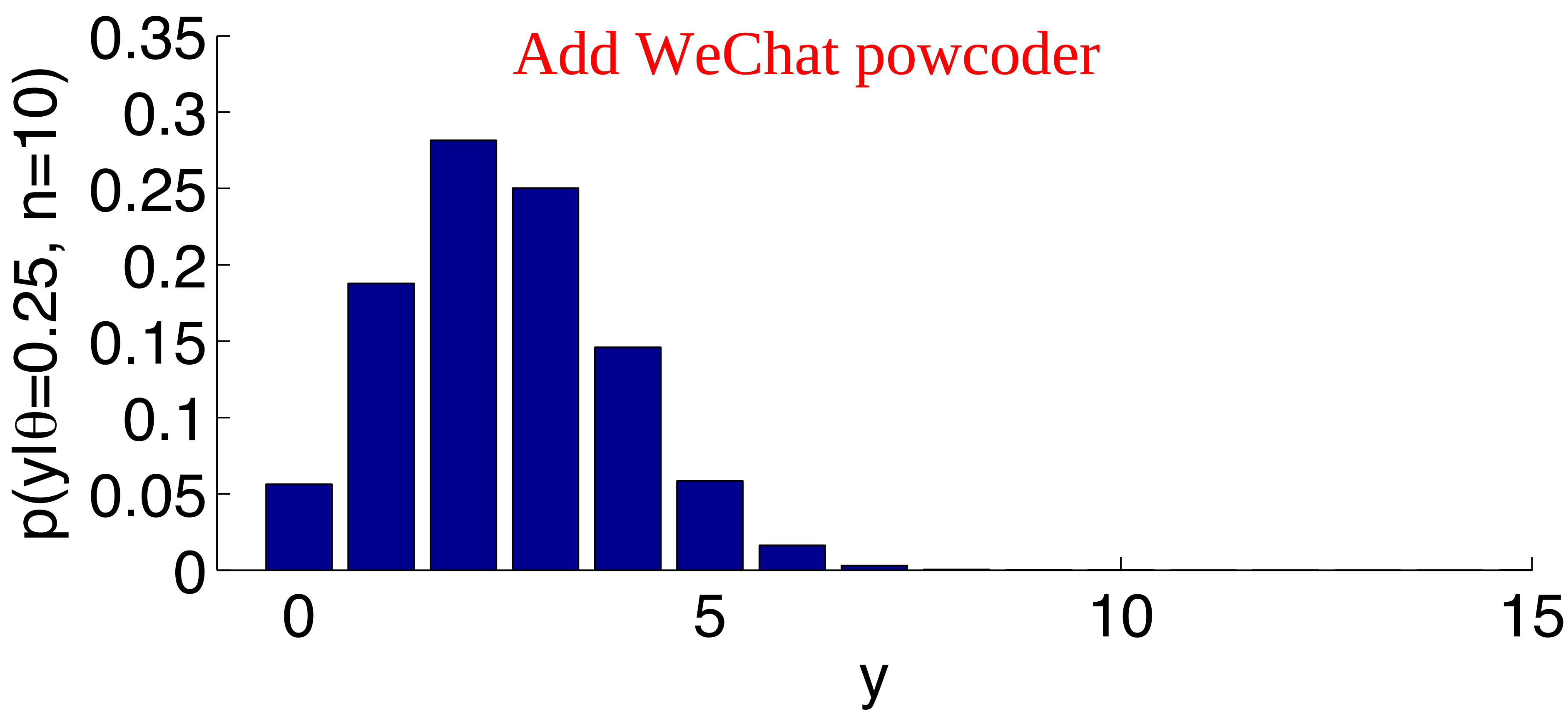
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



The binomial distribution

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

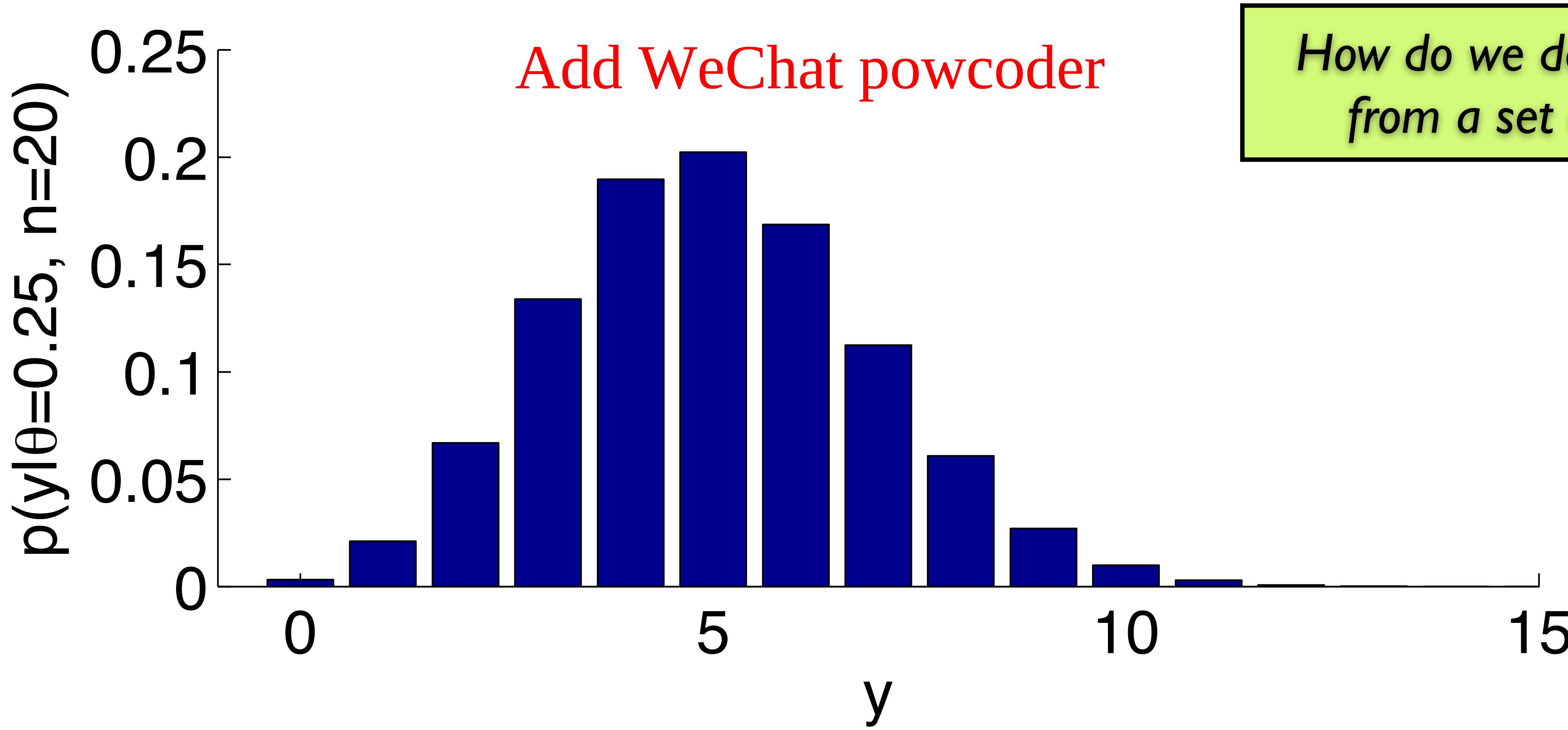
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



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:

likelihood prior

$$\underbrace{p(\theta|y,n)}_{\text{posterior}} = \frac{p(y|\theta,n)p(\theta|n)}{\underbrace{p(y|n)}_{\text{normalizing constant}}} \qquad p(y|n) = \int p(y|\theta,n)p(\theta|n)d\theta$$

- 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:

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

Bayesian inference with continuous variables

likelihood prior

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

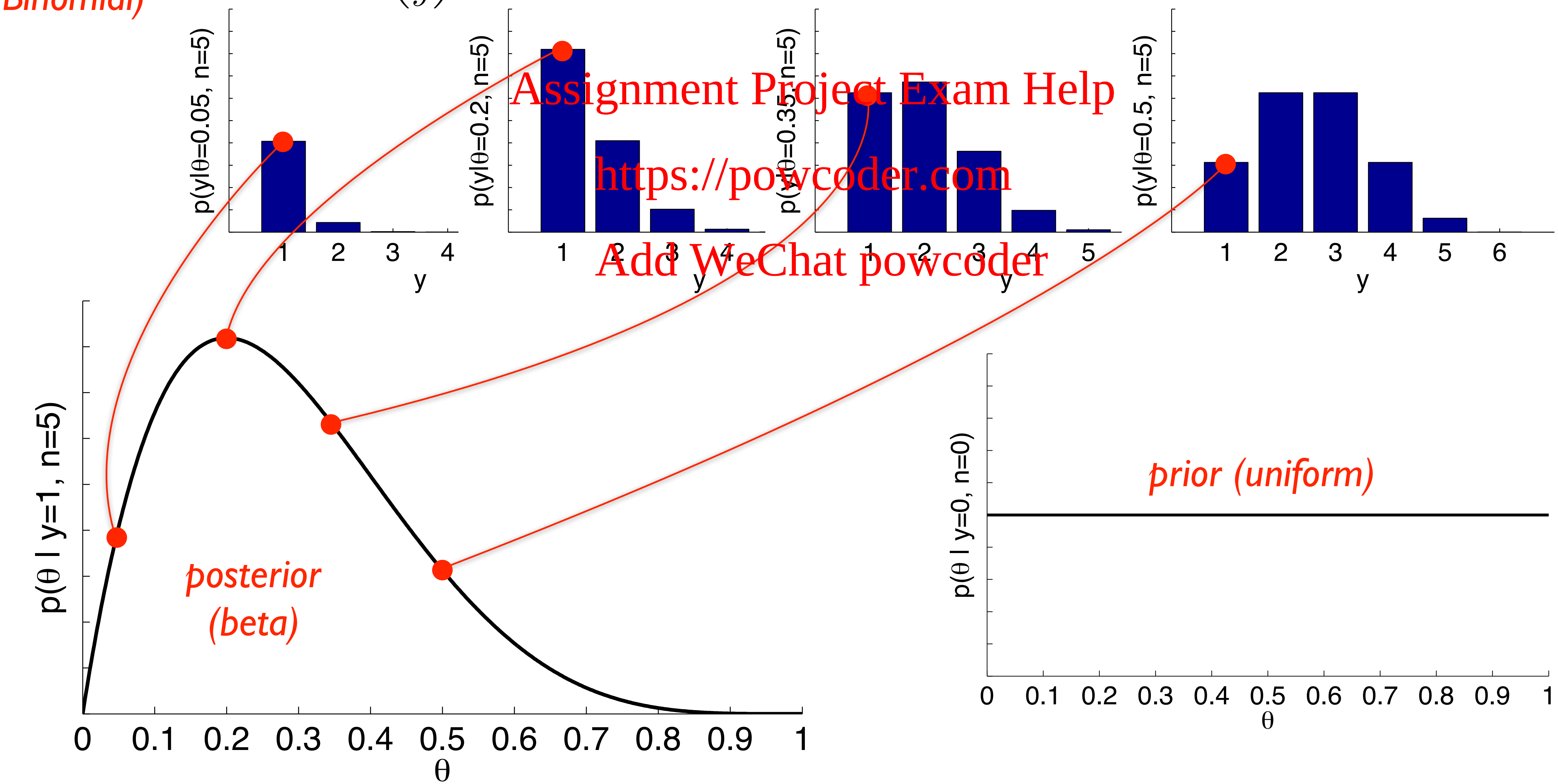
posterior

$$p(y|n) = \int p(y|\theta,n)p(\theta|n)d\theta$$

normalizing constant

likelihood (Binomial)

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

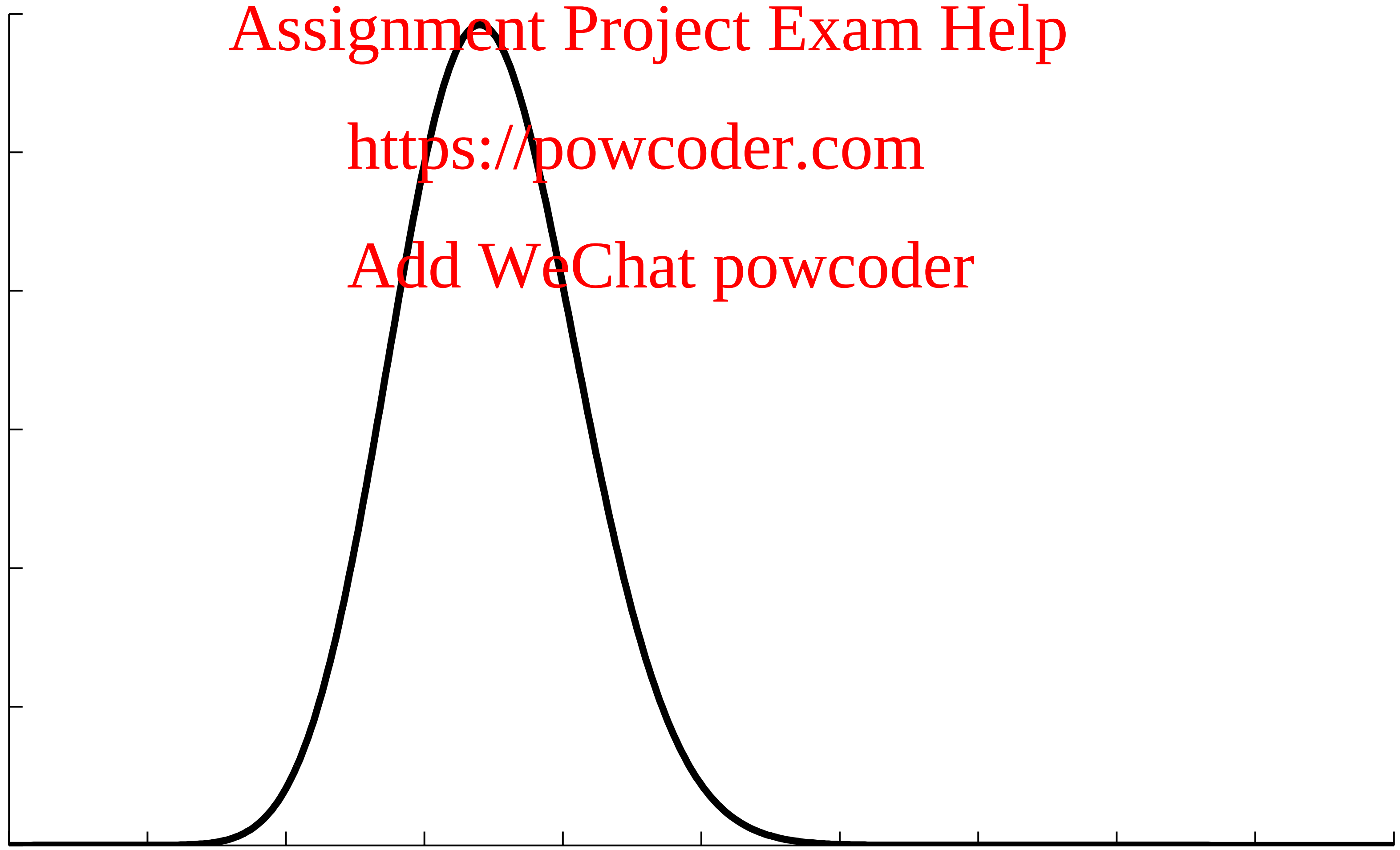


Updating our knowledge with new information

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

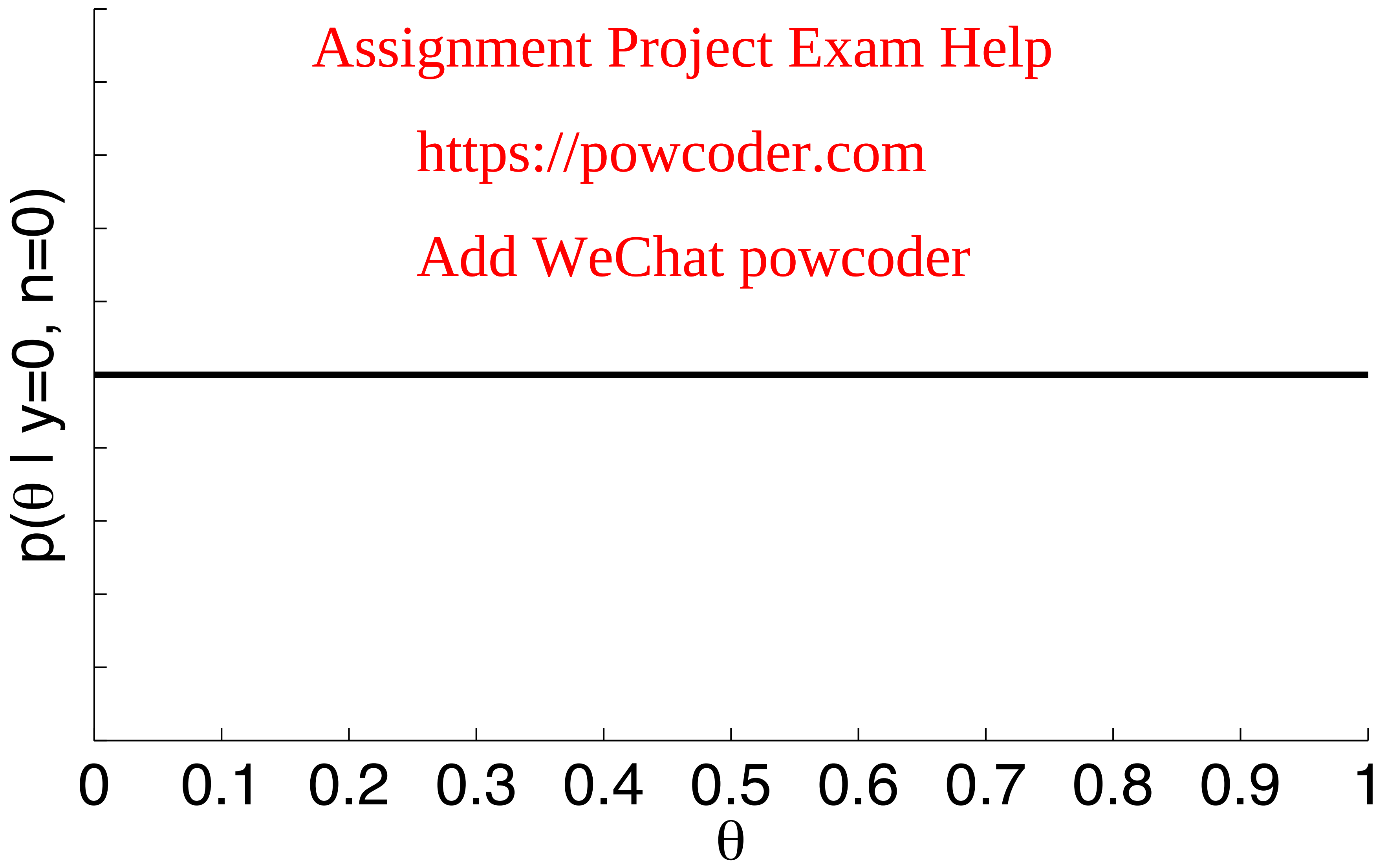
$$p(\theta|y,n) \propto \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Check: What goes on the axes?



Evaluating the posterior

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

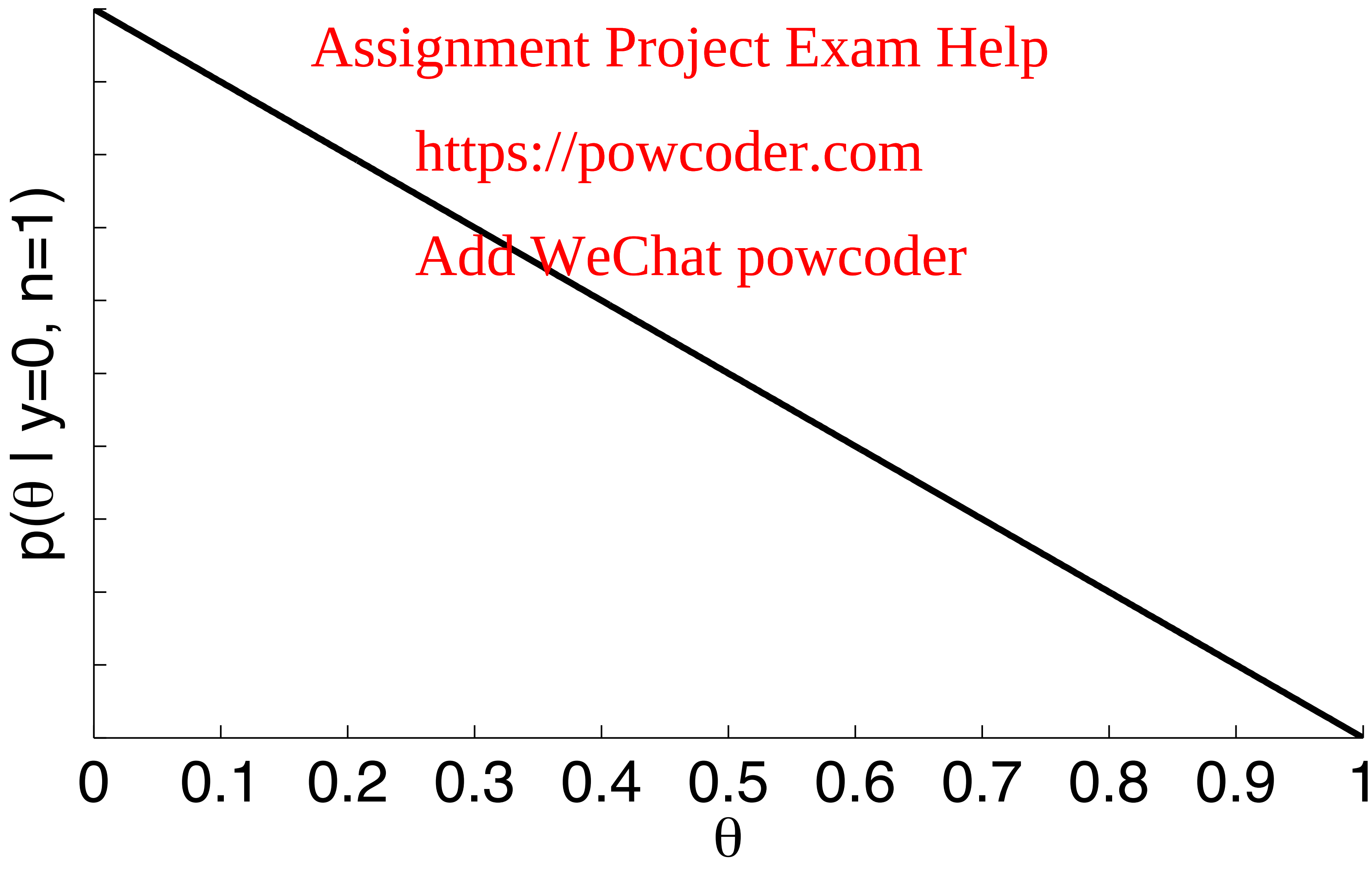


Coin tossing

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

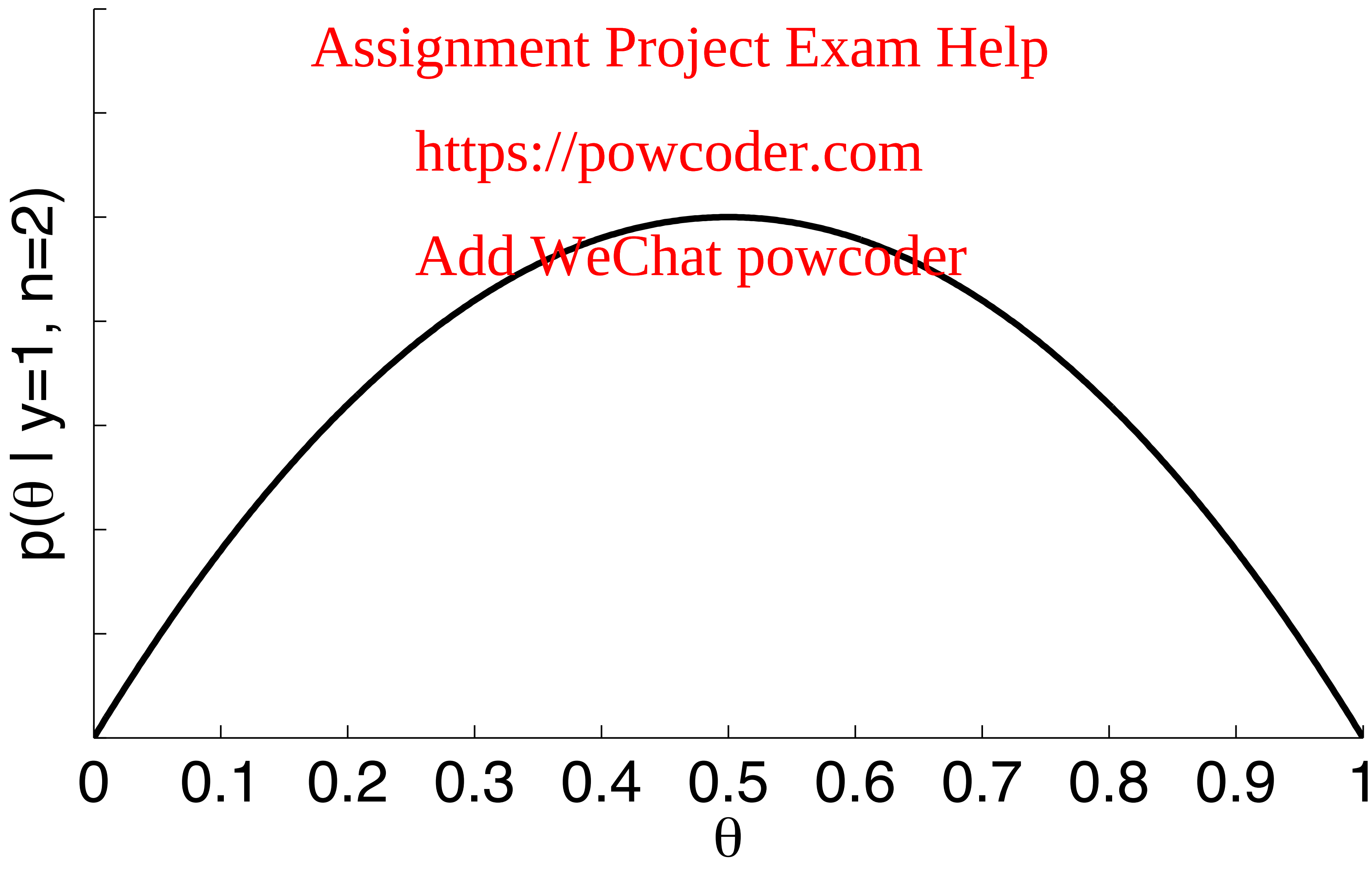
Is the p(θ >0.5) less or greater than 0.5?

What about p(θ >0.3)?



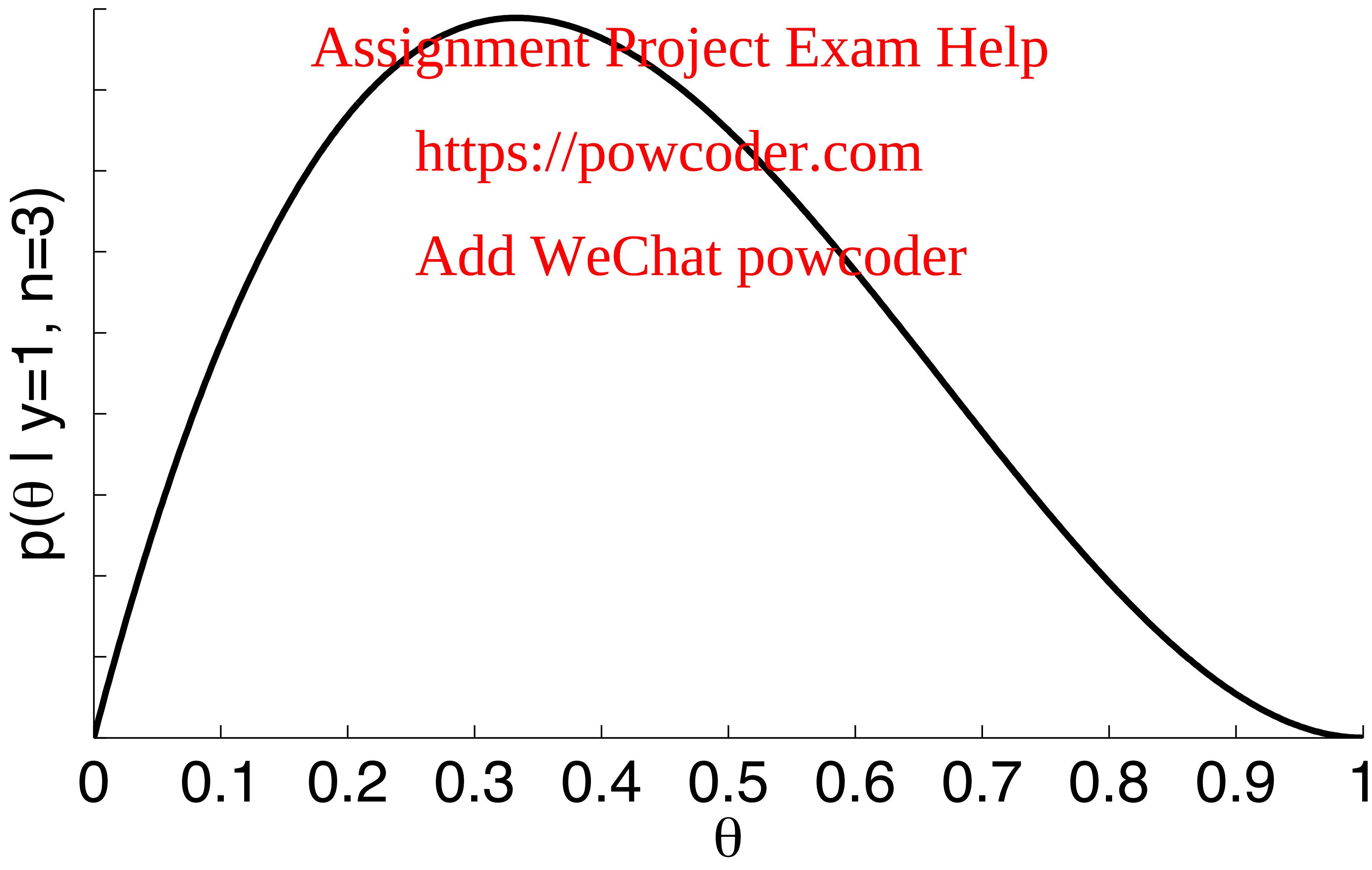
Coin tossing

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



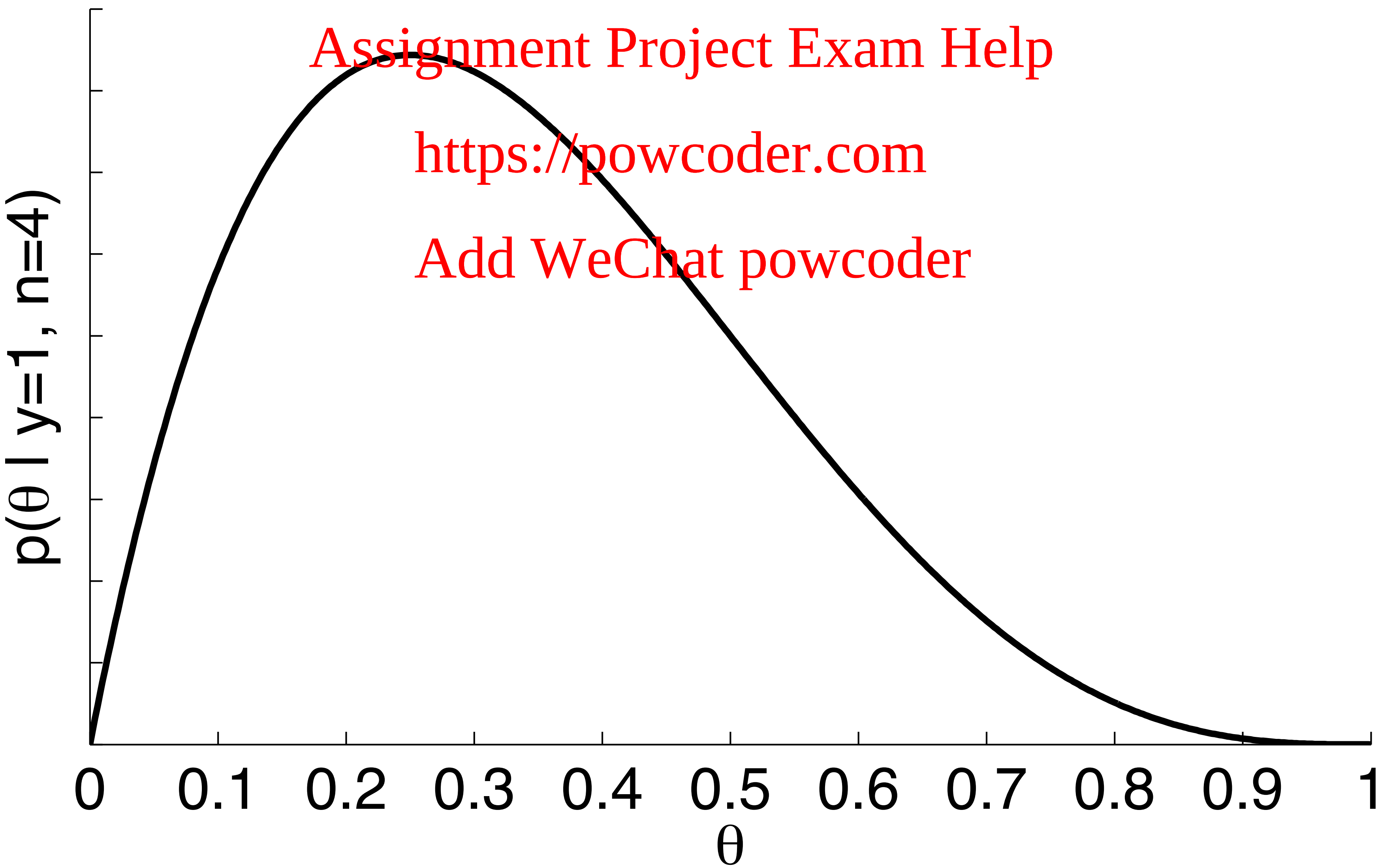
Coin tossing

- 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.

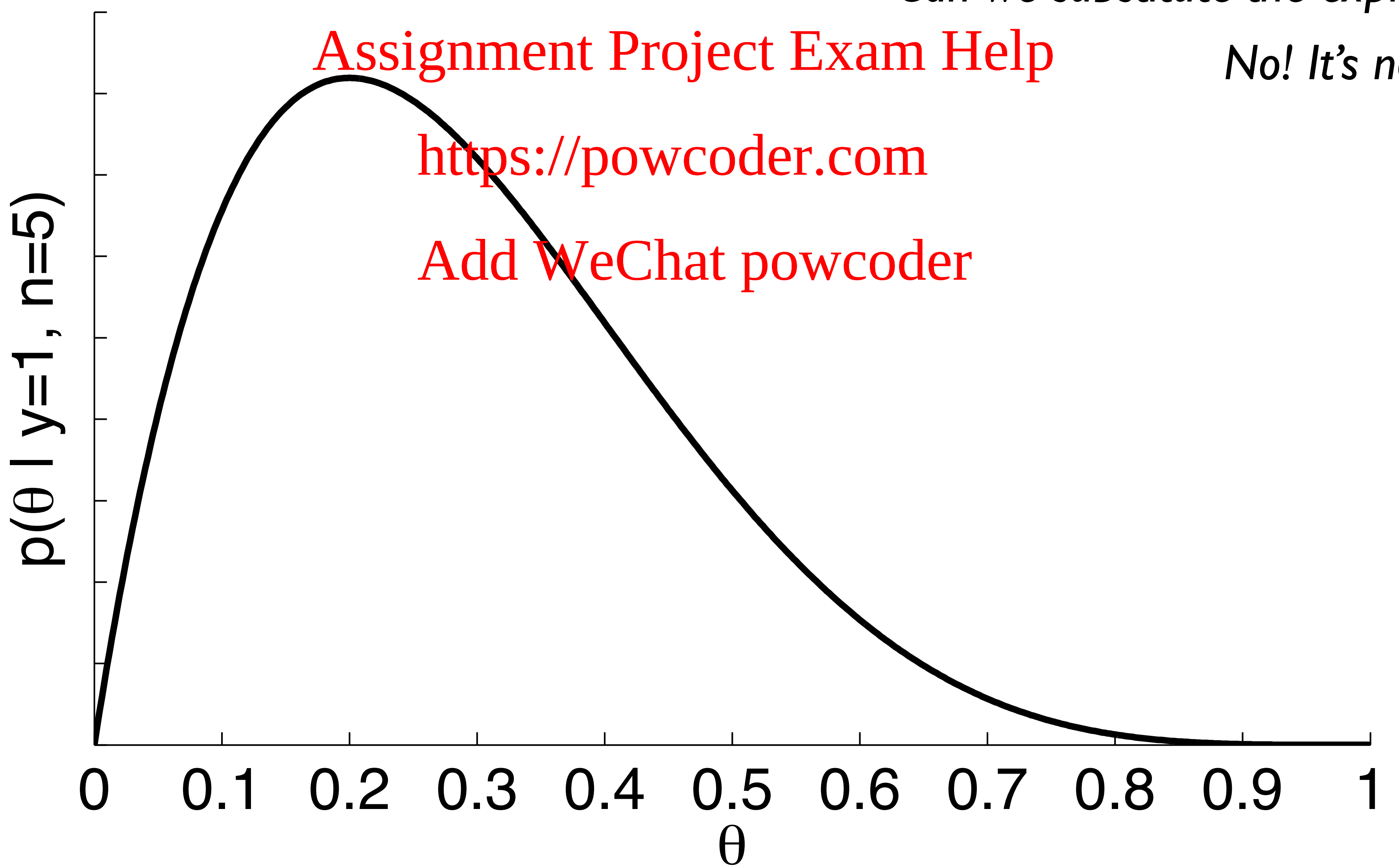
Do we have good evidence that this coin is biased?

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:

$$p(y|n) = \int_0^1 p(y|\theta,n)p(\theta|n)d\theta$$

$$= \frac{1}{n+1}$$

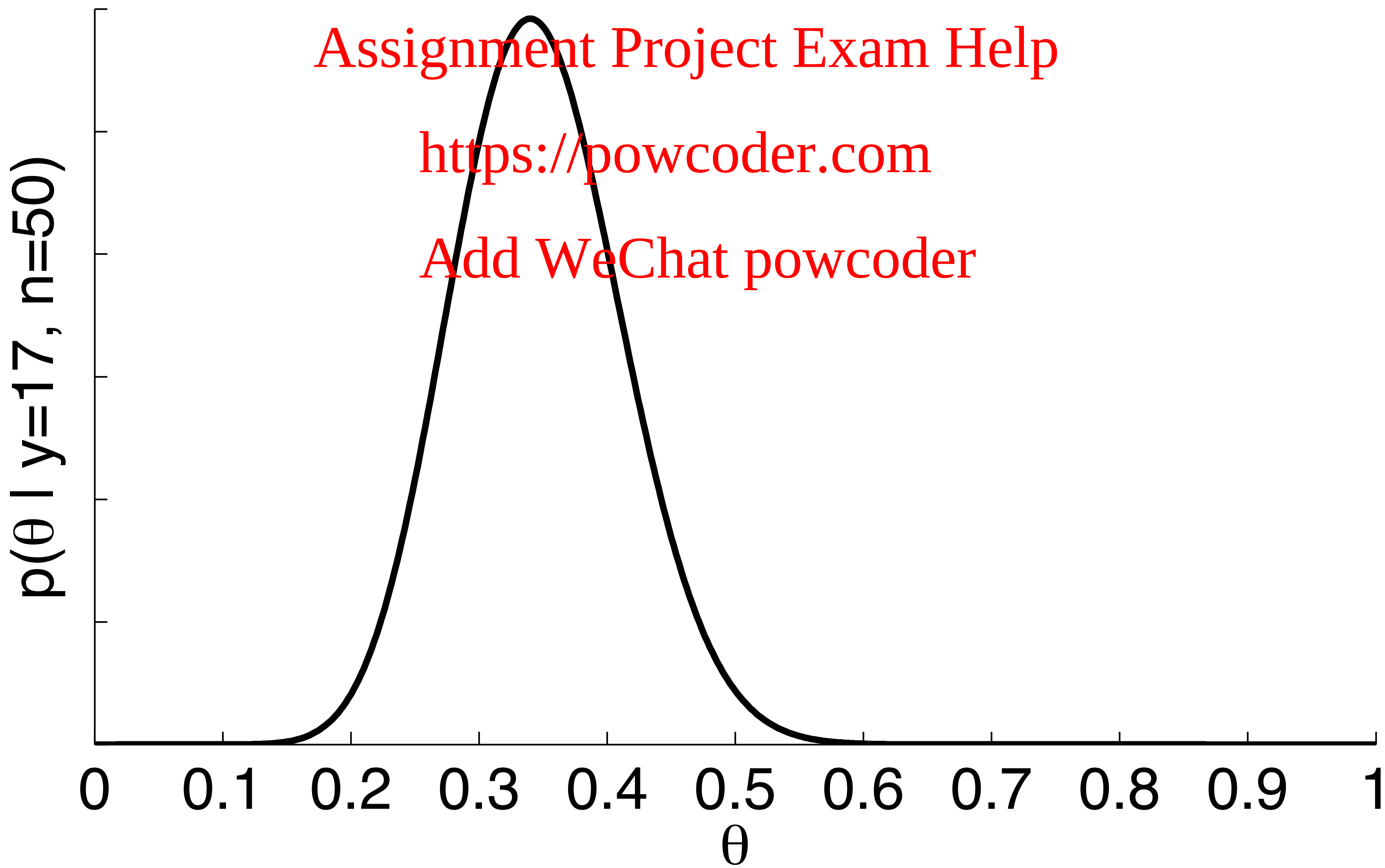
$$\Rightarrow p(\theta|y,n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}(n+1)$$

More coin tossing

- After 50 trials: 17 heads and 33 tails.

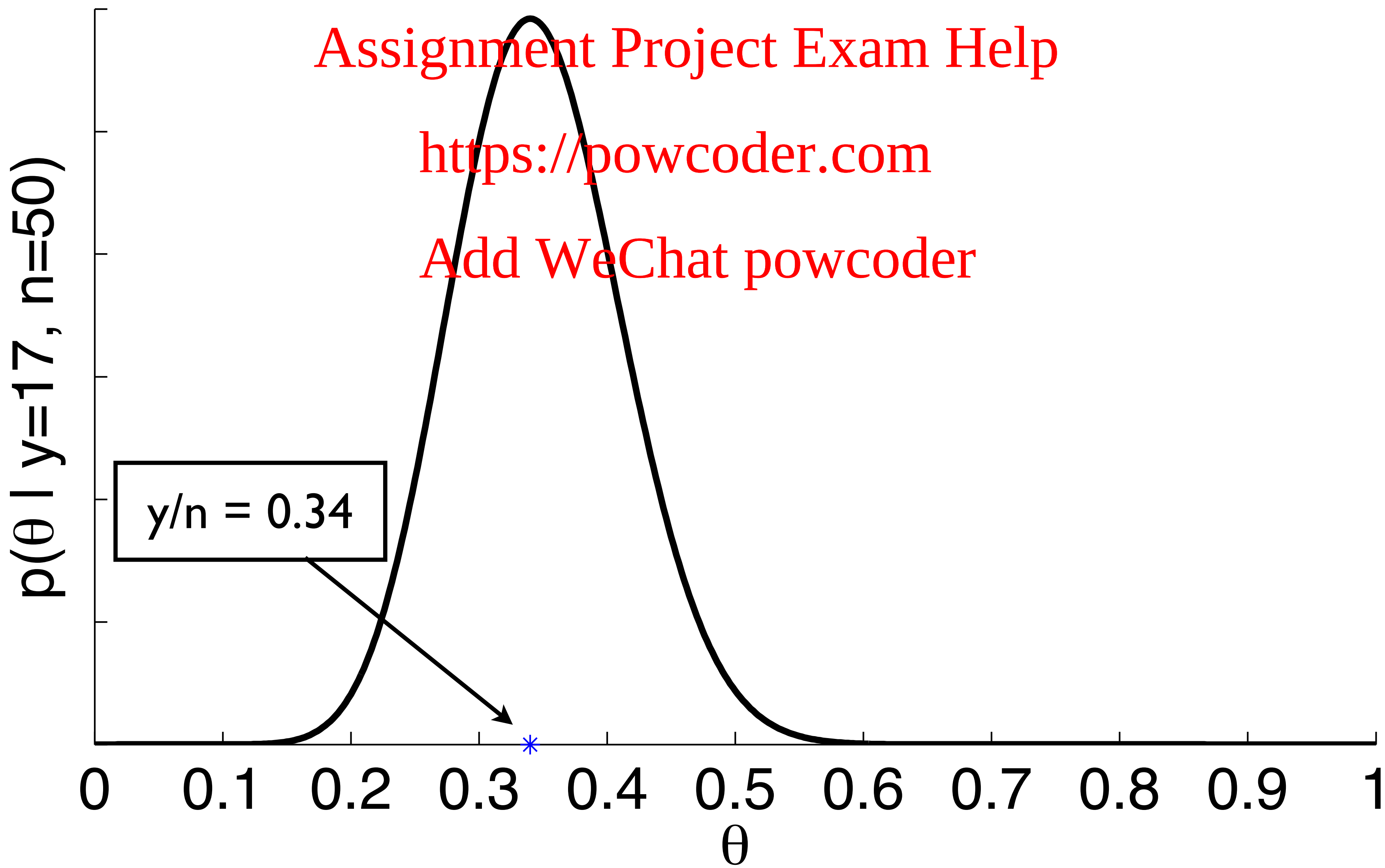
What's a good estimate of θ?

- There are many possibilities.



A ratio estimate

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



Estimates for parameter values

- maximum likelihood estimate (MLE)
  - derive by taking derivative of likelihood, setting result to zero, and solving

$$\frac{\partial \mathcal{L}}{\theta} = \binom{n}{y} \theta^y (1-\theta)^{n-y} = 0$$

  - ignores prior (or assumes uniform prior)

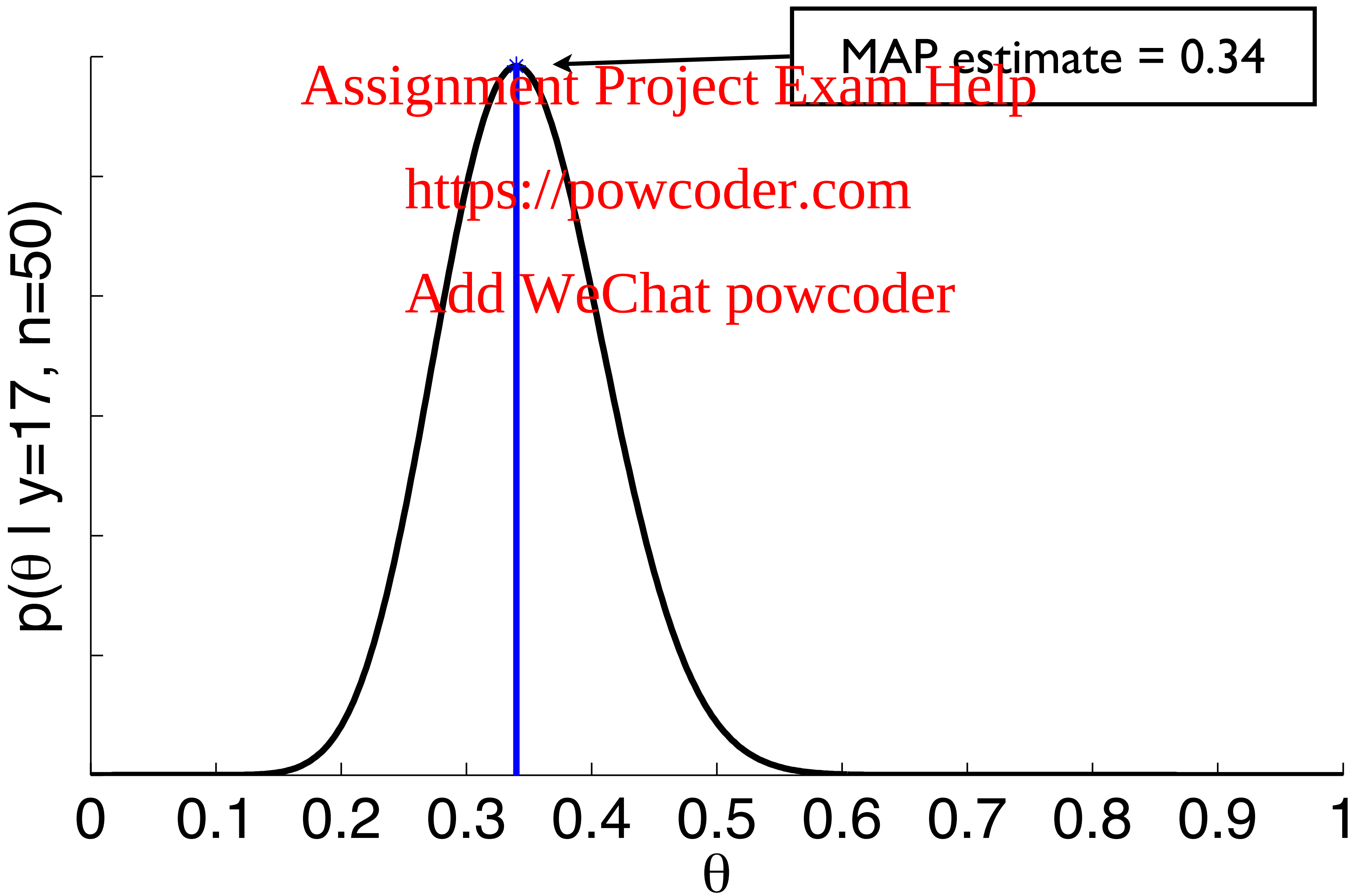
  - $\theta^{\mathrm{ML}} = \frac{y}{n}$ (derived on board)

- Maximum a posteriori (MAP)

  - derive by taking derivative of posterior, setting result to zero, and solving

The maximum a posteriori (MAP) estimate

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

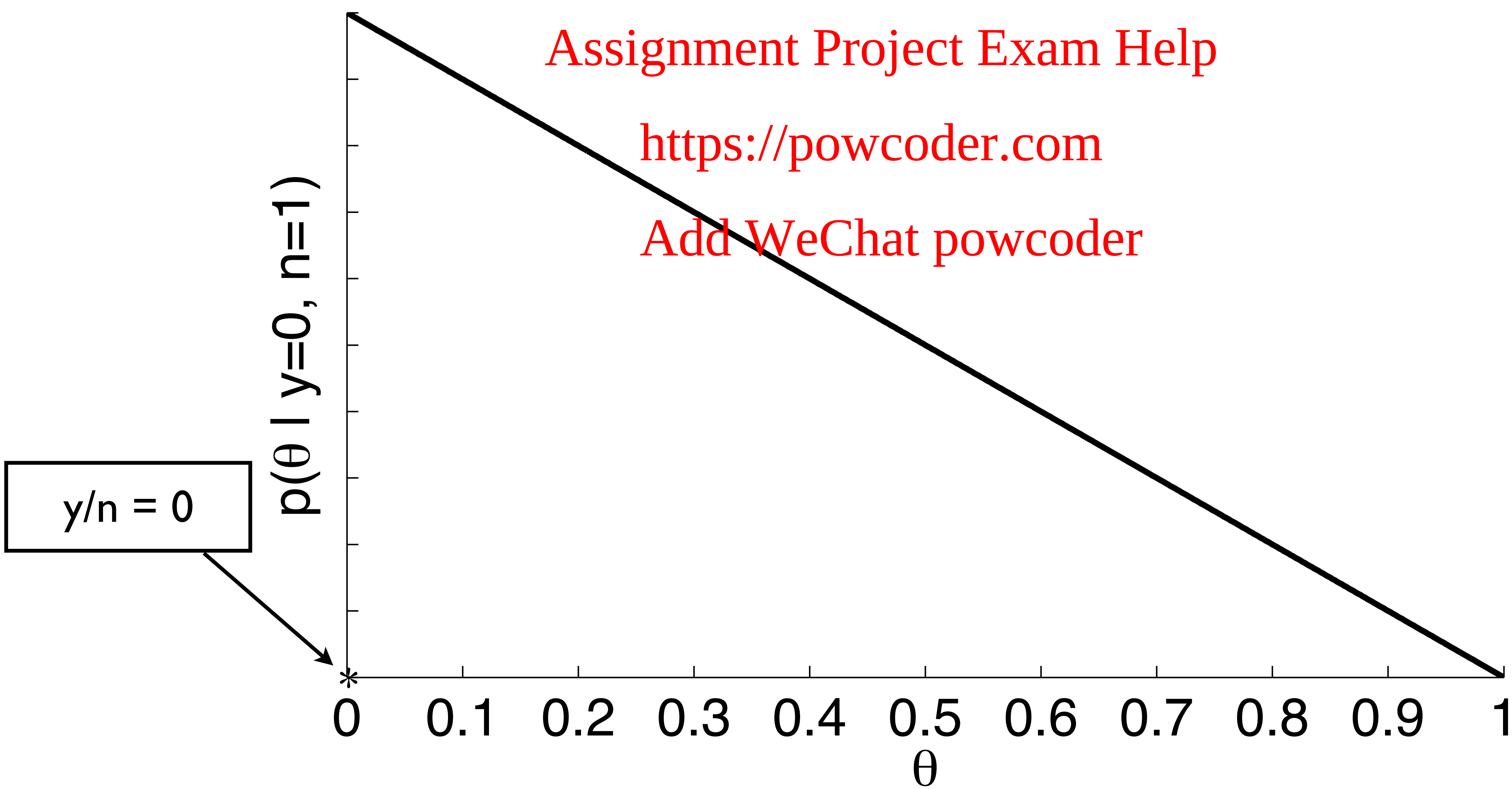


A different case

- 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?

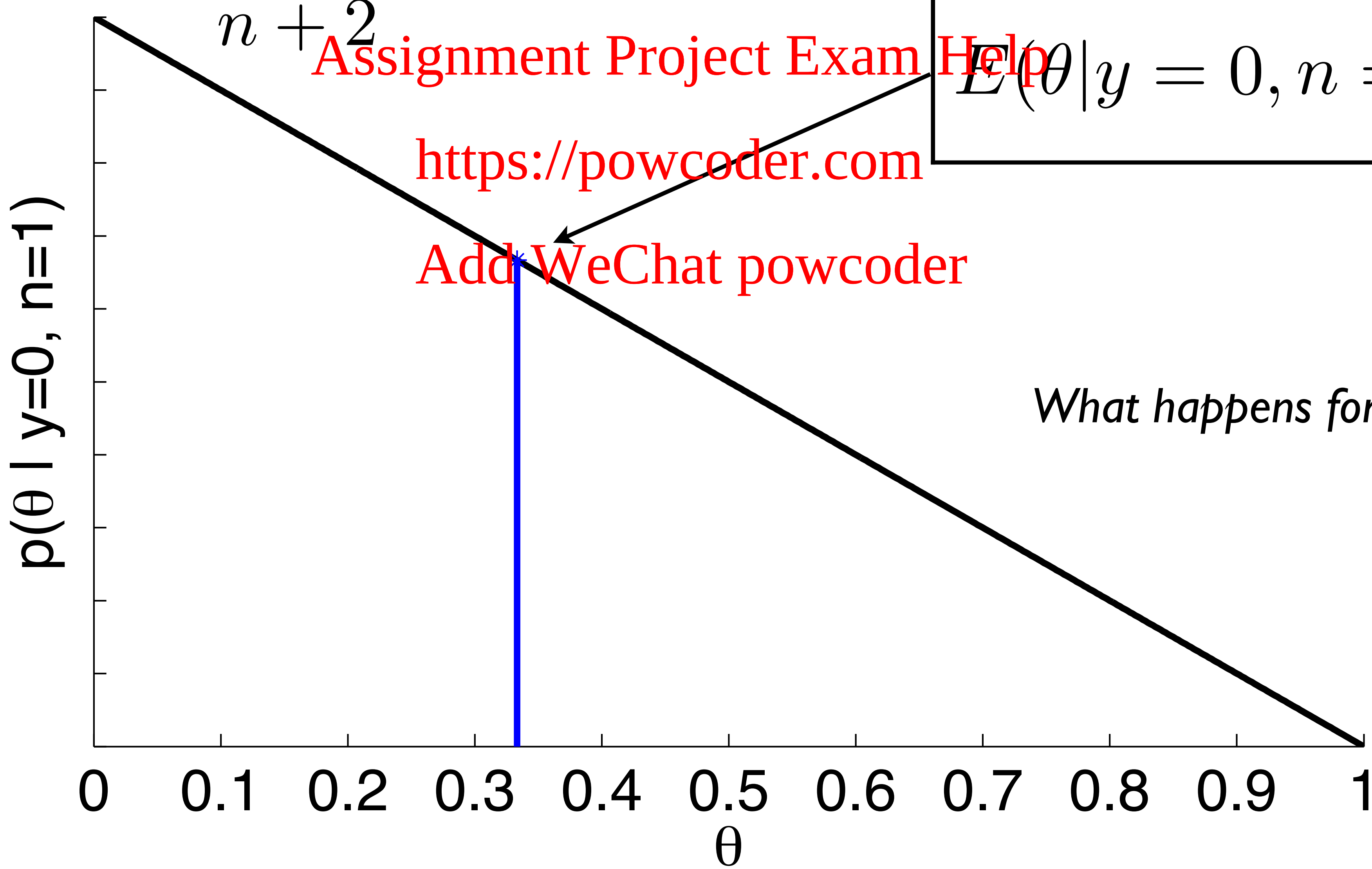


The expected value estimate

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

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

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

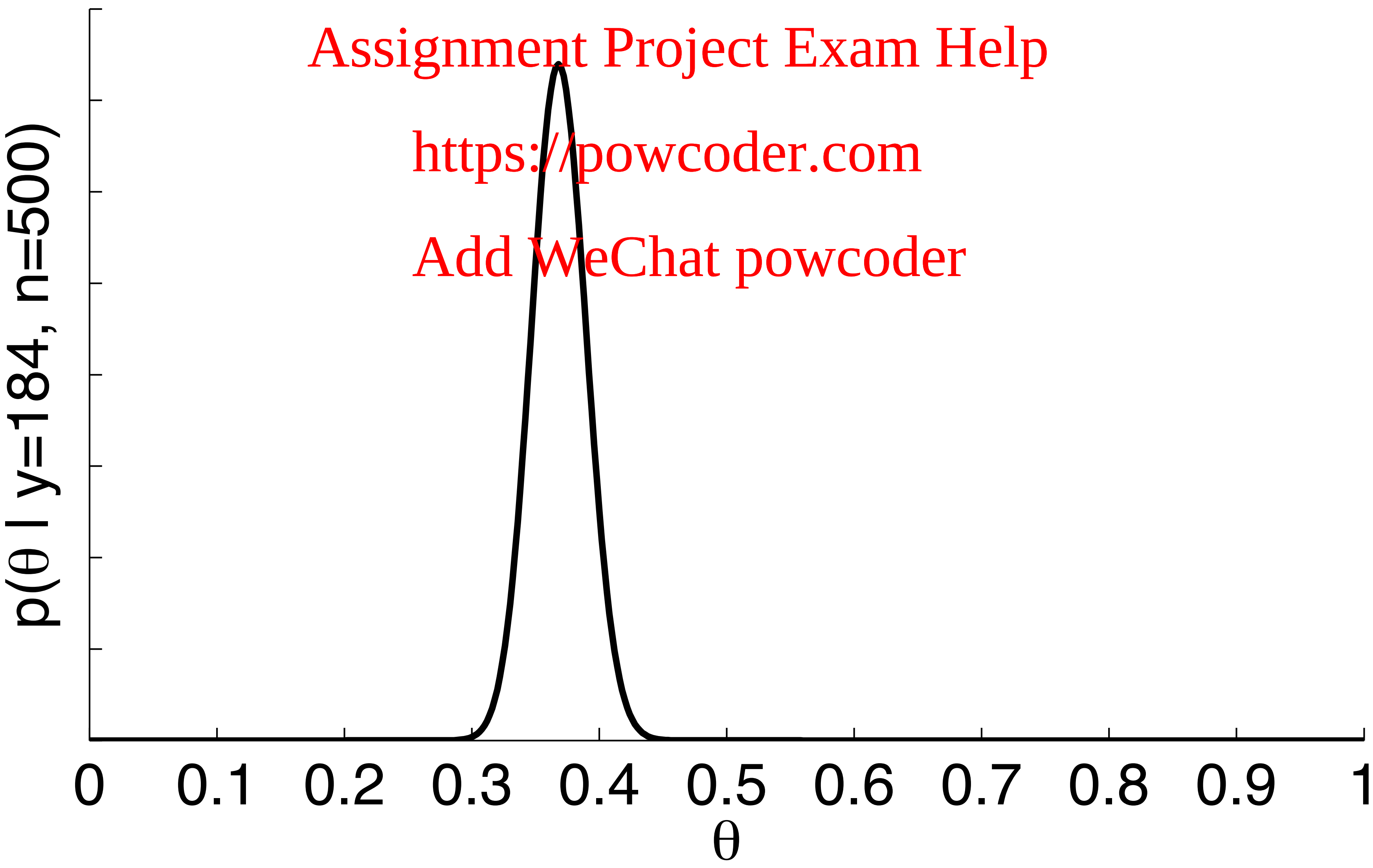


What happens for zero trials?

Much more coin tossing

- After 500 trials: 184 heads and 316 tails.

What's your guess of θ?



Much more coin tossing

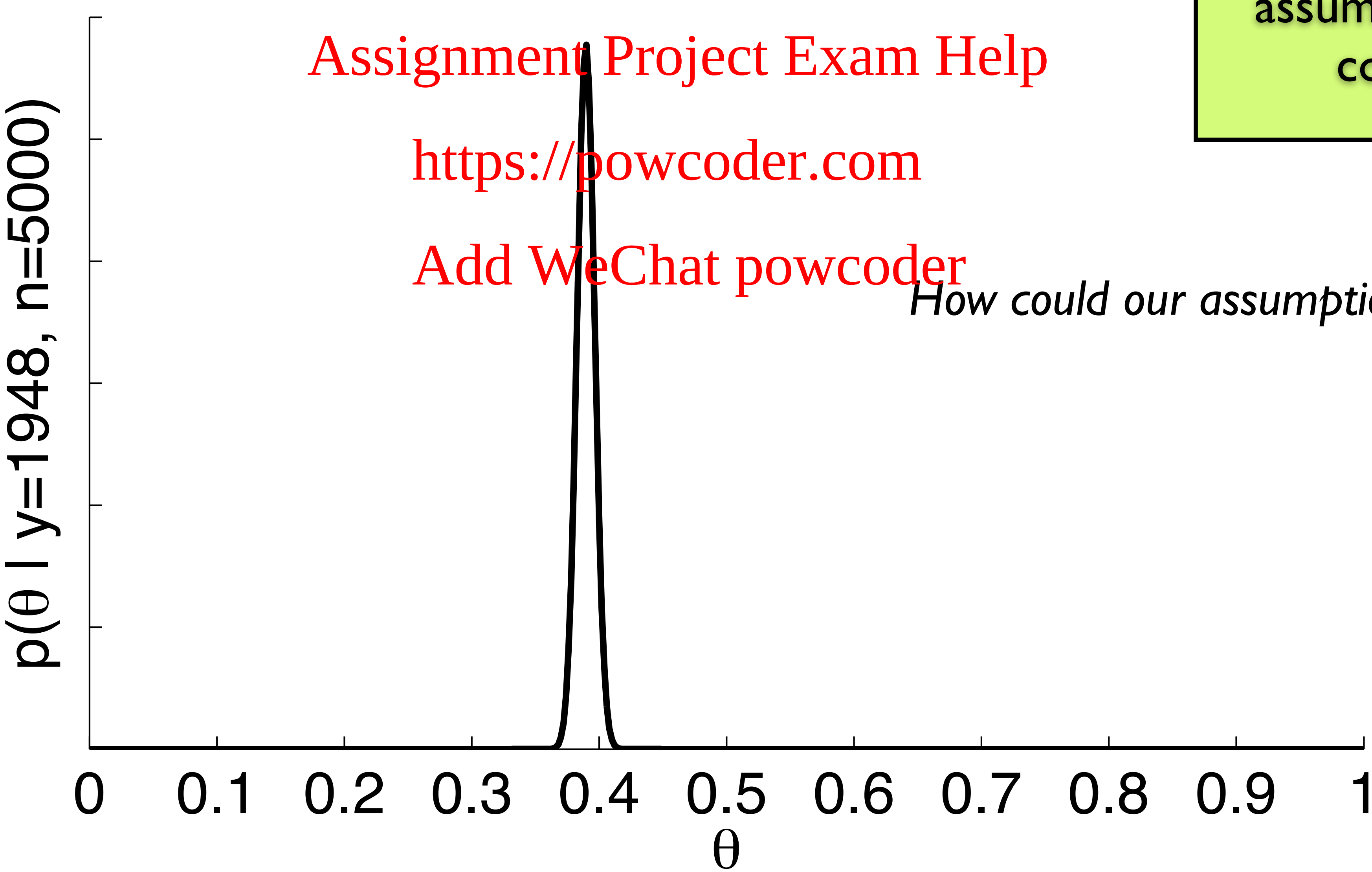
- After 5000 trials: 1948 heads and 3052 tails.

True value is 0.4.

- Posterior contains true estimate.

Is this always the case?

NO! Only if the assumptions are correct.



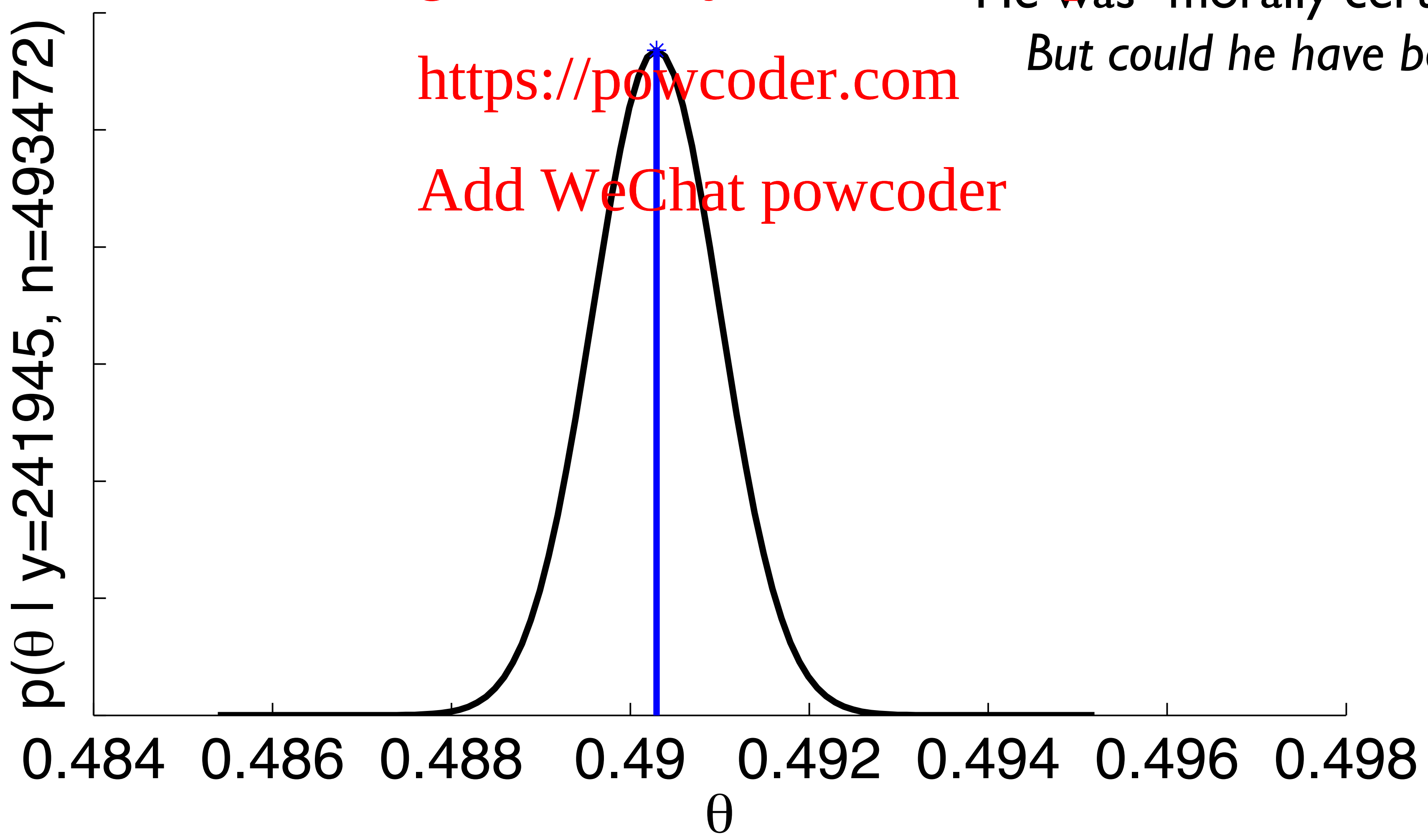
How could our assumptions be wrong?

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

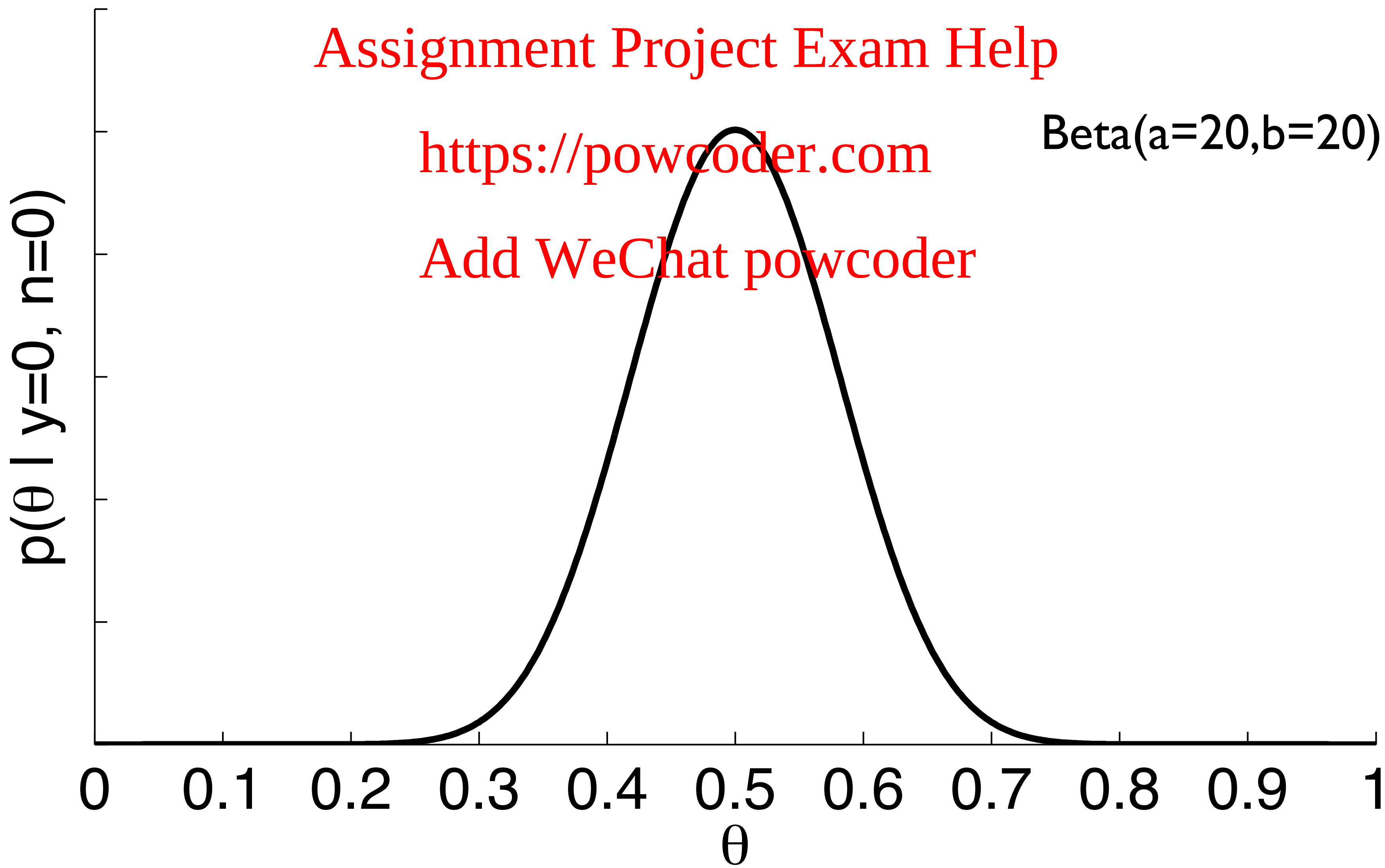
- What if we already know something about θ?
- We can still apply Bayes' rule to see how our knowledge changes as we acquire new observations:

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

- 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

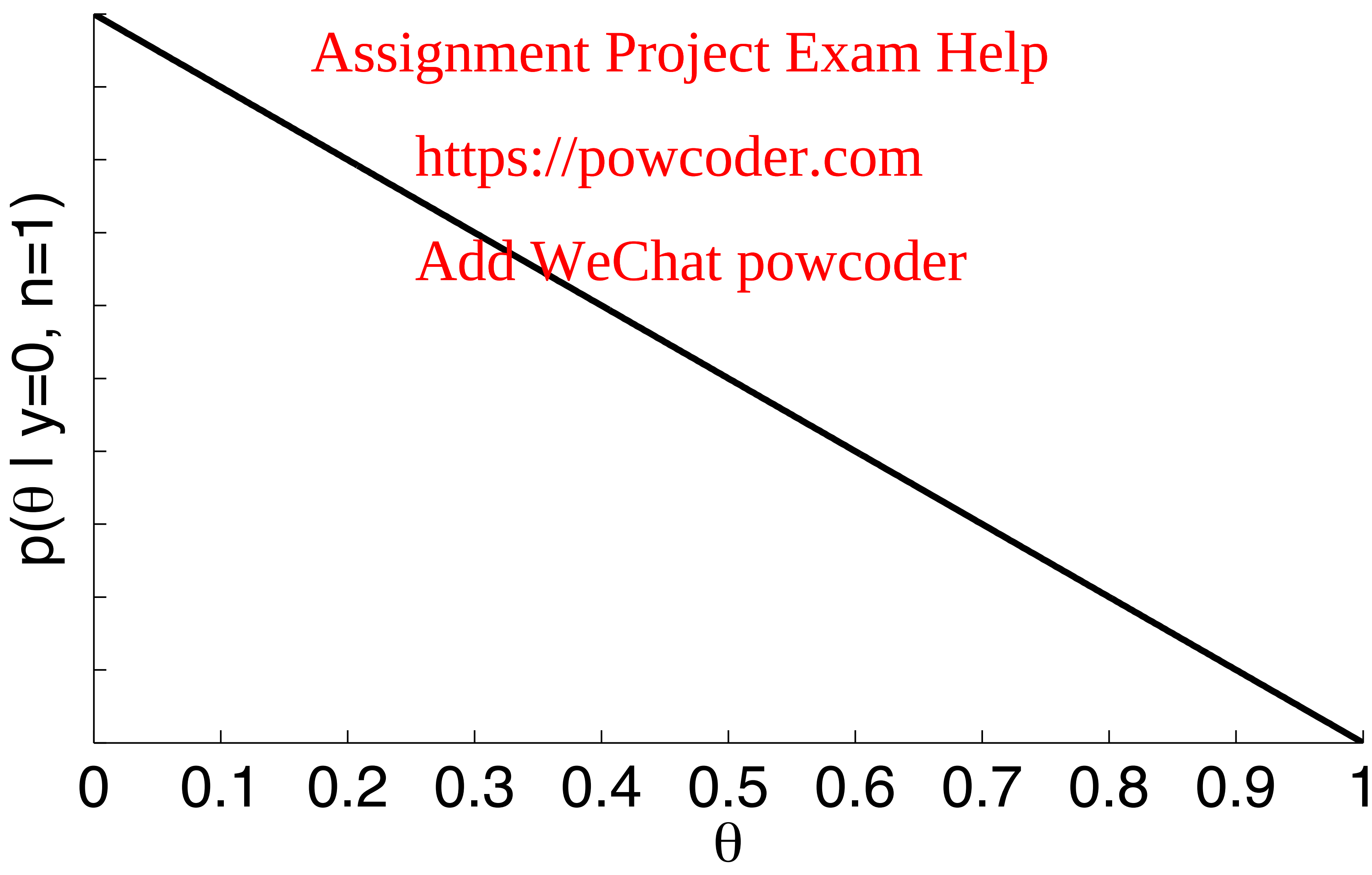
- In this case, before observing any trials our prior is not uniform:



Coin tossing revisited

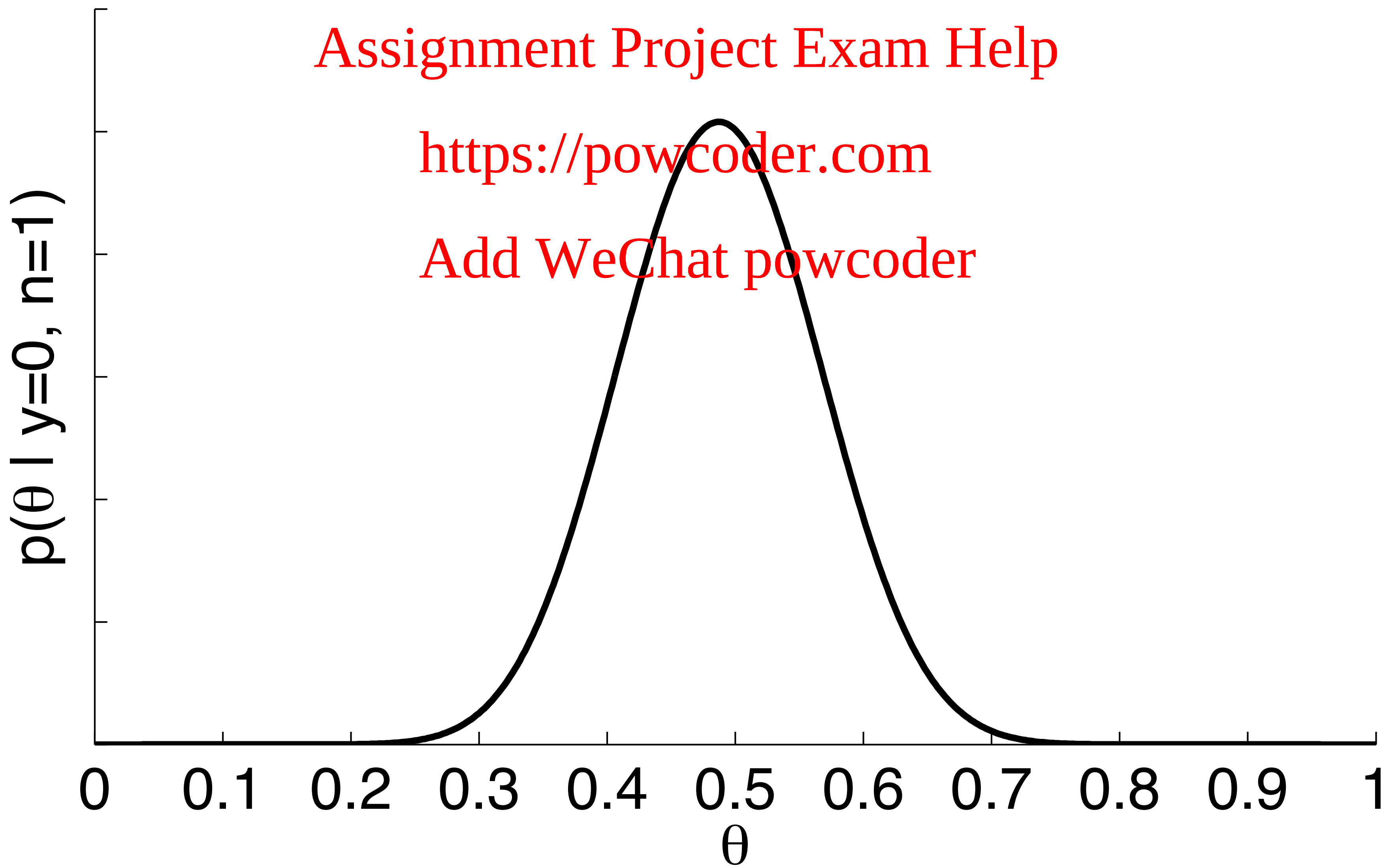
- What is our belief about θ after observing one "tail" ?
- With a uniform prior it was:

What will it look like with our prior?



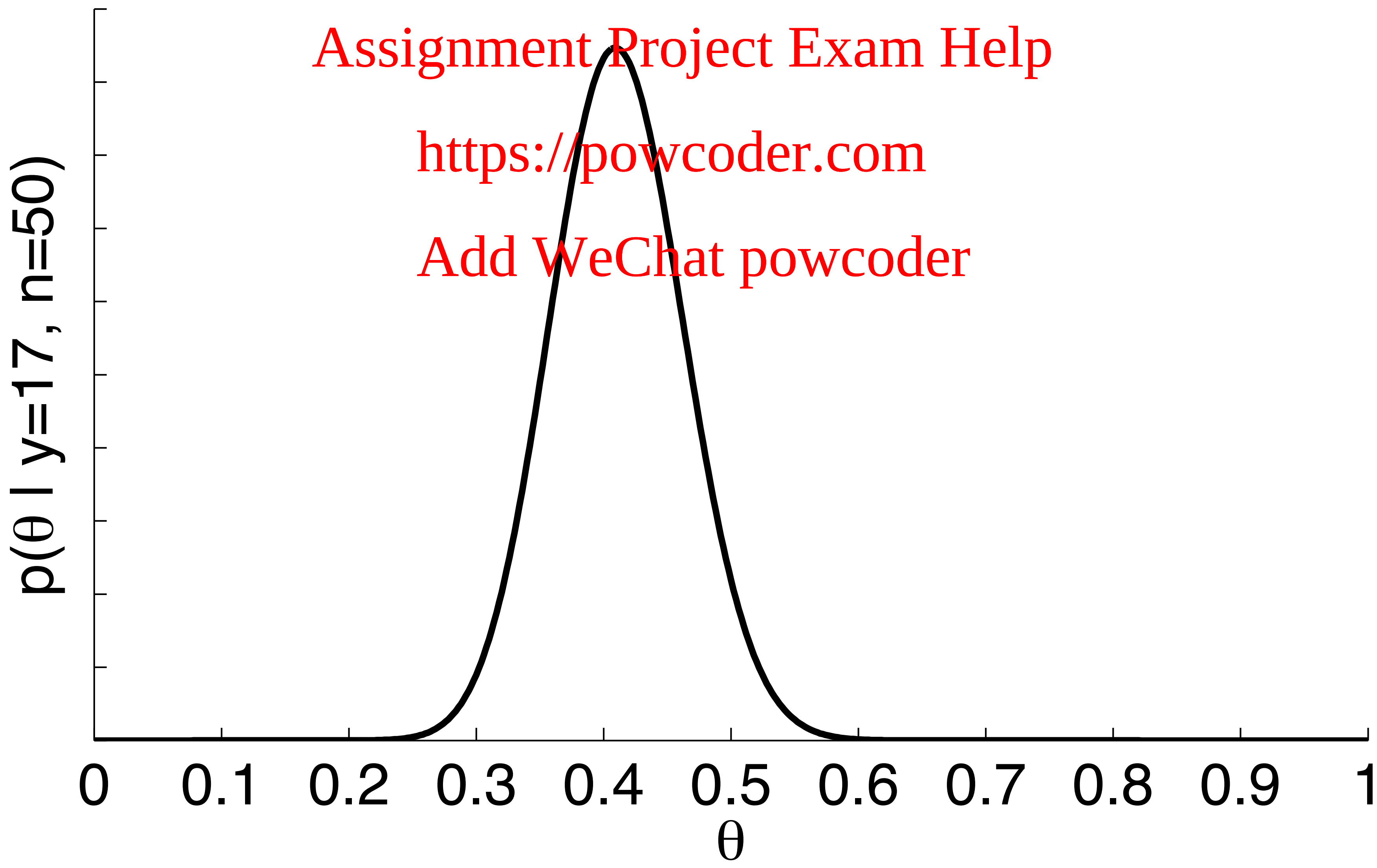
Coin tossing with prior knowledge

- 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?