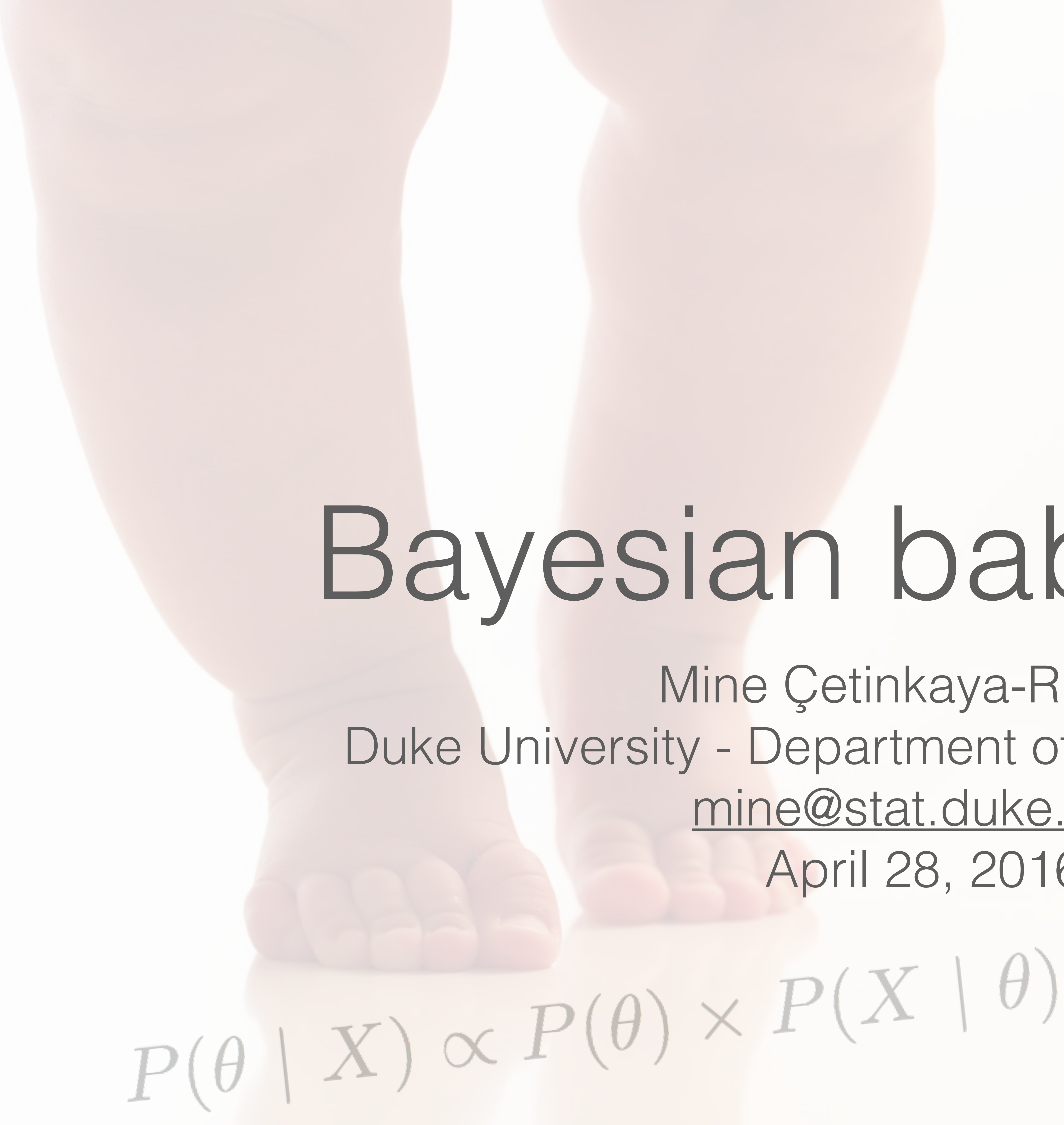

$P(\theta \mid X) \propto P(\theta) \times P(X \mid \theta)$

Bayesian baby steps

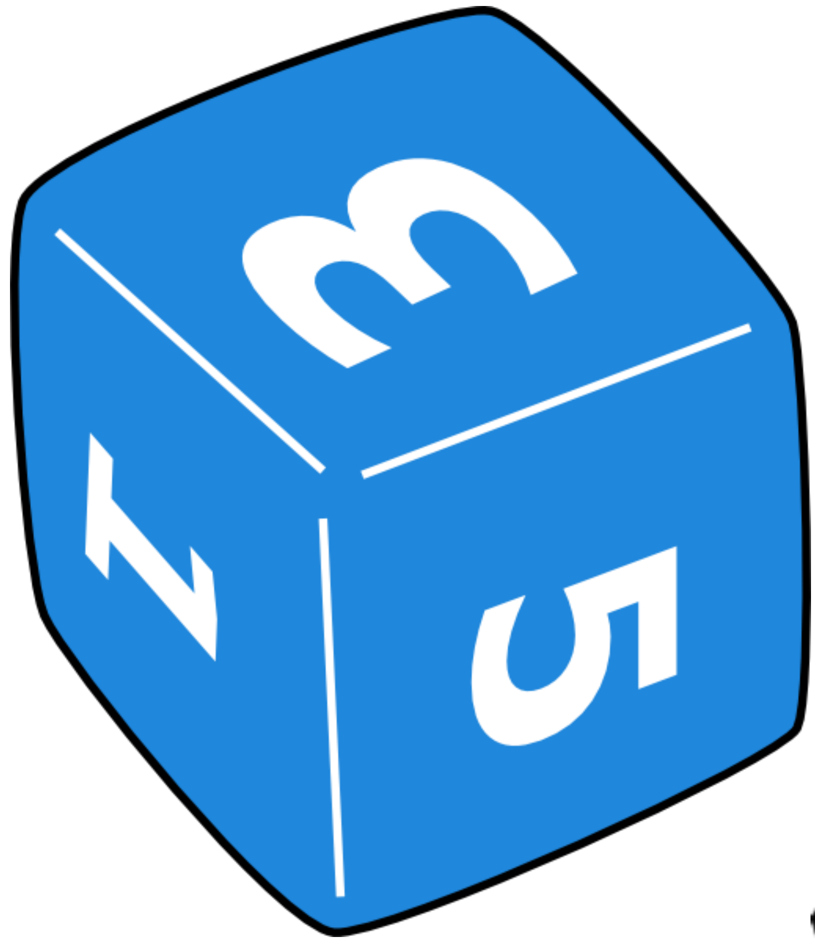
Mine Çetinkaya-Rundel
Duke University - Department of Statistical Science
mine@stat.duke.edu

April 28, 2016

Dice game

Let’s play a game

- I keep one die in my the left and one die in my right hand, and you won’t know which is the 6-sided die and which is the 12-sided.
- You pick die (L or R), I roll it, and I tell you if you win or not, where winning is getting a number ≥ 4.
  - If you win, you get an additional week off this year!
  - If you lose, you “get to” come in to work on the weekends for a year.
- We play this multiple times, and I don’t swap the sides the dice are on at any point.
- The ultimate goal is to come to a consensus about whether the die on the left or the die on the right is the “good” die.
  - If you make the right decision, you get an additional week off.
  - If you make the wrong decision, you work weekends for a year!



6-sided

P(win) = 0.50

“good” die

12-sided

P(win) = 0.75

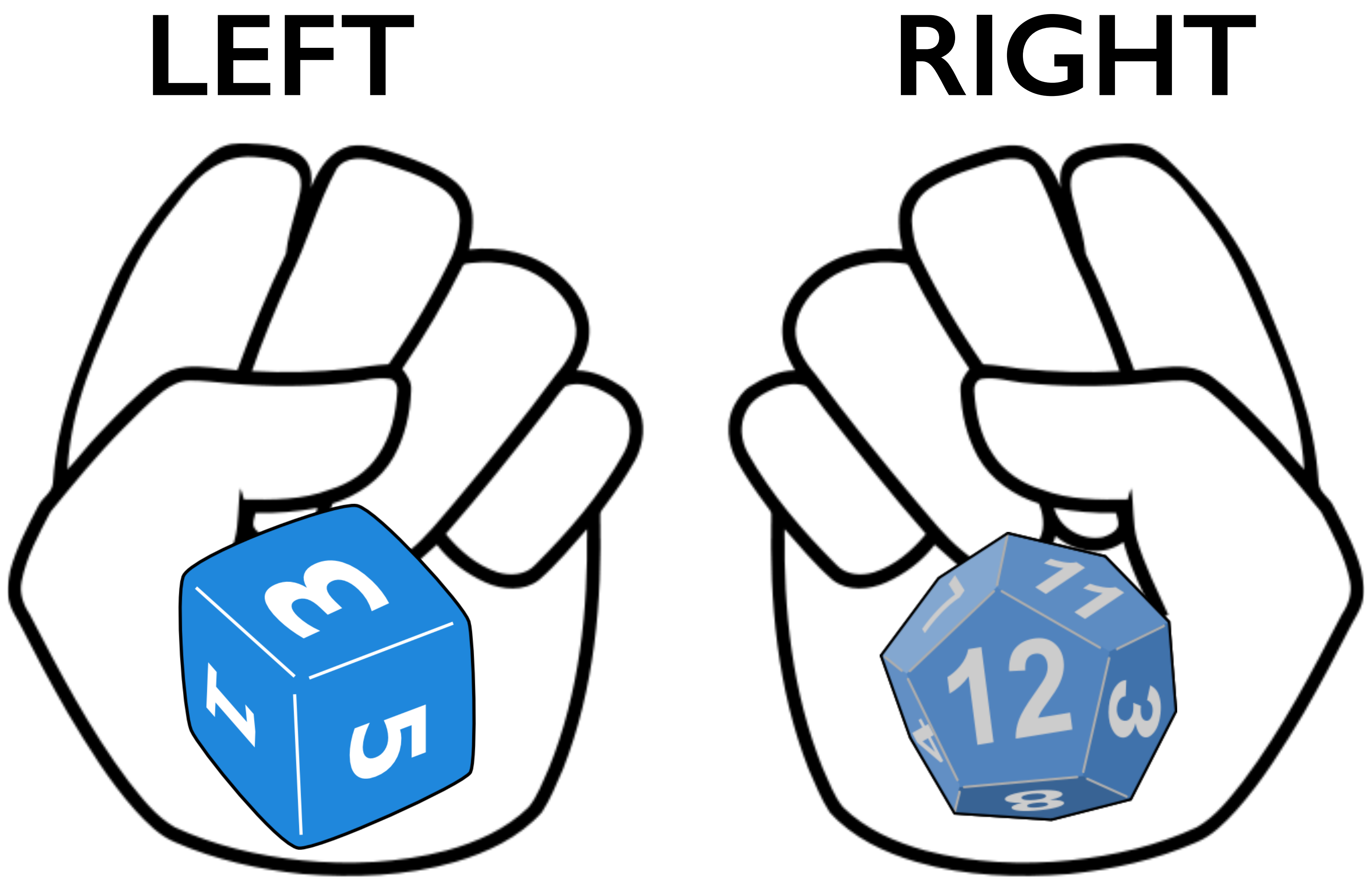
Hypotheses and decisions

		Truth	
		Right good, Left bad	Right bad, Left good
Decision	pick Right	You win the game - additional week off!	You lose - weekends for a year :(
	pick Left	You lose - weekends for a year :(	You win the game - additional week off!

Sampling isn't free!
At each trial you risk having to work an additional weekend (the die comes up < 4). Too many trials means you might be working many weekends. And if we spend too much time rolling dice this talk is going to be bo-ring!

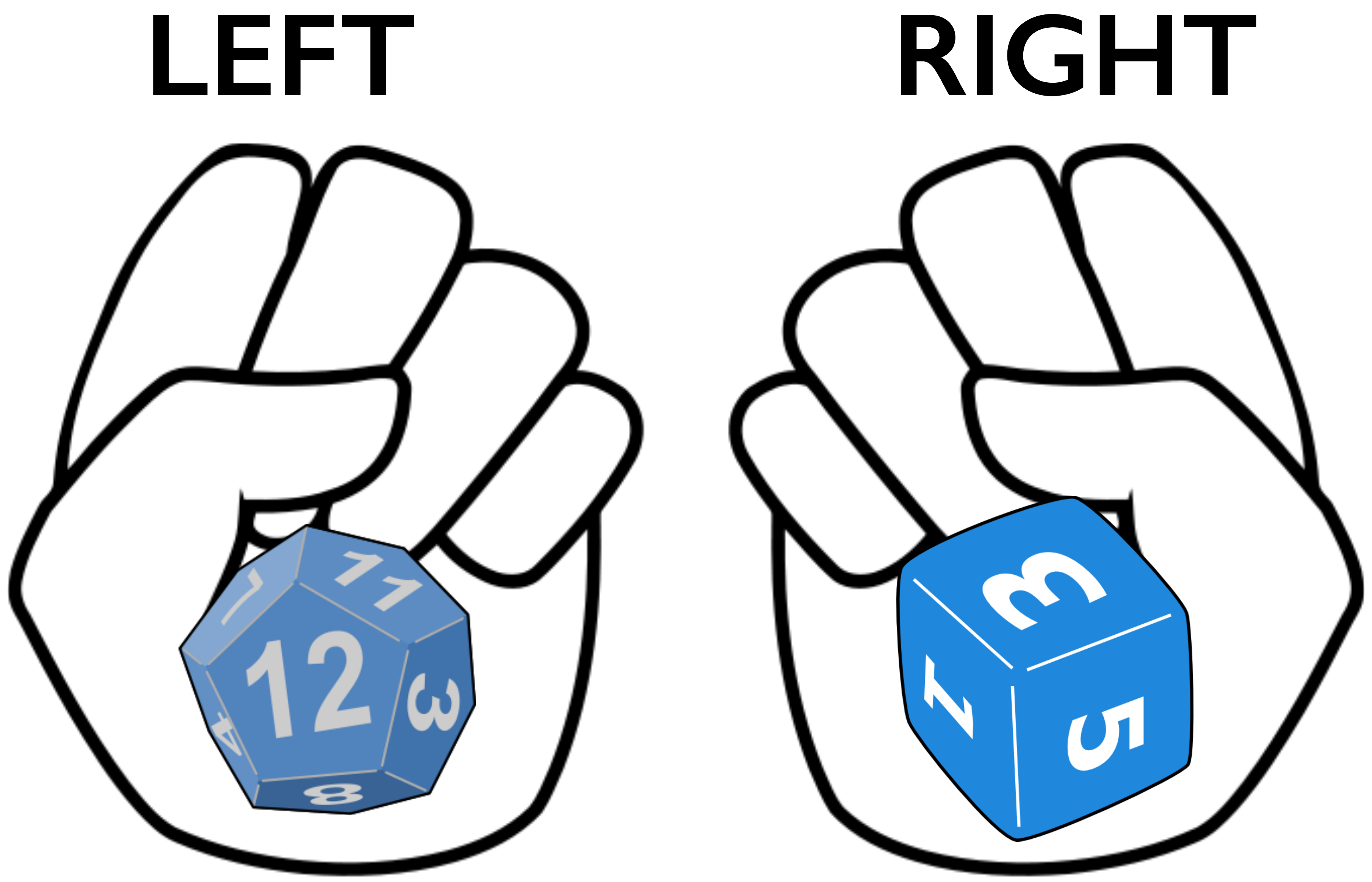
Initial guess

Two possible options:



H_1: **good** die on the Right

$P(H_1) = 0.5$



H_2: **bad** die on the Right

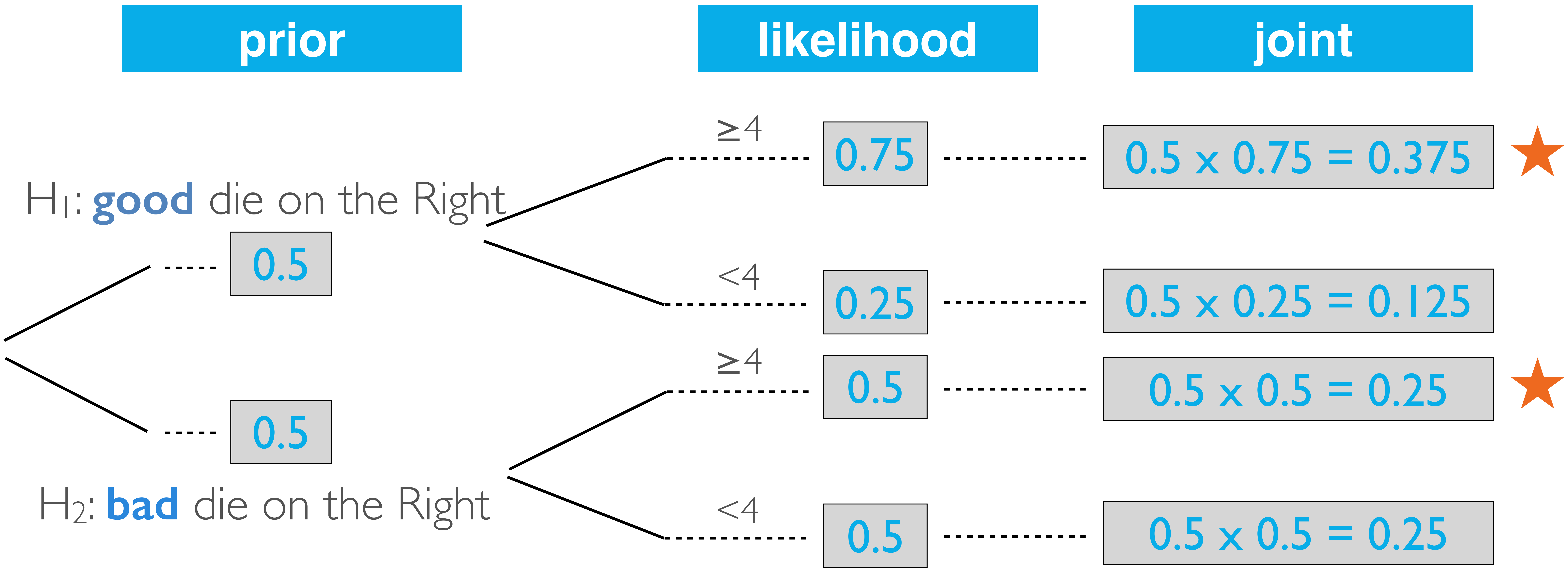
$P(H_2) = 0.5$

Prior probabilities

- These are your **prior probabilities** for the two competing claims:
  - P(H_1: **good** die on the Right) = 0.5
  - P(H_2: **bad** die on the Right) = 0.5
- These probabilities represent what you believe before seeing any data.
- You could have conceivably made up these probabilities, but instead you have chosen to make an educated guess.

Data collection

Round	Choice	Result
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		



prior

likelihood

joint

H1: **good** die on the Right

0.5

≥4 0.75 — 0.5 x 0.75 = 0.375 ★

<4 0.25 — 0.5 x 0.25 = 0.125

0.5

H2: **bad** die on the Right

≥4 0.5 — 0.5 x 0.5 = 0.25 ★

<4 0.5 — 0.5 x 0.5 = 0.25

$$P(H_1 \mid outcome \ \geq \ 4) = \frac{P(H_1 \ and \ outcome \ \geq \ 4)}{P(outcome \ \geq \ 4)}$$

$$= \frac{0.375}{0.375 + 0.25} = 0.6$$

$$P(H_2 \mid outcome \ \geq \ 4) = 1 - 0.6 = 0.4$$

posterior

Posterior probabilities

- The probabilities we just calculated is called a **posterior probabilities**:
  - P(H_1: **good** die on the Right | data) = 0.6
  - P(H_2: **bad** die on the Right | data) = 0.4
- Posterior probabilities are generally defined as P(hypothesis | data).
- These probabilities tell us the probability of a hypothesis we set forth, given the data we just observed.
- They depend on the prior probabilities as well as the likelihood of the observed data.
- Note: This is different than a p-value = P(observed or more extreme outcome | H_0 is true).

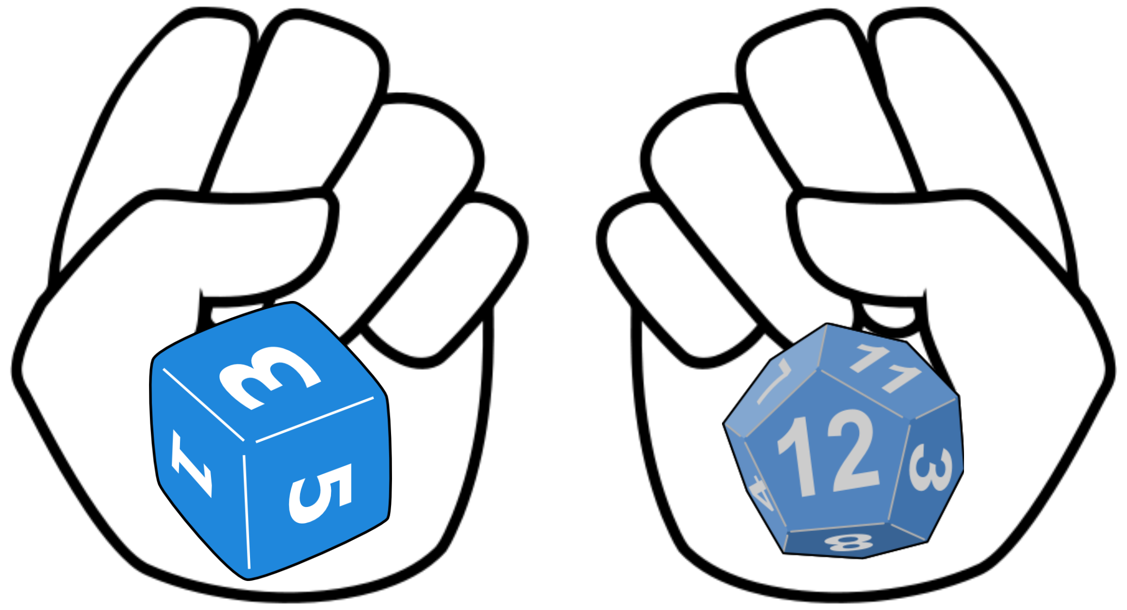
Bayes’ theorem

$$\begin{aligned} P(\theta|X) &= \frac{P(X \ and \ \theta)}{P(X)} \\ &= \frac{P(\theta) \times P(X|\theta)}{P(X)} \\ &\propto P(\theta) \times P(X|\theta) \end{aligned}$$

posterior ∝ prior × likelihood

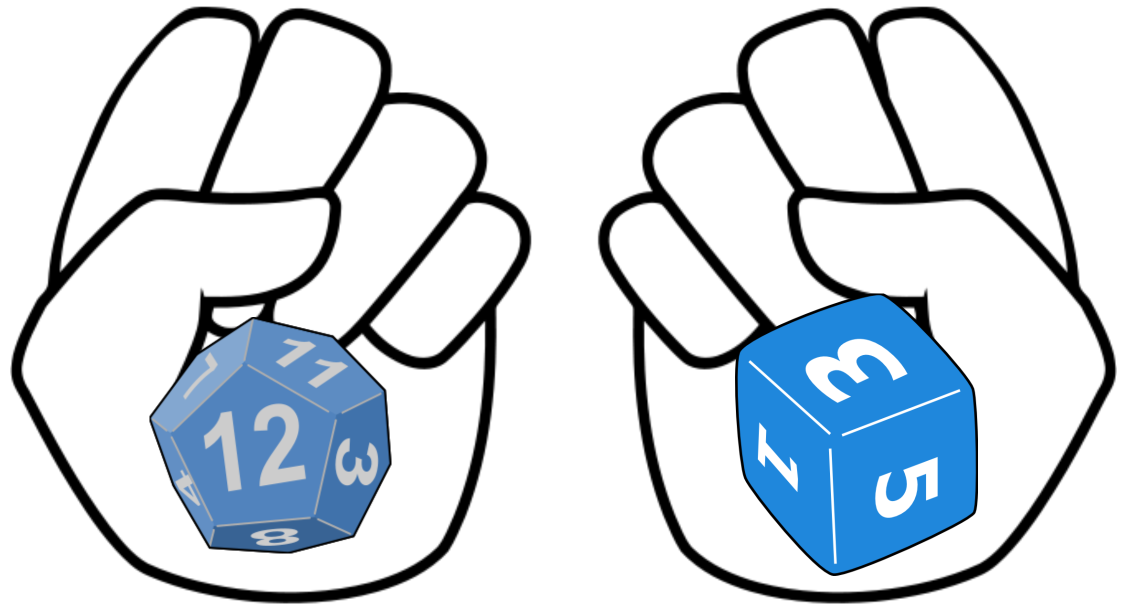
Updating your prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data.
- In the next iteration (roll) we get to take advantage of what we learned from the data.
- In other words, we **update** our prior with our posterior probability from the previous iteration.



H_1: **good** die on the Right

$P(H_1) =$ **0.6**



H_2: **bad** die on the Right

$P(H_2) =$ **0.4**

Recap

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive definition of a p-value: P(observed or more extreme outcome | H_0 is true)
- Instead base decisions on the posterior probability: P(hypothesis is true | observed data).
- A good prior helps, a bad prior hurts, but the prior matters less the more data you have.

Bayesian vs. frequentist probability

Probabilitistic statements

Consider these statements:

- The probability of flipping a coin and getting heads is ½.
- The probability of rolling snake eyes, that is, two 1s on two dice, is 1/36.
- The probability of Apple’s stock price going up today is 0.75.

How you interpret these statements depends on your definition of probability.

Frequentist definition

- If you can repeat flipping a coin indefinitely, and count how many heads you get, and divide that number by the number of flips, the value you obtain should be 0.5.
- In other words, the probability of the event is defined as the number of times the event occurs in *n* trials, when *n* goes to infinity. This is the **frequentist definition** of probability.

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

Bayesian definition

- Suppose now that you are indifferent between winning $1 if event E occurs or winning $1 if you draw a blue chip from a box with 1,000,000 × p blue chips, and 1,000,000 × (1-p) white chips.
- This means you are equating the probability of event E, *P(E)*, to the probability of drawing a blue chip from this box, *p*, in other words

$$P(E) = p$$

- This definition of probability based on your degree of belief is the **Bayesian definition**.

Practical implications

- What are the implications of these two different definitions?
- Example: confidence interval

Confidence interval

Example: Based on a 2015 Pew Research poll on 382 Republicans: *"We are 95% confident that 68% to 77% of voters who identify as Republican think any budget agreement must eliminate funding for Planned Parenthood."*

What does 95% confident mean?

- Correct answer: 95% of random samples of 382 Republicans will produce confidence intervals for the proportion of Republicans who think any budget agreement must eliminate funding for Planned Parenthood that contain the true population proportion.
- Common misconceptions:
  - There is a 95% chance that this confidence intervals includes the true population proportion.
  - The true population proportion is in this interval 95% of the time.

Confidence interval (cont.)

- The frequentist definition of probability allows us to define a probability for the confidence interval procedure, but not for a specific fixed sample.
- In the case of a specific fixed sample, when the data does not change, we will either always capture the true parameter or never capture it.
  - In other words, for a given confidence interval, the true parameter is either in it, or not.
- This is the same as saying that the probability that a given confidence interval captures the true parameter is either 0 or 1.
- The only problem is we can't know whether the probability that this given interval captures the true parameter is 0 or 1.

Credible interval

- The Bayesian definition is a bit more flexible.
- Since it's a measure of belief, it allows us to describe the unknown true parameter not as a fixed value but with a probability distribution.
- This will let us construct something like a confidence interval, except we can make probabilistic statements about the parameter falling within that range.
  - Example: *"The posterior distribution yields a 95% credible interval of 68% to 77% for the proportion of Republicans who think any budget agreement must eliminate funding for Planned Parenthood."*
- These are called credible intervals.

Practical Bayes

Breast cancer screening

- American Cancer Society estimates that about 1.7% of women have breast cancer.

$P(bc) = 0.017$

http://www.cancer.org/cancer/cancerbasics/cancer-prevalence

- Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.

$P(+ \mid bc) = 0.78$

http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html

- An article published in 2003 suggests that up to 10% of all mammograms are false positive.

$P(+ \mid no\ bc) = 0.10$

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940/

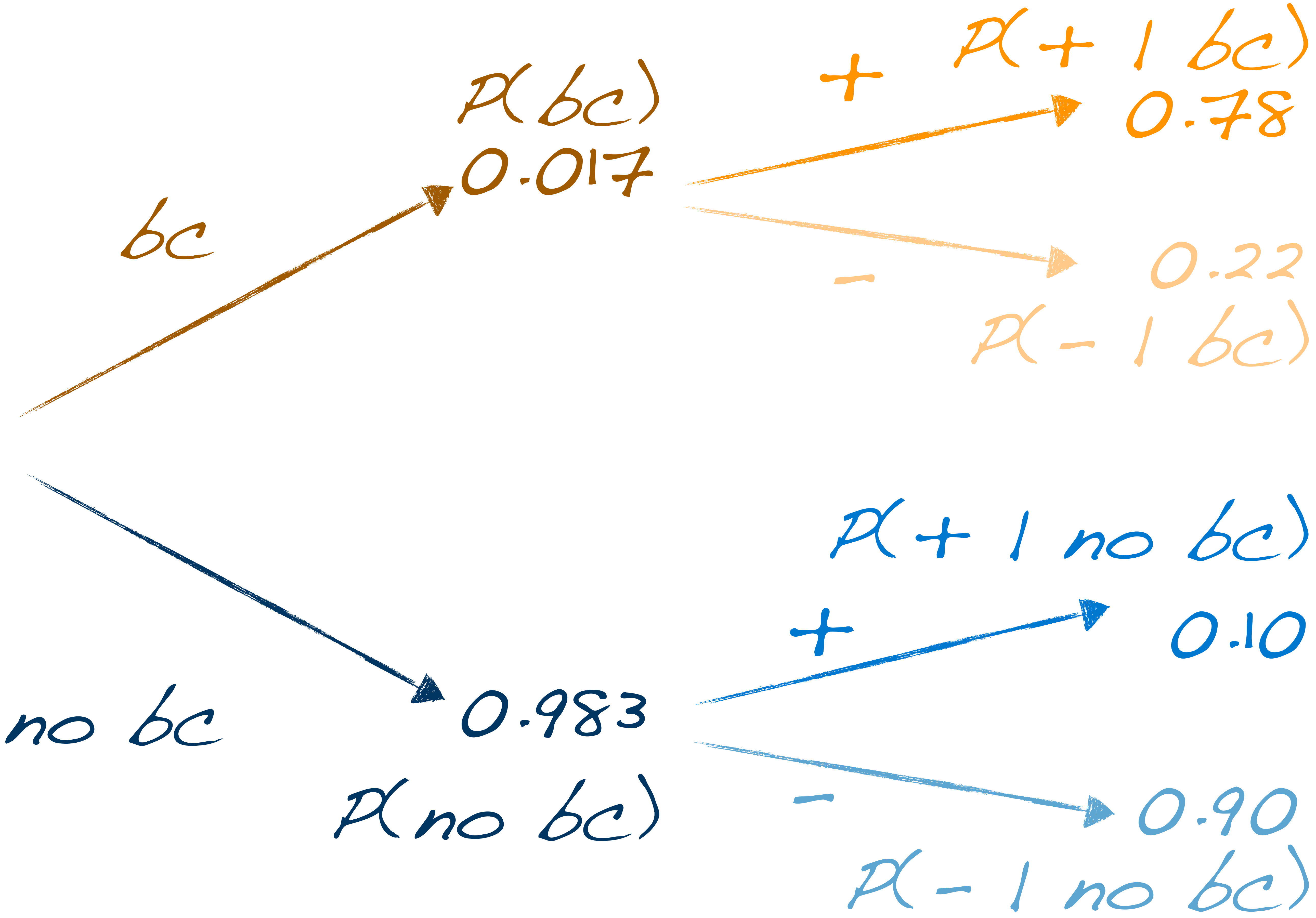
Prior probability

Prior to any testing and any information exchange between the patient and the doctor, what probability should a doctor assign to a female patient having breast cancer?

$$P(bc) = 0.017$$

Posterior probability

When a patient goes through breast cancer screening there are two competing claims: patient has cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer?



P(bc and +)

$$0.017 \times 0.78 = 0.01326$$

$$P(bc \mid +) = \frac{P(bc \text{ and } +)}{P(+)}$$

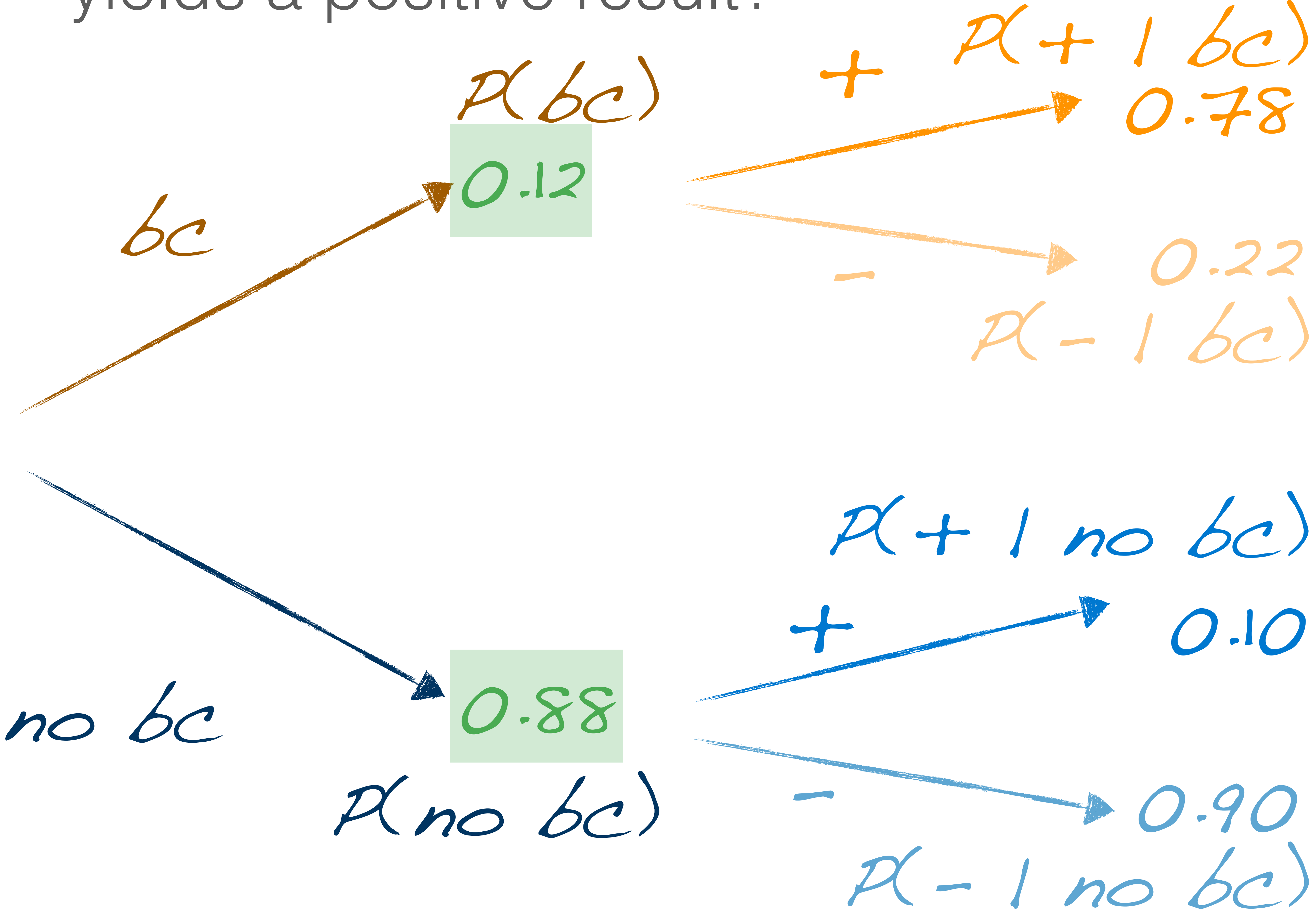
$$= \frac{0.01326}{0.01326 + 0.0983} \approx 0.12$$

$$0.983 \times 0.10 = 0.0983$$

P(no bc and +)

Retesting

Since a positive mammogram doesn't necessarily mean that the patient actually has breast cancer, the doctor might decide to re-test the patient. What is the probability of having breast cancer if this second mammogram also yields a positive result?



$$\overset{P(bc \text{ and } +)}{0.12 \times 0.78 = 0.0936}$$

$$\boxed{\begin{aligned} P(bc \mid +) &= \frac{P(bc \text{ and } +)}{P(+)} \\ &= \frac{0.0936}{0.0936 + 0.088} \approx 0.52 \end{aligned}}$$

$$\underset{P(no\ bc \text{ and } +)}{0.88 \times 0.10 = 0.088}$$

Bayesian vs. frequentist inference

M&Ms

- We have a population of M&Ms.
- The percentage of yellow M&Ms is either 10% or 20%.
- You have been hired as a statistical consultant to decide whether the true percentage of yellow M&Ms is 10%.
- You are being asked to make a decision, and there are associated payoff/losses that you should consider.

Frequentist inference

hypotheses

H_0: 10% yellow M&Ms

H_A: >10% yellow M&Ms

sample

RGYBO

obs. data

$k = 1, n = 5$

p-value

$$P(k \geq 1 | n = 5, p = 0.10)$$
$$= 1 - P(k = 0 | n = 5, p = 0.10)$$
$$= 1 - 0.90^5 \approx 0.41$$

→ Fail to reject H_0

Bayesian inference

hypotheses

H_1: 10% yellow M&Ms

H_2: 20% yellow M&Ms

prior

$$P(H_1) = 0.5$$

$$P(H_2) = 0.5$$

sample

RGYBO

obs. data

$k = 1,\ n = 5$

likelihood

$$P(k = 1 \mid H_1, n = 5) = \binom{5}{1} 0.10 \times 0.90^4 \approx 0.33$$

$$P(k = 1 \mid H_2, n = 5) = \binom{5}{1} 0.20 \times 0.80^4 \approx 0.41$$

posterior

$$\begin{aligned} P(H_1 \mid k = 1, n = 5) &= \frac{P(H_1) \times P(H_1 \mid k = 1, n = 5)}{P(k = 1, n = 5)} \\ &= \frac{0.5 \times 0.33}{0.5 \times 0.33 + 0.5 \times 0.41} \\ &\approx 0.46 \end{aligned}$$

$$\begin{aligned} &P(H_2 \mid k = 1, n = 5) \\ &= 1 - 0.46 = 0.54 \end{aligned}$$

Bayesian vs. frequentist inference

	Frequentist	Bayesian	
obs. data	P(k or more | 10% yellow)	P(10% yellow | n,k)	P(20% yellow | n,k)
$n = 5, k = 1$	0.41	0.44	0.56
$n = 10, k = 2$	0.26	0.39	0.61
$n = 15, k = 3$	0.18	0.34	0.66
$n = 20, k = 4$	0.13	0.29	0.71

Commonly aired issues with p-values

- Fixation with $\alpha = 0.05$
  - it can be far too weak in some cases
  - it can be too strong in others, causing promising lines of inquiry to be abandoned
- Doesn't measure the magnitude or the importance of the effect being investigates
  - use of confidence intervals is often suggested as a solution (but is not a cure-all)
- Commonly misinterpreted as
  - $P(H_0 \text{ true} \mid \text{data})$
  - $P(\text{error is made in rejecting } H_0)$
  - P(replicating experiment would reach the same conclusion)

"What's wrong with [null hypothesis significance testing]? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" - Cohen, 1994

Borrowed from https://stat.duke.edu/~berger/talks.html#p-value

ASA’s statement on p-values

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Source http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN

Generalizing Bayes

Morning after

- A study addressed the question of whether the controversial abortion drug RU 486 could be an effective "morning after" contraceptive.
- The study participants were women who came to a health clinic asking for emergency contraception after having had sex within the previous 72 hours.
- Investigators randomly assigned the women to receive either RU486 or standard therapy consisting of high doses of estrogen and a synthetic version of progesterone.
- Of the women assigned to RU486 (treatment), 4 became pregnant. Of the women who received standard therapy (control), 16 became pregnant.
- How strongly does this information indicate that the treatment is more effective than the control?

Example modified from Don A. Berry’s, Statistics: A Bayesian Perspective, 1995.

Framework

- To simplify matters let's turn this problem of comparing two proportions to a one proportion problem: consider only the 20 total pregnancies, and ask how likely is it that 4 pregnancies occur in the treatment group.
- If the treatment and control are equally effective, and the sample sizes for the two groups are the same, then the probability the pregnancy come from the treatment group is simply $p = 0.5$.
- We'll consider any value of p between 0 and 1, $0 \leq p \leq 1$.

Prior distribution

Assume an **uninformative** uniform distribution, where a = 0 and b = 1

$$f(x) = \begin{cases} \frac{1}{b-a} & \mathrm{for}\ a \le x \le b, \\ 0 & \mathrm{for}\ x < a\ \mathrm{or}\ x > b \end{cases}$$

Prior, likelihood, and posterior

prior

$$f(p) = 1$$

Uniform

likelihood

$$f(k|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial

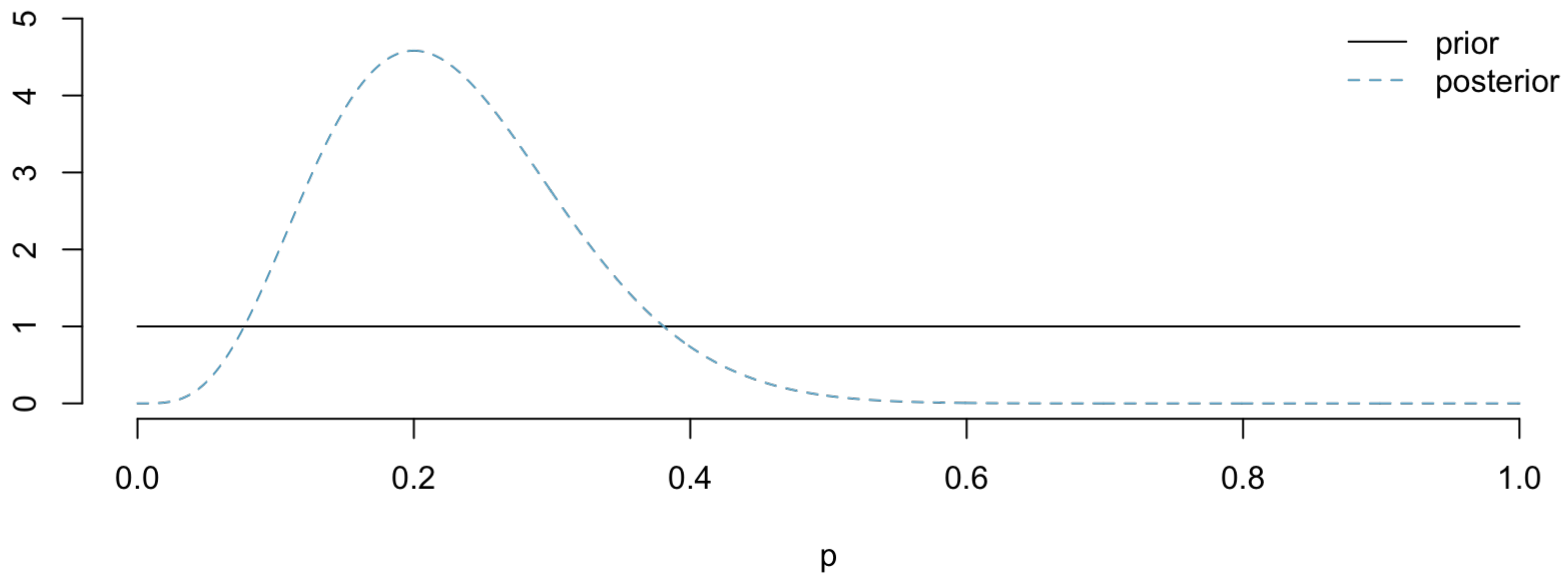
posterior

$$\begin{aligned} f(p|k) &\propto f(p) \times f(p|k) \\ &\propto 1 \times \binom{n}{k} p^k (1-p)^{n-k} \\ &\propto p^k (1-p)^{n-k} \end{aligned}$$

Beta

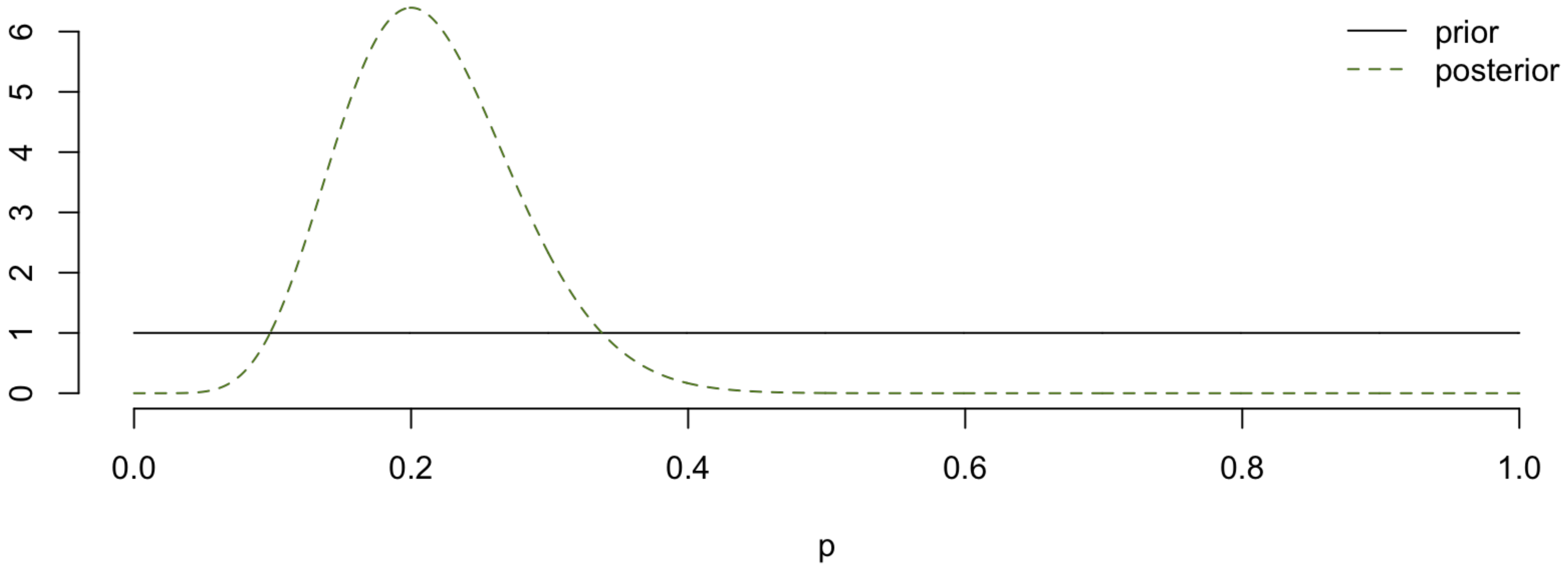
Prior and posterior, visualized

n = 20, k = 4



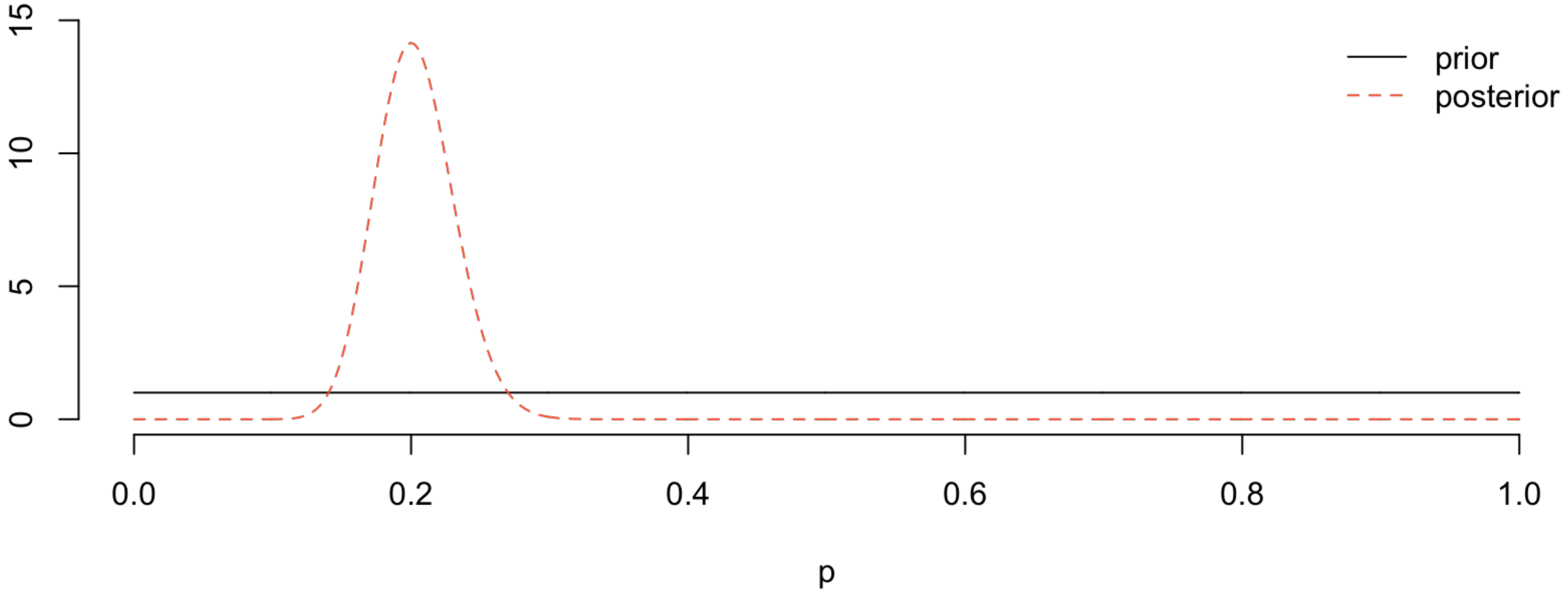
What if we had more data?

n = 40, k = 8

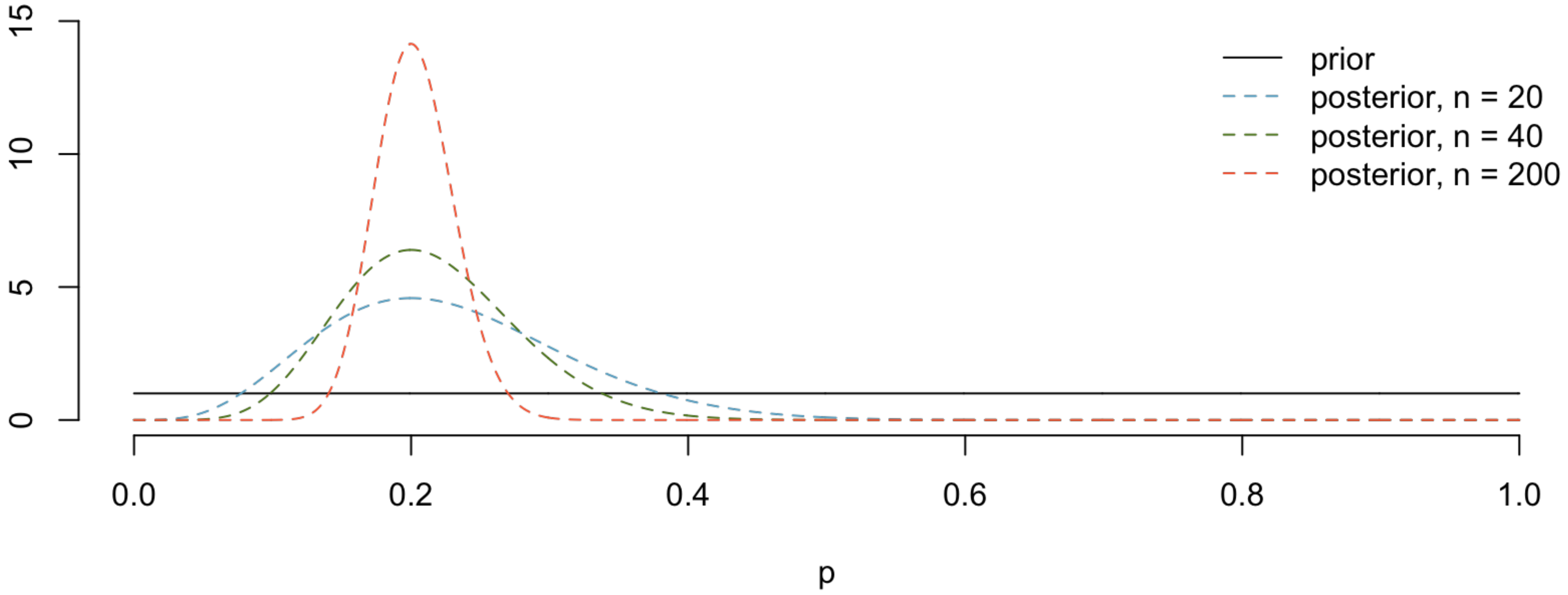


Or even more data?

$n = 200, k = 40$



Putting it all togther



Summary

Bayesian thinking…

- is a natural analog to how we think and learn — updating beliefs based on empirical data
- is useful in practical situations
- brings basic probability into the context of decision making scenarios more naturally than the frequentist p-value
- can be used in the continuous space and when building models

Bayesian approaches…

- rely on the choice of the prior, but the prior matters less the more data you have
- may require heavier use of computational tools when working with complicated models and/or the posterior does not follow a known distribution
  - but things can be complicated with frequentist models as well (e.g. there may not be closed form solution for the MLE)

Thank you!

email	mine@stat.duke.edu
slides	http://bit.ly/bayesian_baby_steps_OPRE