Bayesian Approaches 1

Bayesian

- data are treated as fixed observations
- models (parameters) are treated as random variables
- we compute the probability of all models
- we end up with a richer understanding of relative probability of all models

Frequentist

- data (sample) treated as a random variable
- models (population parameters) are treated as fixed quantities
- we compute the probability of one model (H0)
- we make a decision (reject H0 or not)

Bayes Theorem

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

- probability of B, given A, equals probability of A, given B, times prob of B, divided by probability of A
- p(B|A) is the posterior
- p(A|B) is the likelihood
- p(B) is the prior
- p(A) is the evidence

$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$

Bayes Theorem

$$p(model|data) = \frac{\overbrace{p(data|model)}^{\text{likelihood}}\,\overbrace{p(model)}^{\text{prior}}}{\underbrace{p(data)}_{\text{evidence}}}$$

- probability of model, given data, equals probability of data, given model, times prob of model, divided by probability of data
- p(model|data) is the posterior
- p(data|model) is the likelihood
- p(model) is the prior
- p(data) is the evidence

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Bayes Theorem

$$p(model|data) = \frac{p(data|model)p(model)}{p(data)}$$

- p(data), the marginal probability of data across all models, can be computed as the sum of conditional probabilities of data given each model:

$$p(data) = \sum_{model_i} p(data|model_i)p(model_i)$$

$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$

Bayes Theorem

$$p(model|data) = \frac{p(data|model)p(model)}{p(data)}$$

- We will first look at a discrete probability example, using single-point probabilities, to show how these calculations work
- We will then look at an example of this approach using continuous probability **distributions** instead of point probabilities

Discrete Example

- Let's say you take a home pregnancy test and it comes out positive. What is the probability that you are pregnant?
- Let's say we know the test is 90% accurate
- The "data" we have is
  - p(test+ | preg) = 0.90
  - and test was +
- We want the prob of the "model": preg
- we want to know p(preg | test+)

Discrete Example

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

$$p(preg|test+) = \frac{p(test + |preg)p(preg)}{p(test+)}$$

- p(test+ | preg) = 0.90 (accuracy of test)
- p(test+ | not preg) = 0.50 (false pos rate)
- let's say we also estimate p(preg) ~ 0.15

Discrete Example

$$p(preg|test+) = \frac{\overset{.90}{p(test+|preg)}\,\overset{.15}{p(preg)}}{p(test+)?}$$

- p(test+ | preg) = 0.90
- p(preg) = 0.15
- what is p(test+)? $p(data) = \sum_{model_i} p(data|model_i)p(model_i)$
- $p(test+) = p(test+|preg)p(preg) + p(test+|notpreg)p(notpreg)$
- p(test+) = (.90)(.15) + (.50)(.85) = 0.56

Discrete Example

$$p(preg|test+) = \frac{\overset{.90}{p(test+|preg)}\,\overset{.15}{p(preg)}}{p(test+)\;.56}$$

- p(test+) = (.90)(.15) + (.50)(.85) = .56
- so p(preg | test+) = (.90)(.15) / (.56) = .241
- so probability of pregnant given pos test is 24.1%

Effect of the prior

$$p(model|data) = \frac{\overset{\text{likelihood}}{p(data|model)}\,\overset{\text{prior}}{p(model)}}{p(data)}$$

- the **posterior** p(model | data) is proportional to the **likelihood** p(data | model) multiplied by the **prior** p(model)
- our prior expectation (or previous findings, i.e. data) modulates our prediction of the future
- can be viewed as both a virtue and a shortcoming of the Bayesian approach

Effect of the prior

$$p(preg|test+) = \frac{\overset{.90}{p(test+|preg)}\,\overset{.15}{p(preg)}}{p(test+)\ .56}$$

prior	posterior
0.1	0.17
0.2	0.31
0.3	0.44
0.4	0.55
0.5	0.64

prior	posterior
0.6	0.73
0.7	0.81
0.8	0.88
0.9	0.94
0.99	0.99

Updating the Model

$$p(model|data) = \frac{\overset{\text{likelihood}}{p(data|model)}\,\overset{\text{prior}}{p(model)}}{p(data)}$$

- When you collect new data, you can update your model
- the posterior from the previous model becomes the prior for the new model

Updating the Model

likelihood prior

$$p(model|data) = \frac{p(data|model)p(model)}{p(data)}$$

- Let's say you take another preg test
- We know from our previous calculation:
  - p(preg | test+) = .241
- The other quantities are the same
  - p(test+ | preg) = 0.90 (accuracy of test)
  - p(test+ | not preg) = 0.50 (false pos rate)

Updating the Model

$$p(model|data) = \frac{\overbrace{p(data|model)}^{\text{likelihood}}\overbrace{p(model)}^{\text{prior}}}{p(data)}$$

- Let's say you take another preg test
- We know from our previous calculation:
  - p(preg | test+) = .241
  - this becomes our new prior

Updating the Model

$$p(preg|test+) = \frac{p(test+|preg)p(preg)}{p(test+)}$$

$$p(preg|test+) = \frac{p(test+|preg)p(preg)}{p(test+|preg)p(preg) + p(test+|notpreg)p(notpreg)}$$

- p(preg | test+) =
  (.90)(.241) / ((.90)(.241) + (.50)(1-.241))
  = .364
- so after a second positive test,
  p(preg | test+) is now 36.4%

Test #	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10
p(preg|test+)	0.51	0.65	0.77	0.86	0.92	0.95	0.97	0.98

Updating the Model

- seems like an appropriate thing to do in science
- when new data are gathered, we can re-evaluate a hypothesis
- we do not begin anew (ignorant) each time we ask a question
- previous research provides us information about the merits of the hypothesis

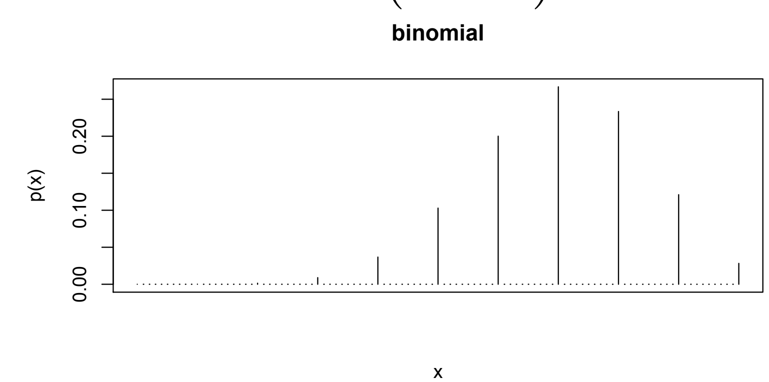
Bayes with Distributions

- in previous example, the likelihood and prior were both single quantities (point probabilities)
- typically Bayesian approaches use full **probability distributions**
- essentially allows us to **evaluate probability of a whole range of possible models**, at once

Bayes with Distributions

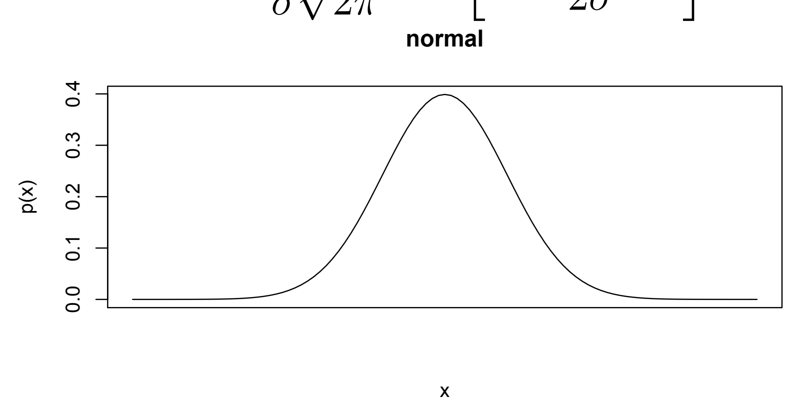
- don’t worry, remember probability distributions are just mathematical functions of a parameter vector
- e.g. binomal prob of k successes in n trials with prob(success) p, is

$$p(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$



- e.g. normal prob of a value x, with mean mu and standard dev sigma, is

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Bayes with Distributions

$$p(model|data) = \frac{\overset{\text{likelihood}}{p(data|model)}\,\overset{\text{prior}}{p(model)}}{p(data)\ \text{evidence}}$$

- let's look at an example: coin flipping
- is my coin fair?
- "data" are 3 flips of the coin: (H, H, T)
- "model" is a proposed process by which the outcome of our coin flip is determined

Bayes with Distributions

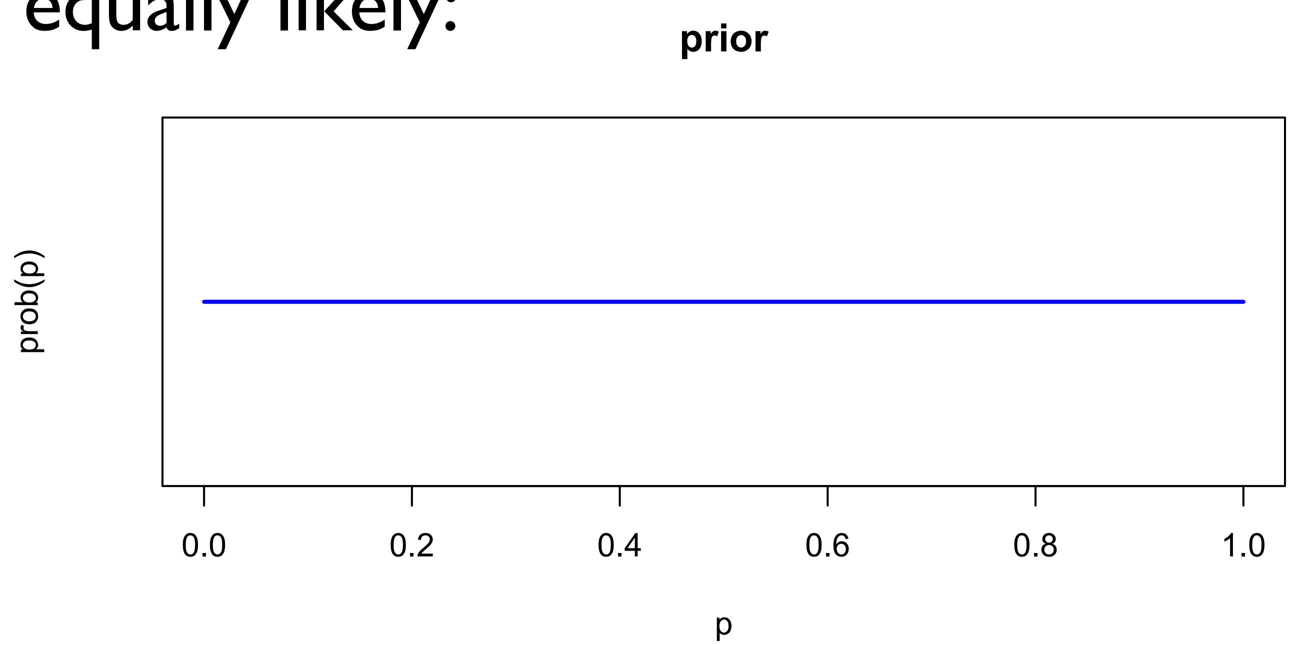
$$p(model|data) = \frac{\overset{\text{likelihood}}{p(data|model)}\,\overset{\text{prior}}{p(model)}}{\underset{\text{evidence}}{p(data)}}$$

- since outcomes are binary (H,T) a natural choice of model is a binomial distribution
- we know the likelihood function for a binomial model is
$$p(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$
- so our "data" are: n=3 tries, k=2 successes (assume Heads=success, Tails=failure)
- likelihood function gives us p(k | n,p) but what we want is the posterior: p(p | n,k) where p is prob(success) (fair is p=0.50)
- according to Bayes theorem this equals likelihood*prior/evidence

Bayes with Distributions

$$p(model|data) = \frac{\overbrace{p(data|model)}^{\text{likelihood}}\overbrace{p(model)}^{\text{prior}}}{\underbrace{p(data)}_{\text{evidence}}}$$

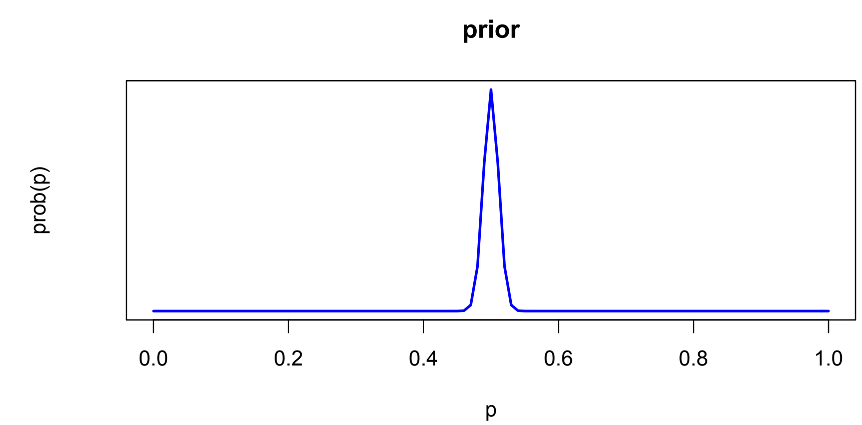
- what should our prior be?
- prior is probability of "model" == probability distribution over possible values of p
- we could decide on an "uninformative prior", postulating that all values of p are equally likely:



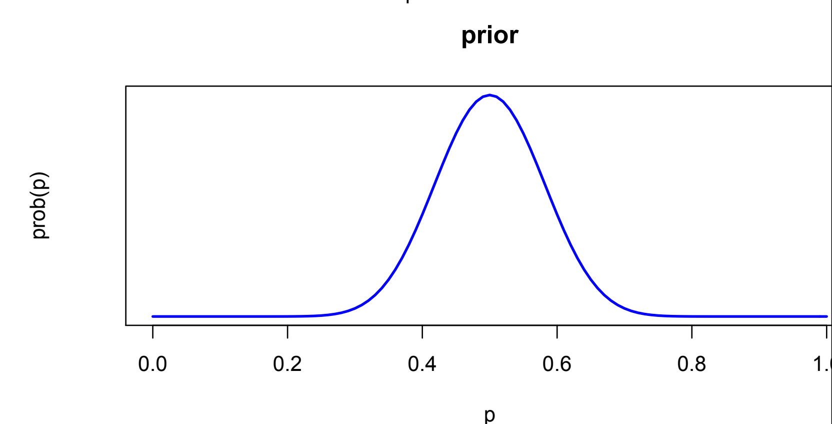
Bayes with Distributions

$$p(model|data) = \frac{\overset{\text{likelihood}}{p(data|model)}\,\overset{\text{prior}}{p(model)}}{\underset{\text{evidence}}{p(data)}}$$

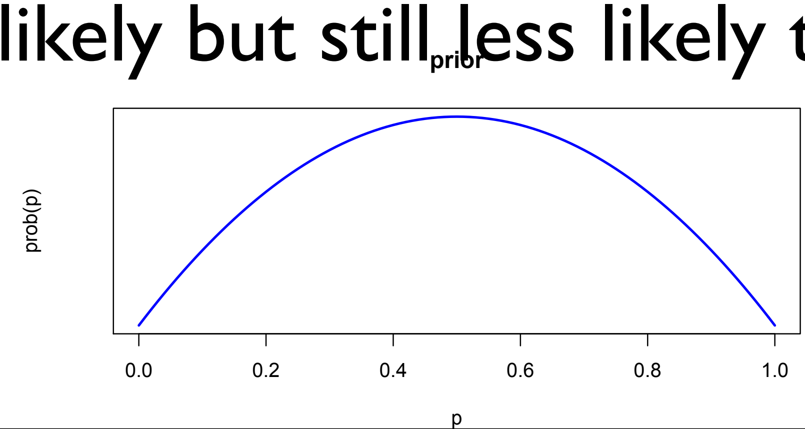
- we could decide that since every coin we have seen in the past has been fair, we expect that this coin will be fair as well and so p will likely be = 0.50



- but how unlikely are values other than p?
- very unlikely?
- moderately unlikely?



- not terribly unlikely but still less likely than .50?



Bayes with Distributions

$$p(model|data) = \frac{\overbrace{p(data|model)}^{\text{likelihood}}\,\overbrace{p(model)}^{\text{prior}}}{\underbrace{p(data)}_{\text{evidence}}}$$

- it's totally up to us to decide on the prior, in several aspects:

1. **scientific/theoretical Q**: in general, what should its shape be?
2. **practical Q**: how do I characterize the prior?
   - "by hand", e.g. as a table (a list) of parameter values & probabilities
   - "algebraically", as a mathematical equation
     A. any old function of our choosing, OR
     B. a specific equation that will help us later in computing the posterior (known as a **conjugate prior**)

Bayes with Distributions

- Two general approaches to computing the posterior:
- **Analytic**: choosing a likelihood model and a conjugate prior from a (relatively short) list of known forms, and taking advantage of clever algebra/calculus that results in a very simple expression for the posterior

Bayes with Distributions

- **Numerical**: you're free to specify your likelihood and your prior as whatever you want, and use iterative computing methods and powerful computers to estimate the posterior distribution
  - grid approximation approach
  - Markov Chain Monte-Carlo (MCMC)

Analytic Approach

- recall our data: 3 coin flips, 2 successes (2 HEADS, one TAILS)
- is the coin fair? === what is prob p in our binomial model
- we want the posterior: $$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$
- likelihood P(D|M) is given by the binomial distribution

$$p(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

- it turns out that a conjugate prior for the binomial, is the **Beta distribution**
- http://en.wikipedia.org/wiki/Conjugate_prior

Conjugate Priors

- If the posterior distribution $p(\theta|x)$ is in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood

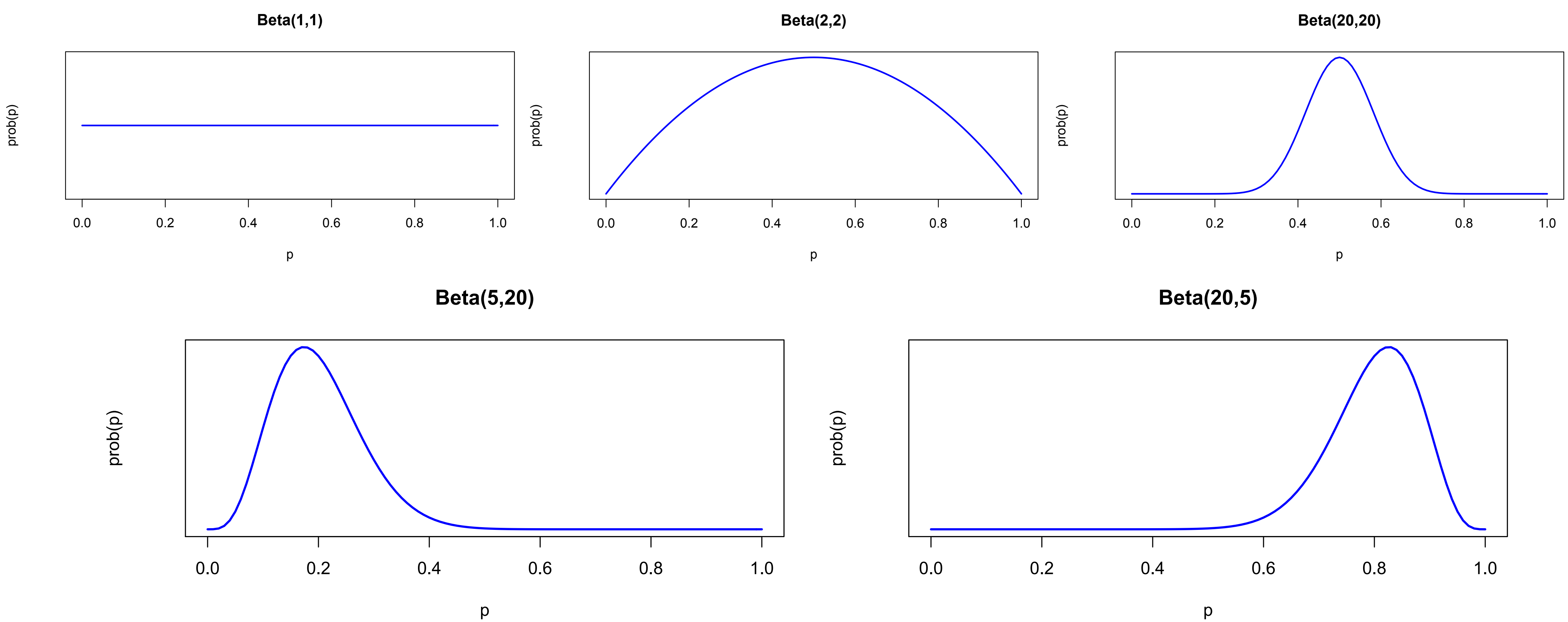
The Beta Distribution

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

http://en.wikipedia.org/wiki/Beta_distribution

- crystal clear, right? :) no of course not
- don't fret though ... this is just a mathematical equation
- it takes in **parameters alpha and beta** and spits out nice looking curves for x values between 0 and 1
- this is convenient for characterizing prior on p, since in our coin, p is betwen 0 and 1

The Beta Distribution



Conjugate Priors

- when you use a prior that is a conjugate for the likelihood, then computing the posterior turns out to be a piece of cake
- clever calculus ninjas have worked out the algebra, and often the posterior can be expressed as a really simple manipulation of the parameters of the likelihood and prior

Conjugate Priors

- for example for the binomial, we have our likelihood function $\mathrm{prob}(k|n,p) = \mathrm{binomial}(k,n,p)$

- and if we specify our prior using a Beta distribution $\mathrm{prob}(p) = \mathrm{beta}(\alpha,\beta)$

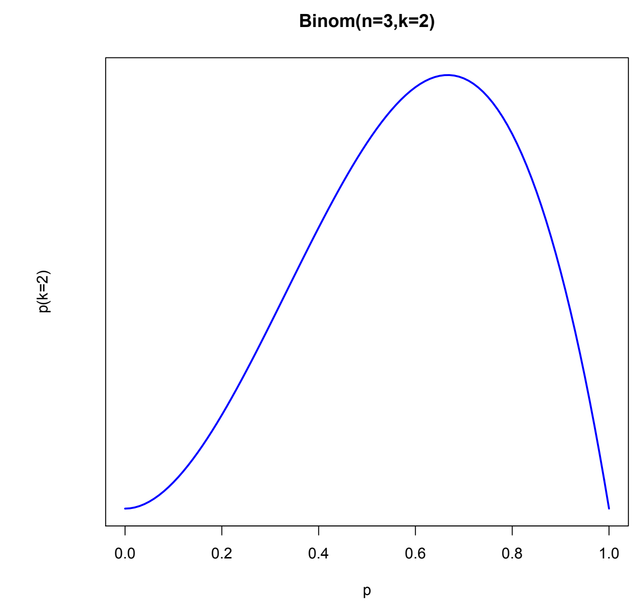
- then the posterior turns out to be equal to another Beta function, with modified alpha and beta parameters: $\mathrm{prob}(p|k,n) = \mathrm{beta}(k+\alpha, N-k+\beta)$

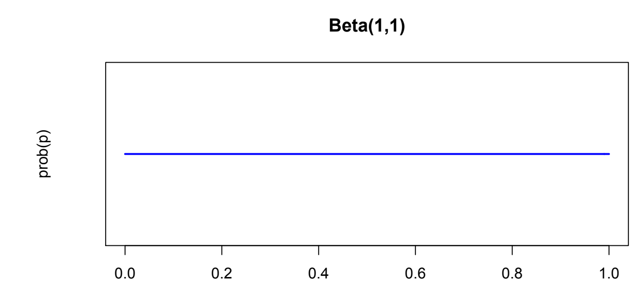
- thank you calculus ninjas!

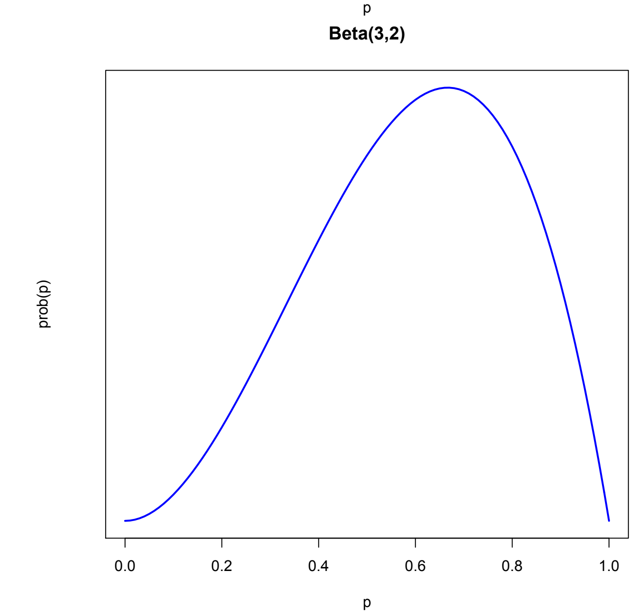
- we don't even need to calculate anything

Back to our example

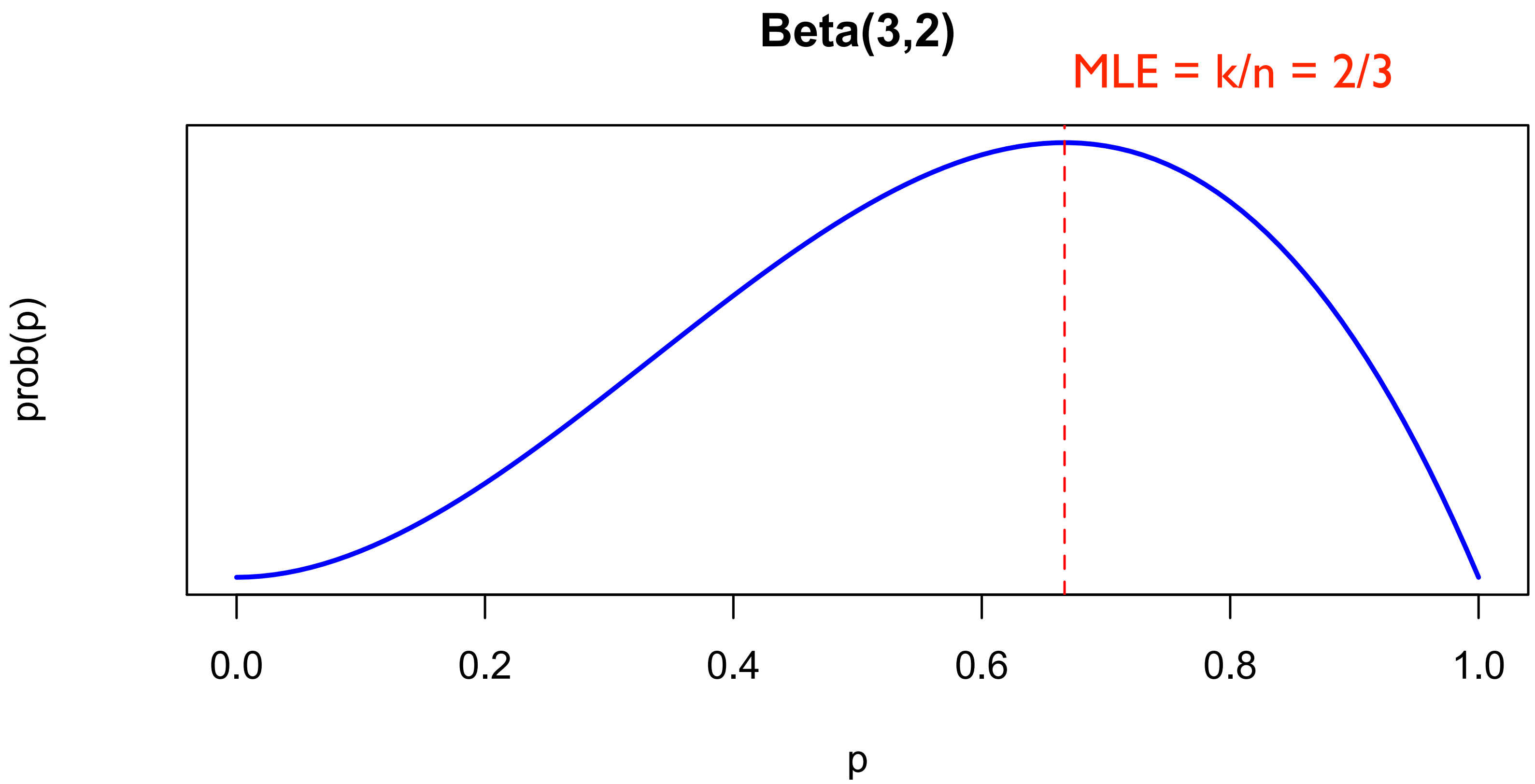
- coin flip: n=3 trials, k=2 success
- likelihood is binomial(n,k,p)
    - n=3, k=2, p is unknown
- prior is Beta(alpha,beta)
    - let's choose a flat prior, alpha=1, beta=1
- our calculus ninjas gave us:
- posterior is Beta(2+1, 3-2+1)







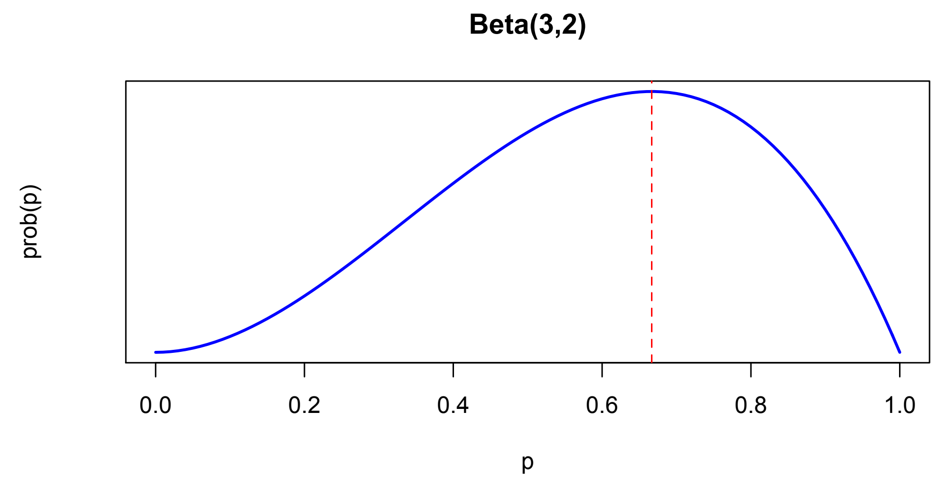
Our curvy posterior



- MLE of p is 0.667
- the posterior also gives us the entire curve

Describing the posterior

- **Graphically**



- **Summary statistics**
    - Analytic
        - well known expressions for mean, variance, mode, MLE, etc...
        - e.g. mean of a Beta dist is $\dfrac{\alpha}{\alpha+\beta}$
        - variance is $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

http://en.wikipedia.org/wiki/Beta_distribution

Describing the posterior

- Numerical
  - use a random number generator to draw a large number of values from the posterior distribution, then compute summary stats from those random draws
- in programs like R we have a whole set of random number generators for lots of probability distributions
- normal, beta, binomial, exponential, poission, etc etc etc...

Numerical Example

- let's compute the **95% credible interval** sometimes called Highest Density Region (HDR)

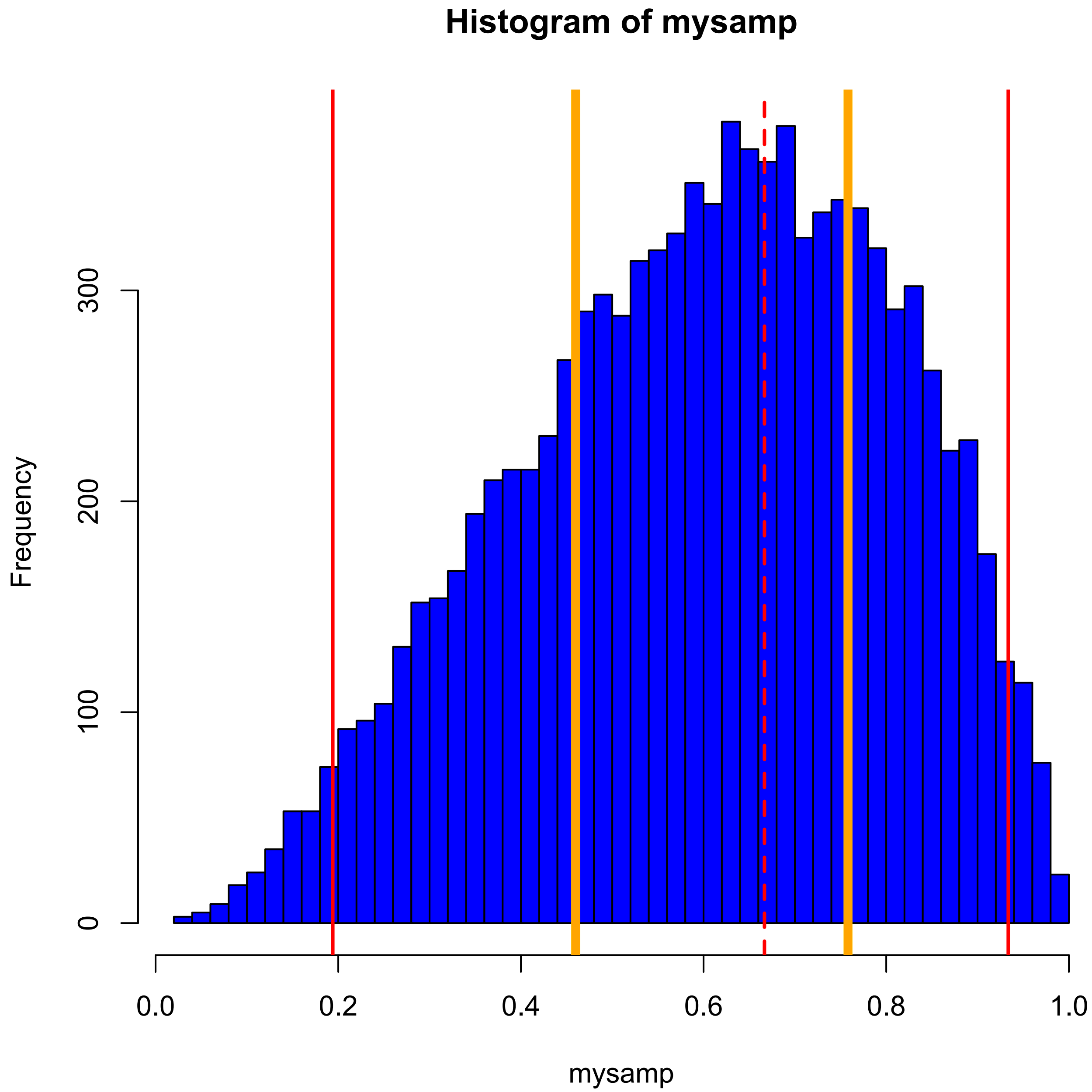
```
> mysamp <- rbeta(10000, 3, 2)
> hist(mysamp, breaks=50, col="blue")
> ci95 = quantile(mysamp, c(.025,.975))
> ci95
     2.5%     97.5%

0.1940172 0.9335989

> abline(v=2/3, col="red", lwd=2, lty=2)
> abline(v=ci95, col="red", lwd=2)
```

- or: **50% credible interval**

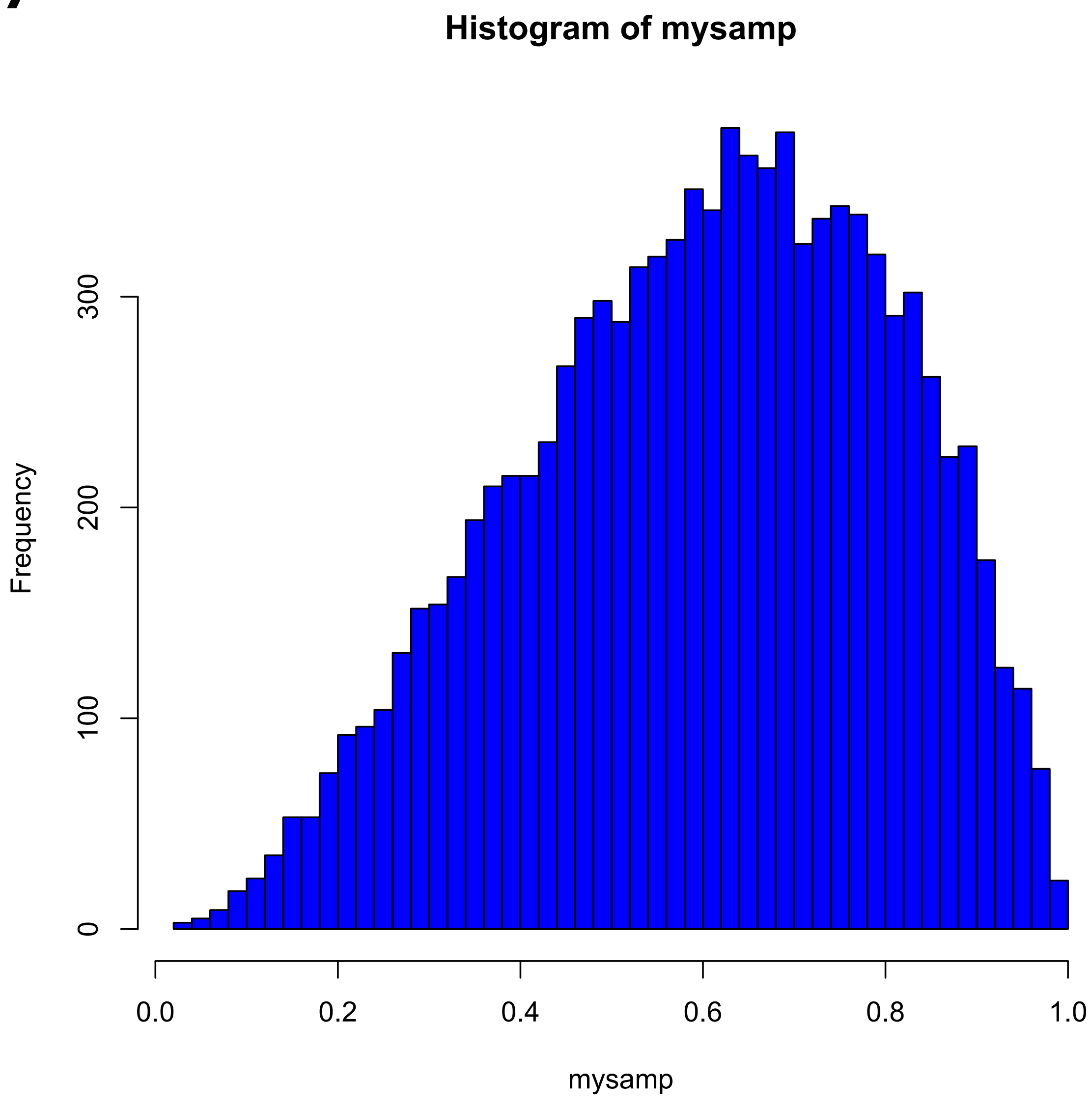
```
> ci50 = quantile(mysamp, c(.25,.75))
> abline(v=ci50, col="orange", lwd=5)
```



Numerical Example

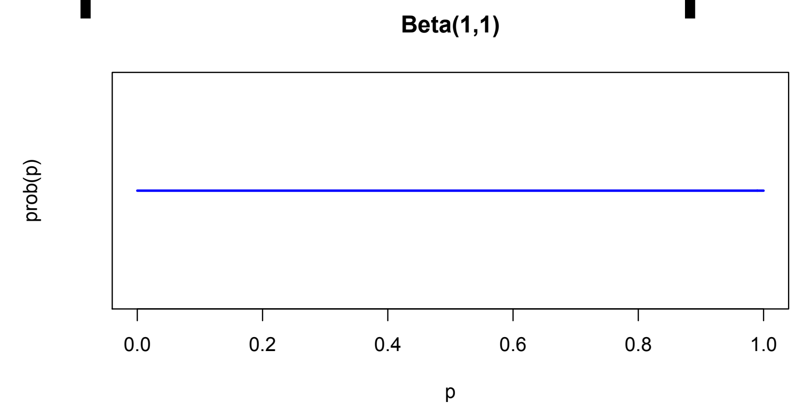
- or any other quantity you could ever want
- after all, you have the ability to sample from the posterior distribution as much as you want
- i.e. you can sample from prob(model | data)

  and characterize its entire shape, over the full range of possible values of the model
- essentially you can evaluate the relative prob of all models

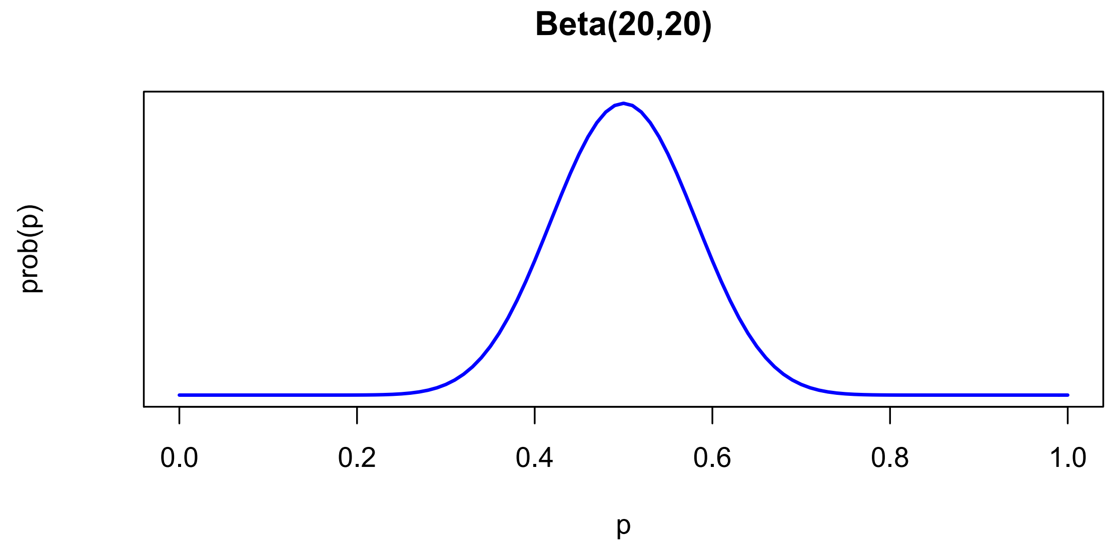


Try a different Prior

- We used a flat prior for the previous example

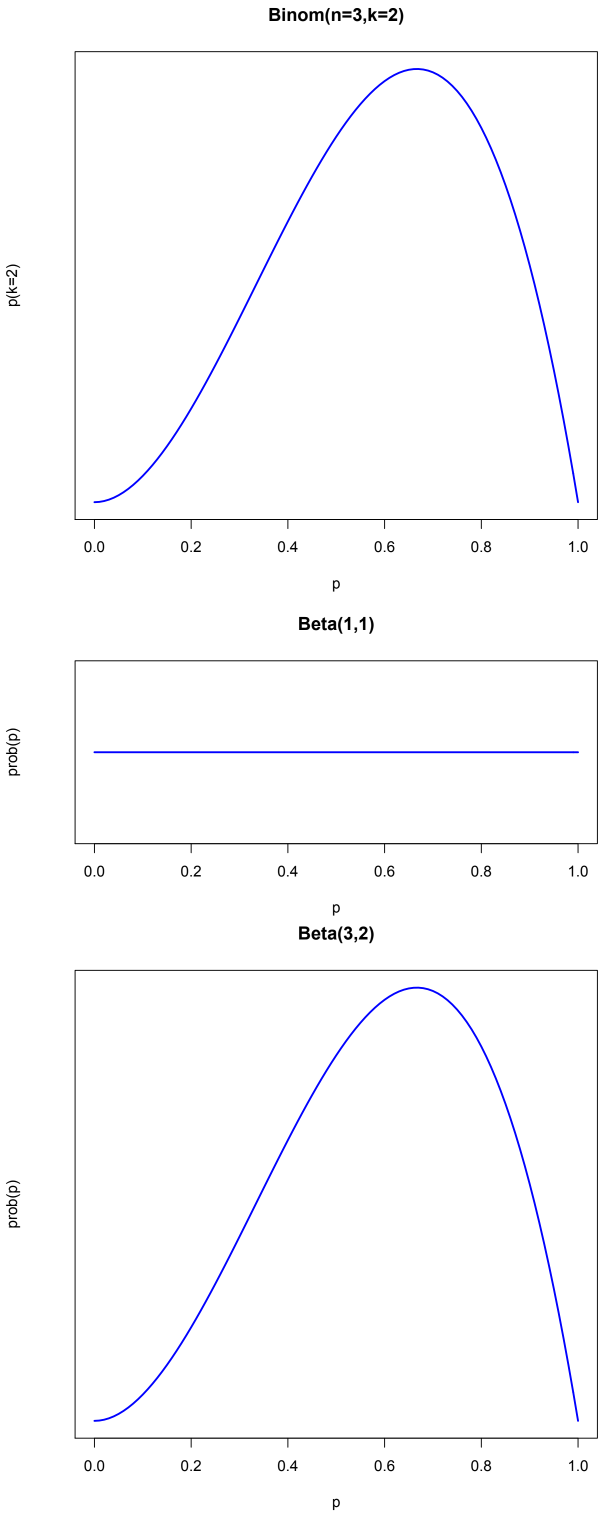


- Let's repeat but use a prior that expresses our evidence to date that coins are in fact fair Beta(20,20)



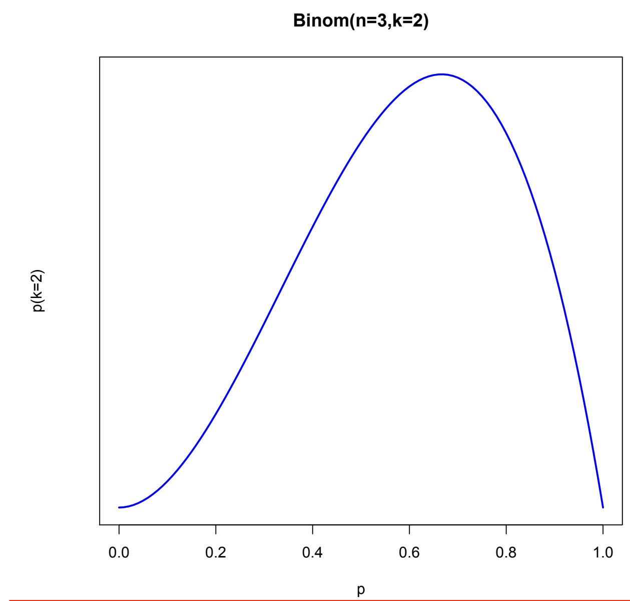
Try a different Prior

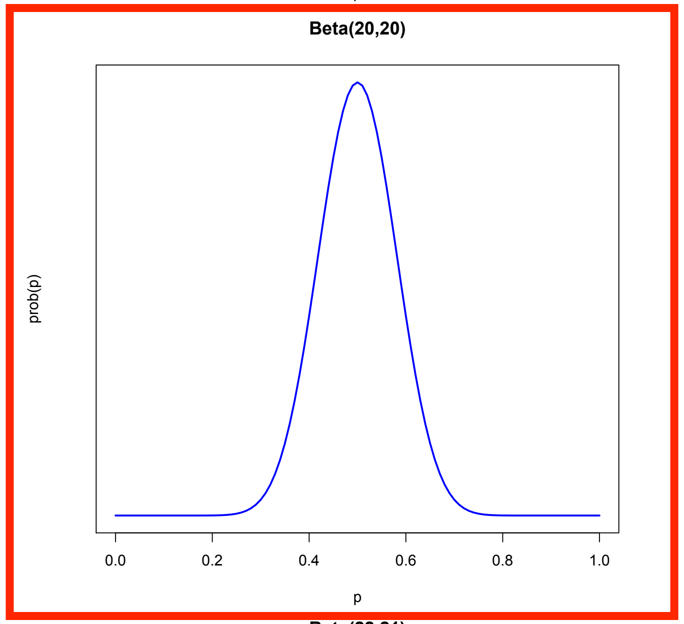
- likelihood
- flat prior didn't really change likelihood
- so posterior essentially equals likelihood

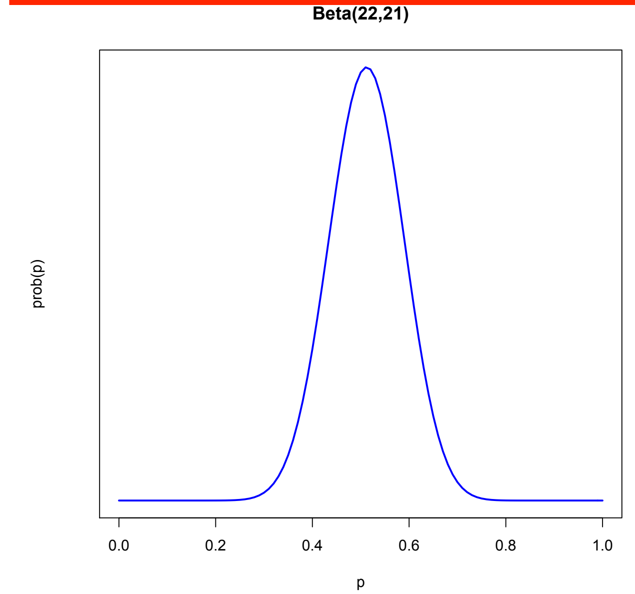


Try a different Prior

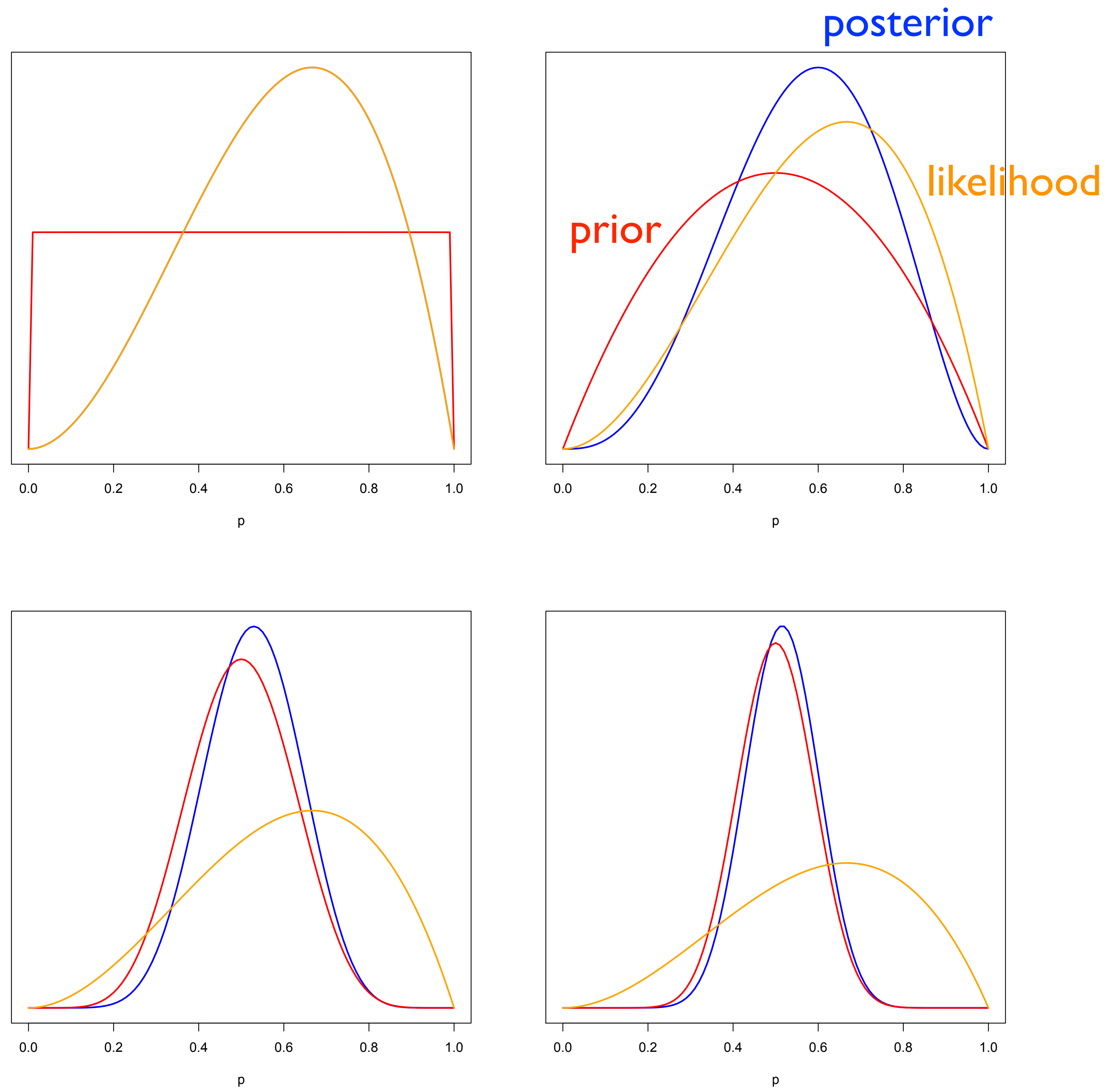
- coin flip: n=3 trials, k=2 success
- likelihood is binomial(n,k,p)
  - n=3, k=2, p is unknown
- prior is Beta(alpha,beta)
  - **let's choose an informative prior, alpha=20, beta=20**
- our calculus ninjas gave us:
- posterior is Beta(2+20, 3-2+20)







Effect of Prior



Criticisms of Bayesian Approach

- the prior: too much "subjectivity"?
- data fixed, models (parameters) random

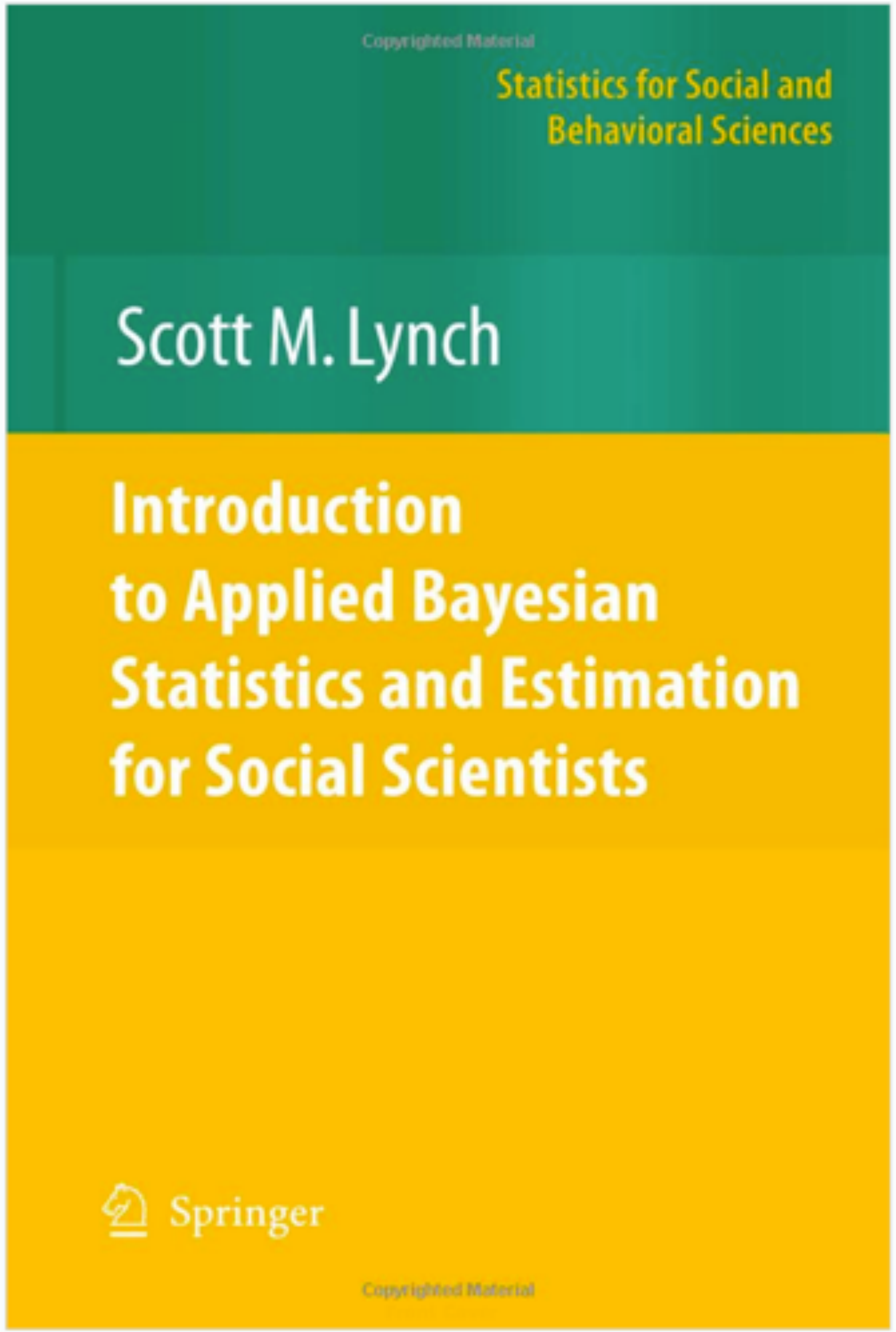
Advantages

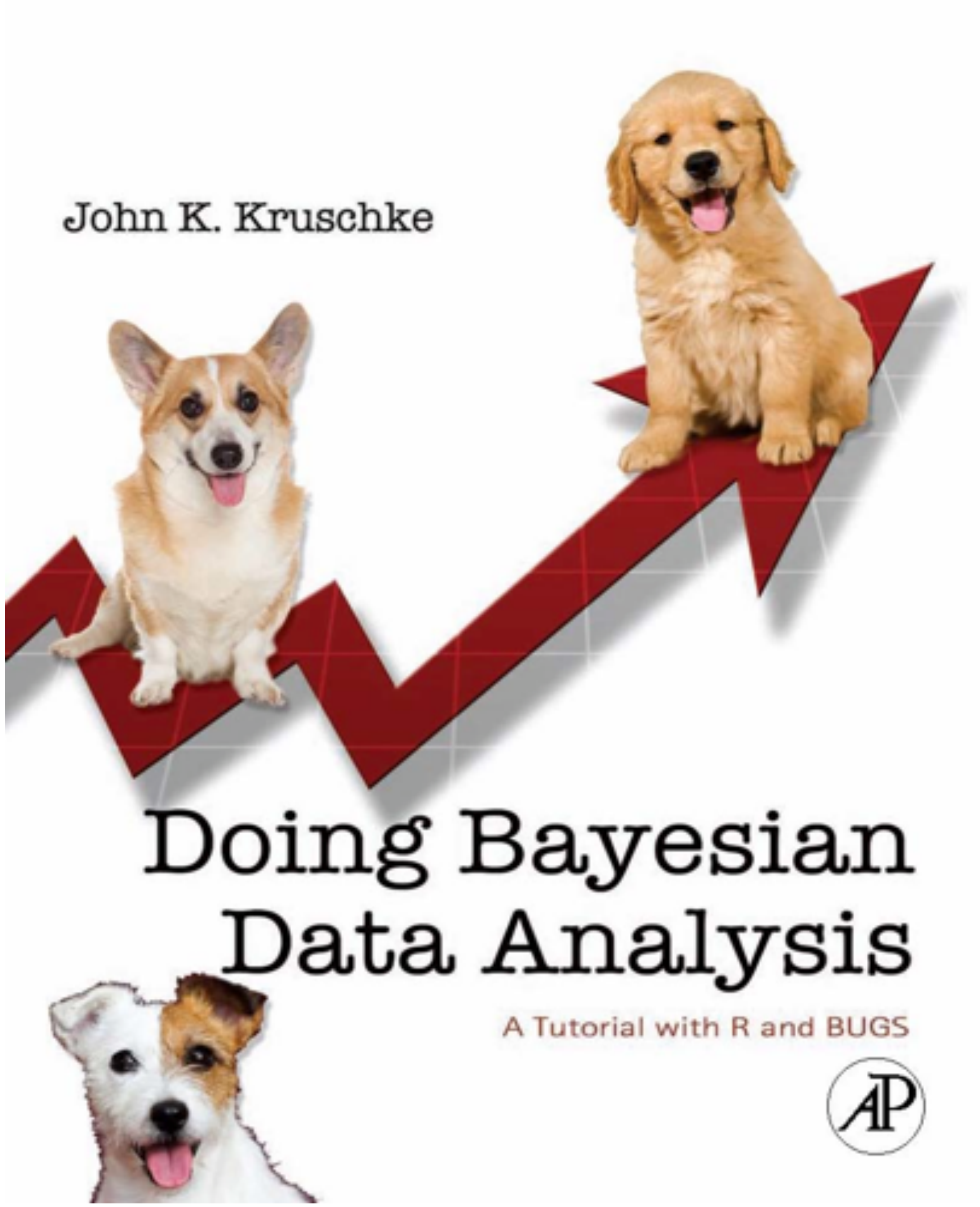
- interval estimates (and other such measures of posterior) have a clearer meaning than CIs in frequentist approaches
- frequentist orientation around "repeated sampling" is unrealistic, we in fact only sample (do our experiment) once
- frequentist involves testing only one hypothesis (model) : the null hypothesis ... Bayesian estimates probability of all models (parameter values)
- in Bayesian approach we get full posterior distribution, a much richer picture than just a mean +/- CI or s.e.
- Bayesian approach allows for incorporating previous findings in a principled way

Next Class

- grid approximation approach (discretizing the prior)
- multidimensional models
- Markov Chain Monte Carlo (MCMC)

gentle books





the full monty