Bayesian Principles

Liliana M. Dávalos
Ecology & Evolution
8 July 2013

Stony Brook University | The State University of New York

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Why use Bayesian?

- The Bayesian approach expands the class of models and easily handles:
  - repeated measures
  - unbalanced or missing data
  - Non-homogenous variances
  - multivariate data
- – and many other settings that are precluded (or much more complicated) in classical settings

From: Finley 2013

Why use Bayesian in geography?

- Simplifies understanding of the processes by factoring complex relationships into simple pieces
- Can incorporate many sources of uncertainty and stochasticity
- Can easily combine ugly data from very different sources
- Can incorporate prior information (expert range maps, previous studies)
- Geographical units are clustered and similar but different (hierarchical models)
- Deals well with large numbers of random effects (e.g. spatial or hierarchical models)

From: Merow & Silander

What makes something Bayesian?

- Bayesian methods contrast with Frequentist/Classical methods
- Bayesian methods have only become popular in the last 20ish years due to their computational demands
- The key difference between classical and Bayesian reasoning is that the Bayesian believes that knowledge is **subjective**
- Bayesian only has beliefs about the parameter value which are updated, based on data

From: Merow & Silander

Today's lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes' theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
    - Gibbs sampling
- Convergence

Probability

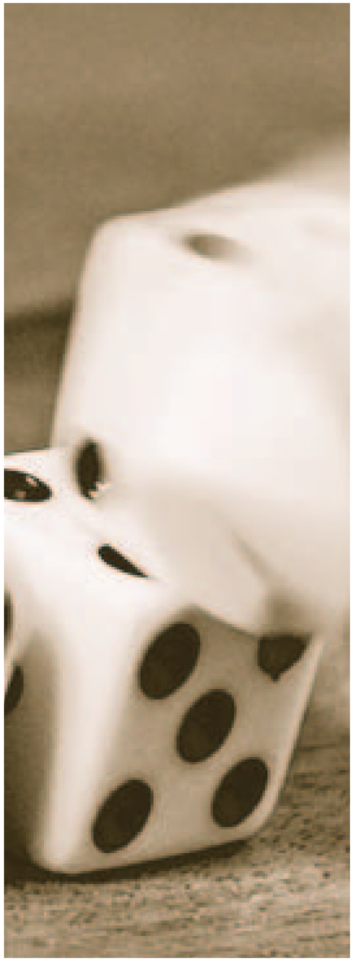
- A way of quantifying uncertainty
- Mathematical theory originally developed to model outcomes in games of chance

Probability

- A way of quantifying uncertainty
- Mathematical theory originally developed to model outcomes in games of chance
  - Many tries
  - Equal probability of different outcomes (e.g., dice)
  - Where does uncertainty come from?

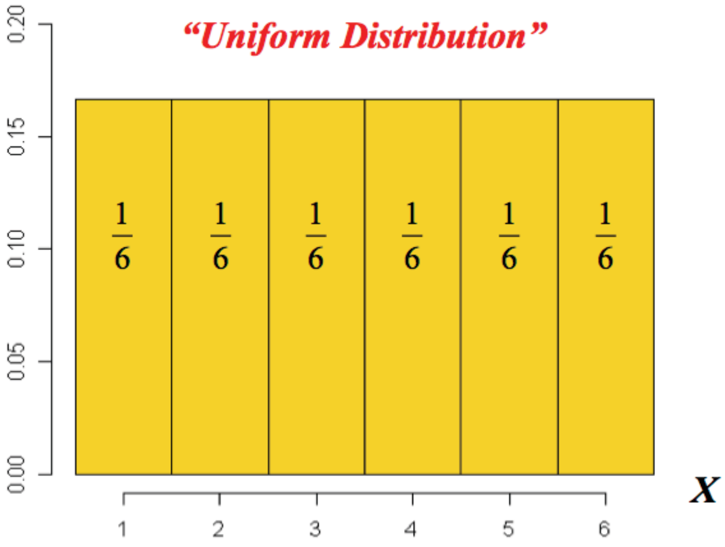
Frequentist definition probability (classical)

- The *probability* of an event = proportion of times that the event would occur if we repeated a random trial over and over again under the same conditions
  - E.g., throw one die
  - E.g., draw a card from a full deck
- A *probability distribution* is a list of all mutually exclusive outcomes of a random trial and their probabilities of occurrence

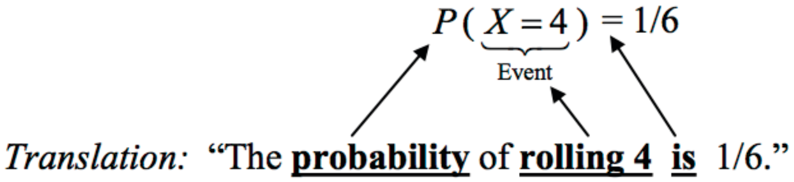


Example: roll 1 die

Event	Probability
x	$f(x) = P(X=x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
	1



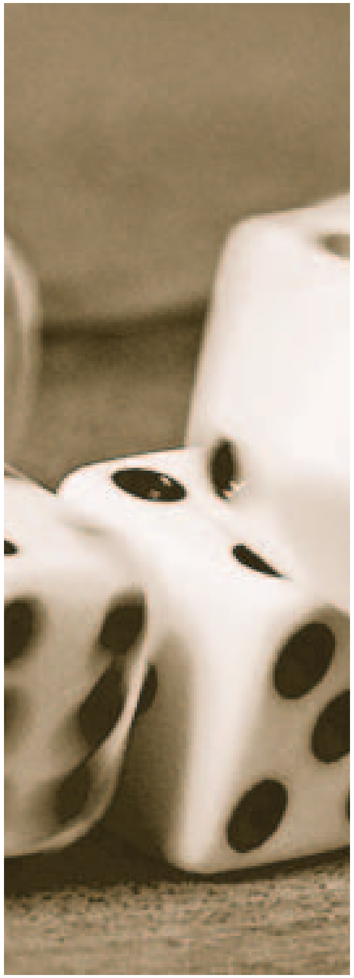
Comment on notation:



$$P(\underbrace{X=4}_{\text{Event}}) = 1/6$$

Translation: "The **probability** of **rolling 4** **is** 1/6."

From: Fischer 2012

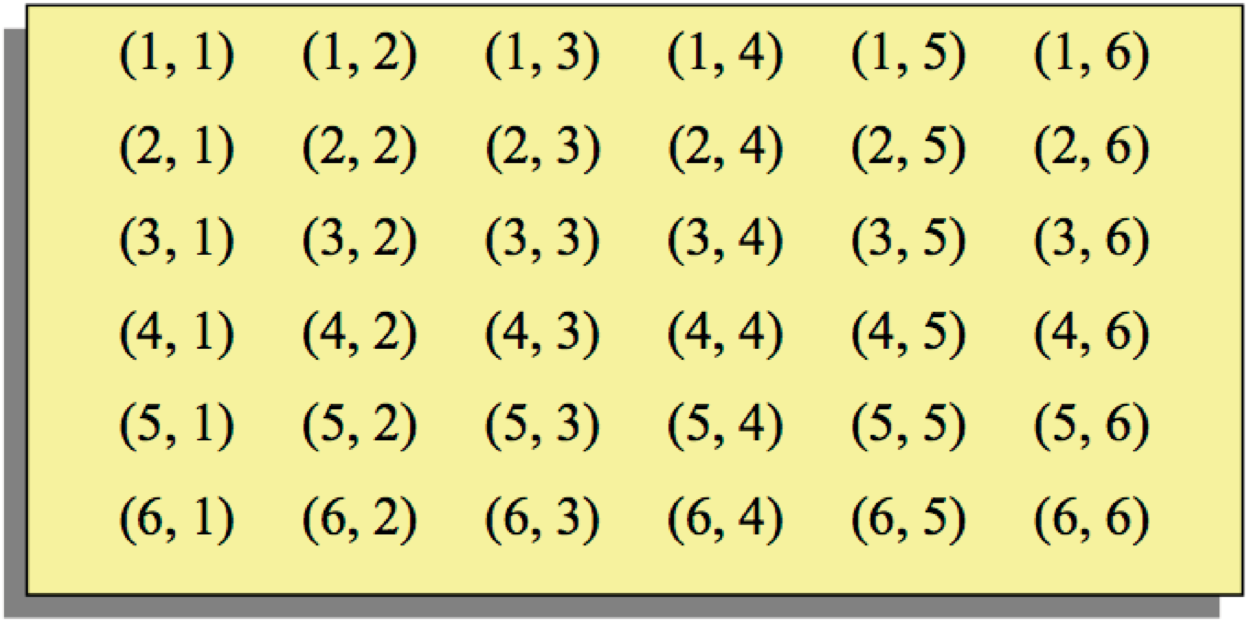


Example: roll 2 dice

- Work with the person next to you and answer the following questions:
  - How many different mutually exclusive results are there for 2 fair dice?
  - Make a table listing the events and their probability (number of times in the list of all possible outcomes that combination is expected to come up)
  - Make an approximate probability histogram
- You have 6 min

How many mutually exclusive results?

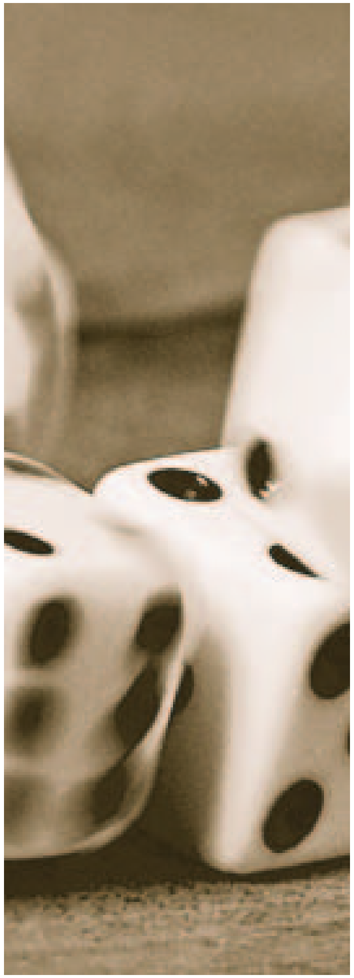
Sample Space: $\boldsymbol{S} = \{(1, 1), \ldots, (6, 6)\}$ $\#(\boldsymbol{S}) = 6^2 = 36$

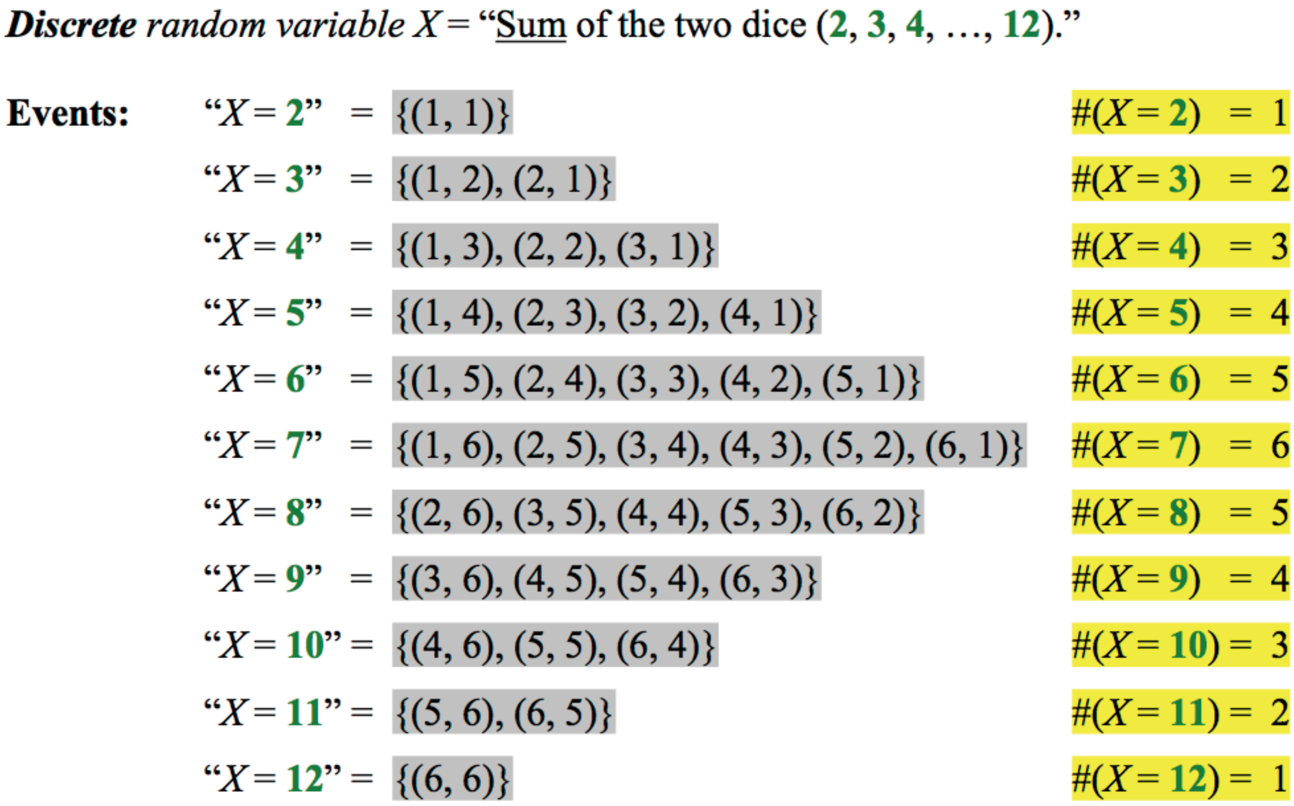


(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

From: Fischer 2012

All possible mutually exclusive events





Discrete *random variable* X = “Sum of the two dice (2, 3, 4, …, 12).”

Events:	“$X = 2$” =	$\{(1, 1)\}$	$\#(X = 2) = 1$
	“$X = 3$” =	$\{(1, 2), (2, 1)\}$	$\#(X = 3) = 2$
	“$X = 4$” =	$\{(1, 3), (2, 2), (3, 1)\}$	$\#(X = 4) = 3$
	“$X = 5$” =	$\{(1, 4), (2, 3), (3, 2), (4, 1)\}$	$\#(X = 5) = 4$
	“$X = 6$” =	$\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$	$\#(X = 6) = 5$
	“$X = 7$” =	$\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$	$\#(X = 7) = 6$
	“$X = 8$” =	$\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$	$\#(X = 8) = 5$
	“$X = 9$” =	$\{(3, 6), (4, 5), (5, 4), (6, 3)\}$	$\#(X = 9) = 4$
	“$X = 10$” =	$\{(4, 6), (5, 5), (6, 4)\}$	$\#(X = 10) = 3$
	“$X = 11$” =	$\{(5, 6), (6, 5)\}$	$\#(X = 11) = 2$
	“$X = 12$” =	$\{(6, 6)\}$	$\#(X = 12) = 1$

From: Fischer 2012

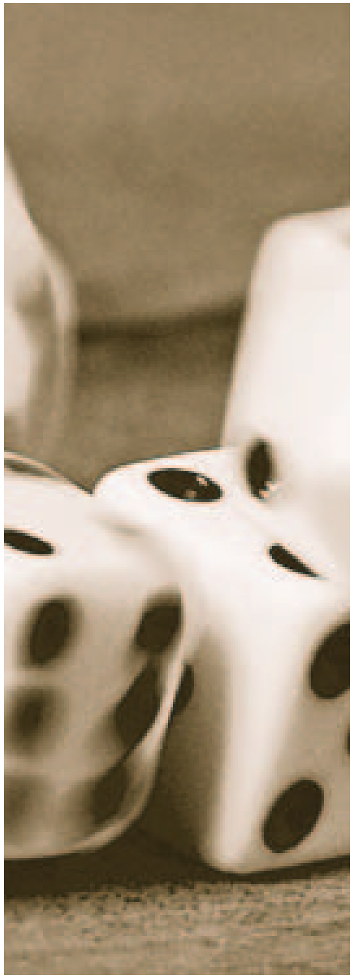
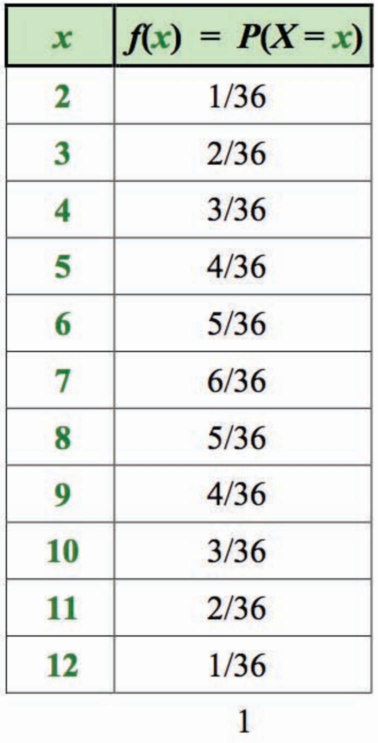
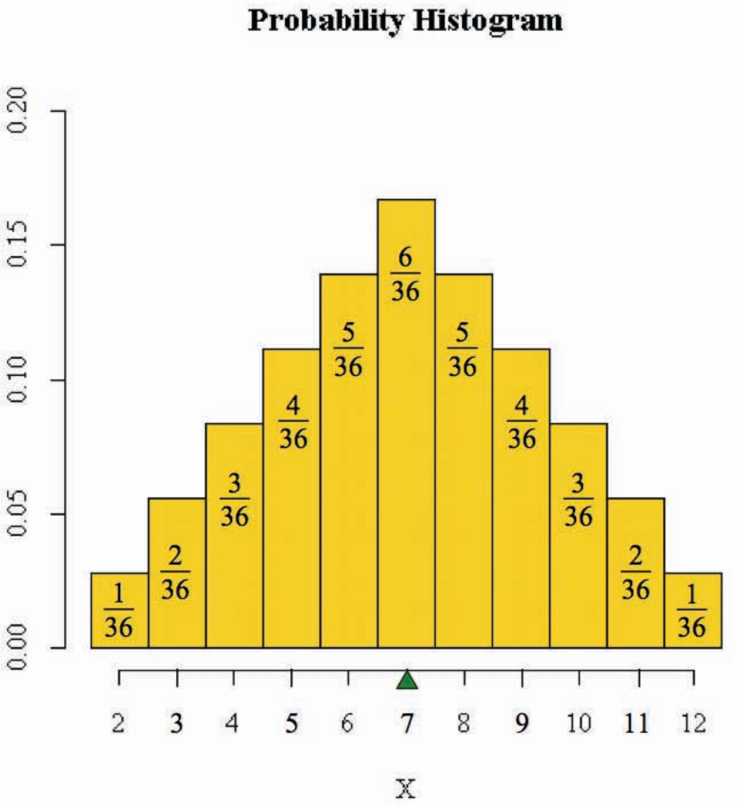


Table and histogram



x	$f(x) = P(X = x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36
	1



From: Fischer 2012

Notice:

- List covers entire universe of outcomes
- Many random trials
- Trials are independent: probability of one result in one trial does not affect another result in another trial

Some examples that make sense in this framework

- If we toss a fair coin, what is the probability of 10 heads in a row
- If we assign treatments randomly to subjects, what is the probability that a sample mean difference between treatments will be greater than 10%?
- Under a process of genetic drift in a finite population, what is the probability of fixation of a rare allele?
- What is the probability of a result at least as extreme as that observed if the null hypothesis is true?

Notice

- If we toss a fair coin, what is the probability of 10 heads in a row
- If we assign treatments randomly to subjects, what is the probability that a sample mean difference between treatments will be greater than 10%?
- Etc.

- Sampling error is the source of uncertainty

Some examples that do NOT make sense

- What is the probability that Iran is building nuclear weapons?
- What is the probability that Guaviare will increase its coca cultivation next year?
- What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?
- What is the probability that polar bears will be extinct in the wild in 40 years?

Why don’t they make sense?

  - What is the probability that Iran is building nuclear weapons?
  - What is the probability that Guaviare will increase its coca cultivation next year?
  - What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?
  - What is the probability that polar bears will be extinct in the wild in 40 years?

- Discuss with your neighbor and explain why these don’t make sense

- You have 5 min

Why they don’t make sense

  - What is the probability that Iran is building nuclear weapons?

- [either it is or it isn’t – no random trial]
  - What is the probability that Guaviare will increase its coca cultivation next year?

- [either it will increase or will not – no random trial]
  - What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?

- [either there is one or there are two – no random trial]
  - What is the probability that polar bears will be extinct in the wild in 40 years?

- [difficult to cast this as a frequency of occurrence]

Why they don’t make sense

  - What is the probability that Iran is building nuclear weapons?
  - What is the probability that Guaviare will increase its coca cultivation next year?
  - What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?
  - What is the probability that polar bears will be extinct in the wild in 40 years?

- In these examples there is no random trial, so no sampling error. Information is the source of uncertainty

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Bayesian definition of probability

- *Probability* is a measure of a degree of belief associated with the occurrence of an event
  - A measure of *uncertainty associated with the occurrence of a particular event*, given the *available information* and the accepted *assumptions*
- A *probability distribution* is a list of all mutually exclusive events and the degree of belief associated with their occurrence
- Bayesian statistics applies the mathematics of probability to subjective degree of belief

Bayesian methods are increasingly used in ecology and evolution

- Is this good?

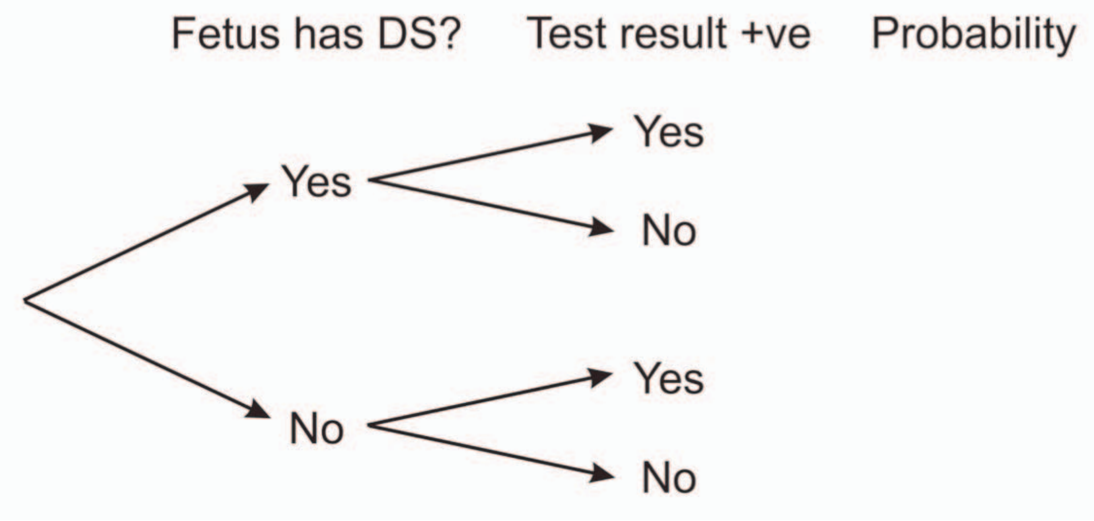
- "*Ecologists should be aware that Bayesian methods constitute a radically different way of doing science. Bayesian statistics is not just another tool to be added into the ecologists' repertoire of statistical methods. Instead, Bayesians categorically reject various tenets of statistics and the scientific method that are currently widely accepted in ecology and other sciences.*" B. Dennis, 1996, Ecology

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

How Bayes' theorem works

- Example: detection of Down syndrome (DS)
- DS occurs in about 1 in 1000 pregnancies. The most accurate test requires amniocentesis, which carries a small risk of miscarriage
- The triple test is used first
- It has not risk, but is less accurate



From: Schluter

Notice

- Tests are NOT the event. We have a DS test, separate from the event of actually having DS. We have a HIV test, separate from the event of actually having HIV

- Tests are flawed. Tests detect things that don't exist (*false positive*), and miss things that do exist (*false negative*)

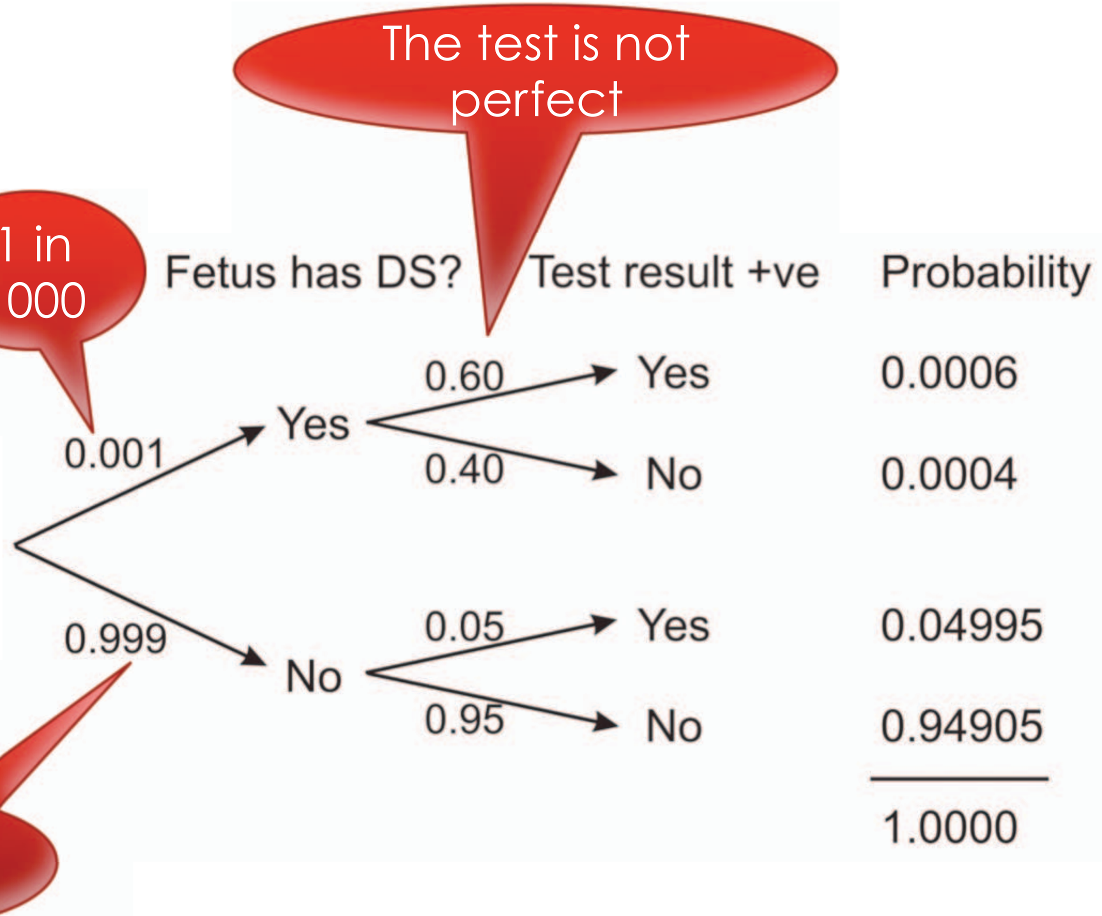
- Tests give us test probabilities, not the real probabilities. People often consider the test results directly, without considering the errors in the tests

- False positives skew results. Suppose you are searching for something really rare (1 in 1 million). Even with a good test, it's likely that a positive result is really a false positive on somebody in the 999,999

From: Azad @ http://betterexplained.com

Conditional probability

- The *conditional probability* of an event is the probability of that event occurring given that a condition is met
- E.g.: the probability of a +ve test result from the triple test, given that a fetus has DS, is 0.6



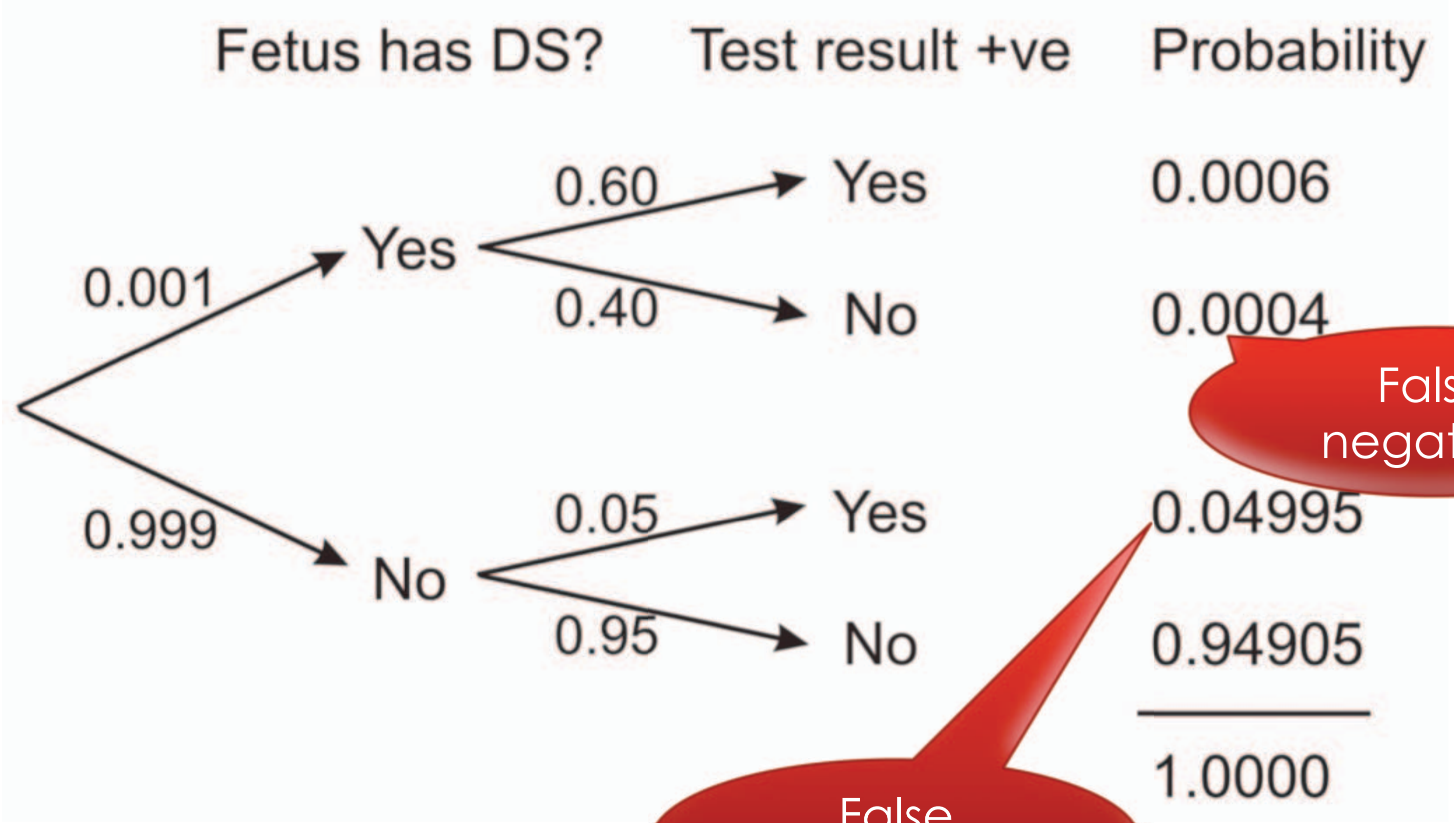
The test is not perfect

1 in 1000

999 in 1000

From: Schluter

Conditional probability



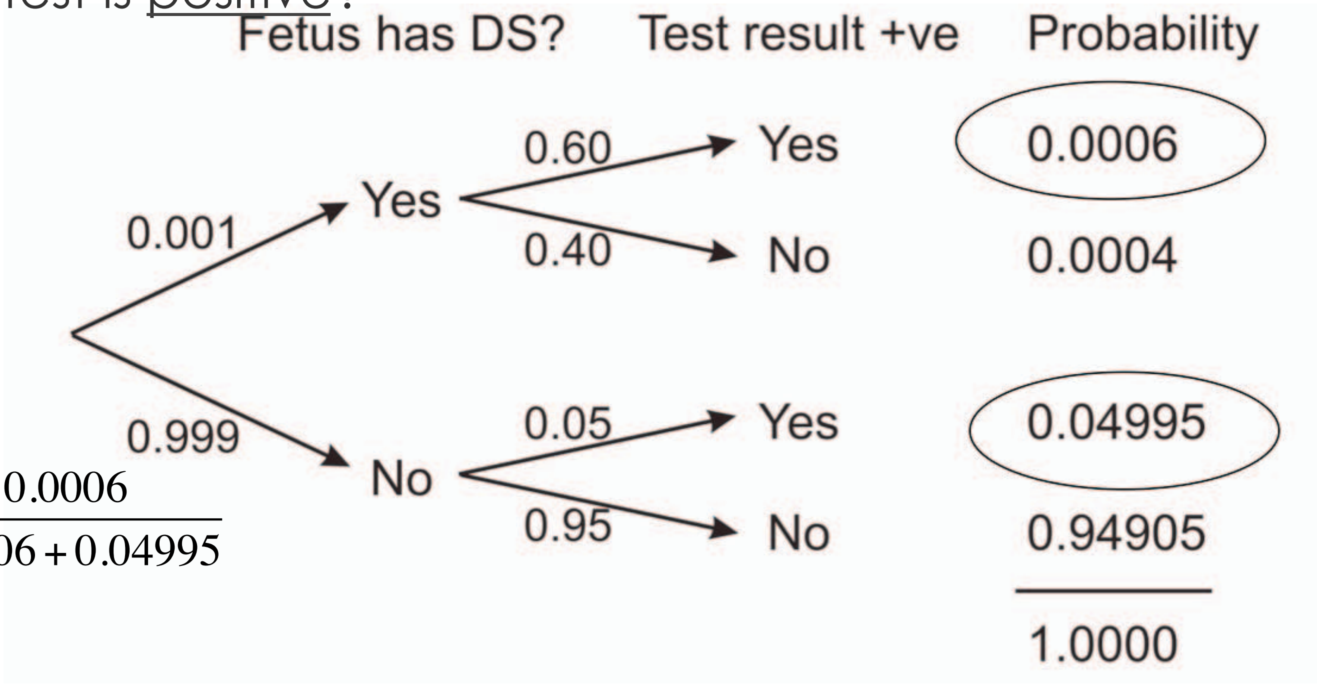
False negatives

False positives

From: Schluter

Conditional probability calculation

- What is the probability that a fetus has DS if the test is positive?

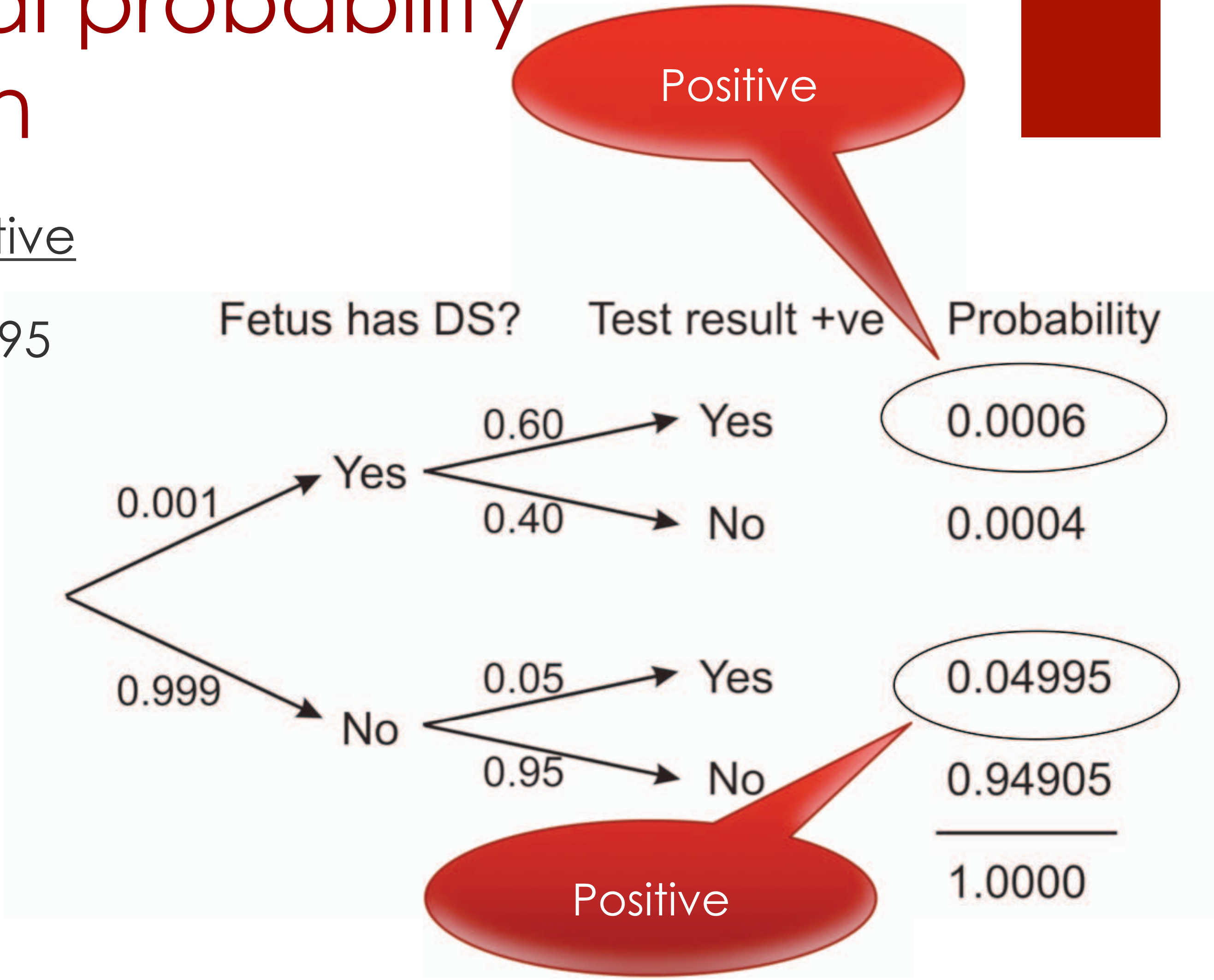


- $\Pr[DS \mid positive] = \dfrac{0.0006}{0.0006 + 0.04995}$
- = 0.012
- = 1.2%

From: Schluter

Conditional probability calculation

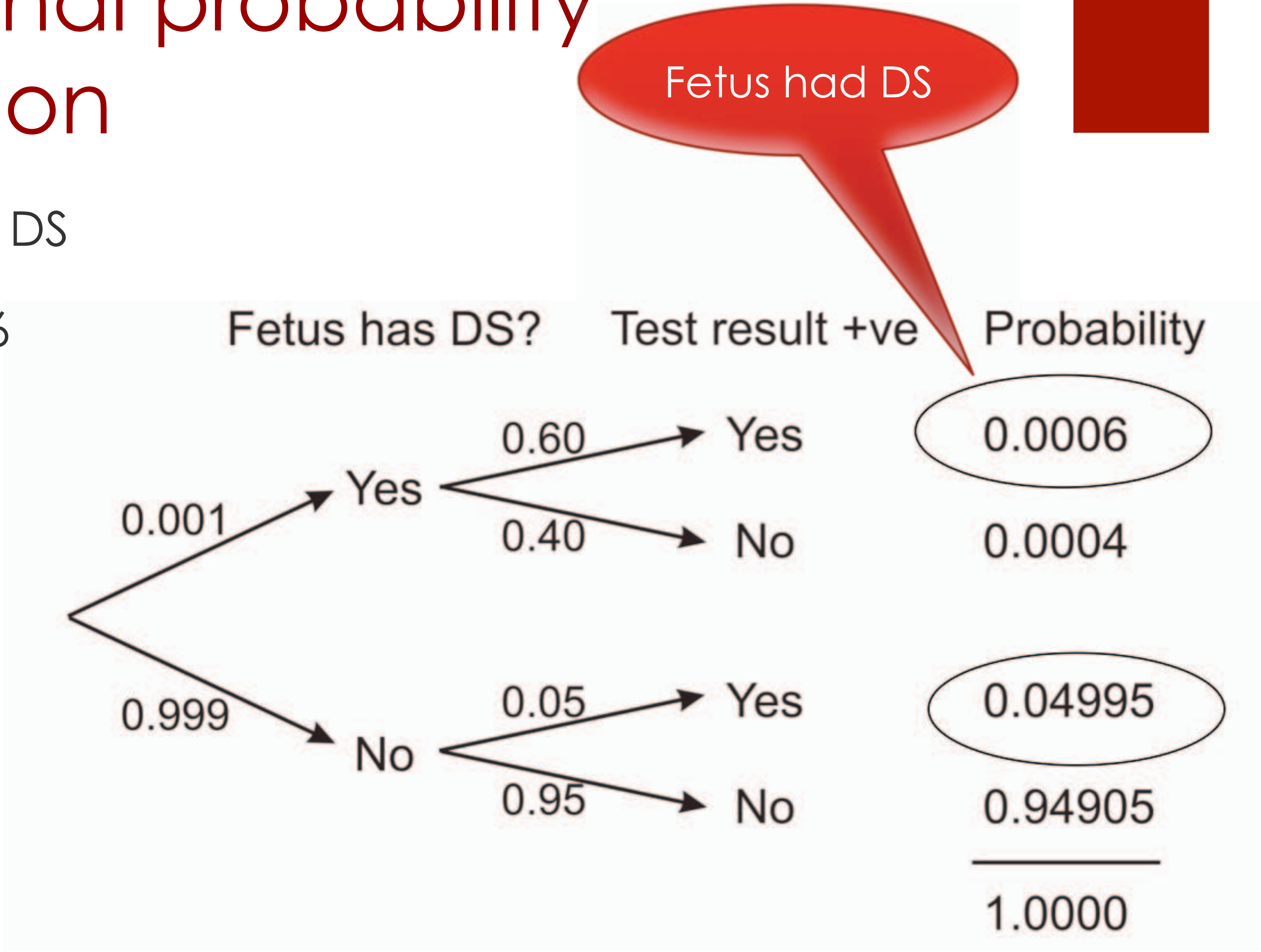
- The test was positive
- => 0.0006 + 0.04995
- = 0.05055



From: Schluter

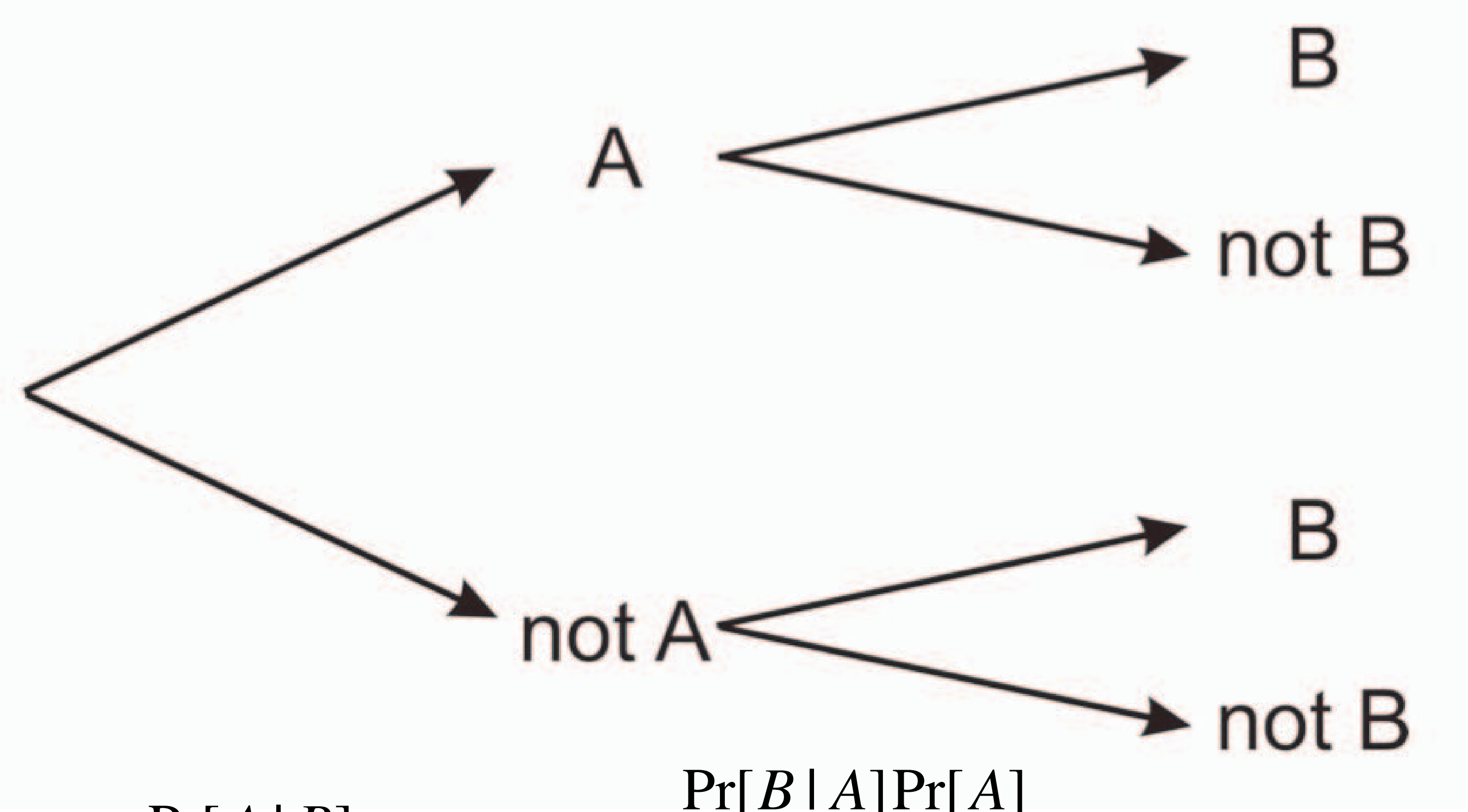
Conditional probability calculation

- The fetus had DS
- => only 0.0006



From: Schluter

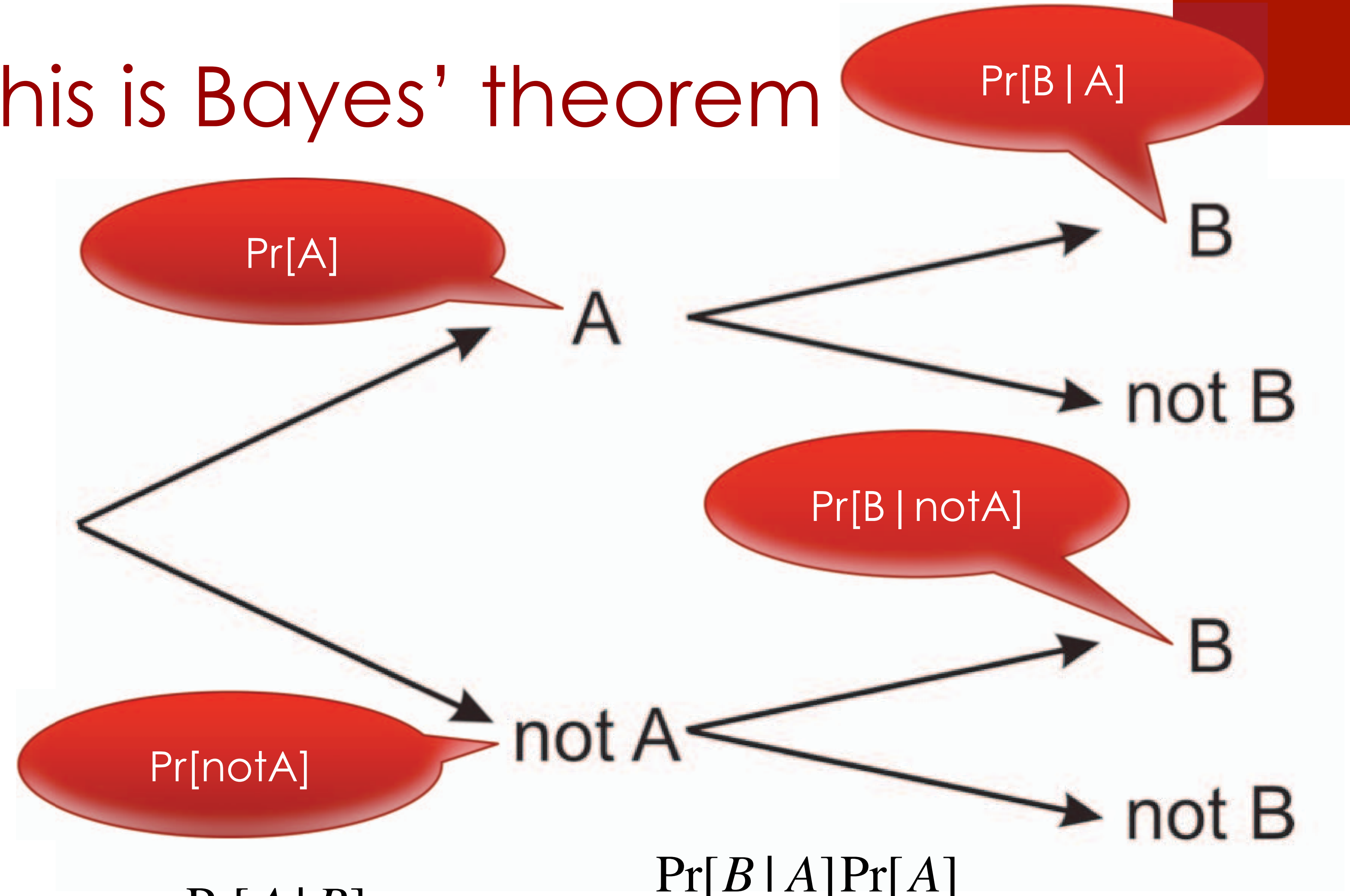
This is Bayes' theorem



$$\Pr[A \mid B] = \frac{\Pr[B \mid A]\Pr[A]}{\Pr[B \mid A]\Pr[A] + \Pr[B \mid notA]\Pr[notA]}$$

From: Schluter

This is Bayes' theorem



$$\Pr[A \mid B] = \frac{\Pr[B \mid A]\Pr[A]}{\Pr[B \mid A]\Pr[A] + \Pr[B \mid notA]\Pr[notA]}$$

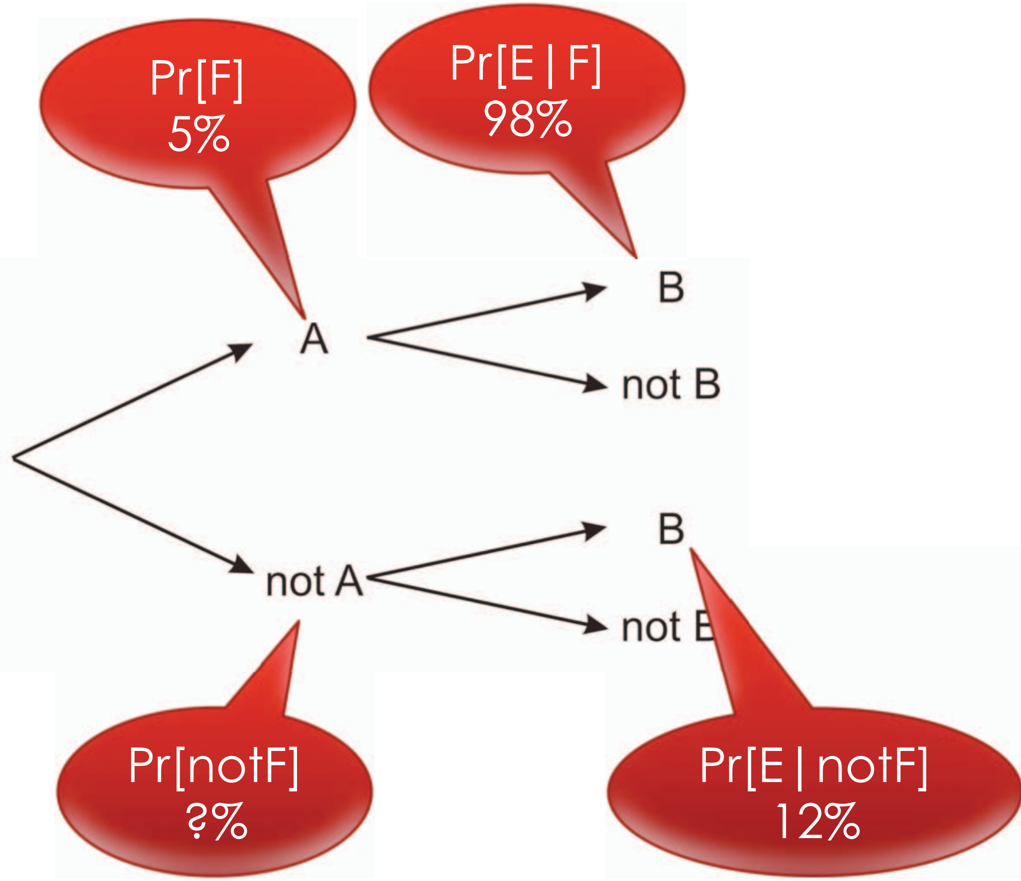
Let’s solve this problem

- When a test for steroids is given to football players, 98% of the players taking steroids test positive and 12% of the players not taking steroids test positive
- Suppose that 5% of football players take steroids
- What is the probability that a soccer player who tests positive takes steroids?

From: Rincon 2012

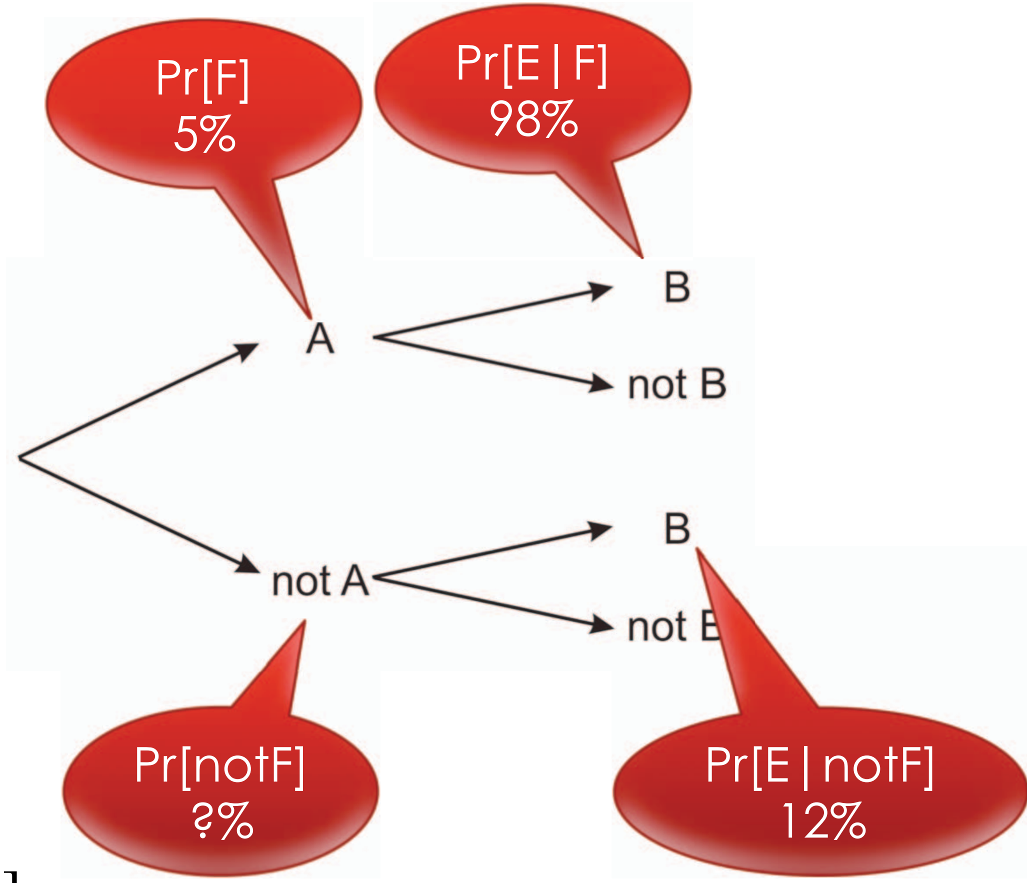
What do we know?

- Let *E* be the event that a soccer player selected at random tests positive, and *F* the event that a player selected at random takes steroids
- 5% of football players take steroids
- 98% of the players taking steroids test positive
- 12% of the players not taking steroids test positive



Let's solve this problem

- What is the probability that a soccer player who tests positive takes steroids?



Pr[F]
5%

Pr[E | F]
98%

Pr[notF]
?%

Pr[E | notF]
12%

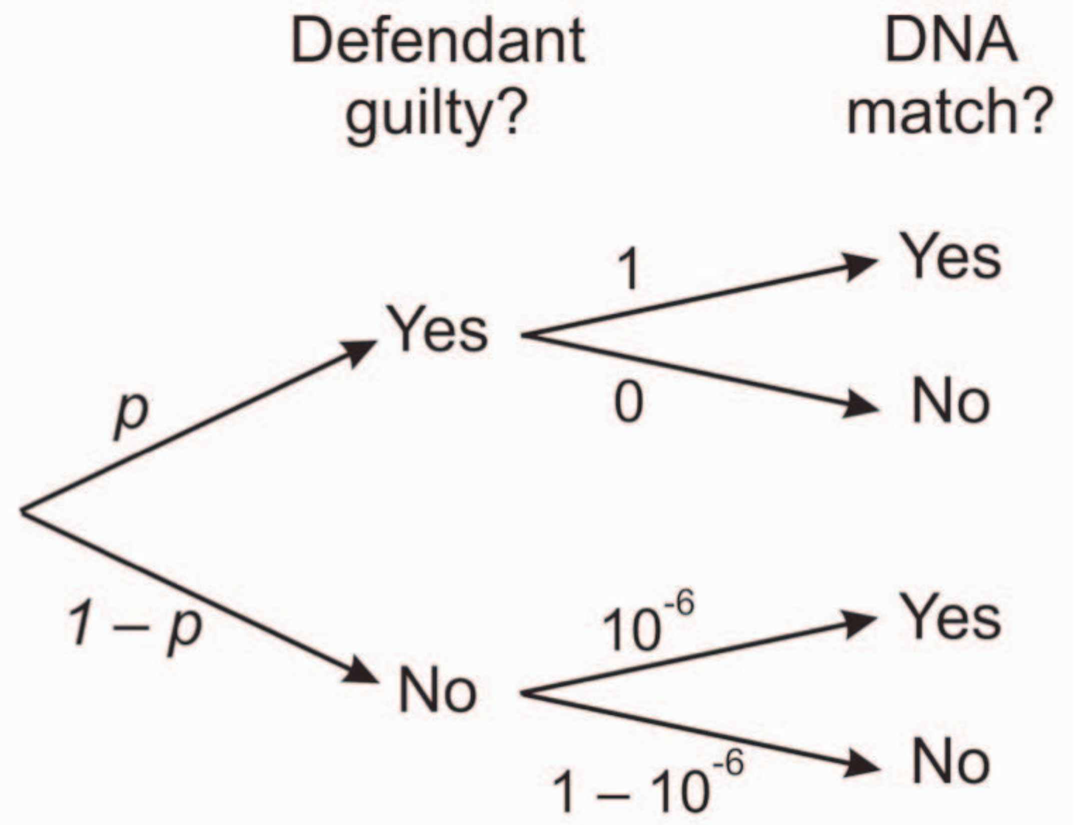
$$\Pr[F \mid E] = \frac{\Pr[E \mid F]\Pr[F]}{\Pr[E \mid F]\Pr[F] + \Pr[E \mid notF]\Pr[notF]}$$

From: Rincon 2012

Why Bayesian is controversial

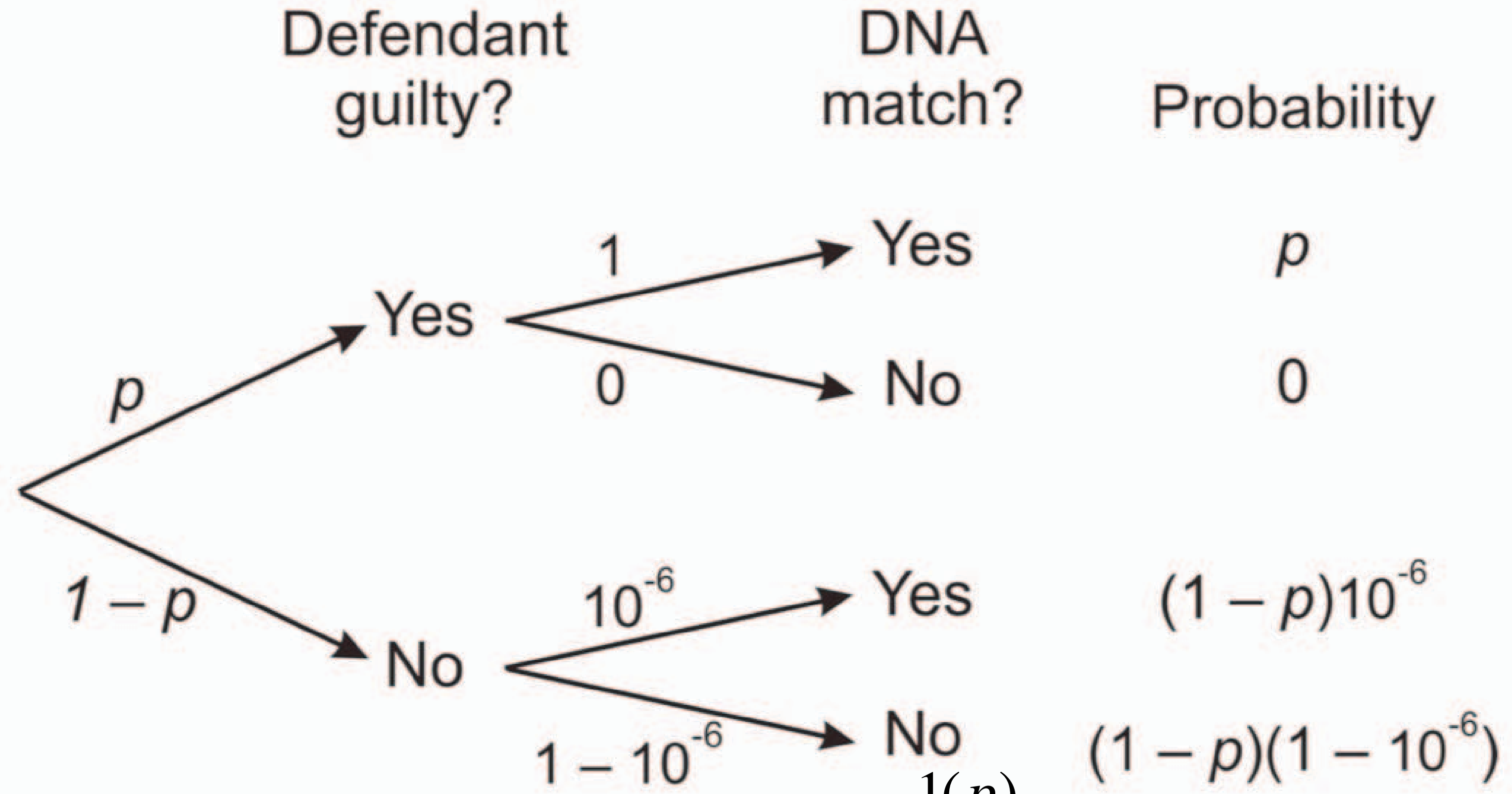
- For example: forensic evidence. Bayesian inference can be used in a court to quantify the evidence for and against the guilt of the defendant based on a match with DNA evidence left at the scene of the crime

- What is the probability of guilt given a positive DNA match (assuming no contamination of samples)?



Using Bayes' theorem

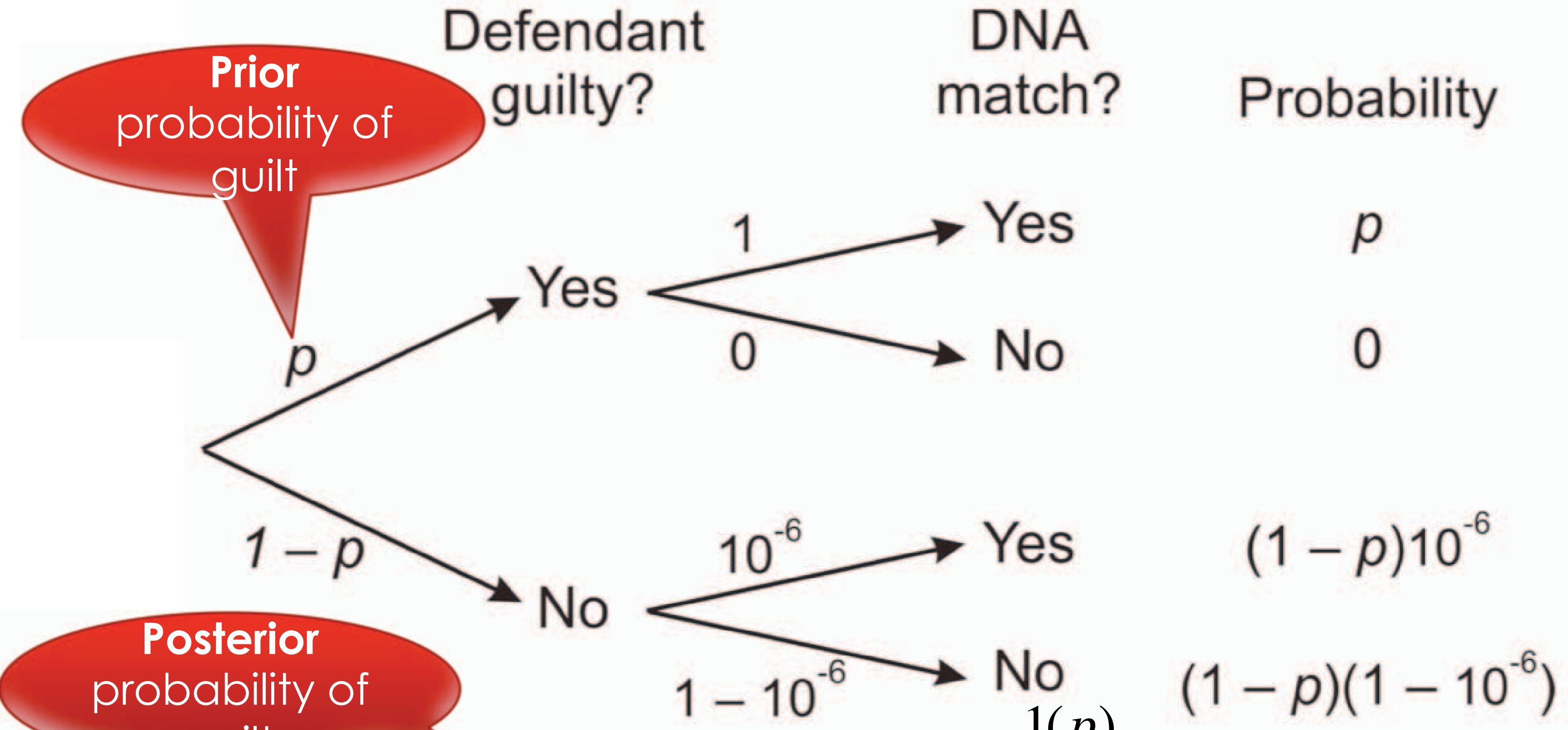
What is the probability of guilt given a positive DNA match?



$$\Pr[guilt \mid match] = \frac{1(p)}{1(p) + 10^{6}(1 - p)}$$

From: Schluter

Prior and **posterior** probability



Prior probability of guilt

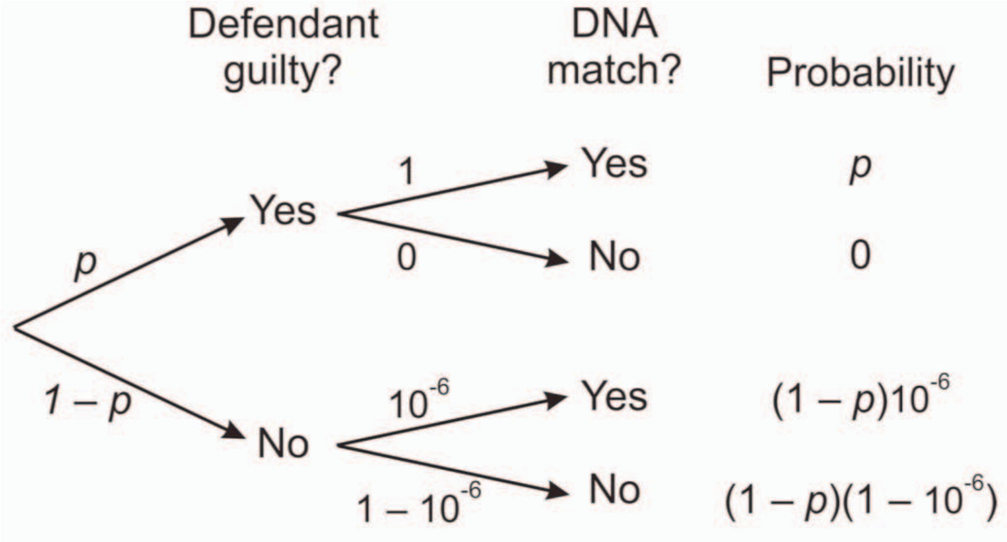
Posterior probability of guilt

$$\Pr[guilt \mid match] = \frac{1(p)}{1(p) + 10^6(1-p)}$$

From: Schluter

Bayesian inference at work

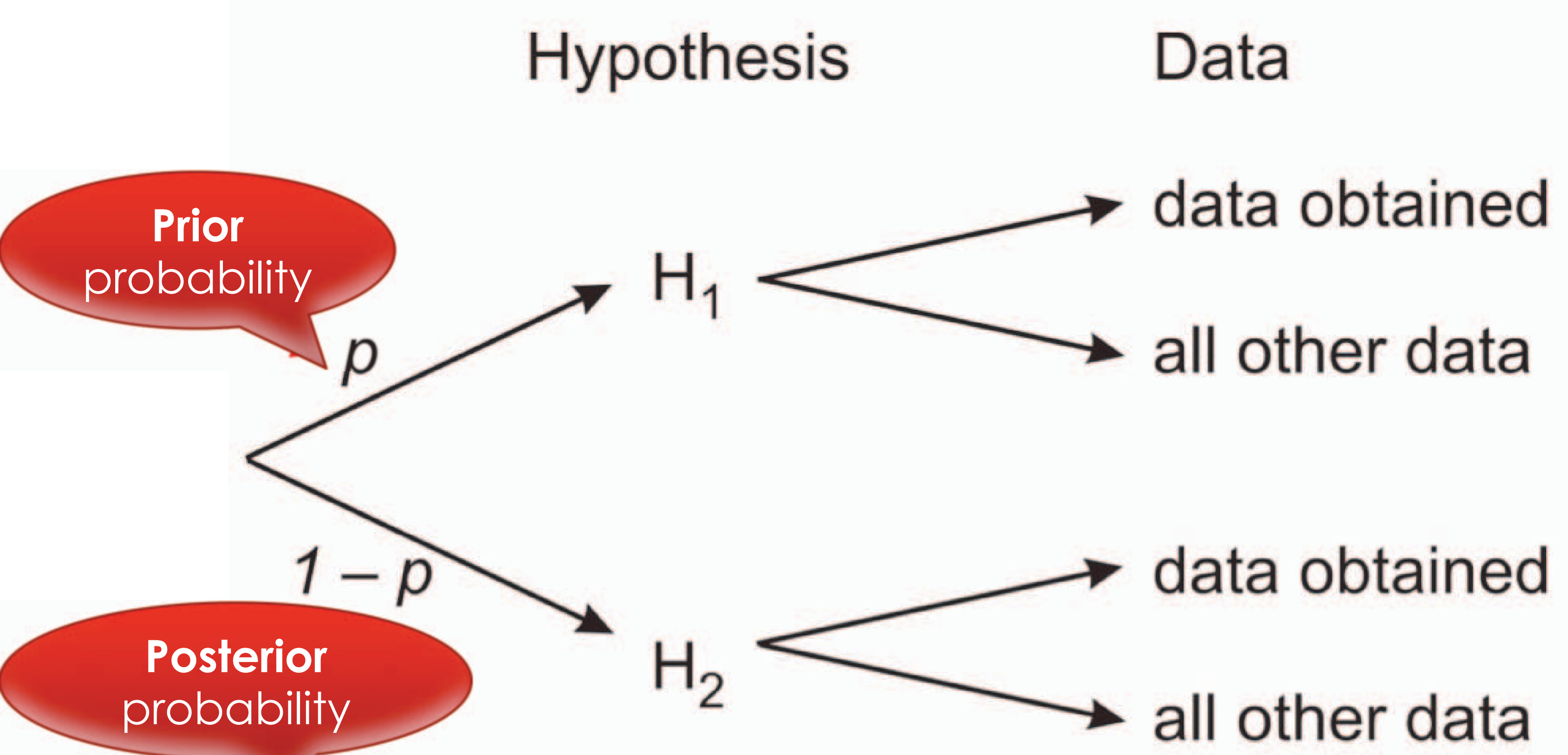
- Influence of the prior
- If p = 10$^{□6}$ then Pr[guilt | match] = 0.5
- If p = 0.5 then Pr[guilt | match] = 0.999999
- So, is the defendant guilty or innocent?



$$\Pr[guilt \mid match] = \frac{1(p)}{1(p) + 10^{6}(1 - p)}$$

From: Schluter

Bayesian inference with data



Hypothesis

Data

p

$1 - p$

H_1

H_2

data obtained

all other data

data obtained

all other data

Prior probability

Posterior probability

$$\Pr[H_1 \mid data] = \frac{\Pr[data \mid H_1]\Pr[H_1]}{\Pr[data \mid H_1]\Pr[H_1] + \Pr[data \mid H_2]\Pr[H_2]}$$

From: Schluter

Some terminology

Posterior probability H_1

Likelihood of data conditional on H_1

Prior probability H_1

$$\Pr[H_1 \mid data] = \frac{\Pr[data \mid H_1]\Pr[H_1]}{\Pr[data \mid H_1]\Pr[H_1] + \Pr[data \mid H_2]\Pr[H_2]}$$

How Bayesian inference is different in statistics

- Mathematically, the **hypothesis** or **parameter** is treated as though it is random, not fixed
  - Random as in "has a **probability density function** (**PDF**)"
- Classical statistics: model parameters are ***fixed*** and unknown
- Bayesian statistics: parameters are **random**
  - Parameters have **distributions** (just like the data)

How Bayesian inference is different in general

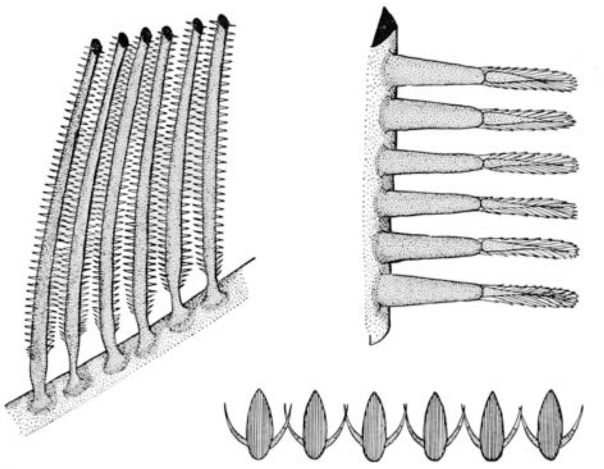
- The **prior** probability represents the investigator’s strength of belief about the value of the **fixed parameter** or **hypothesis**
- The **posterior** probability expresses how the investigator’s beliefs have been altered by the **data**
- The only data considered are the **data observed**, not the data that “might have been”; other possible outcomes of the experiment are not considered
- The key to Bayesian inference is “**learning**” or “**updating**” of prior beliefs. Thus, **posterior information ≥ prior information**

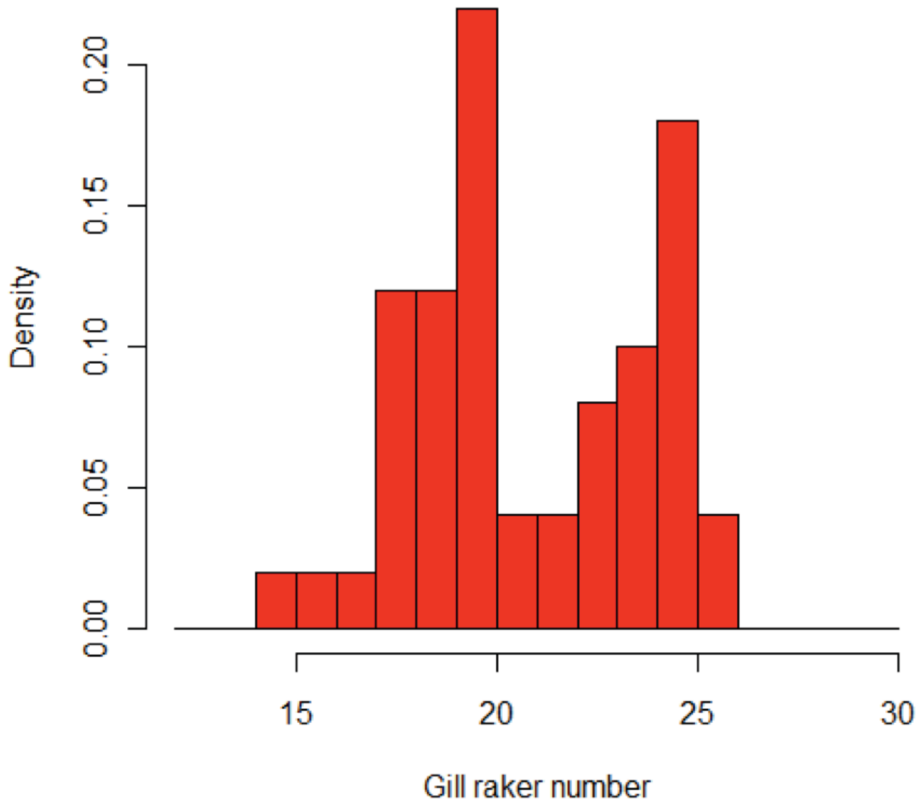
Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Example Bayesian hypothesis

- Data: Gill raker number of 50 fish collected from a new lake
- What is the probability that they represent 2 species rather than 1?

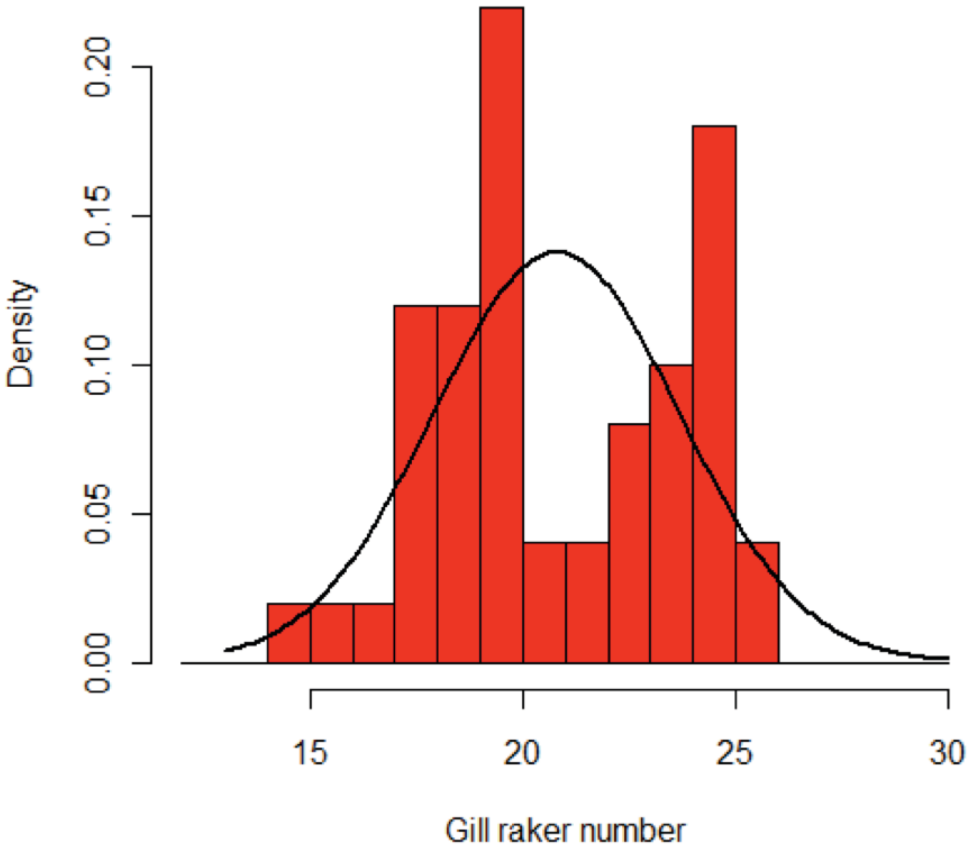




From: Schluter

H_1: one species

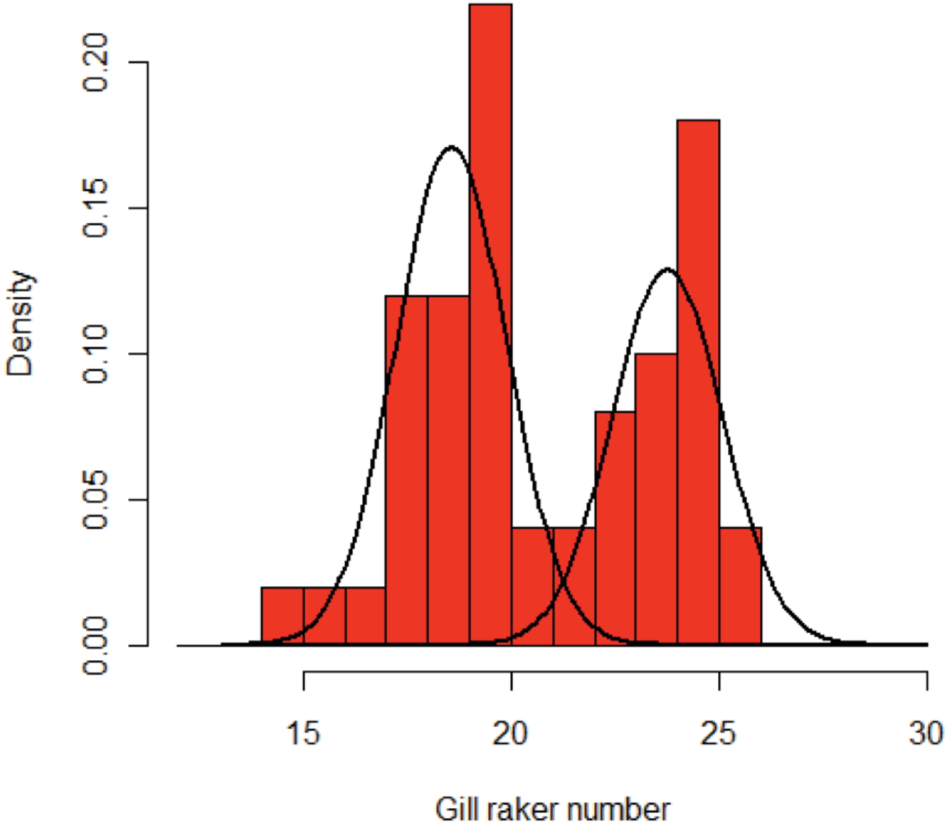
- Assume a normal distribution of measurements
- Pr[data | H_1] = log *L*[H_1 | data] = –124.06



From: Schluter

H_2: two species

- Assume normal distributions with equal variance in both groups
- Pr[data | H_2] = log L[H_2 | data] = –116.51



From: Schluter

Posterior model probabilities

- Plug the likelihoods into Bayes Theorem to calculate the posterior probabilities of each hypothesis given the data
- Posterior probability depends on the prior probability
- Here is the probability that H_2 is correct (two species are present):

Prior probability $Pr[H_2]$	Posterior probability $Pr[H_2 \mid data]$
0.500	0.99
0.005	0.91
0.001	0.66

If prior is small, need more data to increase posterior probability

Solving a Bayesian problem

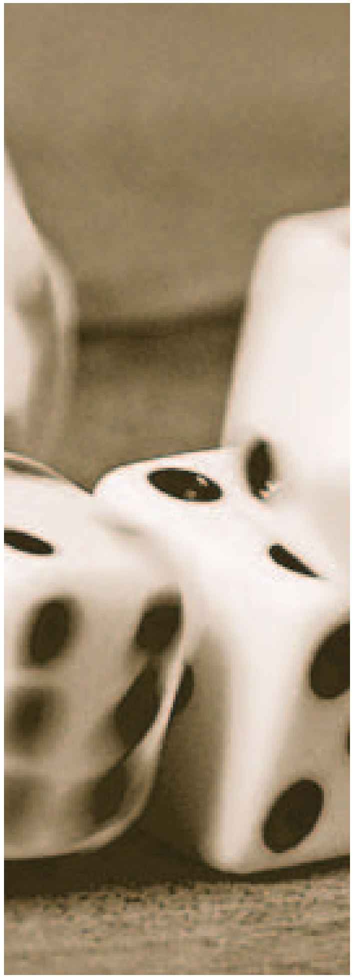
- Suppose there are two bowls of cookies
  - Bowl 1 contains 30 vanilla cookies and 10 chocolate cookies
  - Bowl 2 contains 20 of each
- Now suppose you choose one of the bowls at random and, without looking, select a cookie at random
- The cookie is vanilla
- What is the probability that it came from Bowl 1?
- This is a conditional probability; we want p(Bowl 1 | vanilla)

From: Downey 2012

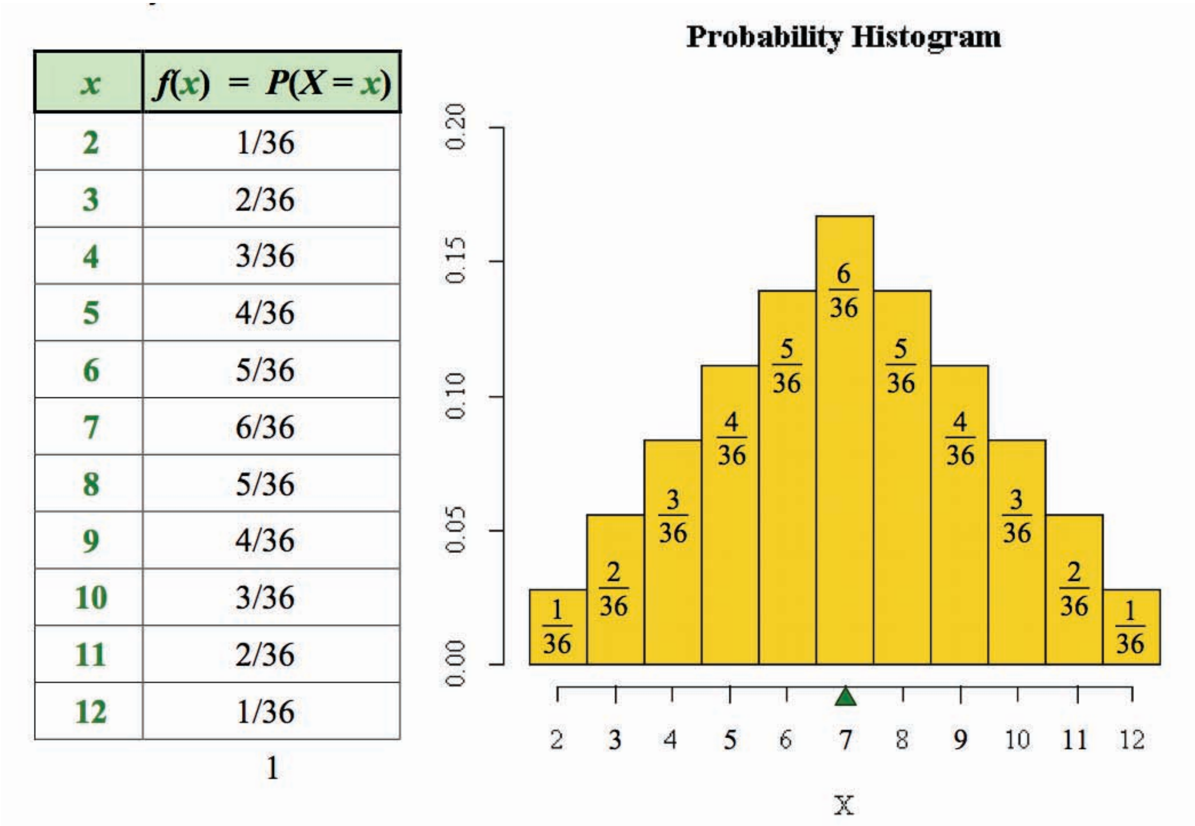
Break

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence



Remember this?



x	$f(x) = P(X = x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36
	1

From: Fischer 2012

Let’s do an experiment

- Team up with your neighbor
- Flip a coin 2 times
- 2 min
- What is the universe of results?
- How much probability does each result have?
- Which result did you get?

Some questions

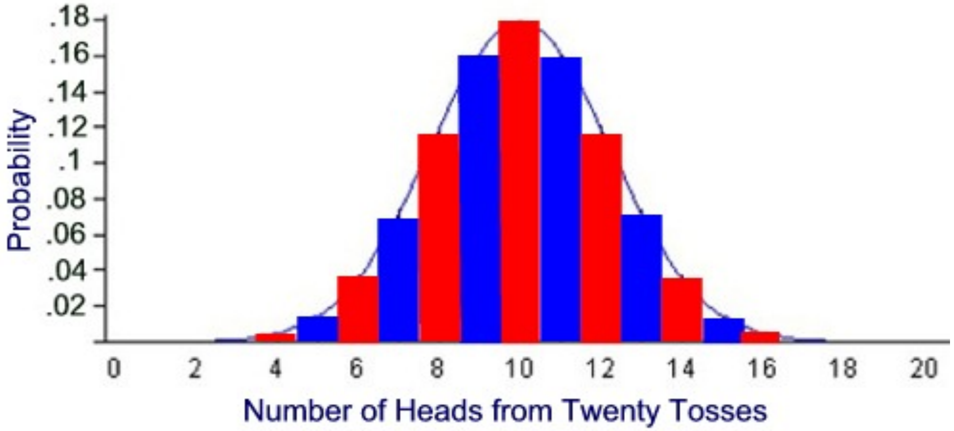
- Are there repeated trials in this experiment?
- How many possible outcomes are there for each try?
- What is the probability of getting heads each time?
- Are the trials dependent or independent?

Answers

- Are there repeated trials in this experiment?
- Yes
- How many possible outcomes are there for each try?
- Only 2: heads or tails
- What is the probability of getting heads each time?
- 0.5 and it is constant
- Are the trials dependent or independent?
- Trials are independent

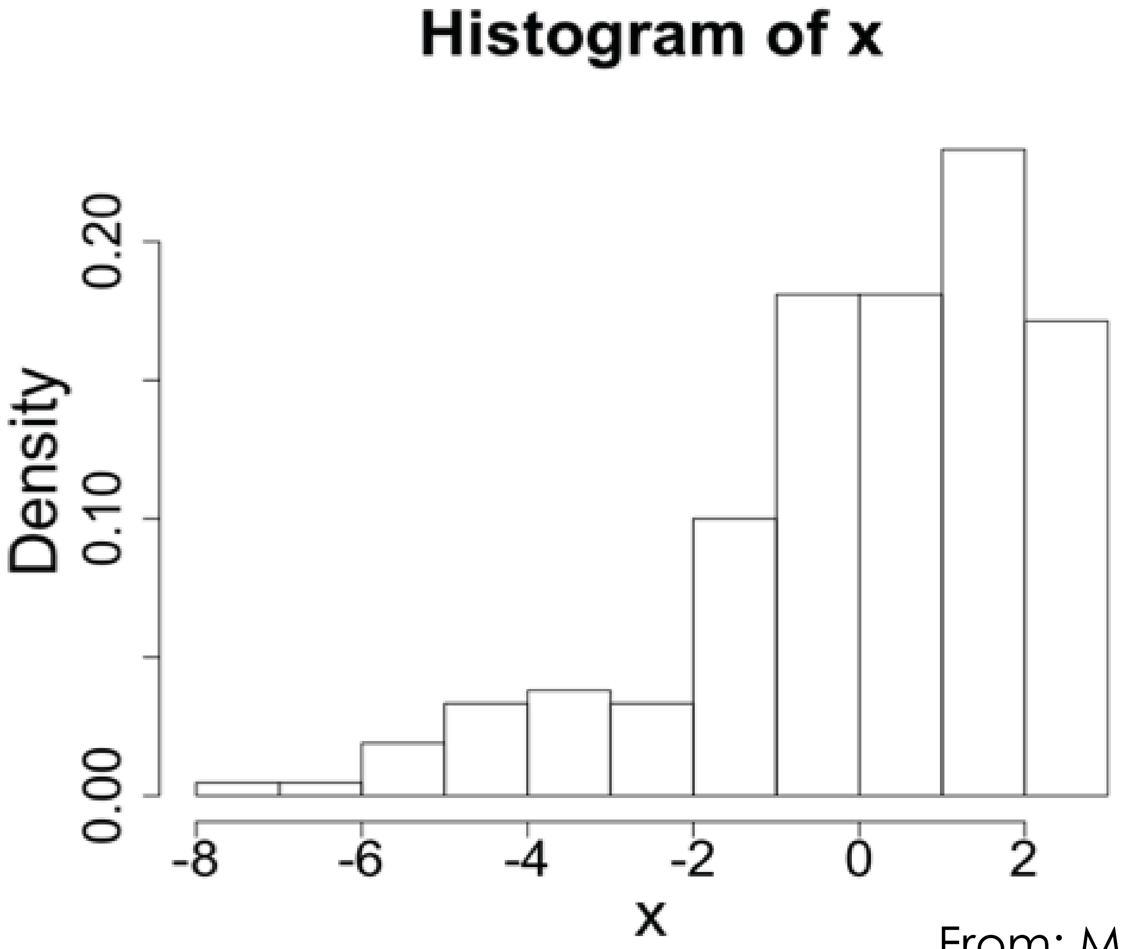
Binomial probability

- b(*x*; *n*, *P*)
- the probability that an *n*-trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is *P*
- We shall revisit later today and on Day 3



Probability density functions

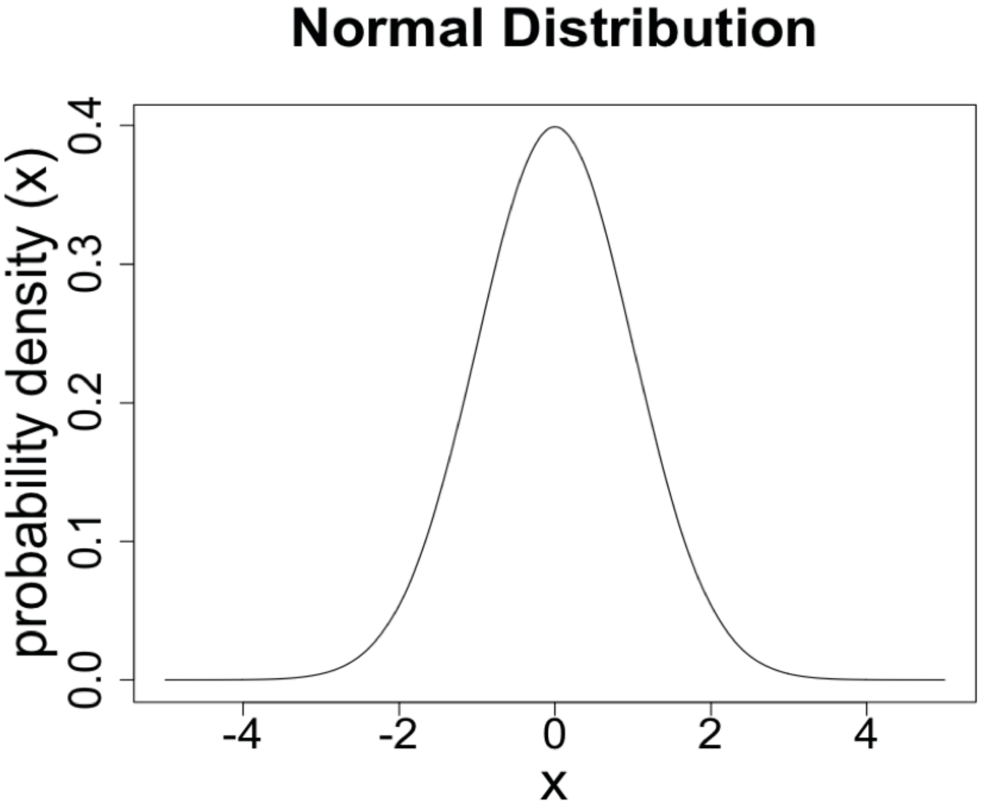
You can think of a PDF as a histogram (normalized)



From: Merow & Silander

Probability density functions (PDF)

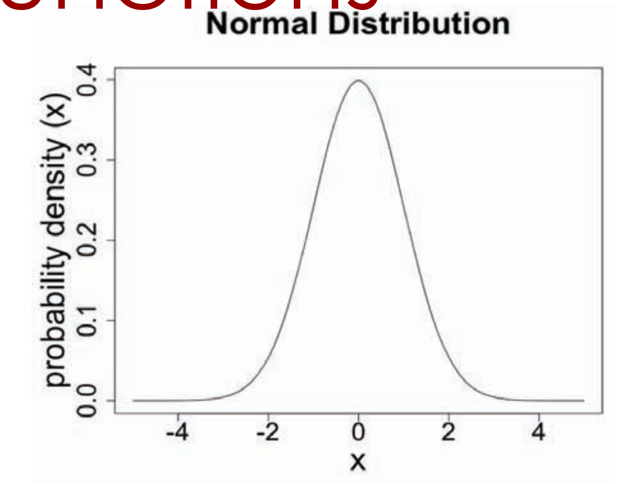
- A pdf, *f*(x), assigns a probability to x taking a value between x and x+ *d*x based on the area under the curve
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$

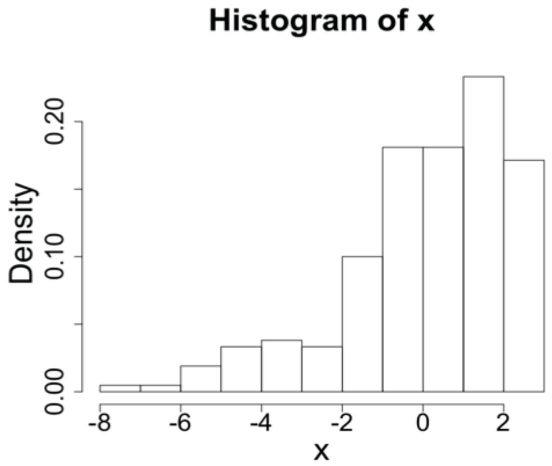


Probability density functions

- "Densities" can occur in Bayesian models because you specify a distribution, like a Normal, Beta, Poisson, etc.

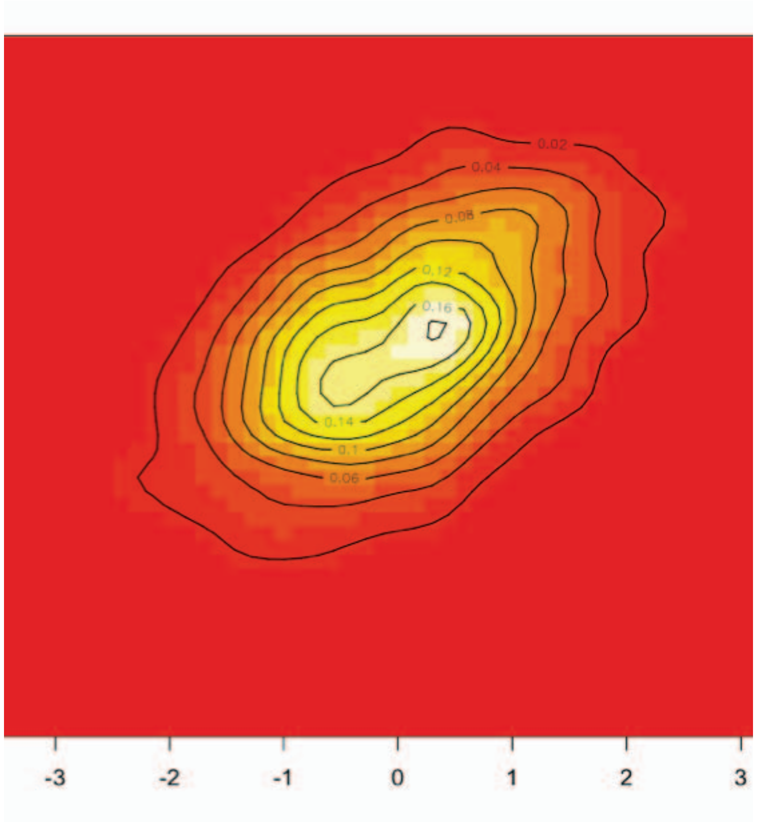
- Or because you're sampling from an unknown distribution that doesn't necessarily have a pretty form

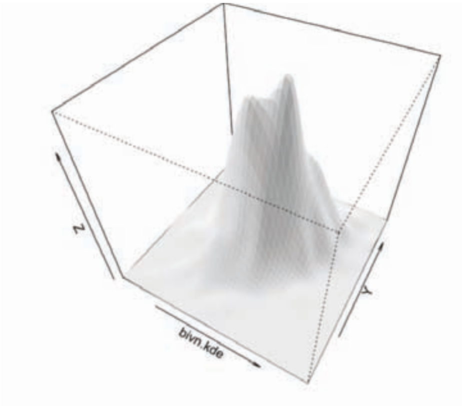




From: Merow & Silander

Probability density functions





- Joint pdfs describe the distribution of 2 or more variables which may or may not be independent of one another

From: Merow & Silander

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Remember likelihood?

Posterior probability H_1

Likelihood of data conditional on H_1

Prior probability H_1

$$\Pr[H_1 \mid data] = \frac{\Pr[data \mid H_1]\Pr[H_1]}{\Pr[data \mid H_1]\Pr[H_1] + \Pr[data \mid H_2]\Pr[H_2]}$$

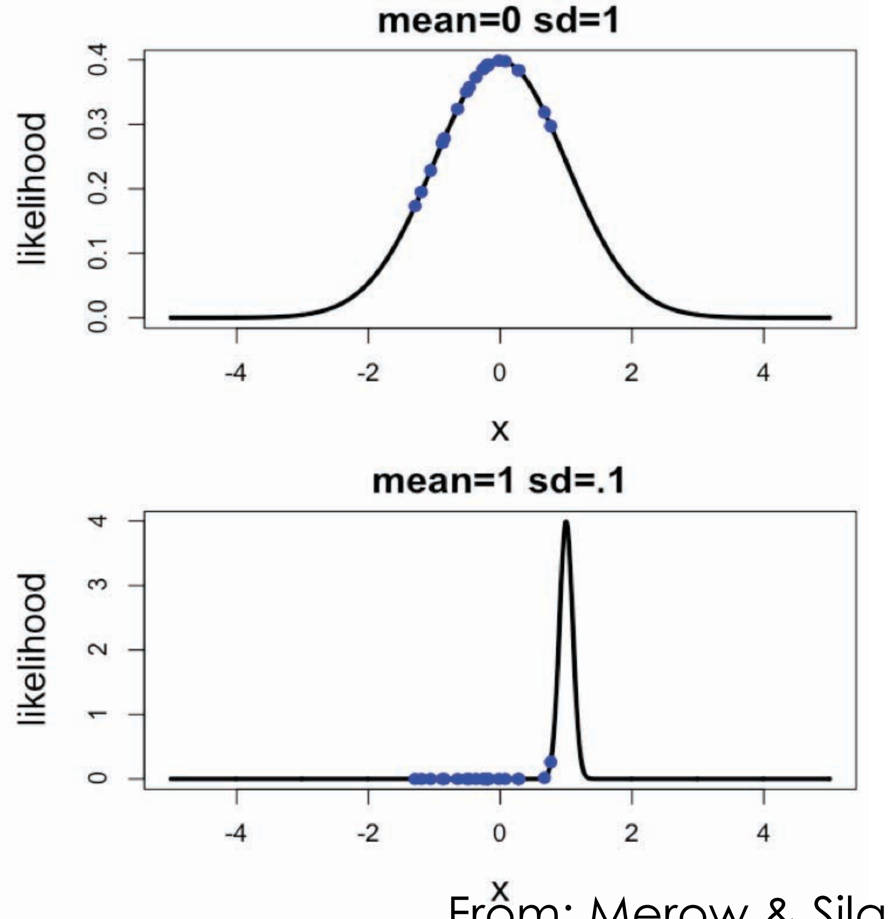
Likelihood and pdfs

- Likelihood can be thought of as the opposite of probability
- Probability (pdfs) allow us to predict an ***unknown outcome*** from *known parameters*
- Likelihood allows us to predict **unknown parameters** from a *known outcome*
- Likelihoods do not integrate to 1, while pdfs do
- We are not going to work on the functions directly, but it is important to remember the relationship between likelihood and pdfs

From: Merow & Silander

Likelihood

- Likelihood allows us to estimate *unknown parameters* with **known data**



From: Merow & Silander

Back to Bayesian framework

Likelihood, can estimate unknown parameters from known data

Prior probability H_1

$$\Pr[H_1 \mid data] = \frac{\Pr[data \mid H_1]\Pr[H_1]}{\Pr[data \mid H_1]\Pr[H_1] + \Pr[data \mid H_2]\Pr[H_2]}$$

Posterior probability H_1 After considering data

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Prior distribution

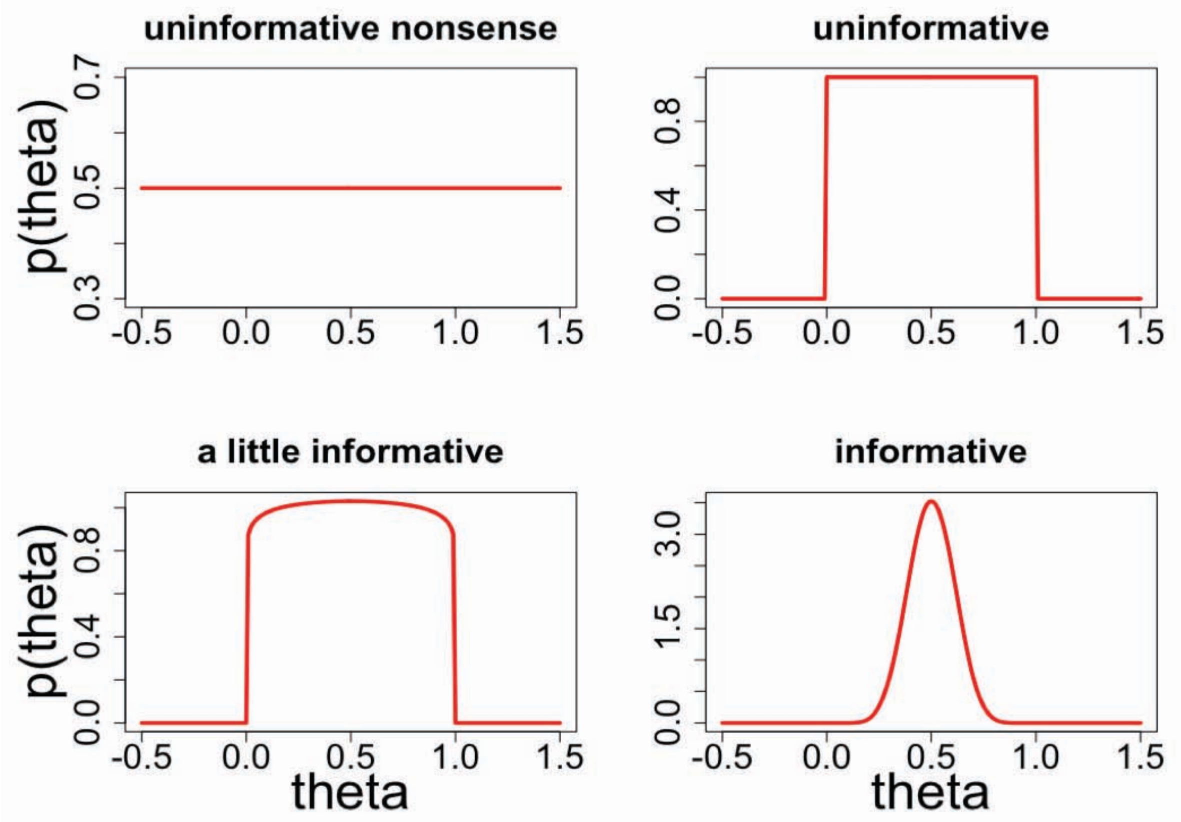
- This describes our previous belief/knowledge about a hypothesis or parameter value (□)
- It can incorporate information or be vague / uninformative
- Let's consider a parameter □ and data *D* for a single outcome (unlike previous examples that always had 2 outcomes)

Prior

$$\Pr[\theta \mid D] = \frac{\Pr[D \mid \theta]\Pr[\theta]}{\Pr[D]}$$

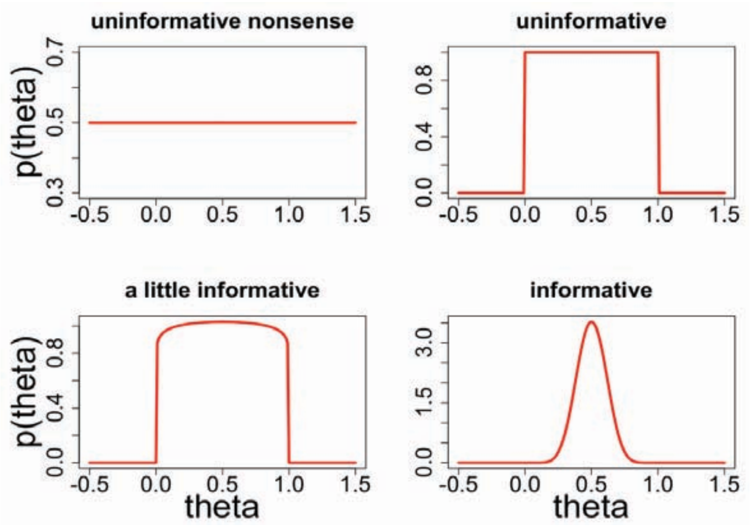
Prior distributions

Consider a prior for the probability parameter from a binomial distribution



From: Merow & Silander

Think, pair, share (5 min)



- Why is the top left distribution nonsense?
- Why is the top right uninformative?
- Why is the bottom left distribution a little informative?
- Why is the bottom right informative?

Notice

- Priors are represented by probability density functions (PDFs)

Where do priors come from?

- Uninformative / vague
  - Chosen to have minimal information content, allows the likelihood to dominate the analysis
- Previous analyses
  - Must be equivalent
- “The literature”
  - Meta-analysis
- Expert knowledge

Where do priors come from?

- PRIOR SPECIFICATION MUST BE BLIND TO THE DATA IN THE ANALYSIS
- IF NOT, LEADS TO FALSELY OVERCONFIDENT RESULTS

Likelihood

$$\Pr[\theta \mid D] = \frac{\Pr[D \mid \theta]\Pr[\theta]}{\Pr[D]}$$

- The exponential distribution likelihood: $L(\rho) = \rho e^{-\rho x}$
- With mean: $\dfrac{1}{\rho}$

- 1. Write the likelihood function for 5 data points
- 2. Write the log likelihood function from 1
- 3. Generate some random exponential data using *rexp*
- 4. Evaluate **3** for different values of□
- 5. Plot **3** as a function of □

From: Merow & Silander

Likelihood

- The exponential distribution likelihood: $L(\rho) = \rho e^{-\rho x}$
- With mean: $\dfrac{1}{\rho}$

- 1. Write the likelihood function for 5 data points n=5
- 2. Write the log likelihood function from 1

$$L(\rho) = \prod_{i=1}^{n} \rho e^{-\rho x_i} = \rho^n [e^{-x_1} e^{-x_2} e^{-x_3} \ldots e^{-x_n}] = \rho^n e^{-\rho \sum x_i}$$

From: Merow & Silander

Likelihood

- The exponential distribution likelihood: $L(\rho) = \rho e^{-\rho x}$
- With mean: $\dfrac{1}{\rho}$

- 1. Write the likelihood function for 5 data points n=5
- 2. Write the log likelihood function from 1

$$L(\rho) = \prod_{i=1}^{n} \rho e^{-\rho x_i} = \rho^n [e^{-x_1} e^{-x_2} e^{-x_3} \ldots e^{-x_n}] = \rho^n e^{-\rho \sum x_i}$$

From: Merow & Silander

$$L(\rho) = \prod_{i=1}^{n} \rho e^{-\rho x_i} = \rho^n [e^{-x_1} e^{-x_2} e^{-x_3} \ldots e^{-x_n}] = \rho^n e^{-\rho \sum x_i}$$

Likelihood

- 1. Write the likelihood function for 5 data points n=5
- 2. Write the log likelihood function from 1

$$\ln(L) = \ln\left(\rho^n e^{-\rho \sum x_i}\right) = \ln \rho^n + \ln\left(e^{-\rho \sum x_i}\right)$$

$$= n \ln \rho \left(-\rho \sum x_i\right) = n \ln \rho - \rho \sum x_i$$

From: Merow & Silander

Likelihood

- 1. Write the likelihood function for 5 data points
- 2. Write the log likelihood function from 1
- 3. Generate some random exponential data using *rexp*
- 4. Evaluate **3** for different values of ρ
- 5. Plot **3** as a function of ρ

From: Merow & Silander

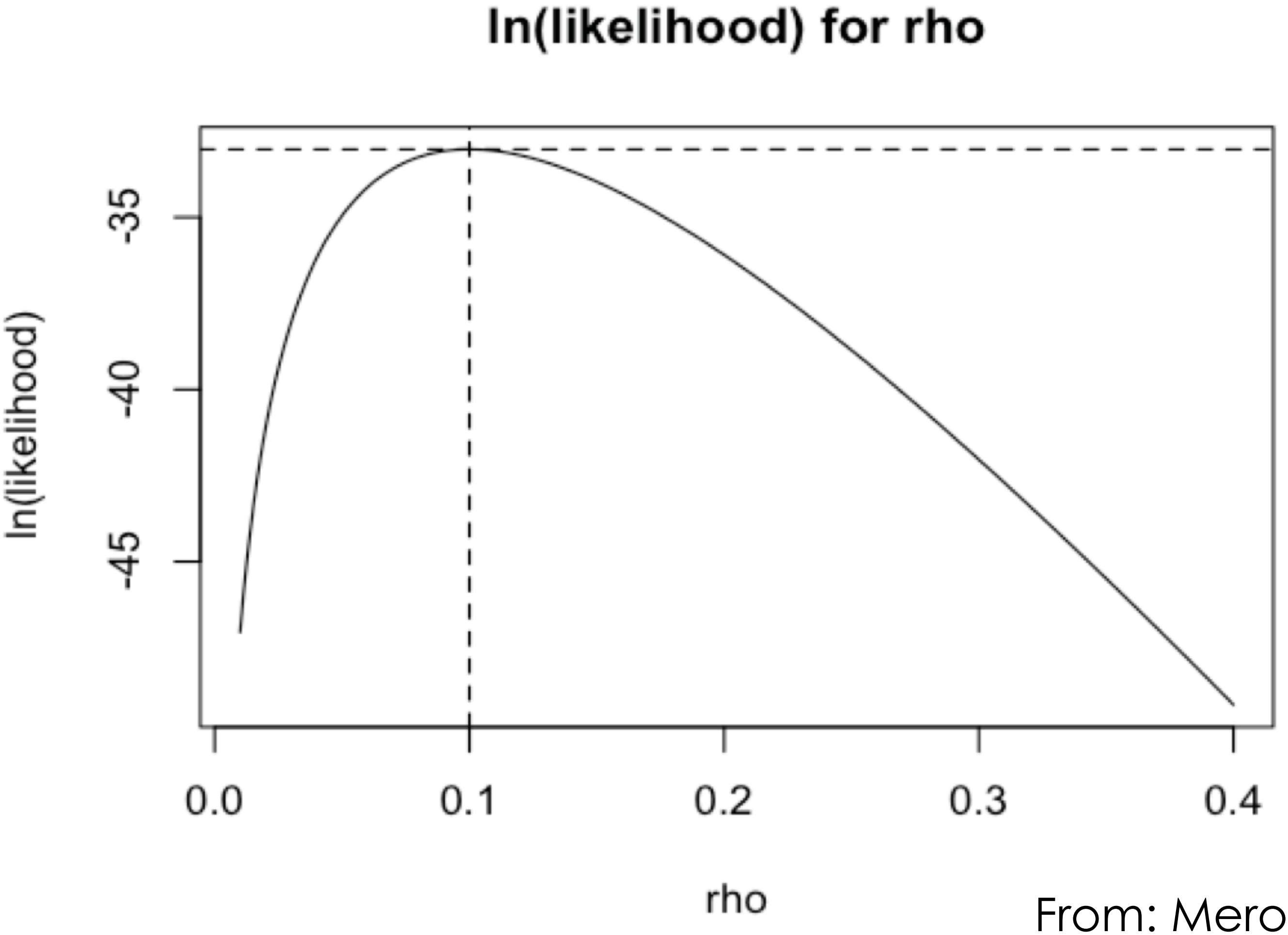
Likelihood

```
data=rexp(10,.1) #simulate data

      #explore likelihood function numerically
rho.seq=seq(.01,.4,length=300) #define a sequence of rho

lexp=function(rho,data){ #write likelihood function
      n=length(data)
      n*log(rho) - rho*sum(data)
      }

plot(rho.seq,lexp(rho.seq,data),type='l') #plot L vs. rho
```

Directions:
1. Paste this code in to R
2. Try some different values for the mean of the exponential distribution (the second argument of **rexp**) to see how the likelihood changes
3. Try sampling different numbers of points by changing the first argument of **rexp** to see how sharply peaked the likelihood function becomes with more data

From: Merow & Silander

Likelihood



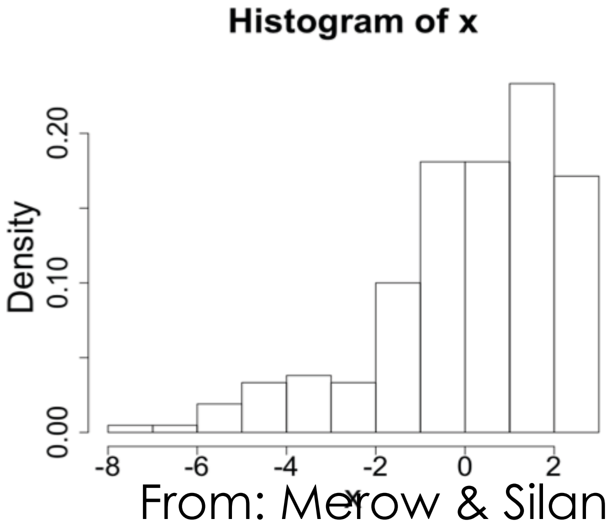
From: Merow & Silander

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Posterior Distribution

- To get posterior distributions for the parameters, we use a search algorithm to find the best values for each parameter
- The search algorithm proceeds by obtaining **samples** from the posterior
- We can summarize the results of all these steps using histograms



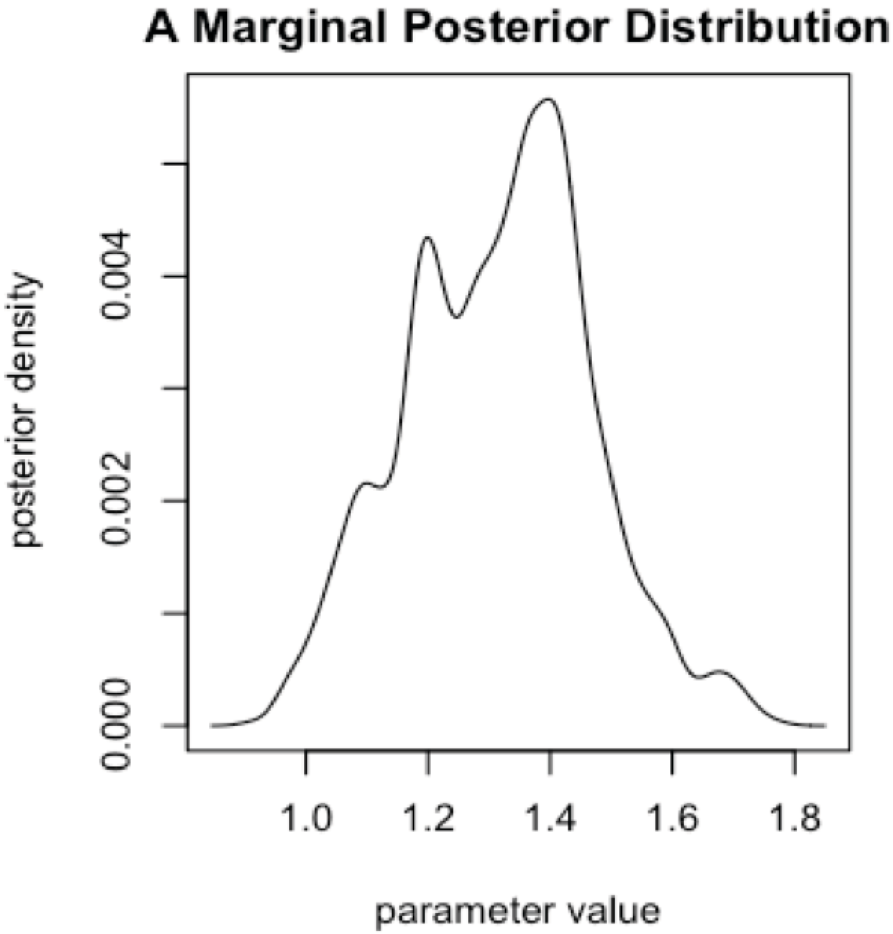
From: Merow & Silander

$$\Pr[\theta \mid D] = \frac{\Pr[D \mid \theta]\Pr[\theta]}{\Pr[D]}$$

Posterior

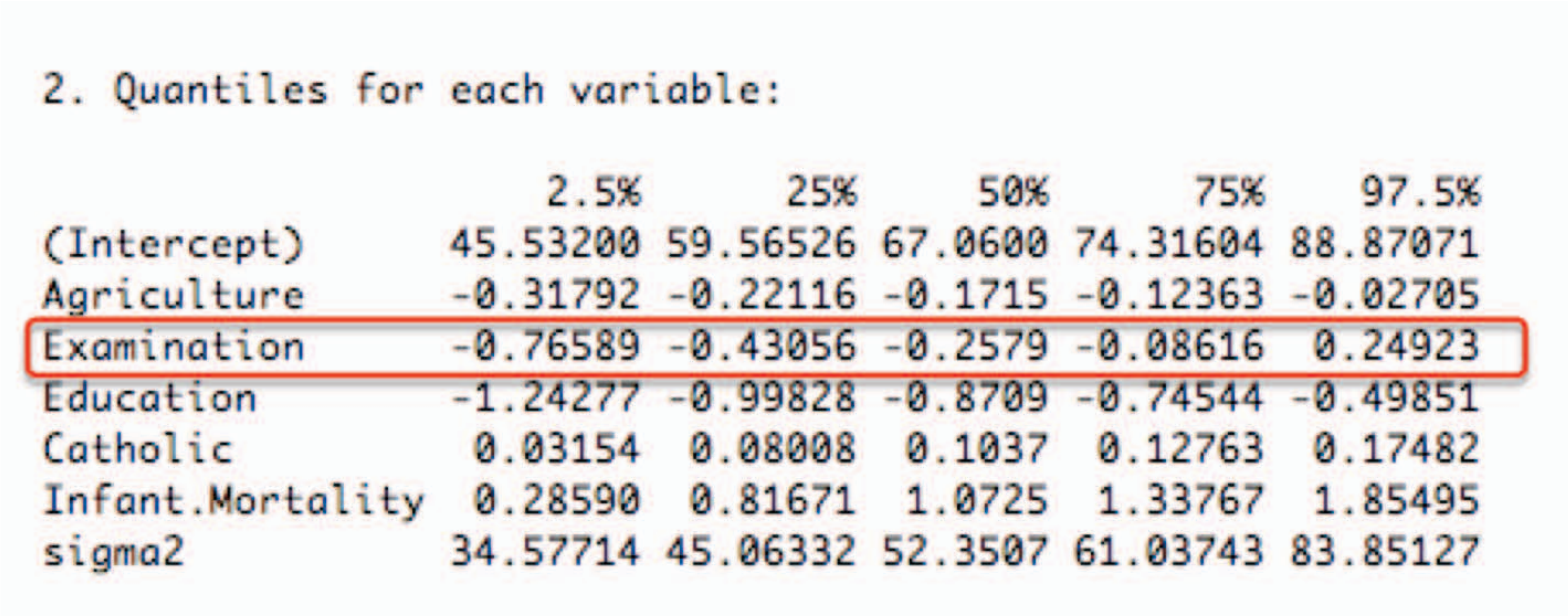
Posterior Distribution

- The posterior distribution summarizes all that we know after analyzing the data
- We get a posterior distribution for each parameter in the model
- The posterior distribution gives the probability that the parameter takes a certain value



From: Merow & Silander

Posterior Quantile Summaries



```
2. Quantiles for each variable:

                    2.5%      25%     50%      75%    97.5%
(Intercept)     45.53200 59.56526 67.0600 74.31604 88.87071
Agriculture     -0.31792 -0.22116 -0.1715 -0.12363 -0.02705
Examination     -0.76589 -0.43056 -0.2579 -0.08616  0.24923
Education       -1.24277 -0.99828 -0.8709 -0.74544 -0.49851
Catholic         0.03154  0.08008  0.1037  0.12763  0.17482
Infant.Mortality 0.28590  0.81671  1.0725  1.33767  1.85495
sigma2          34.57714 45.06332 52.3507 61.03743 83.85127
```

From: Merow & Silander

Today’s lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes’ theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Direct Monte Carlo

- **Monte Carlo** means ‘wandering around’ sort of randomly
- Some algorithms can generate independent samples exactly from the posterior distribution
- In these situations there are **NO** convergence problems or issues
- Sampling is called exact

Markov chain Monte Carlo (MCMC)

- **Markov chain** means that a the sample at *t*+1 depends only on the sample at *t*
  - e.g. $\mu_{t+1} \sim N(\mu_t$, some variance you set)
- In general, exact sampling may not be possible/feasible
- MCMC is a far more versatile set of algorithms that can be invoked to fit more general models
- Note: anywhere where direct Monte Carlo applies, MCMC will provide excellent results too

Convergence issues

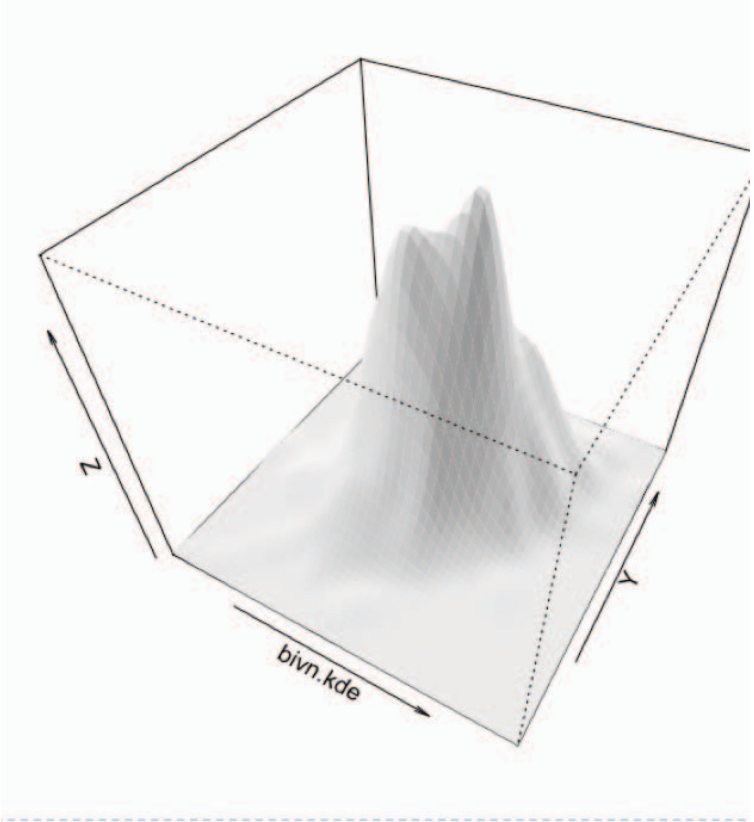
- There is no free lunch!
- The power of MCMC comes at a cost
- The initial samples do not necessarily come from the desired **posterior** distribution
- Rather, they need to converge to the true posterior distribution
- Therefore, one needs to assess convergence, discard output before the convergence and retain only post-convergence samples
- The time of convergence is called **burn-in**

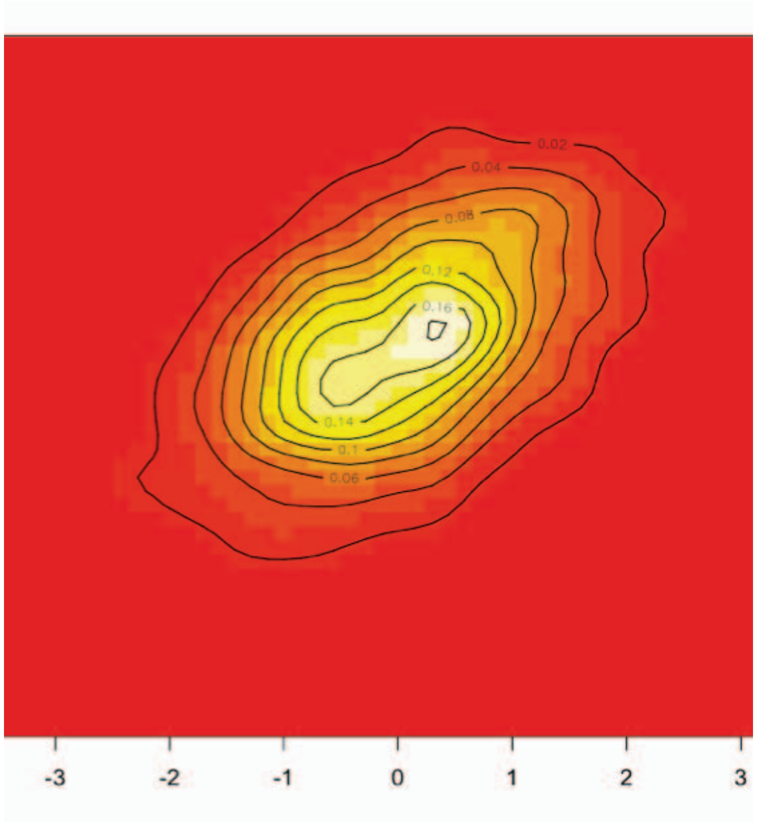
From: Finley 2012

Today's lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes' theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Joint probability density functions





From: Merow & Silander

Gibbs sampling (a type of MCMC)

- Gibbs sampling is an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables
  - Random, as in having a pdf
- The purpose of such a sequence is to to approximate the **marginal** distribution of each of the variables
  - Marginal = probabilities of various values of the without reference to the values of the other variables
- The point of Gibbs sampling is that given a **joint distribution** it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution

From: Merow & Silander

Gibbs Sampling - Algorithm

- If $y_i \sim N(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \tau)$, and
- Then $f(\beta, \tau \mid y) \propto \prod_{i=1 \text{ to } n} f(y_i \mid \beta, \tau) * \prod_{j=1 \text{ to } k} f(\beta_j) * f(\tau)$
  - Posterior $\propto$ likelihood $* \ \beta_j$ priors $* \ \tau$ prior
- The Gibbs Sampler works by repeatedly sampling each of the conditional distributions
  - 1st iteration
    - $\beta_1^{t+1} \sim p(\beta_1^t \mid \beta_2^t, \ldots, \beta_k^t, \tau^t, Y)$
    - $\beta_2^{t+1} \sim p(\beta_2 \mid \beta_1^{t+1}, \beta_3^t, \ldots, \beta_k^t, \tau^t, Y)$
    - …
    - $\beta_k^{t+1} \sim p(\beta_k^t \mid \beta_1^{t+1}, \beta_2^{t+1} \ldots, \beta_{k-1}^t, \tau^t, Y)$
    - $\tau^{t+1} \sim p(\tau^t \mid \beta_{1t+1}, \ldots, \beta_k^{t+1}, Y)$

From: Merow & Silander

Gibbs Sampling - Algorithm

- The Gibbs Sampler works by repeatedly sampling each of the conditional distributions
  - 1st iteration
    - $\beta_1^{t+1} \sim p(\beta_1^t | \beta_2^t, \ldots., \beta_k^t, \tau^t, Y)$
    - $\beta_2^{t+1} \sim p(\beta_2 | \beta_1^{t+1}, \beta_3^t, \ldots., \beta_k^t, \tau^t, Y)$
    - …
    - $\beta_k^{t+1} \sim p(\beta_k^t | \beta_1^{t+1}, \beta_2^{t+1} \ldots., \beta_{k-1}^t, \tau^t, Y)$
    - $\tau^{t+1} \sim p(\tau^t | \beta_{1t+1}, \ldots., \beta_k^{t+1}, Y)$
  - 2nd iteration
    - $\beta_1^{t+2} \sim p(\beta_1^{t+1} | \beta_2^{t+1}, \ldots., \beta_k^{t+1}, \tau^{t+1}, Y)$
    - $\beta_2^{t+2} \sim p(\beta_2^{t+1} | \beta_1^{t+2}, \beta_3^{t+1}, \ldots., \beta_k^{t+1}, \tau^{t+1}, Y)$
    - …
    - $\beta_k^{t+2} \sim p(\beta_k^{t+1} | \beta_1^{t+2}, \ldots., \beta_{k-1}^{t+1}, \tau^{t+1}, Y)$
    - $\tau^{t+2} \sim p(\tau^{t+1} | \beta_1^{t+1}, \ldots., \beta_k^{t+2}, Y)$
  - 3rd iteration
    - Etc.

From: Merow & Silander

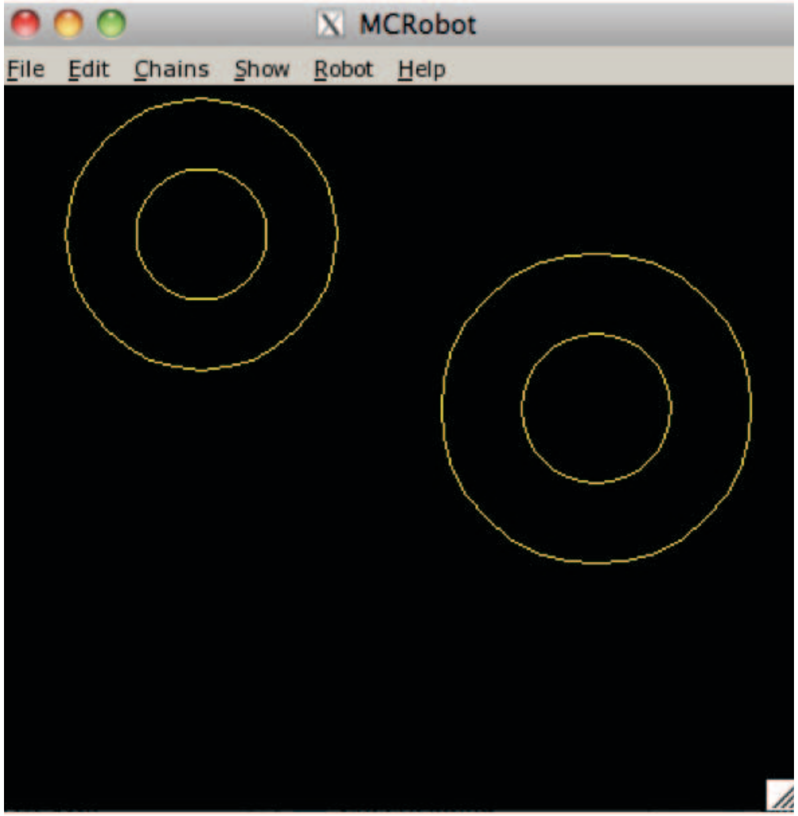
Gibbs Sampling

- Target distribution is the posterior
- Proposal (Jump, Tuning) distribution decides where to go next
- In Gibbs sampling, the proposal distribution is a random walk
- In Gibbs sampling, the proposed value is always accepted

From: Merow & Silander

Gibbs sampling

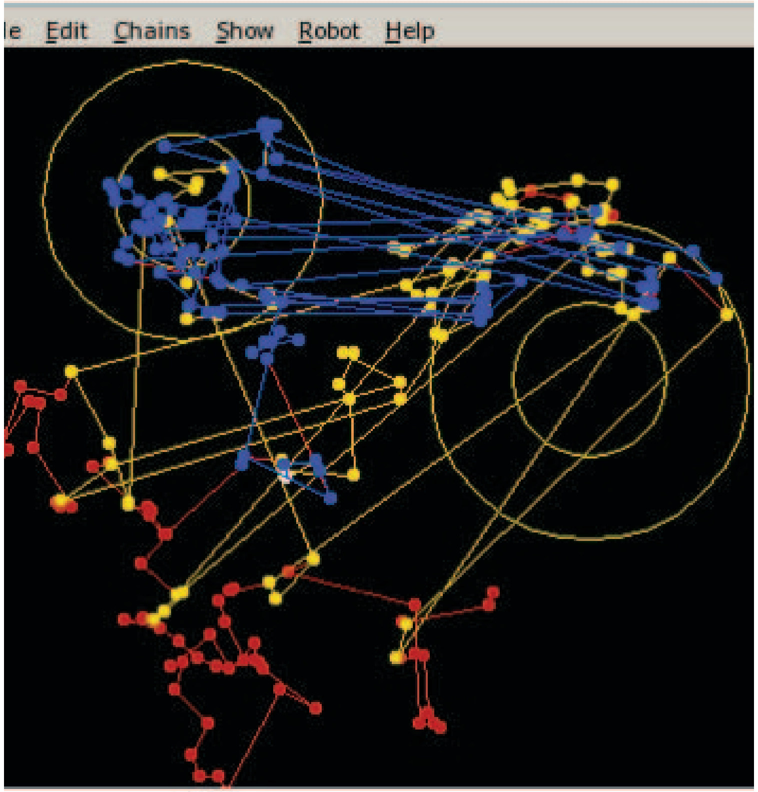
- Example: MCMC Robot – a program by Paul Lewis
  - Installed on the computers
- Input: a joint posterior distribution for all parameters
- Output: the trajectory of a random MCMC walk
- Most commands you need are under the 'Robot' menu
- Clicking on the black area and dragging your mouse creates normal distributions with means at your starting point



From: Merow & Silander

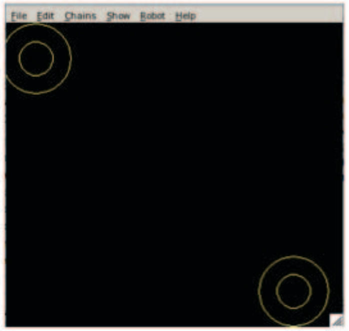
MCMC Robot

- To start a chain, choose ‘start walking’ from the ‘robot’ menu
- Explore the options on the robot menu
  - Note that you can set the number of steps per run, which allows you to see more of each chain
- Try making a more complex sample space
- ‘Clear’ erases the walk but not the landscape; you have to close and reopen to get new landscapes
- Use the ‘chains’ menu to enable multiple chains
- Use the ‘show’ menu to show these different chains
- You can read the very short help file if these settings are unclear under the help menu
- See questions on next slide

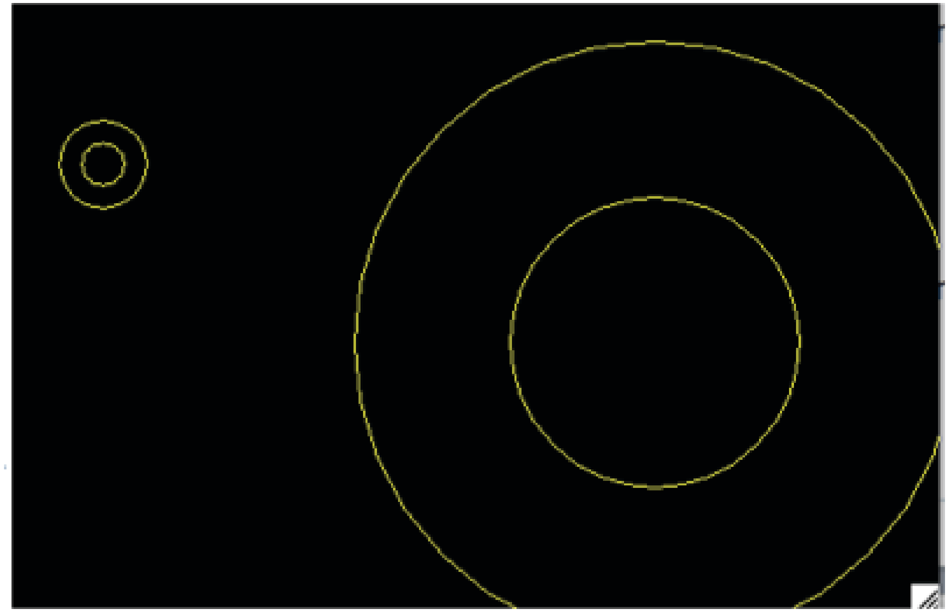


From: Merow & Silander

MCMC Robot



- QUESTIONS:
- What similarities and differences do you notice among different chains?
- What happens to the chain if you make 2 small islands in opposite corners of the window?
- What happens to the chain if you make 1 large and 1 small island?



From: Merow & Silander

Today's lecture

- Why use Bayesian?
- Probability
- Alternative definition probability
- Bayes' theorem
- An example from ecology
- Probability density function
- Likelihood
- Prior
- Posterior
- Estimating posteriors
  - Gibbs sampling
- Convergence

Convergence and MCMC

- General questions to ask :
  - At what point have we converged to the stationary distribution?
  - (i.e. how long should our "burn-in" period be?)
  - After we have reached the stationary distribution, how many iterations will it take to summarize the posterior distribution?
- The answers to both of these questions remain a bit ad hoc because the desirable results that we depend on are only true asymptotically
- When there are island of high-probability states with no paths between them, the chain can get stuck
- Given infinite time it will emerge, but you probably don't have that long

From: Merow & Silander

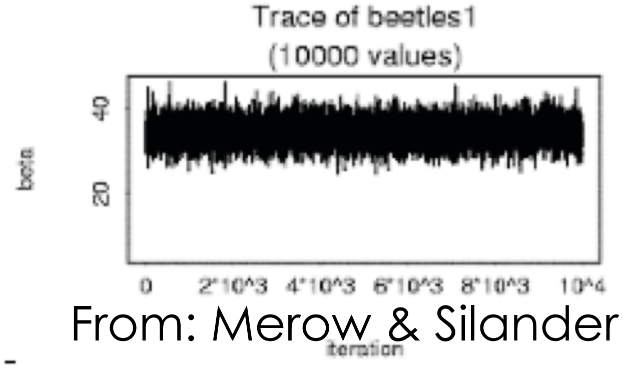
Diagnosing convergence

- Usually a few parallel chains are run from rather different starting points
- The sample values are plotted (called trace-plots) for each of the chains
- The time for the chains to “mix” together is taken as the time for convergence

From: Finley 2012

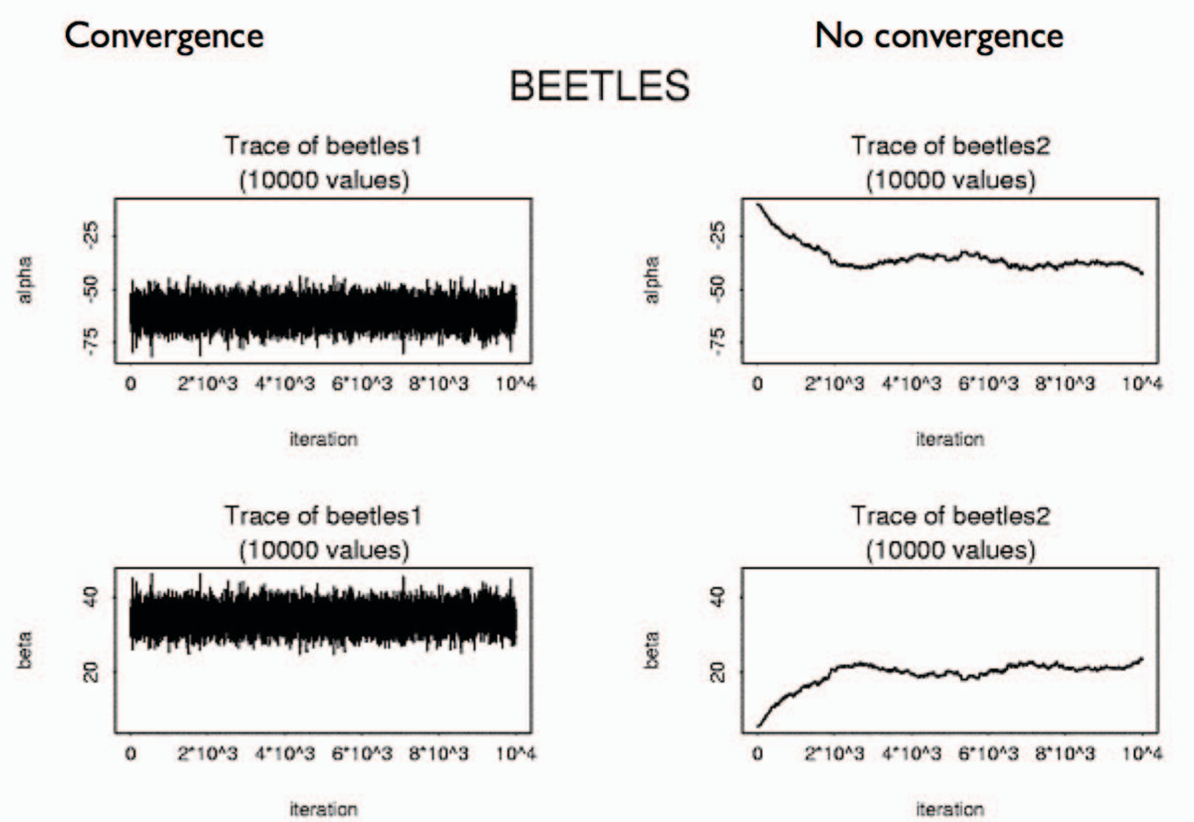
Posterior Trace Plots

- Plots the parameter value at time t against the iteration number.
- If the model has converged, the trace plot will move around the mode of the distribution.
- A clear sign of non-convergence with a trace plot occurs when we observe some trending in the sample space.
- The problem with trace plots is that it may appear that we have converged, however, the chain is trapped in a local region rather exploring the full posterior.



From: Merow & Silander

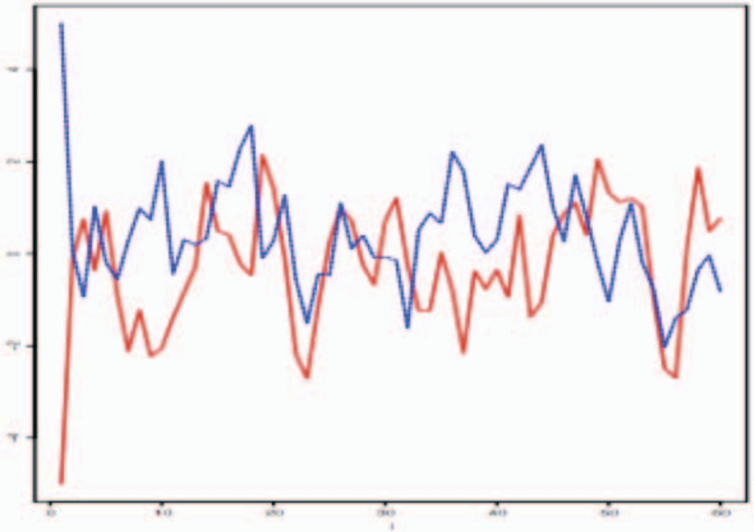
Apparent Convergence and Non- Convergence



From: Merow & Silander

MCMC convergence

- To check for convergence, multiple starting points can be used
- Convergence diagnostics
  - Geweke time series diagnostic
  - Gelman-Rubin diagnostic

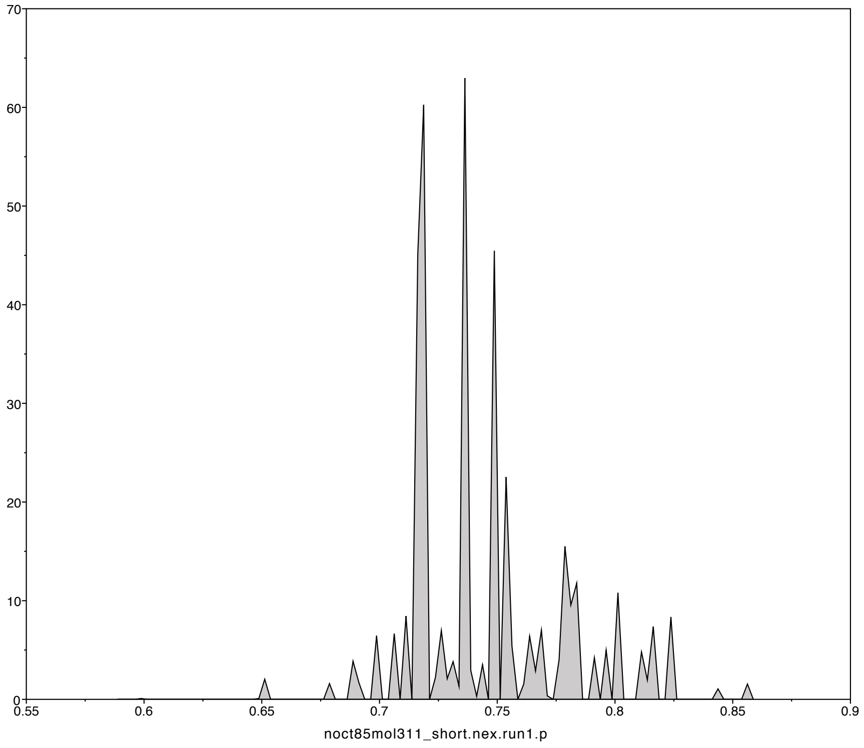


After about 5–7 iterations the chains seemed to have forgotten their starting positions.

From: Merow & Silander

Posterior Parameter Density Plots

- Sometimes non-convergence is reflected in lumpy distributions so search longer
- BAD!



Example: Trace plots and posterior densities from a regression

```
install.packages('MCMCpack’)

library(MCMCpack)

data(swiss)

posterior1 = MCMCregress(Fertility ~ Agriculture + whatever you want,
data=swiss, mcmc=200, burnin=0)

summary(posterior1) plot(posterior1)
```

- Try different values for `mcmc`, which sets the number of posterior samples, and for burnin, which sets the number of samples to throw out before collecting samples
  - What is a sufficient burnin length and number of samples to make sure that there’s not a long term trend in the trace plot and that the density plot is pretty smooth?
- Use different combinations of predictors to find the best model
- Look at the posterior summary: How can you tell if a predictor should be dropped?
- Have a look at the help file for MCMCregress if you get bored and alter more settings, like the starting values for each chain.

From: Merow & Silander

- THE END