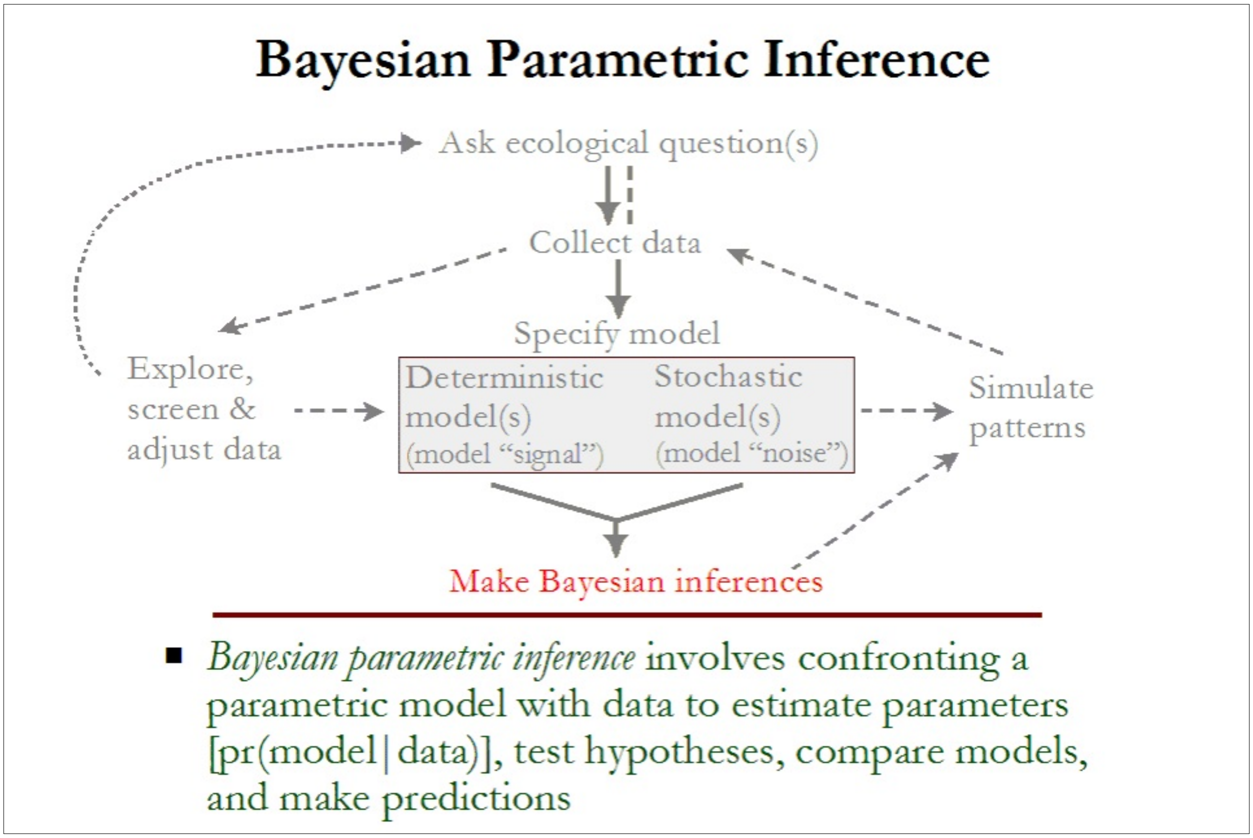
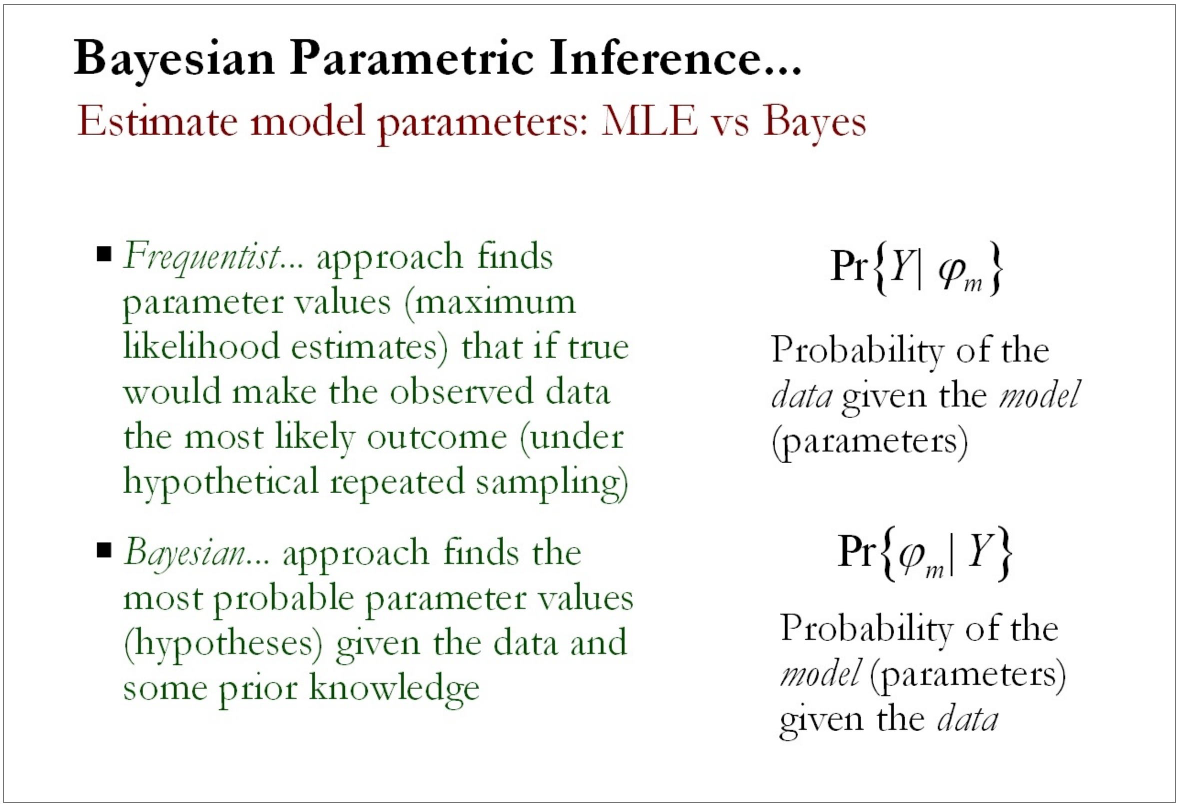
# Analysis of Environmental Data
## Conceptual Foundations:
### *Bayesian Inference*




## 1. Bayesian parametric inference

As we have seen, the method of ordinary least squares can be used to find the best fit of a model to the data under minimal assumptions about the sources of uncertainty and the method of maximum likelihood can be used to find the best fit of a model the data when we are willing to make certain assumptions about how uncertainty enters into the system. When a parametric approach can be justified, Bayesian inference offers an alternative to Maximum Likelihood and allows us to determine the probability of the model (parameters) given the data – something we cannot do with maximum likelihood. Bayesian inference is especially useful when we want to combine new data with prior knowledge of the system to make inferences that better reflect the cumulative nature of scientific inference.




## 2. Parameter estimation the Bayesian way

The next step is to fit the model; i.e., estimate the model parameters. Recall from our earlier chapter on inference frameworks that the Bayesian framework says that the experimental outcome – what we actually saw happen – is the truth, while the parameter values or hypotheses have probability distributions. The Bayesian framework solves many of the conceptual problems of frequentist statistics: answers depend on what we actually saw and not on a range of hypothetical outcomes, and we can legitimately make statements about the probability of different hypotheses or parameter values. The major catch is that to accomplish this, we have to specify our prior beliefs about the probability of different hypotheses, and these prior beliefs actually affect our answers. More on the pros and cons of Bayesian inference later; first, let's see how it works.

What we need is an objective method of computing parameter estimates from the data that are in some sense the 'best' estimates of the parameters for the data and this particular model. Thus, the goal is the same as in the frequentist maximum likelihood approach. The difference is in how they go about trying to achieve this goal. Whereas the frequentist approach seeks to find parameter values (maximum likelihood estimates) that if true would make the observed data the most likely outcome (under hypothetical repeated sampling), the Bayesian approach seeks to find the most probable parameter values (hypotheses) given the data and some prior knowledge. The Bayesian approach produces estimates of the probabilities of alternative parameter values (or hypotheses) based on all the available data, including previously collected data – which we might argue is the goal of science. Bayes' Theorem provides a way of doing this.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

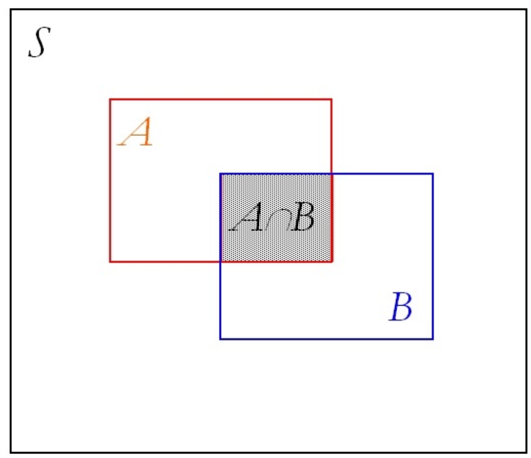
1. Conditional probability:

$\Pr\{B|A\}$ = (area common to $A$ and $B$) / (area of $A$)

$\Pr\{B|A\} = \Pr\{A \cap B\} / \Pr\{A\}$

$\Pr\{A|B\}$ = (area common to $A$ and $B$) / (area of $B$)

$\Pr\{A|B\} = \Pr\{A \cap B\} / \Pr\{B\}$




*Conditional Probability:*

To understand Bayes' Theorem and how it applies to parameter estimation, we first have to understand some basic probability theory. Consider an experiment that can produce a variety outcomes. We will denote the entire collection of outcomes as $S$ and represent it as a polygon of some "area". A smaller collection of outcomes, $A$, contained within $S$ has probability defined as the "area" of $A$ divided by the "area" of $S$, with "area" suitably defined. Different probability models give different definitions of what "area of $S$" really means. Suppose we know that event $A$ occurred. What is the probability that a different event $B$ occurred, given (conditioned on) the knowledge about $A$? If $A$ occurred, then the collection of all possible outcomes of the experiment is no longer $S$, but must be $A$. Thus, the conditional probability of $B$ given $A$ is:

$$\begin{aligned} \Pr\{B|A\} &= \text{probability of event } B \text{ given that } A \text{ occurred} \\ &= (\text{area common to } A \text{ and } B) \,/\, (\text{area of } A) \end{aligned} \qquad \text{(Eq. 1)}$$

The vertical line in the probability statement on the left-hand side of the equation means "given" or "conditioned on". Dividing the numerator and denominator of the right-hand side of equation 1 by the area of $S$, to convert "area" to probability, we have:

$$\Pr\{B|A\} = \Pr\{A \cap B\} \,/\, \Pr\{A\} \qquad \text{(Eq. 2)}$$

where "$\cap$ " stands for "and" or "intersection". By analogy, since A and B are fully interchangeable here, we must also have:

$$\Pr\{A|B\} = \Pr\{A \cap B\} \;/\; \Pr\{B\} \qquad \text{(Eq. 3)}$$

Equation 3 is how we define "conditional probability": the probability of something given (conditioned on) something else.

# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

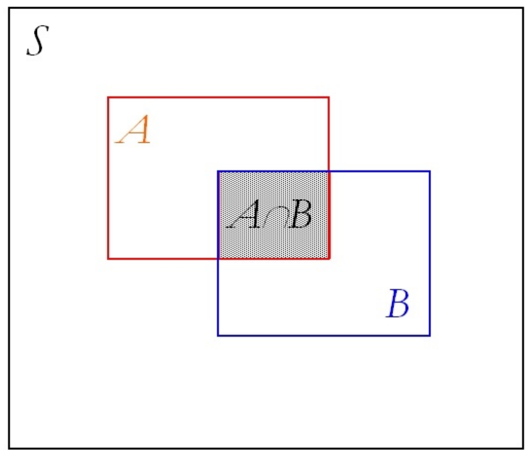
2. Bayes' Theorem/Bayes' Rule:

$\Pr\{A|B\} = \Pr\{A \cap B) \;/\; \Pr\{B\}$

$\boxed{\Pr\{A \cap B\}} = \Pr\{A|B\}\ \Pr\{B\}$

$\Pr\{B|A\} = \boxed{\Pr\{A \cap B\}} \;/\; \Pr\{A\}$

Bayes' Rule:
$\Pr\{B|A\} = \Pr\{A|B\}\ \Pr\{B\} \;/\; \Pr\{A\}$




*Bayes Theorem*:

With this simple conditional probability rule, we can now derive *Bayes' Theorem* or *Bayes' Rule*, discovered by Reverend Thomas Bayes and published two years after his death in 1763. Bayes' Rule can be described a variety of ways, ranging from a very generalized form to a very specific form as used in model inference, as follows:

(1) By rearranging equation 3 to isolate $\Pr\{A \cap B\}$ on one side of the equation and substituting the result into equation 2, we get the following:

$$\Pr\{B|A\} = \Pr\{A|B\}\Pr\{B\} \;/\; \Pr\{A\} \qquad \text{(Eq. 4)}$$

which is called Bayes' Rule in its most general form, where $A$ and $B$ can represent anything. In this form, Bayes' Rule states that the probability of event $B$ given the occurrence of event $A$ is equal to the probability of event $A$ given that event $B$ has occurred times the probability of event $B$ occurring, divided by the probability of event $A$ occurring.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

2. Bayes' Theorem/Bayes' Rule:

$\Pr\{B|A\} = \Pr\{A|B\}\ \Pr\{B\}\ /\ \Pr\{A\}$

Substitute "hypothesis" for B and "data" for A:
Pr{hypothesis | data} = Pr{data | hypothesis} Pr{hypothesis} / Pr{data}

Abbreviated:

$$\Pr\{H|\ D\} = \frac{\Pr\{D|\ H\}\ \Pr\{H\}}{\Pr\{D\}}$$

(2) If we substitute "hypothesis" (or "model") for *B* and "data" for *A* in equation 4, we get the following:

$$\Pr\{\text{hypothesis}|\text{data}\} = \Pr\{\text{data}|\text{hypothesis}\}\Pr\{\text{hypothesis}\}\ /\ \Pr\{\text{data}\}$$

or simply:

$$\Pr\{H|D\} = \Pr\{D|H\}\Pr\{H\}\ /\ \Pr\{D\} \qquad \text{(Eq. 5)}$$

In this context, Bayes' Rule states that the probability of the hypothesis (parameter values) given the data is equal to the probability of the data given the hypothesis (the likelihood associated with H), times the probability of the hypothesis, divided by the probability of the data. The left-hand side of the equation gives the probability of any particular parameter value(s) given the data.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

2. Bayes' Theorem/Bayes' Rule:

$$\Pr\{H|\, D\} = \frac{\Pr\{D|\, H\}\,\Pr\{H\}}{\Pr\{D\}} \;\Rightarrow\; \Pr\{\varphi_m|\, Y\} = \frac{\Pr\{Y|\, \varphi_m\}\,\Pr\{\varphi_m\}}{\Pr\{Y\}}$$

- $\Pr\{Y|\varphi_m\}$ is the *likelihood* of observing the data for different values of $\varphi$
- $\Pr\{\varphi_m\}$ is the "unconditional" prior probability of $\varphi_m$ expressed as a pdf
- $\Pr\{Y\}$ is the "unconditional" probability of the data, this standardization means that the area under the posterior pdf equals one
- $\Pr\{\varphi_m|Y\}$is the posterior probability of $\varphi_m$ conditional on the data being observed, expressed as a pdf

(3) Following the notation we used to describe likelihood, we can write Bayes' Rule as follows:

$$\Pr\{\varphi_m|\, Y\} = \frac{\Pr\{Y|\, \varphi_m\}\,\Pr\{\varphi_m\}}{\Pr\{Y\}} \qquad \text{(Eq. 6)}$$

where:

- $\varphi_m$ is the population parameter(s) to be estimated; we are looking for the value(s) with the maximum probability given the data,
- $\Pr\{Y|\varphi_m\}$ is the likelihood of observing the data for different values of $\varphi$, expressed as a likelihood function,
- $\Pr\{\varphi_m\}$ is the "unconditional" prior probability of $\varphi_m$ expressed as a probability distribution summarizing our prior views about the probability of $\varphi$ taking on different values,
- $\Pr\{Y\}$ is the "unconditional" probability of the data, which is equal to the expected value (mean) of the likelihood function; this standardization means that the area under the posterior probability distribution (below) equals one, and
- $\Pr\{\varphi_m|Y\}$is the posterior probability of $\varphi_m$ conditional on the data being observed, expressed as a probability distribution summarizing the probability of $\varphi$ taking on different values by combining the prior probability distribution and the likelihood function. Note, the "posterior" refers to the fact that it reflects the probability of the parameter values *after* observing the data, whereas the "priors" reflect the probability of the parameter values *before* observing the data.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

2. Bayes' Theorem/Bayes' Rule:

$$\Pr\{H|\,D\} = \frac{\Pr\{D|\,H\}\,\Pr\{H\}}{\cancel{\Pr\{D\}}}$$

Posterior probability $\boxed{\propto}$ likelihood x prior probability

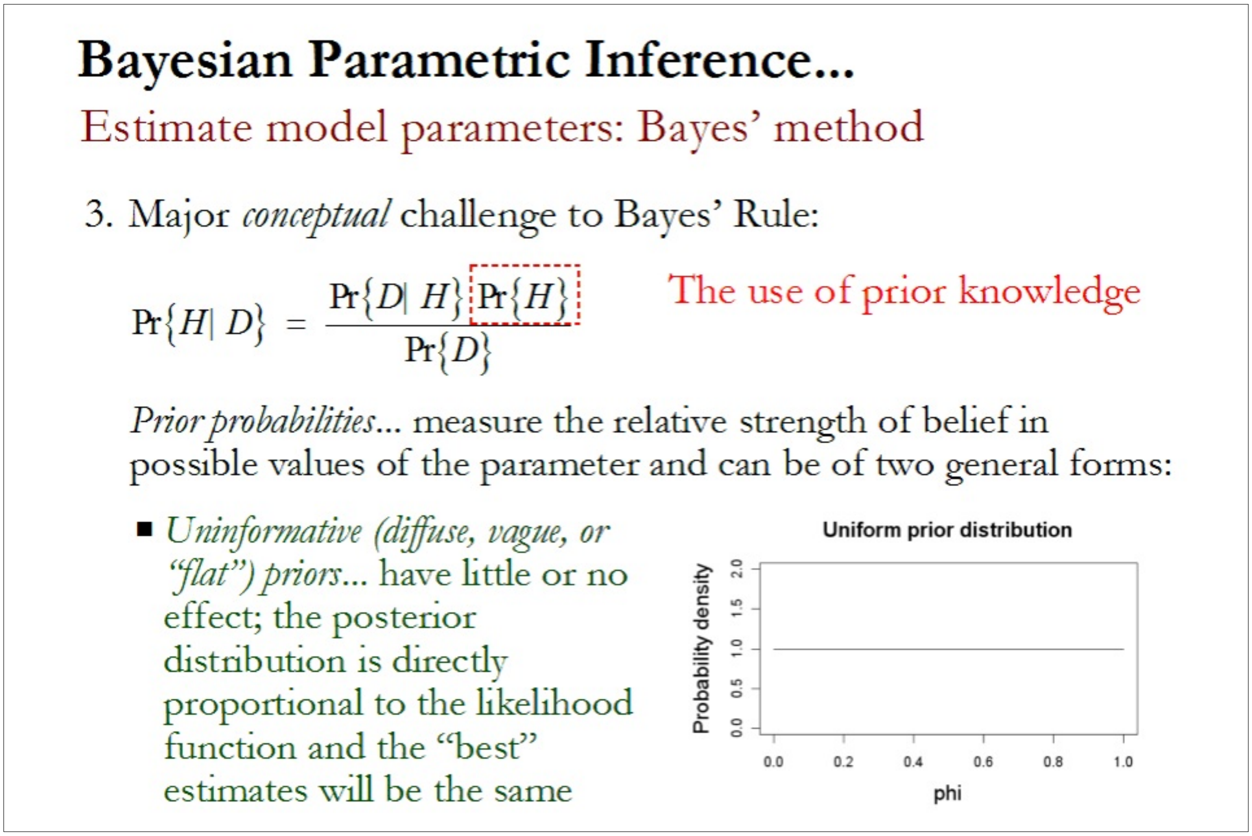
- $\Pr\{D\}$ is just a normalizing constant, which means the area under the posterior distribution equals 1 (true probability)
- Thus, Bayes' posterior probability is essentially just likelihood modified by prior knowledge

(4) We can re-express Bayes' Rule even more simply as follows:

$$\text{Posterior probability} \propto \text{likelihood x prior probability} \qquad \text{(Eq. 7)}$$

where "∝" means proportional to, because the denominator $\Pr\{Y\}$in equation 6 is simply a normalizing constant that makes the posterior probability a true probability. Note that likelihood is where our data comes in and is the same likelihood we used in maximum likelihood estimation. Thus, the major difference between Bayesian estimation and maximum likelihood estimation is whether prior knowledge is taken into account.

If Bayes' Rule seems too good to be true, it's because it might just be. There are a couple of major problems with the application of Bayes' Rule in statistical inference (that frequentists are keen to point out): (1) we don't know the unconditional probability of the data $\Pr\{Y\}$, and (2) we don't know the unconditional prior probability of the hypothesis $\Pr\{\varphi_m\}$– isn't that what we were trying to figure out in the first place. These problems have received considerable attention among statisticians and have important implications, so let's briefly consider each of these problems in turn.




*Prior probabilities*:

Once of the major sources of controversy with Bayesian statistics is the use of prior probability distributions in Bayes' Rule – which is a necessity to make it work. Prior probabilities measure the relative strength of belief in possible values of the parameter and can be of two general forms:

(1) <u>Uninformative or "flat" priors</u>.–In many cases, we have no (or only vague) prior knowledge to suggest what value a parameter might take. Bayesian's might argue that we always have some knowledge and that it should be incorporated at what ever level it exists. However, others argue that assuming prior ignorance is a conservative approach and helps overcome the criticism of Bayesian statistics that subjectively determined prior opinion can have too much influence on the inferential process. This gets at one of the major sources of philosophical contention between Bayesian's and frequentist's. Nevertheless, we can represent prior ignorance with a uninformative prior distribution, sometimes called a diffuse distribution because such a wide range of values of $\varphi$ is considered possible. The most typical diffuse prior is a rectangular (uniform or flat) probability distribution, which says that each value of the parameter is equally likely. Not surprisingly, when we use a flat prior, the posterior distribution of the parameter is directly proportional to the likelihood function and the "best" estimates will be the same. A flat prior can be considered a reference prior, a class of priors designed to represent weak prior knowledge and let the data, and therefore the likelihood, dominate the posterior distribution.

# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

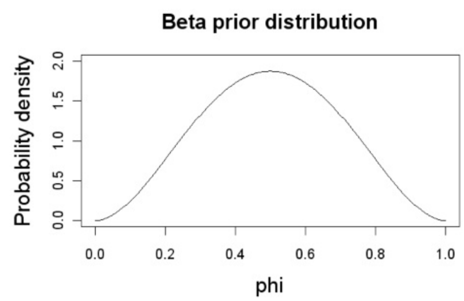
3. Major *conceptual* challenge to Bayes' Rule:

$$\Pr\{H|\,D\} = \frac{\Pr\{D|\,H\}\boxed{\Pr\{H\}}}{\Pr\{D\}}$$

The use of prior knowledge

*Prior probabilities*... measure the relative strength of belief in possible values of the parameter and can be of two general forms:

- *Informative priors*... can effect the posterior and thus the estimates, depending on sample size; their construction is one of the most contentious aspects of Bayesian inference




(2) Informative priors.–The alternative is an informative prior, representing cases where we have substantial prior knowledge about the value of the parameter. Here we might specify, for example, a beta distribution about our prior expectation of the parameter value. The construction of informative priors is one of the most contentious aspects of Bayesian inference, especially if they are constructed from subjective opinion as opposed to previous empirical estimates. Bayesian's claim the use of informative priors is logical and consistent with the way we typically do science: we down-weight observations that are inconsistent with our current beliefs, while using those in line with our current beliefs to strengthen and sharpen those beliefs. Frequentist's oppose any form of weighting based on prior beliefs.

# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

4. Major *technical* challenge to Bayes' Rule:

$$\Pr\{H|\,D\} = \frac{\Pr\{D|\,H\}\,\Pr\{H\}}{\boxed{\Pr\{D\}}}$$

Determining the probability of the data

*Probability of the data...* measures the unconditional probability of the data; normalizing constant which makes the area under the posterior distribution equal 1 (true probability); two situations to consider:

- *Discrete hypotheses...* set of exhaustive, mutually exclusive hypotheses (parameter values)
- *Continuous hypotheses...* set of continuous hypotheses (parameter values)

*Probability of the data*:

The other major challenge with Bayesian statistics, which is purely technical and does not raise any deep conceptual issues, is the problem of figuring out the unconditional probability of the data $\Pr\{Y\}$, the denominator in Bayes' Rule. There are two situations we need to consider: (1) a set of exhaustive, mutually exclusive discrete hypotheses (parameter values), and (2) a continuous set of hypotheses.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

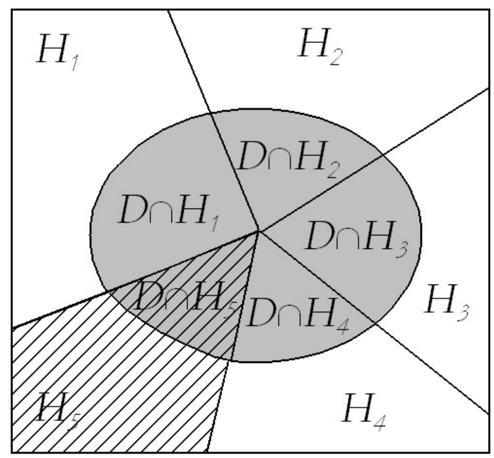
4. Major *technical* challenge to Bayes' Rule:

$$\Pr\{H|\, D\} \;=\; \frac{\Pr\{D|\, H\}\,\Pr\{H\}}{\boxed{\Pr\{D\}}}$$

(1) Discrete hypotheses:

$$\Pr\{D\} \;=\; \sum_{j=1}^{N} \Pr\{D \cap H_j\} \;=\; \sum_{j=1}^{N} \Pr\{D|\, H_j\}\,\Pr\{H_j\}$$

$$\Pr\{H_i|\, D\} \;=\; \frac{\Pr\{D|\, H_i\}\,\Pr\{H_i\}}{\sum_{j=1}^{N} \Pr\{D|\, H_j\}\,\Pr\{H_j\}}$$



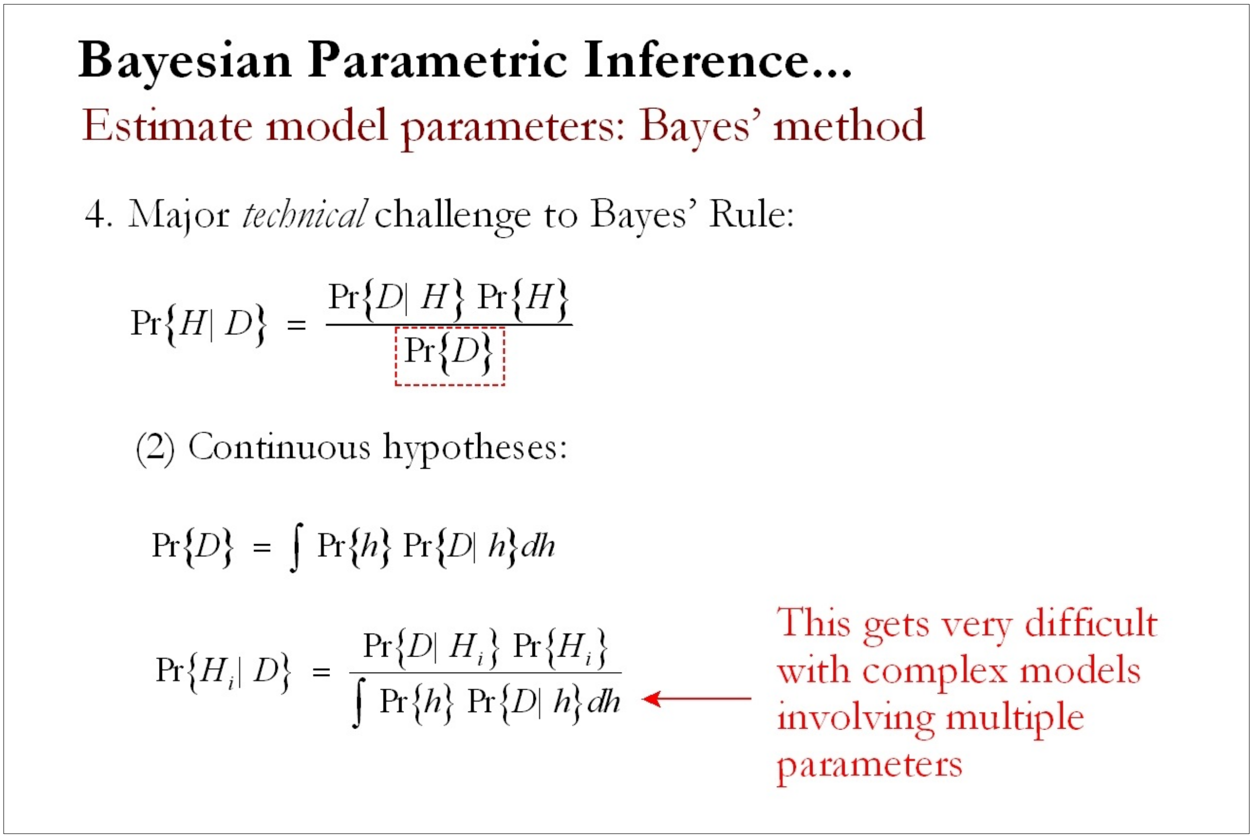

From Bolker (2008)

(1) <u>Discrete hypotheses</u>.–If we have a set of exhaustive and mutually exclusive hypotheses, we can assume that one, and only one, of our hypotheses is true. Thus, in the geometric representation shown here, the shaded ellipse represents *D*, the set of all possibilities that could lead to the observed data. If one of the hypotheses must be true, then the unconditional probability of observing the data is the sum of the probabilities of observing the data under any of the possible hypotheses. Referring to equation 3 rearranged, for *N* different hypotheses, $H_1$ to $H_n$,

$$\Pr\{D\} \;=\; \sum_{j=1}^{N} \Pr\{D \cap H_j\} \;=\; \sum_{j=1}^{N} \Pr\{D|\, H_j\}\,\Pr\{H_j\} \qquad \text{(Eq. 8)}$$

Substituting equation 8 into equation 5 gives the full form of Bayes' Rule for a particular hypothesis $H_i$ when it is one of a mutually exclusive set of hypotheses ($H_j$). The probability of the truth of $H_i$ in light of the data is:

$$\Pr\{H_i|\, D\} \;=\; \frac{\Pr\{D|\, H_i\}\,\Pr\{H_i\}}{\sum_{j=1}^{N} \Pr\{D|\, H_j\}\,\Pr\{H_j\}} \qquad \text{(Eq. 9)}$$




(2) Continuous hypotheses.–If the set of hypotheses (parameter values) is continuous, then the denominator must be obtained through integration, as follows:

$$\Pr\{D\} = \int \Pr\{h\}\Pr\{D|h\}dh \qquad \text{(Eq. 10)}$$

where *h* is a particular parameter value. Substituting equation 10 into equation 5 gives the full form of Bayes' Rule for a particular hypothesis $H_i$ when it is one of a continuous range of possible values, *h*. The probability of the truth of $H_i$ in light of the data is:

$$\Pr\{H_i|D\} = \frac{\Pr\{D|H_i\}\Pr\{H_i\}}{\int \Pr\{h\}\Pr\{D|h\}dh} \qquad \text{(Eq. 11)}$$

It turns out that for relatively simple cases it's easy to sum or integrate the denominator either analytically or numerically, and if we care only about the relative probability of different hypotheses, which is usually the case, then we don't need to integrate the denominator at all because it has the same constant value for every hypothesis and we can therefore simply ignore it. This is good news, because calculating the unconditional probability of the data can be extremely complex for more complicated problems. Consequently, most current methods for estimating the posterior distribution are based on a sophisticated technique known as Markov Chain Monte Carlo (MCMC) which does not require us to calculate the unconditional probability of the data. The details of MCMC are beyond the scope of this chapter, but suffice it to say that MCMC methods involve sampling the posterior distribution many times in order to estimate the posterior distribution, from which we can extract useful summaries such as the mean of the sample posterior distribution.

*Posterior probability*:

All conclusions from Bayesian inference are based on the posterior probability distribution of the parameter or an estimate of the posterior derived from MCMC sampling. This posterior distribution represents our prior probability distribution modified by the current data through the likelihood function. Bayesian inference is usually based on the shape of the posterior distribution, particularly the range of values over which most of the probability mass occurs. The best estimate of the parameter is usually determined from the mean of the posterior distribution.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

Example: *known priors* – bird use of wildlife openings

$C_1$ = unoccupied
$C_2$ = present but not breeding
$C_3$ = present & breeding

N = #nearby (<10km) occupied openings

| N | $C_1$ | $C_2$ | $C_3$ | Total |
|---|---|---|---|---|
| 0 | 14 | 1 | 0 | 15 |
| 1 | 9 | 4 | 1 | 14 |
| 2 | 7 | 6 | 3 | 16 |
| 3 | 3 | 6 | 7 | 16 |
| 4 | 0 | 3 | 15 | 18 |
| 5 | 1 | 2 | 18 | 21 |
| Total | 34 | 22 | 44 | 100 |
| Pr($C_i$) | 0.34 | 0.22 | 0.44 | 1 |

| Pr(N\|$C_1$) | Pr(N\|$C_2$) | Pr(N\|$C_3$) |
|---|---|---|
| 0.412 | 0.045 | 0.000 |
| 0.265 | 0.182 | 0.023 |
| 0.206 | 0.273 | 0.068 |
| 0.088 | 0.273 | 0.159 |
| 0.000 | 0.136 | 0.341 |
| 0.029 | 0.091 | 0.409 |
| 1.000 | 1.000 | 1.000 |

What's the probability of a proposed opening remaining unoccupied if it has 3 neighboring occupied openings?

$$\Pr\{C_i|\, N\} = \frac{\Pr\{N|\, C_i\}\Pr\{C_i\}}{\sum_{j=1}^{3}\Pr\{N|\, C_j\}\Pr\{C_j\}}$$

$$\Pr\{C_1|3\} = \frac{0.088 \cdot 0.34}{(0.088 \cdot 0.34) + (0.273 \cdot 0.22) + (0.159 \cdot 0.44)} = 0.19$$

$$\Pr\{C_2|3\} = 0.38 \qquad \Pr\{C_3|3\} = 0.44$$

Example 1: Known priors – bird use of wildlife openings

Let's begin with a simple example where we know the prior probabilities of a set of discrete hypotheses. Let's say we want to know the probability that a newly created wildlife opening will be used as breeding habitat by chestnut-sided warblers in the following classes: $C_1$ = not used, $C_2$ = present, but not breeding, and $C_3$ = present and breeding. This probability depends on the number of wildlife openings nearby (<10 km) that are already used by chestnut-sided warblers (*N*). We have measurements on chestnut-sided warbler use of wildlife openings from the field, and for each opening we know the number of nearby openings that are used. Thus, if we calculate the fraction of openings in each use class $C_i$ with *N* nearby occupied openings, we get an estimate of $\Pr\{N|C_j\}$. We also know the overall fractions of openings in each class, $\Pr\{C_j\}$; in other words, we know the unconditional prior probability of each "hypothesis". If we create a new opening, we know the neighborhood occupancy level *N* from the spatial data on existing openings. Thus, we can use Bayes' Rule to assign the probability of occupancy and breeding to a new opening, as follows:

$$\Pr\{C_i|\, N\} = \frac{\Pr\{N|\, C_i\}\Pr\{C_i\}}{\sum_{j=1}^{3}\Pr\{N|\, C_j\}\Pr\{C_j\}}$$

For example, given a proposed location for a wildlife opening, say we find that there are 3 nearby occupied openings. Given the likelihoods of an opening being in each use class given 3 nearby occupied openings $\Pr\{3|C_j\} = \{0.088, 0.273, 0.159\}$and the overall fraction of openings in each class $\Pr\{C_j\} = \{0.34, 0.22, 0.44\}$, we can calculate the probability of the proposed opening being unoccupied ($C_1$) as follows:

$$\Pr\{C_1|3\} = \frac{0.088 \cdot 0.34}{(0.088 \cdot 0.34) + (0.273 \cdot 0.22) + (0159 \cdot 0.44)} = 0.19$$

The probability of the proposed opening being occupied but not for breeding $\Pr\{C_2|3\} = 0.38$ and occupied for breeding $\Pr\{C_3|3\} = 0.44$. Thus, we expect that proposed opening has a much higher probability of being occupied for breeding than remaining unoccupied. Bayes' Rule allows us to quantify this guess. We could evaluate Bayes' Rule for every potential wildlife opening and use the results to prioritize those with the greatest chance for success.

## Bayesian Parametric Inference...

### Estimate model parameters: Bayes' method

Example: *unknown priors* – reproductive sucess rate

Model: Y ~ binomial($p$, $N$); Prior: $p$ ~ beta($a$, $b$)

Posterior: $$\Pr\{H_i | D\} = \frac{\Pr\{D | H_i\} \Pr\{H_i\}}{\int \Pr\{h\} \Pr\{D | h\} dh}$$

Likelihood of data given 'p' — *Binomial pdf* — *Beta pdf* — Beta prior for 'p'

$$= \frac{\boxed{\left(\prod_{i=1}^{n} \binom{N}{k} p^{k_i} (1-p)^{N-k_i}\right)} \cdot \boxed{\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}\right)}}{\boxed{\int_0^1 \Pr\{Y | p\} \Pr\{p\} dp}}$$

Normalizing constant: *L**prior, integrated over all possible values of 'p'

Example 2: Unknown priors – reproductive success in marbled salamanders

Let's try a slightly more complex example in which we do not know the priors and we have a continuous set of hypotheses. Here, we will focus on the mechanics of Bayesian estimation. In this example, we have observed marbled salamander reproductive success in 13 vernal pools over a period of seven years. We classified a pool as successful in a particular year if ≥1 juvenile salamanders emerged from the pool per breeding female. Pond-years in which there was no breeding effort were not included, thus the number of years in which we recorded success or failure varied among pools for a total sample size of 61. Let's say we are interested in determining the probability of reproductive success among pools and we have prior knowledge about reproductive success that we would like to incorporate into the analysis. Bayes' Rule offers an elegant solution.

The first thing we need to do is specify a model for the data. For simplicity, we will assume a purely stochastic model in which the number of successes is distributed binomially with parameters $p$=constant probability of success and $N$=trial size per pool (number of years with reproductive effort). To complete the Bayesian model we need to specify a prior distribution on the parameter $p$; $N$ is specified by the data. Since $p$ is a proportion, it makes sense to use the beta distribution for the prior on $p$; there are other reasons for using the beta but we will ignore them here.

Given this model, we can use Bayes' Rule to determine the posterior probability distribution; i.e., the probability that $p$ is any particular value given the data and the prior.

# Bayesian Parametric Inference…

## Estimate model parameters: Bayes' method

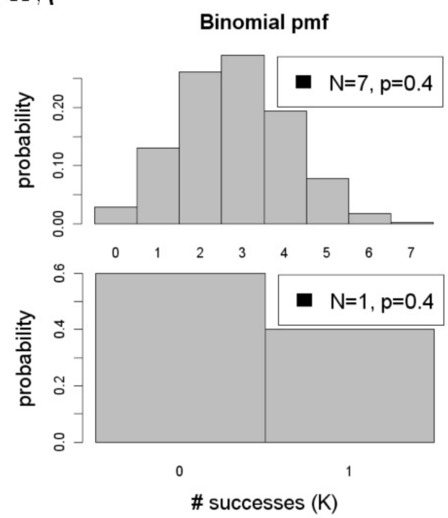
<u>Example</u>: *unknown priors* – reproductive success rate

Compute Likelihood: $L\{D|H_i\} \equiv \Pr\{D|H_i\}$

Trial estimate: $\boxed{p=0.4}$

| Obs | k | N |
|---|---|---|
| 1 | 5 | 7 |
| 2 | 5 | 7 |
| 3 | 6 | 7 |
| … | | |
| 13 | 0 | 1 |

$L(y_1)=0.0774144$
$L(y_2)=0.0774144$
$L(y_3)=0.0172032$
…
$L(y_{13})=0.6$




$L(Y|p=0.4)=$ 6.662139e-11
$LL(Y|p=0.4)=$ -23.432
$NLL(Y|p=0.4)=$ 23.432

Next, we need to compute the likelihood of the data. We do this the same way we did before with maximum likelihood estimation. Briefly, we start with a trial value for the parameter $p$, say 0.4 Then, for each data point we compute its likelihood from the binomial probability mass function. For example, the likelihood of the first point, given $p = 0.4$, is the probability of observing $k = 5$ successes from a binomial distribution with parameters $p = 0.4$ and $N = 7$ years with reproductive effort, which equals 0.0774144.

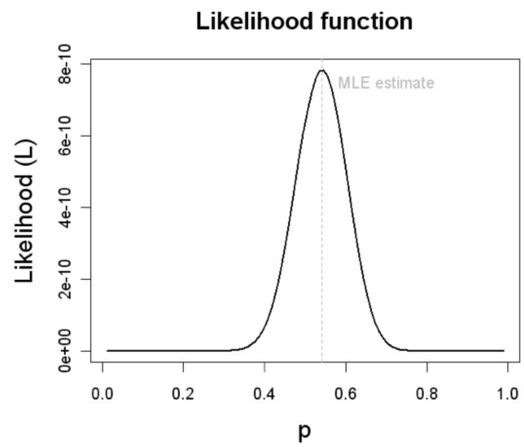
# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

Example: *unknown priors* – reproductive success rate

Likelihood function: $L\{D|H_i\} \equiv \Pr\{D|H_i\}$

$L(Y|p=0.1)$= 7.694657e-26
$L(Y|p=0.2)$= 2.442916e-17
$L(Y|p=0.3)$= 3.759633e-13
$L(Y|p=0.4)$= 6.662139e-11
$L(Y|p=0.5)$= 6.376307e-10
$L(Y|p=0.6)$= 5.059062e-10
$L(Y|p=0.7)$= 2.600336e-11
$L(Y|p=0.8)$= 2.501546e-14
$L(Y|p=0.9)$= 4.543618e-21




We repeat this for each data point and take the product of the likelihoods (or the sum of the log-likelihoods), which equals 6.662139e-11. This gives us the likelihood of the data given a single hypothesized value of *p*. We need to repeat this process for other possible values of *p* – which is the basis for the likelihood function. The likelihood function simply depicts how the likelihood changes as we vary the value of *p*. The peak of this curve is the value of *p* that makes the data the most likely and gives us the maximum likelihood estimate of the parameter. If we were Frequentists, we would be done. However, as Bayesians we are interested in the posterior distribution, which we get by combining the likelihood function with the prior probability distribution.

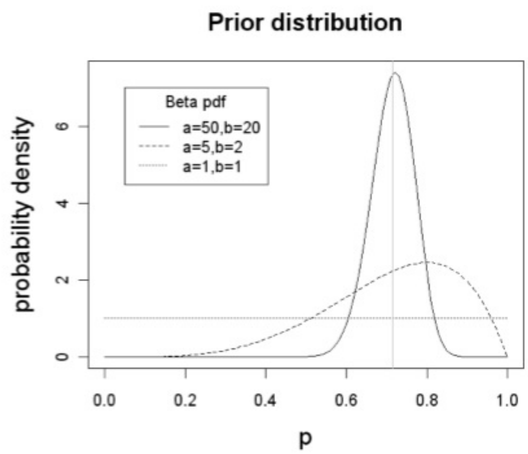
# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

Example: *unknown priors* – reproductive success rate

Prior probability function: $\Pr\{H_i\}$

- *Informative priors...*
  - *Strong...* beta($a$=50, $b$=20); mean=0.71, mode=0.72
  - *Weak...* beta($a$=5, $b$=2); mean=0.71, mode=0.8
- *Uninformative priors...*
  - *Flat...* beta($a$=1, $b$=1); mean=0.5, mode=NA




Next, we need to compute the prior probability of each possible value of the parameter $p$. Here, we have several options. We can choose informative priors in order to incorporate prior knowledge into the analysis. Since the parameter of interest, $p$, is a proportion, a logical choice of probability distributions is the beta, which varies between 0-1. The nice thing about the beta is that it can take on a wide variety of shapes depending on the size and ratio of its two shape parameters, $a$ and $b$. The beta has other advantages as well in this case because it is the *conjugate prior* for the binomial which makes it easy to do the math for Bayesian analysis. Let's assume that we have prior data on the topic. Specifically, let's say that another researcher studying the same phenomenon observed a total of 68 separate breeding efforts, of which 49 were successful and 19 were failures. We can specify this as a relatively strong prior by using a beta with $a$ - 1 successes ($a$ =50) and $b$ - 1 failures ($b$=20), which has a mean probability of success ($p$) of 0.71 and a mode (peak) of 0.72. If, on the other hand, our only prior knowledge came from a pilot study that we conducted in which we observed 4 successes and 1 failures, we can specify this as a relatively weak prior using a beta with beta with $a$ - 1 successes ($a$ =5) and $b$ - 1 failures ($b$=2), which has a mean probability of success ($p$) of 0.71 and a mode (peak) of 0.8. Note here that our two informative priors have the same mean (expected) value of $p$, but the strong prior is based on ten times as many data points as the weak prior. We should probably have more confidence in the strong prior than the weak prior. It should take more convincing in our new data to move away from the strong prior than the weak prior. Alternatively, let's assume that we have no prior knowledge. We can specify a "flat" prior using the beta by setting $a$ and $b$ equal to 1. By doing so, the prior probability of $p$ will be equal to 1 for every value of $p$ between 0-1.

## Bayesian Parametric Inference…

### Estimate model parameters: Bayes' method

Example: *unknown priors* – reproductive success rate

Probability of the data: $\int \Pr\{h\}\,\Pr\{D|\,h\}dh$

- Probability of the data ($D$) is a normalizing constant that makes the posterior distribution a probability distribution
- It is only sometimes useful to be able to express the posterior as a true probability; thus, we can usually ignore the denominator of Bayes' Rule

Let's not worry about the probability of the data!

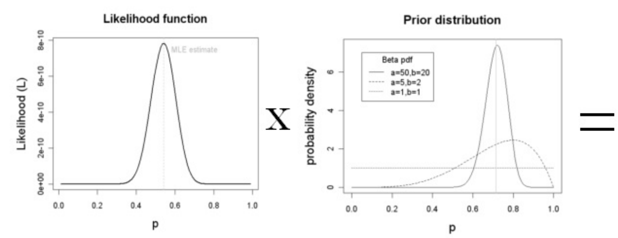
Next, to complete the full Bayesian analysis, we need to compute the probability of the data, the denominator in the Bayes' equation. This is necessary to make the posterior distribution a true probability distribution – area equal to 1. However, it is only sometimes useful to be able to express the posterior as a true probability – our first example on wildlife openings was such a case. More often than not, however, we can ignore the denominator because our interest lies in finding the best estimate of the parameter not in determining its absolute probability. For our purposes, we will simply ignore the denominator. Calculating the denominator will not change the shape of the posterior distribution at all; it will only change the scale of the y-axis. In simple models like the one we are considering here, it is actually quite simple to calculate the denominator using either analytical or numerical integration, but the procedures for doing this go beyond the scope of this chapter.

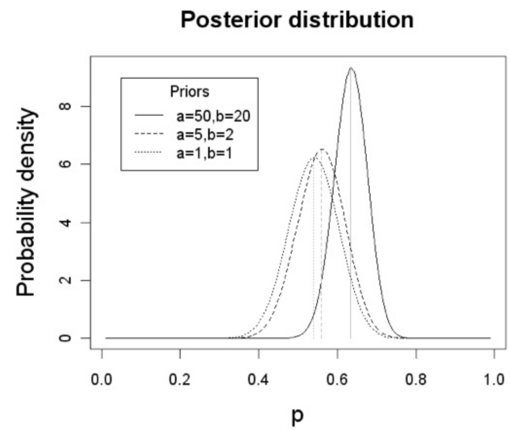
# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

<u>Example</u>: *unknown priors* – reproductive success rate

Posterior: $\Pr\{H_i|\, D\} = \dfrac{\Pr\{D|\, H_i\}\ \Pr\{H_i\}}{\int \Pr\{h\}\ \Pr\{D|\, h\}dh}$




Bayesian estimates:
Strong prior: $p = 0.63$
Weak prior: $p = 0.56$
Flat prior: $p = 0.54$ = MLE estimate

The last step is to compute the posterior distribution by combining the likelihood and the prior. Remember, the posterior is simply the prior distribution modified by the data as expressed through the likelihood. Specifically, we simply multiply the likelihood distribution by the prior distribution to get the unnormalized posterior distribution. The result shown here reveals how the prior distribution is modified by the data under three different priors. If the prior is a strong one and differs from the likelihood distribution, then the posterior distribution will be strongly influenced by the prior or, conversely, the prior will be weakly influence by the data. On the other hand, if the prior differs strongly from the likelihood, but is a weak one reflecting very little data (as is the case with our weak prior, which is based on only 5 data points), then the posterior will largely reflect the likelihood. Of course, if the prior is flat, and thus offers no information, the posterior will look exactly like the likelihood. In our specific example, the Bayesian best estimate of the probability of success – which is based on the mean of the posterior – is 0.63 in the case of the strong prior. In this case, the number of data points in the prior data set ($N_{prior}$=68) is slightly more than in our data ($N_{prior}$=61), so the posterior is weighted more towards the mean of the prior (0.71) than the mean of the data (0.54). In the case of our weak prior representing only 5 data points, the posterior mean of 0.56 largely reflects the mean of the data 0.54.

# Bayesian Parametric Inference...

## Estimate model parameters: Bayes' method

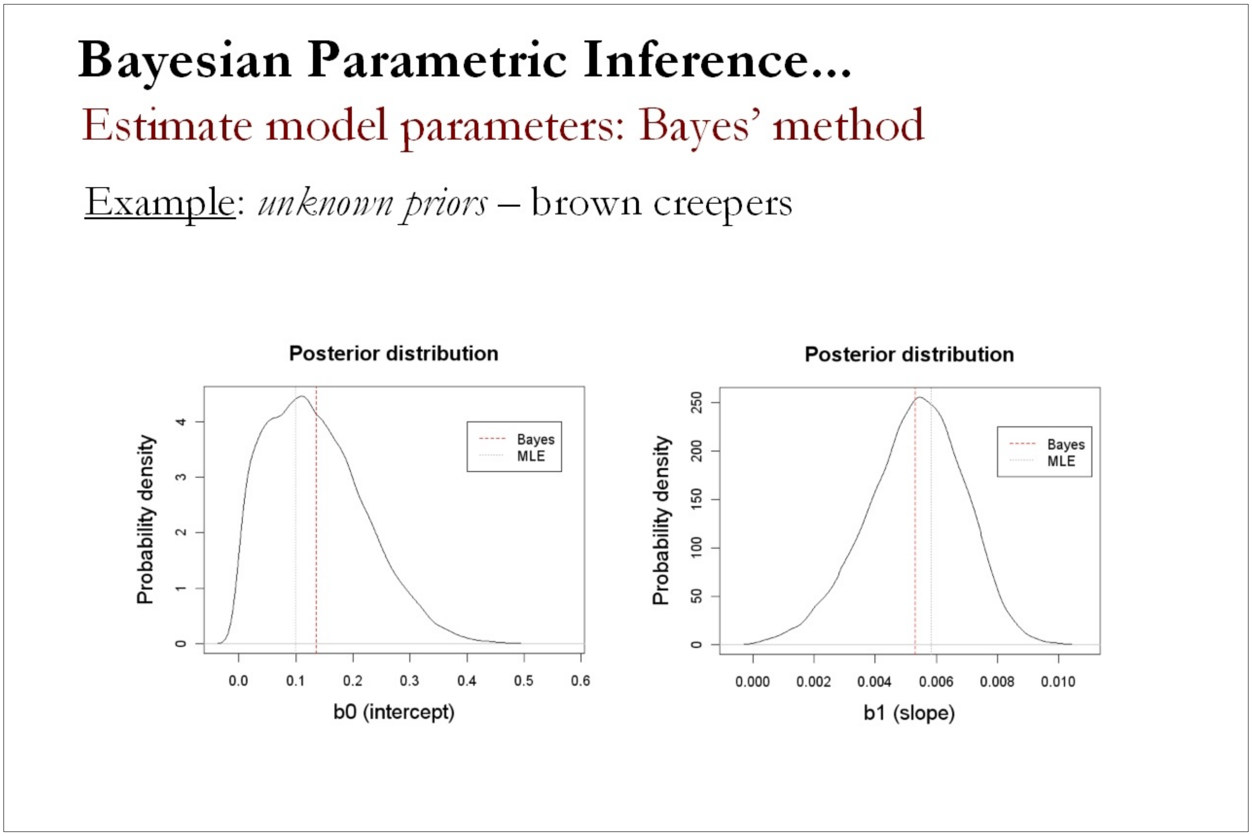
Example: *unknown priors* – brown creepers

Model: Y ~ Normal($a + bx$, $\sigma$);
Priors: $b_0$,$b_1$ ~ beta(1,1); $1/\sigma^2$ ~ gamma(0.01, 1)

Posterior: $$= \frac{\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - b_0 + b_1 x)^2}{2\sigma^2}\right)\right) \cdot \left(\Pr\{b_0, b_1, \sigma\}\right)}{\iiint \Pr\{Y|\, b_0, b_1, \sigma\} \Pr\{b_0, b_1, \sigma\} db_0 db_1 d\sigma}$$

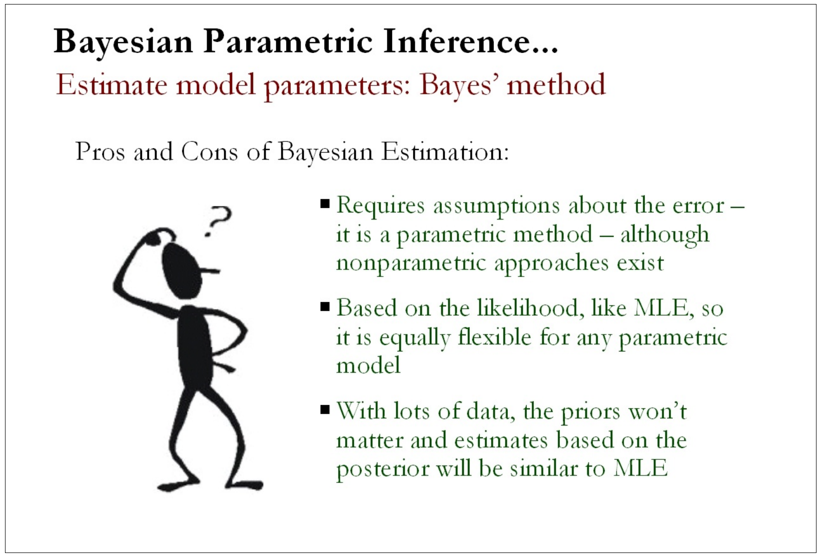
Example: Let's continue with the familiar brown creeper example. Here, we are proposing a linear model for the deterministic component (i.e., a linear relationship between brown creeper abundance and late-successional forest extent) and normally distributed errors for the stochastic component (for simplicity). For now, let's say that prior to our study we were ignorant of the relationship between brown creepers and late-successional forest. As a result, we specify a flat beta prior distribution for the slope and intercept, giving an equal probability to values between 0-1 for both parameters (note, here we are constraining the slope to be positive, which is reasonable in this case based on preliminary examination of the data), and a gamma prior distribution for $1/\sigma^2$. The posterior for this model is partially given as follows:

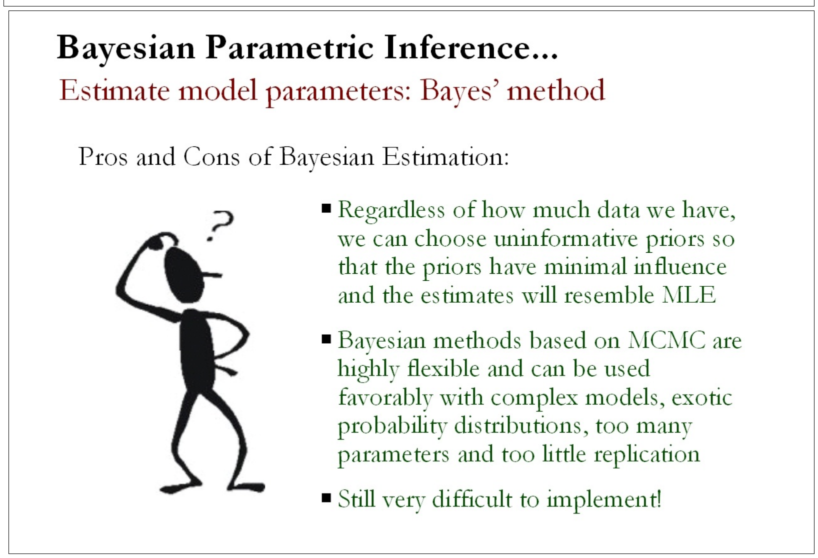
$$= \frac{\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - b_0 + b_1 x)^2}{2\sigma^2}\right)\right) \cdot \left(\Pr\{b_0, b_1, \sigma\}\right)}{\iiint \Pr\{Y|\, b_0, b_1, \sigma\} \Pr\{b_0, b_1, \sigma\} db_0 db_1 d\sigma}$$




The likelihood is the same as before. The priors include distributions for the three model parameters. And the denominator is now a triple integral since we have to integrate the numerator over all three parameters. Believe it or not, an analytical solution is probably available for this problem. Alternatively, we can use numerical methods to figure out the posterior distribution of each parameter. Methods based on Markov Chain Monte Carlo (MCMC) procedures allow us to estimate the posterior distribution by sampling from the posterior throughout the parameter space and eliminate the need to solve the denominator of the Bayes' equation. The details of how MCMC algorithms work is beyond the scope of this chapter. The posterior distributions shown here were obtained using such an approach. One thing to notice is that the Bayesian posterior point estimates of the parameters are usually based on the mean of the posterior distribution, which is often not the same as the peak or mode of the distribution – which is the basis for maximum likelihood methods. Here, the Bayes' estimates and MLE estimates vary slightly because the posterior distributions are slightly skewed.

*Pros and cons of Bayesian estimation*:



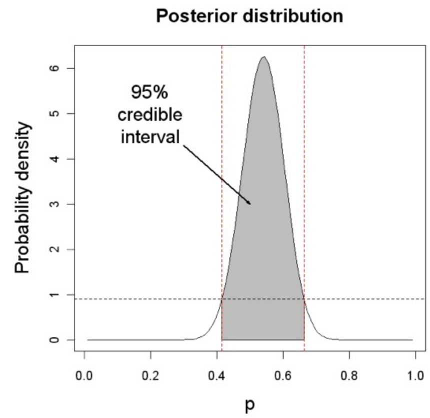

- Bayesian estimation is (typically) a parametric procedure; thus, it requires that we make assumptions about the stochastic component of the model; nonparametric Bayesian procedures exits but are not widely used as of yet.
- Like maximum likelihood estimation, Bayesian analysis also depends on the Likelihood, which can be constructed for virtually any problem.
- With lots of data the choice of prior won't matter much and the parameter estimates based on the posterior distribution will resemble parameter estimates based on maximum likelihood.
- Regardless of how much data we have, we can choose to carry out an "objective" Bayesian analysis by using uninformative (vague, flat, diffuse) priors so that the priors have minimal impact on determining the posterior distribution and the parameter estimates will resemble those based on maximum likelihood.
- Bayesian estimation based on MCMC methods can be used to obtain parameter estimates for distributions when it is far too complicated to obtain these estimates using ordinary frequentist optimization methods; for example, when we are dealing with complex models with many random effects, exotic probability distributions, too many parameters and too little replication.
- However, compared to maximum likelihood estimation conducting Bayesian estimation is extremely complicated and requires considerable technical expertise, even for relatively simple models. Thus, it remains to be seen whether Bayesian approaches will ever become mainstream for ecologists.

## Bayesian Parametric Inference...

Credible intervals for model parameters

Posterior distributions:

- *Credible intervals* are defined as the region in the center of the posterior distribution containing x% of the probability of the distribution
- Bayesian credible interval is symmetrical in *height*, whereas frequentist confidence interval is symmetrical in *area*




## 3. Credible intervals

Rarely are we interested in just point estimates of the parameters, which are generally useless without measures of uncertainty. We really want to know the uncertainty associated with the parameter estimates. Instead of drawing likelihood profiles for parameters, Bayesians draw the posterior distribution (proportional to likelihood x prior). Instead of calculating confidence limits using the (frequentist) Likelihood Ratio Test (LRT), they define the *credible interval*, which is the region in the center of the posterior distribution containing 95% (or some other standard proportion) of the probability of the distribution, bounded by values on either side that have the same probability (or probability density). Technically, the credible interval is the interval $[x_1, x_2]$ such that $Pr(x_1) = Pr(x_2)$ and $C(x_1) - C(x_2) = 1 - \alpha$, where Pr() is the probability (density) and C is the cumulative density. Note, the credible interval is slightly different from the frequentist confidence interval, which is defined as the interval $[x_1, x_2]$ such that $C(x_1) = \alpha/2$ and $C(x_2) = 1- \alpha/2$. Note the difference between a credible interval and a confidence interval. The credible interval is symmetrical in height; the cutoff value on either end of the distribution has the same posterior probability (density). The confidence interval is symmetrical in area; the likelihood of more extreme values in either direction is the same. The credible interval's height symmetry leads to a uniform probability cutoff; we never include a less probable value on one boundary than on the other. The confidence interval's area symmetry leads to a potentially uneven probability cutoff, so long as the cumulative probability of a more extreme value (i.e., area of the tail) is the same. To a Bayesian, holding the probability the same on both boundaries makes more sense than insisting (as the Frequentists do in defining confidence intervals) that the cumulative probabilities of more extreme values in either direction are the same.

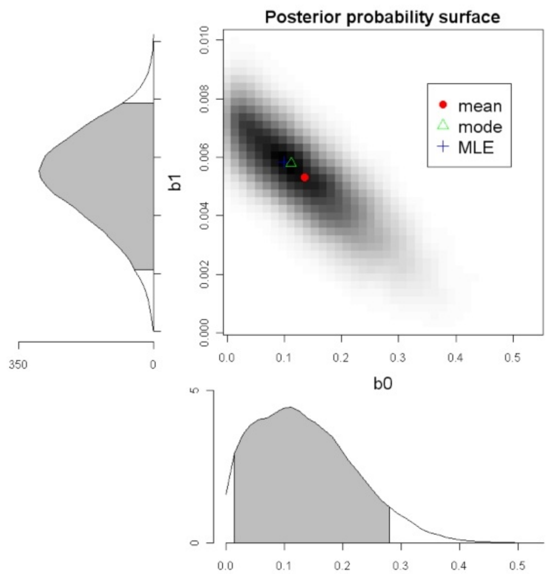
Example: Continuing with the salamander reproductive success example, the Bayesian 95% credible interval for the parameter $p$ (the probability of reproductive success) is 0.416-0.663.

# Bayesian Parametric Inference...

## Credible intervals for model parameters

Marginal probability density:

- For multiparameter models, the *posterior probability surface* is analogous to the likelihood surface
- The *marginal probability density* is the Baysian analogue to the likelihood profile; the probability for a value of a focal parameter integrated over all other parameters




For multiparameter models, the likelihood surface is analogous to a bivariate or multivariate probability distribution, which for two dimensions (parameters) can be depicted as a posterior probability surface. The *marginal probability density* is the Bayesian analogue of the likelihood profile. Where Frequentists use likelihood profiles to make inferences about a single parameter while taking the effects of the other parameters into account, Bayesians use the marginal posterior probability (density), the overall probability for a particular value of a focal parameter integrated over all the other parameters.

Example: Continuing with the brown creeper example, the posterior probabilities for the two key parameters, the slope $b_1$ and intercept $b_0$ are plotted here on a posterior probability surface. The marginal probability distributions depict the probability density for each parameter integrated over all other parameters. Thus, the marginal distribution for $b_1$ is simply the probability density evaluated over varying values of $b_1$. On the two-parameter posterior probability surface shown here, this is equivalent to finding the area under the surface along a horizontal belt transect and expressing this as a proportion of the area under the entire surface. In practice this entails taking a horizontal slice of the surface for each value of $b_1$ (i.e., summing the values of the posterior surface along the horizontal transect) and dividing it by the area under the entire surface.

## Bayesian Parametric Inference...

### Hypothesis testing

- Bayesians generally have no interest in hypothesis testing per se!
- However, procedures do exist for hypothesis testing based on the posterior probabilities of alternative models
- Bayesian methods are probably better suited for parameter estimation than hypothesis testing

## 4. Model comparison

Bayesians generally have no interest in hypothesis testing (although it can be done in the Bayesian framework) and little interest in formal methods of model selection. Dropping a variable from a model is often equivalent to testing a null hypothesis that the parameter is exactly zero, and Bayesians consider such *point* null hypotheses silly. They would describe a parameter's distribution as being concentrated near zero rather than saying its value is exactly zero. Nevertheless, Bayesians do compute the relative probability of different models, in a way that implicitly recognizes the bias-variance trade-off and penalizes more complex models. Bayesians prefer to make inferences based on averages rather than on the most likely values; for example, they generally use the posterior mean values of parameters rather than the posterior mode. This preference extends to model selection.

# Bayesian Parametric Inference...

## Model comparison

Marginal likelihood and Bayes factor:

- *Marginal likelihood...* the probability of observing the data (likelihood), averaged over the prior distribution of the parameters

$$\bar{L} = \int L(\varphi) \cdot prior(\varphi) d\varphi$$

$$B_{12} = \frac{\bar{L}_1}{\bar{L}_2}$$

- *Bayes factor...* the ratio of two marginal likelihoods, or the odds in favor of model 1 (numerator model), generally assuming equal prior probabilities for both models

Rule of thumb:

| $2\log B_{12}$ | Evidence in favor of 1 |
|---|---|
| 0-2 | weak |
| 2-6 | positive |
| 6-10 | strong |
| >10 | very strong |

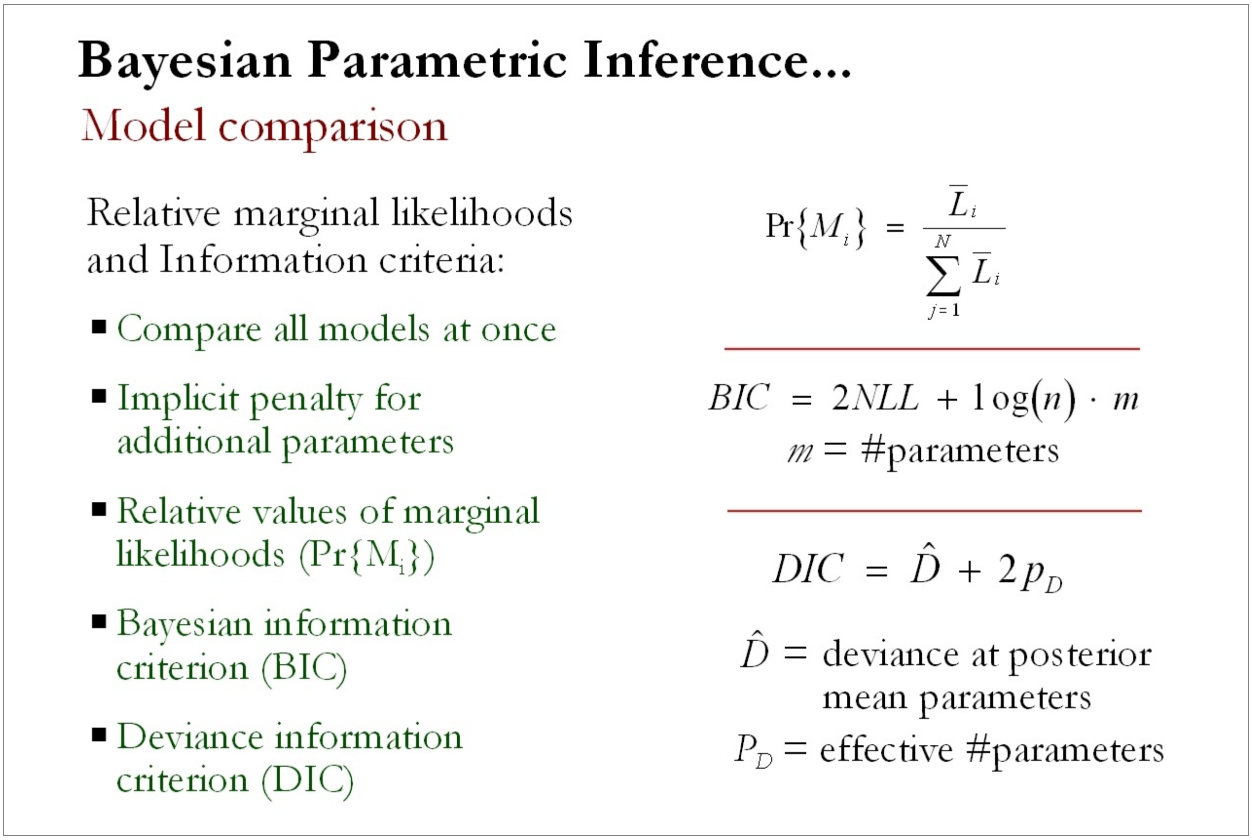
The *marginal likelihood* of a model is the probability of observing the data (likelihood), averaged over the prior distribution of the parameters:

$$\bar{L} = \int L(\varphi) \cdot prior(\varphi) d\varphi$$

where $\varphi$ represents a parameter or set of parameters (if a set, then the integral would be a multiple integral). The marginal likelihood (the average probability of observing a particular data set exactly) is often very small, and we are really interested in the relative probability of different models. The *Bayes factor* is the ratio of two marginal likelihoods, or the odds in favor of model 1 (numerator model), generally assuming equal prior probabilities for both models:

$$B_{12} = \frac{\bar{L}_1}{\bar{L}_2}$$

In principle, using Bayes factors to select the better of two models is simple. If we compare twice the logarithm of the Bayes factors (thus putting them on the deviance scale), the generally accepted rules of thumb for Bayes factors are similar to those for AIC, where 0-2 is weak evidence in favor of model 1, 2-6 is positive evidence in favor of model 1, 6-10 strong evidence in favor of model 1, and >10 very strong evidence in favor of model 1.




If we want to compare several different (not necessarily) nested models, we can look at the pairwise Bayes factors or compute a set of posterior probabilities for the set of models – assuming that all of the models have the same prior probability – by computing the relative values of the marginal likelihoods:

$$\Pr\{M_i\} = \frac{\bar{L}_i}{\sum_{j=1}^{N} \bar{L}_i}$$

Bayes factors and marginal likelihoods incorporate an implicit penalty for overparameterization. When you add more parameters to a model, it can fit better – the maximum likelihood and the maximum posterior probability increase – but at the same time the posterior probability distribution spreads out to cover more less-well-fitting possibilities. Since marginal likelihoods express the mean and not the maximum posterior probability, they will actually decrease when the model becomes too complex. Unfortunately, computing the marginal likelihood for a particular model can be tricky, involving either complicated multidimensional integrals or some kind of stochastic sampling from the prior distribution. There are approximation methods however, but we will not describe them here. A simpler alternative to the Bayes factor and marginal likelihoods is the Schwarz information criterion (or Bayes Information Criterion, BIC), which approximates the log of the Bayes factor and is easy to calculate:

$$BIC = 2NLL + 1\text{og}(n) \cdot m$$

where $n$ is the number of observations and $m$ is the number of parameters. When $n$ is greater than $e^2 \approx 8$ observations (so that log $n > 2$), the BIC is more conservative than the AIC, insisting on a greater improvement in fit before it will accept a more complex model. Note, while the BIC is derived from a Bayesian argument, it is not inherently a Bayesian technique. Nevertheless, it is commonly used and reported as such.

A more recent criterion is the DIC, or deviance information criterion, which is similar in concept to AIC – penalized goodness-of-fit – only the goodness-of-fit metric is based on deviance instead of the negative log-likelihood:

$$DIC = \hat{D} + 2p_D$$

where $D$ hat is the deviance calculated at the posterior mean parameters and $p_D$ is an estimate of the effective number of parameters.

# Bayesian Parametric Inference...

## Model comparison

Information criteria:

- Not suitable for significance testing
- Rule of thumb:
  - ΔBIC/DIC <5: equivalent
  - ΔBIC/DIC 5-10: distinguishable
  - ΔBIC/DIC >10: definitely different
- BIC model weight: relative likelihood of a model – probability of the model – given the data

Alternative models:

BRCR = ls
BRCR = ls + p.contag
BRCR = ls + p.contag + s.sidi

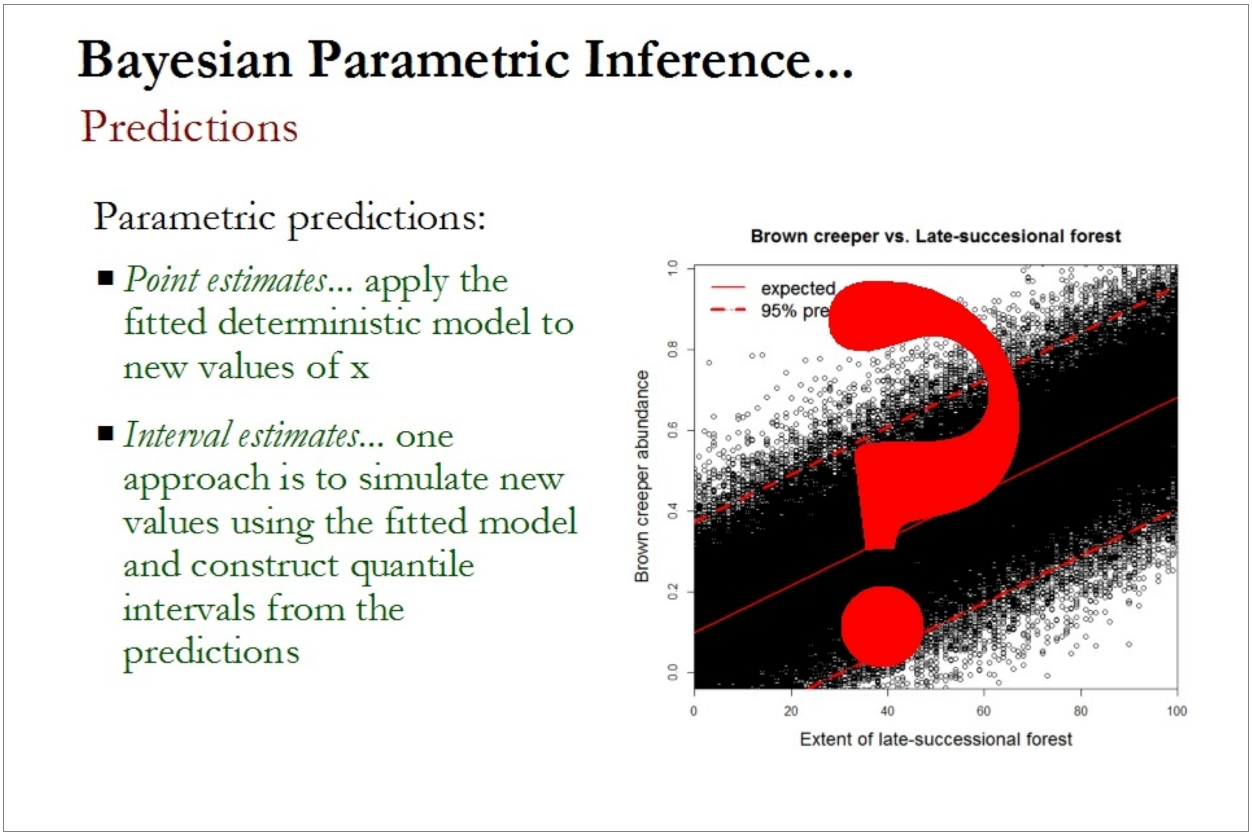
BIC/DIC Table:

| | dBIC | df | weight | DIC |
|---|---|---|---|---|
| fit1 | 0.0 | 3 | 0.524 | -5.444 |
| fit3 | 1.0 | 5 | 0.314 | -3.616 |
| fit2 | 2.3 | 4 | 0.162 | -3.614 |

The information criteria such as BIC and DIC are useful for comparing multiple models all at once and are not suitable for hypotheses testing – which is in line with Bayesian philosophy. Like all information criteria, e.g., AIC, BIC and DIC have a general rule of thumb for their interpretation. Differences <5 are considered equivalent, 5-10 distinguishable, and >10 definitely different. Thus, they provide a quick and useful way to compare the weight of evidence in favor of alternative models. Indeed, model weights can be calculated from the IC values which give the relative likelihood or probability of a model given the data.

Unfortunately, delving into these measures further is beyond the scope of this chapter and beyond my complete understanding. Suffice it to say that there are several methods for comparing models in a Bayesian framework, but the tradeoffs among approaches is still being hotly debated.

<u>Example</u>: Continuing with the brown creeper example, recall the three alternative models we considered in the maximum likelihood setting. Recall that based on AIC, model 3 was selected as the “best” model based, with model 1 and 2 close behind. BIC gives model 1 the most weight with model 3 and 2 close behind. DIC gives a similar ranking. Not surprisingly, both BIC and DIC were more conservative than AIC and chose model 1 over the more complex model 3.




## 5. Predictions

The final goal of statistical inference is to make predictions. In many cases, once we confirm that our model is a good one, say by confirming that it is significantly better than the null model (e.g., of no relationship between x and y) or that it is the best among competing models considered, we might want to use this model to predict values for future observations or for sites not sampled.

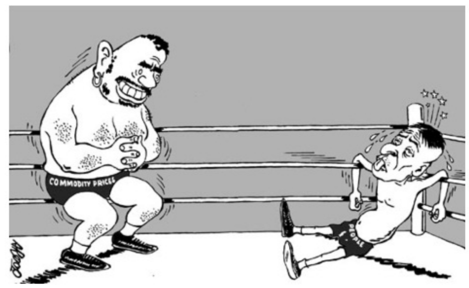
Unfortunately, I know next to nothing about model predictions in a Bayesian framework. Prediction is certainly doable in the Bayesian framework, but the methods are not familiar to me. So we will skip this topic, but recognize that prediction is possible in a Bayesian framework.

## Bayesian Parametric Inference...

### Pros and cons of Bayesian inference

- Philisophical distinction may be as important as methodological uniqueness
- Methods for estimation are still too complex and technically challenging for mainstream ecology
- Current buzz in ecological modeling – will it last?

$Y \sim \mathrm{Normal}(a + bx, \sigma^2)$



## 6. Pros and cons of Bayesian likelihood inference

The frequentist Bayesian inference framework is hugely appealing in many respects but it is not without some serious drawbacks. We already touched base on many of the pros and cons of Bayesian inference in the section on estimation. Here we will finish with a few additional overarching thoughts.

1. *Philosophy*... Bayesian inference is as much a philosophy as a method. Its distinctiveness arises as much from its philosophical underpinnings as the uniqueness of the methods themselves.
2. *Technical challenges*... Despite the many conceptual advantages of Bayesian statistics, the methods for estimation are largely still too complex and technically challenging for mainstream ecology. It remains to be seen whether enough progress is made in facilitating Bayesian analysis to make it accessible to vast major of ecologists. Until then, Baysian statistics will likely remain out of reach for most individuals not willing or able to delve into the gory details of Bayesian anlaytical methods.
3. *Modern statistical inference*... Without question Bayesian statistics is the current buzz in ecological modeling and gaining rapidly in popularity over maximum likelihood. However, for the reasons mentioned above it remains to be seen whether it will end up being a temporary fad or persist.