Introduction to Bayesian analysis

Marc Kéry

USGS Patuxent Wildlife Research Center,
Laurel/MD, 8–12 July 2013

Outline of talk

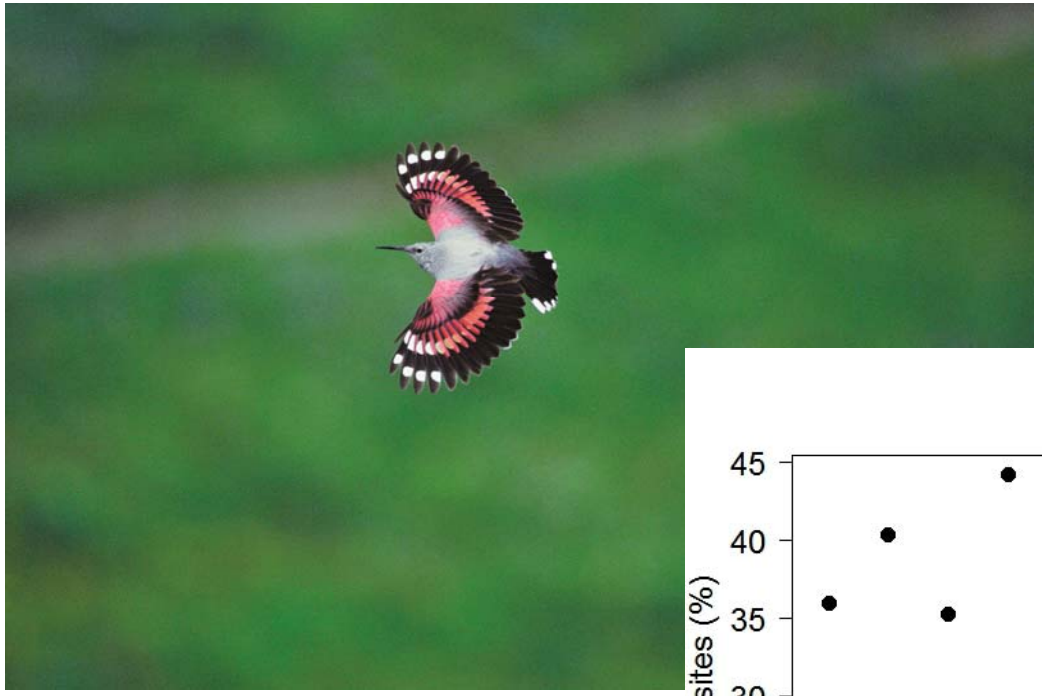
- Intro: What's the fuss ?
- Role of models in science
- Analysis of models: frequentist and Bayesian
- Bayesian computation
- WinBUGS
- Hierarchical models
- Concluding remarks on Bayesian/frequentist choice
- BUGS frees the hierarchical modeler in you

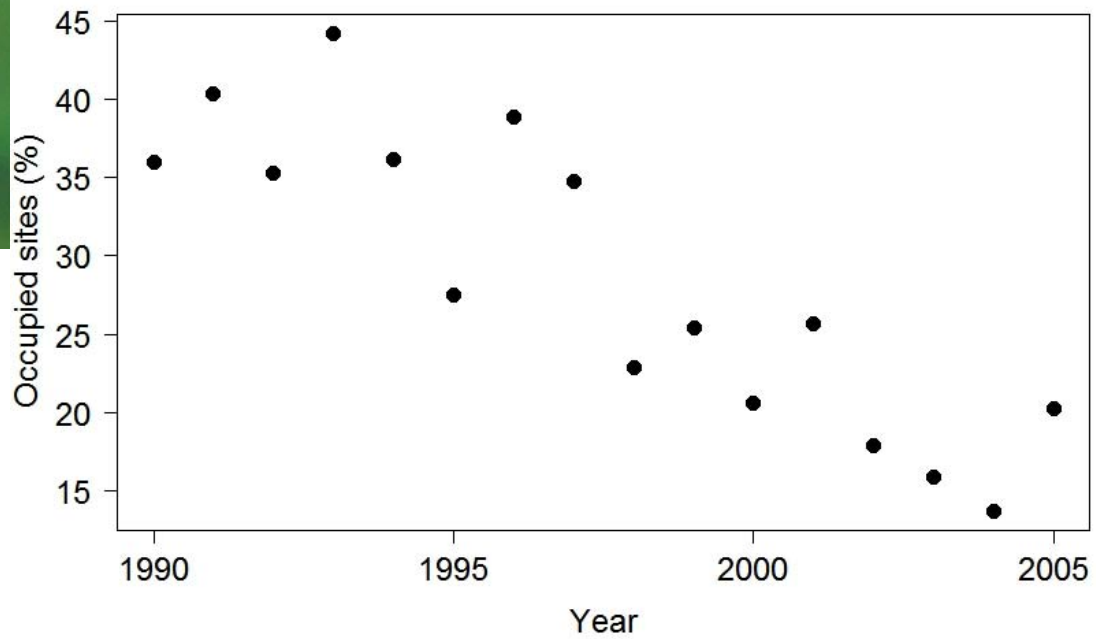
What's the fuss ?

- A simple example

What's the fuss ?

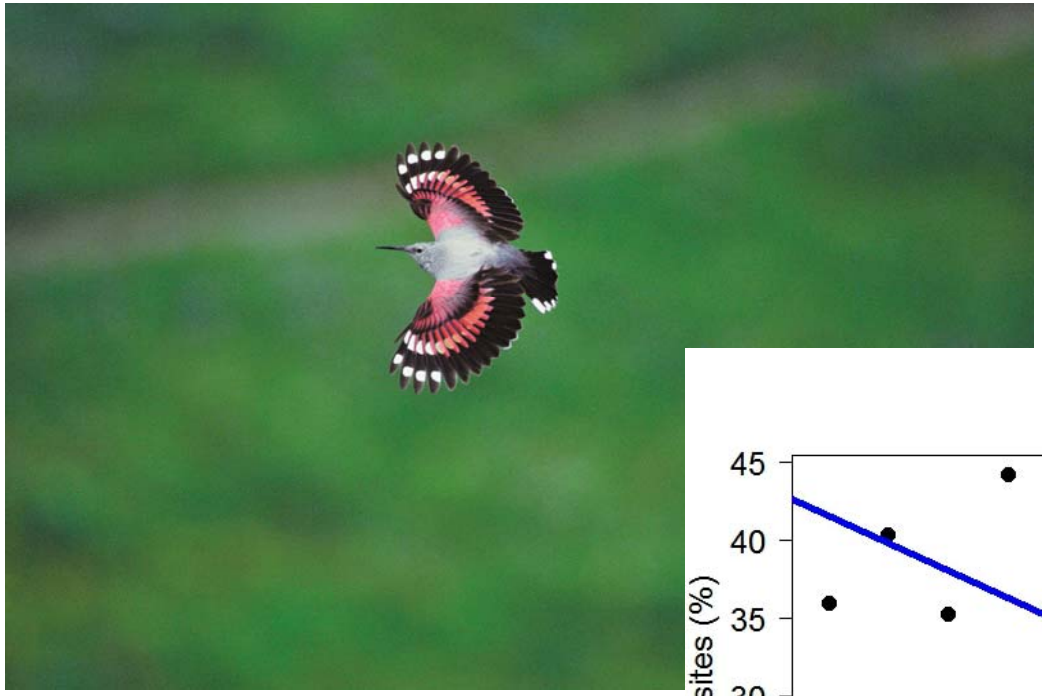
- A simple example





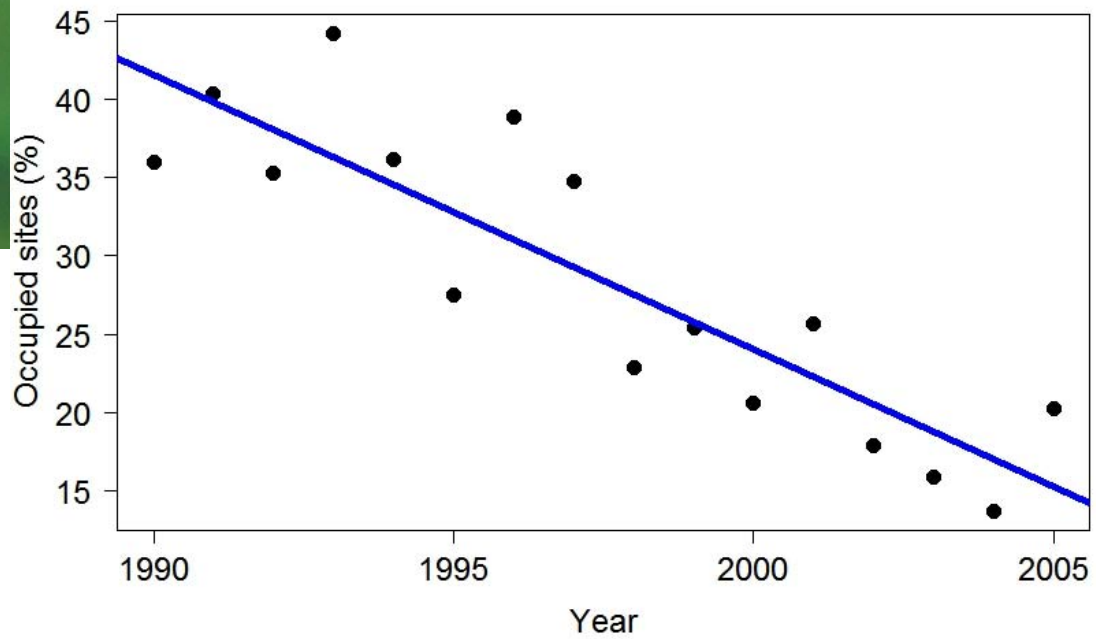
What's the fuss ?

- A simple example



Trend estimate

b = -1.754

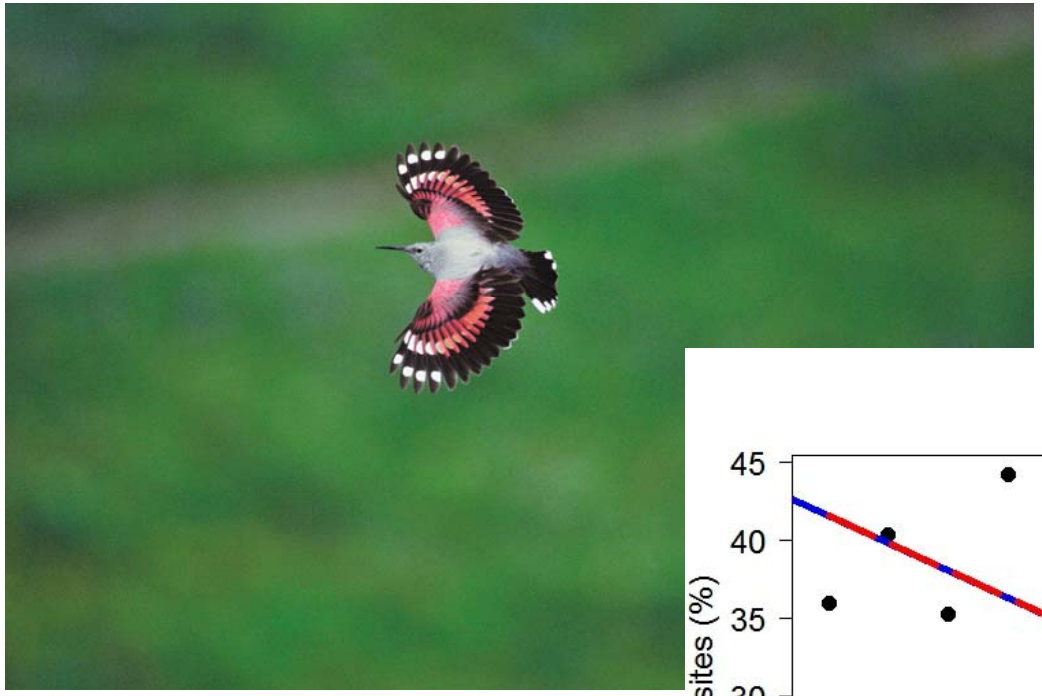


$$y = a + b * X + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

What's the fuss ?

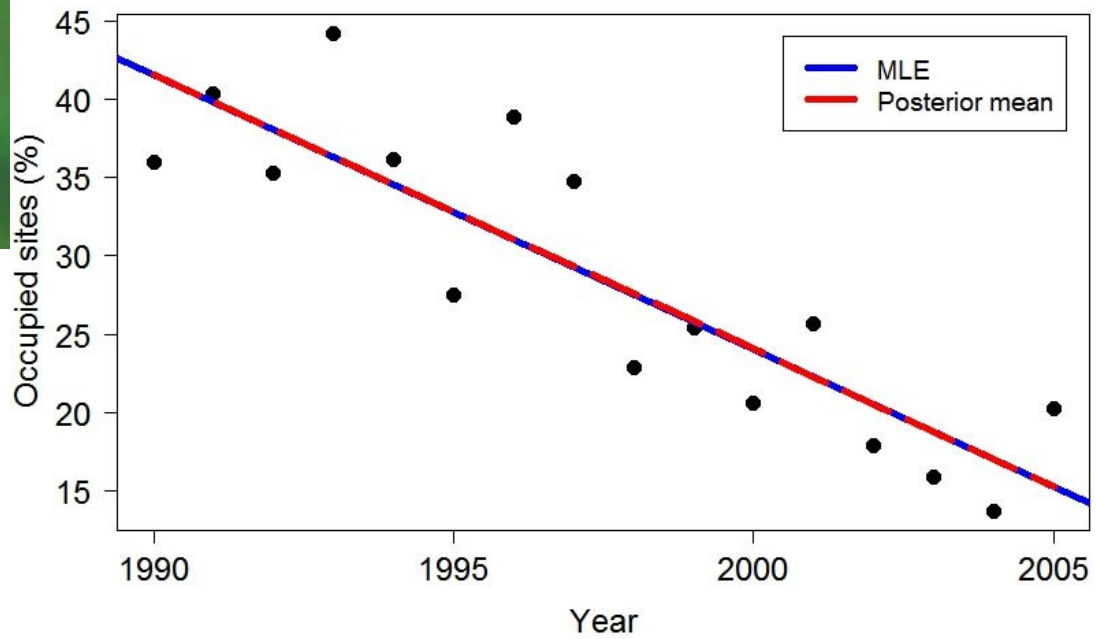
- A simple example



Trend estimate

b = -1.754

b = -1.756



$$y = a + b * X + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

What's the fuss ?

- Statistical models exist independently from method of statistical analysis !
- There are no "Bayesian models" or "frequentist models"
- May choose to analyse a model (e.g., linear regression) in Bayesian way
- Typically, Bayesian and frequentist analyses yield numerically very similar estimates

Role of models in science

- Science: explain nature, so you can better **understand** and/or **predict**
- Management (e.g., conservation): … so you can better manage Nature
- Nature too complex to understand
- Must reduce complexity
- A model (broadly): greatly simplified version of nature, should help understand/predict
- Every model has an objective:

  e.g. understanding ≈ mechanism

  e.g. predicting ≈ description

Everybody is a modeler !

- Model = set of assumptions
- Description of model: words, graphs, algebra, ...
- Any explanation is based on a model, stated or unstated

To make sense of an observation,
everybody needs a model ...
Whether he knows it or not !

- Interpretation of data without a model is impossible
- [or is it ? ..... what about data mining ?]
- Explicit models are better than implicit models (e.g., assumptions more transparent, can test them)

Mathematical and statistical models

- Mathematical models: written in algebra, e.g.,

$$y = \alpha + \beta * x$$

- Advantage: clarity greatly increased over description in words
- Algebraic model descriptions enforce clarity of thought

Mathematical and statistical models

- Mathematical models: written in algebra, e.g.,

$$y = \alpha + \beta * x$$

- Advantage: clarity greatly increased over description in words
- Algebraic model descriptions enforce clarity of thought
- Statistical models: acknowledge stochasticity in systems, e.g.

$$y = \alpha + \beta * x + \varepsilon$$

$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$

Statistical models

- Statistical models:

$$y = \alpha + \beta * x + \varepsilon$$

$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$

Response = systematic part + random part

- Parametric statistical model: Description of the processes thought to have produced y

Statistical models

- Statistical models:

$$y = \alpha + \beta * x + \varepsilon$$

$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$

Response = systematic part + random part

- Parametric statistical model: Description of the processes thought to have produced y
- Generalized linear model (GLM):
  quintessential statistical model

Statistical models

Three most frequent GLMs:

- Normal response:

  Random part: $y \sim \text{Normal}(\mu, \sigma^2)$

  Systematic part: $\mu = \alpha + \beta * x$

- Poisson response:

  Random part: $y \sim \text{Poisson}(\lambda)$

  Systematic part: $\log(\lambda) = \alpha + \beta * x$

- Binomial response:

  Random part: $y \sim \text{Binomial}(p, N)$

  Systematic part: $\text{logit}(p) = \alpha + \beta * x$

Analysis of a statistical model

- Sketch of a model

$$x \longrightarrow \boxed{\theta} \longrightarrow y$$

- Data viewed as result of random process(es)
- Input x, output y, parameters θ
- Parameters (θ) fixed and **unknown** constants
- How should we guess at value(s) of θ ?
- ... at missing covariates (x) ? ... at missing response (y) ?

--> Statisticians devise many procedures for guessing, e.g.,

- method of moments
- least-squares
- maximum likelihood, maximum partial likelihood, pseudo-likelihood, penalized likelihood, ...
- Bayesian analysis
- ...

Frequentist analysis of a model

- Example: Estimate probability of detection (θ) of tadpoles

  -> Release n=50 in artificial pond, later resight y=20

Frequentist analysis of a model

<u>(One) Frequentist way of guessing at θ: maximum likelihood</u>

- Parametric model describes data-generating probabilistic mechanism: sampling distribution p(y|θ)
- "*probability of observing data y, given fixed param. value θ*"
- **Note:** probability statement about the data, **not** about parameter *θ*
- Probability defined as long-run frequency in hypothetical replicate data sets
- E.g., binomial sampling distribution:

$$p(y|\theta) = \frac{n!}{y!(n-y)!}\theta^{y}(1-\theta)^{n-y}$$

Frequentist analysis of a model

Maximum likelihood

- **Idea:** good choice of θ is that which maximises function value of sampling distribution for my data set

$$p\left(20\middle|\theta\right)=\frac{50!}{20!\left(50-20\right)!}\theta^{20}\left(1-\theta\right)^{50-20}$$

- **Likelihood function:** read sampling distribution "in reverse", i.e., as a function of θ

$$L(\theta\,|\,y)=\frac{n!}{y!\left(n-y\right)!}\theta^{y}\left(1-\theta\right)^{n-y}$$

- Call maximiser of L the Maximum Likelihood estimate (MLE)
- MLE makes actual, observed data most probable

Frequentist analysis of a model

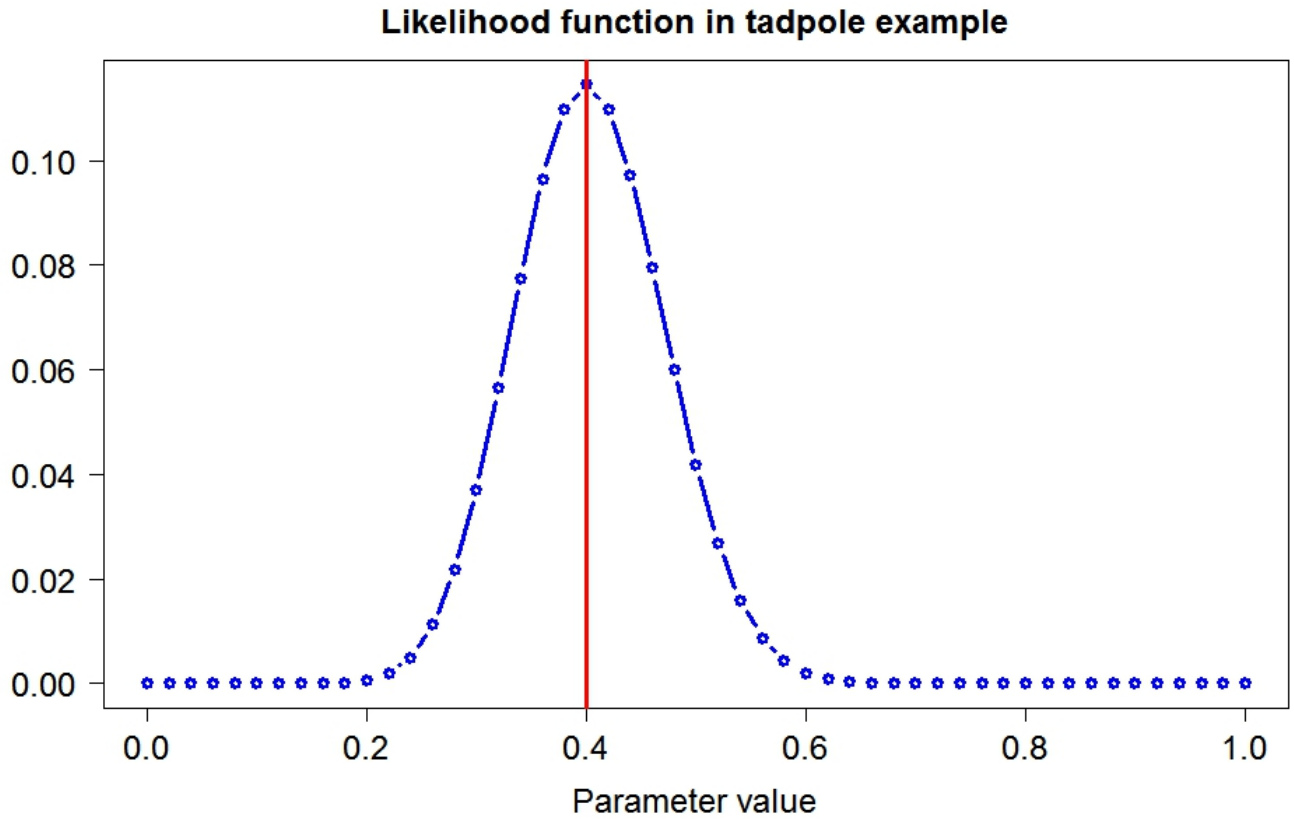
<u>Maximum likelihood</u>

- How to find the MLE ?
  - Analytically (sometimes)
  - Numerically (most of the times)

Frequentist analysis of a model

Maximum likelihood

- Numerical estimation by brute force:
  try out and plot large number of values for θ



vogelwarte

Frequentist analysis of a model

Maximum likelihood

- Numerical estimation by function minimisation: e.g. `optim()` in R (also `nlm()` and others)

```
> # Define the data
> r <- 20
> N <- 50
>
> # Define negative log-likelihood function
> nll <- function(p) -dbinom(r, size = N, prob = p, log = TRUE)
>
> # Minimize function for observed data and return MLE
> fit <- optim(par = 0.5, fn = nll, method = "BFGS")

Maximum likelihood estimate of p:  0.4000000

>
> fit
$par
[1] 0.4000000

$value
[1] 2.166669
```

Frequentist analysis of a model

Maximum likelihood

- Numerical estimation using special functions: R `glm()`

```
> # Estimate parameter on link scale
> fm <- glm(cbind(20,30) ~ 1, family = binomial)
> summary(fm)

Call:
glm(formula = cbind(20, 30) ~ 1, family = binomial)

Deviance Residuals:
[1]  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4055     0.2887  -1.405     0.16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 0  degrees of freedom
Residual deviance: 4.4409e-15  on 0  degrees of freedom
AIC: 6.3333

Number of Fisher Scoring iterations: 2
```

Frequentist analysis of a model

Some characteristics of maximum likelihood

- Long history (Fisher, 1920s)
- Much theory, well studied and understood
- "Automatic inference": simply define likelihood function and then find parameter values that maximise it
- Produces "good estimates", e.g., asymptotically unbiased, consistent, transformation invariant
- "Gold standard" in statistics
- Most statistical modeling in ecology is based on MLE

Frequentist analysis of a model

BUT:

- MLEs can be hard or impossible for complex models
- SEs and CIs asymptotic (valid for infinite sample size), unknown how good for *your* ecological data set
- Functions of parameters difficult to obtain, i.e., error propagation can be hard
- "Indirect" probability statements about data, rather than about params: $p(y|\theta)$
- 95% CI does *not* contain θ with $P=0.95$
- Impossible in principle to say things like "*I am 95% certain that this population is declining*"
- Appeal to large number of hypothetical replicate data unsatisfactory in many practical cases: e.g., what does a "replicate population of the Florida panther" mean ?

Bayesian analysis of a model

- Sketch of model

$$x \longrightarrow \boxed{\theta} \longrightarrow y$$

- Data viewed as result of random process(es)
- Input x, output y, parameters θ
- Parameters (θ) fixed and **unknown** constants
- How should we guess at values of θ ? … or missing x ?
  … or predict y ?

Bayesian analysis of a model

- Sketch of model

$$x \longrightarrow \boxed{\theta} \longrightarrow y$$

- Data viewed as result of random process(es)
- Input x, output y, parameters θ
- Parameters (θ) fixed and **unknown** constants
- How should we guess at values of θ ? … or missing x ? … or predict y ?
- **Bayesian approach:** in the face of uncertainty about magnitude of θ use conditional probability, $p(\theta|y)$
- "Guess" at θ conditions on what is *certain* or what we *know* (i.e., data x and y)

Bayesian analysis of a model

Recipe of every Bayesian analysis:

1. What is known ? The data ($y=20$, $n=50$)
2. What is unknown ? Prob. of detection (θ)
3. What to do ? Calculate $p(\theta|y)$

   "*Prob. of parameter, given data*"

- **Note:** probability statement about the parameter
- Data, once collected, are fixed
- Degree-of-belief concept of probability:
  Express imperfect knowledge (about θ) using probability distribution
- Hence, parameters treated **as if** they were random variables
- How should $p(\theta|y)$ be computed ?

Bayesian analysis of a model

- Bayes rule

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

- Mathematical fact of probability
- E.g., can be deduced from p(A,B) = p(B | A) * p(A)
  (joint prob. = conditional prob. * marginal/unconditional prob.)
- Can be applied in non-Bayesian probability calculations for observable quantities, e.g., clinical testing

Bayesian analysis of a model

- Example: football and birdwatching (from Pigliucci)

	Good weather (g)	Bad weather (b)	
Go birdwatching (B)	**0.5**		**0.7**
Watch football (F)			
	0.6		

- What is p(b|F) ?

Bayesian analysis of a model

- Example: football and birdwatching (from Pigliucci)

	Good weather (g)	Bad weather (b)	
Go birdwatching (B)	**0.5**	0.2	**0.7**
Watch football (F)	0.1	0.2	0.3
	0.6	0.4	1.0

- What is p(b|F) ?

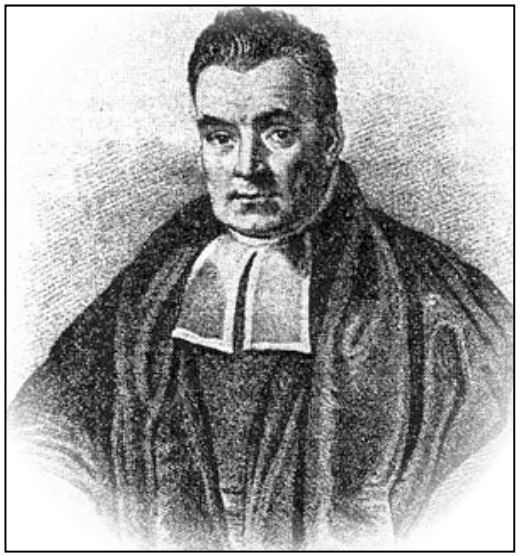
Bayesian analysis of a model

- Bayes rule

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

- Thomas Bayes, English minister/mathematician (1702-1761)

- Thomas Bayes applied the rule to unobservables such as parameters, i.e., for parameter estimation



Bayesian analysis of a model

Bayes rule for statistical inference:

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

- Posterior distribution: $p(\theta \mid y)$
- Likelihood function: $p(y \mid \theta)$
- Prior distribution: $p(\theta)$
- Prob. of data: $p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta$

- **NOTE:** Use probability to express imperfect knowledge
- Direct probability statements about unknown quantites:
  Can say "*... I am 95% certain that prob of detection > 0.2*" !

Bayesian analysis of a model

Formal steps in Bayesian analysis

- Use probability as a universal measure of uncertainty about unknown quantities (here: θ)
- Treat all statisical inference (estimation, testing, ...) as a simple probability calculation
- Express your knowledge about parameter θ (excluding information contained in y) by a probability distribution: the prior p(θ)
- Use Bayes rule to *update* that knowledge with the information contained in the data y and embodied by the likelihood function, p(y|θ)
- Result is probability distribution, p(θ|y), for every unknown
- Unlike ML, where result is single value

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

vogelwarte.ch

Bayesian analysis of a model

<u>Heuristic appeal of Bayes rule as model for inference</u>

- "Human" concept of probability ("*I am 95% certain that ...*")
- $p(\theta|y) \propto p(y|\theta) \times p(\theta)$
- "Posterior = Likelihood x prior"
- Like human learning:
  - Conclusion is combination of experience and new information (e.g., 3 m tall man)
  - New information changes ("updates") my previous state of knowledge to my current state of knowledge
  - Every analysis could be a meta-analysis: synthesizes *all* existing knowledge

Bayesian analysis of a model

Advantage of prior distribution:

- Bayesian inference allows formal incorporation of external knowledge into estimation via prior distribution
- Strength of Bayesian analysis !
- E.g., small sample sizes (ecology of rare species)
- Advantage of 'informative priors':
  - Don't feign to be stupid
  - More precise estimates
  - Can estimate additional parameters

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

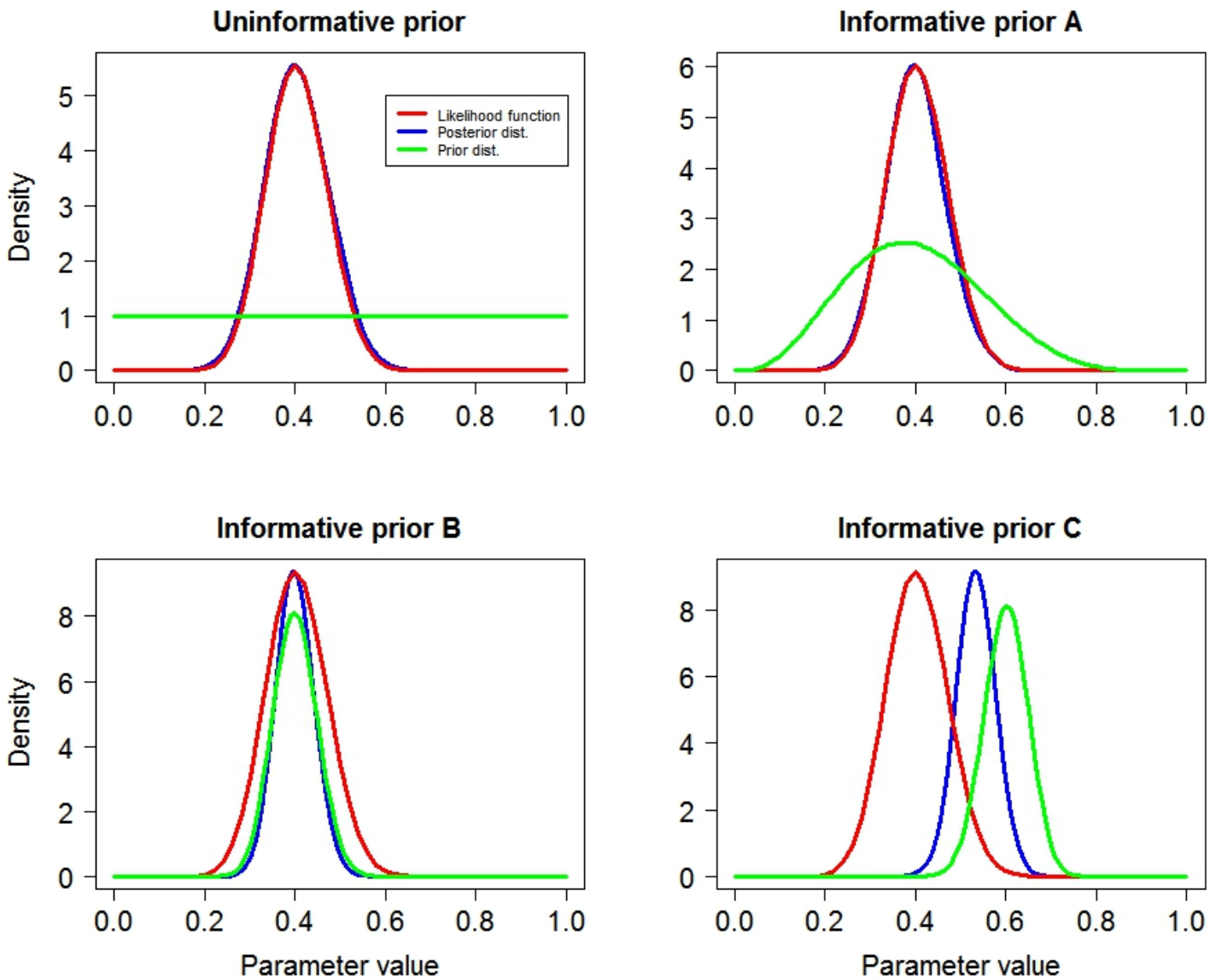
Bayesian analysis of a model

Disadvantage of prior distribution (?):

- 'Results' (i.e., estimates) always depend on priors !
- Have to choose priors --> analysis 'subjective'
- But can specify 'non-informative' (vague etc.) priors
- (though may be difficult to specify "non-information")
- Must report priors for every analysis
- Justify choice of informative priors
- Here (as Royle & Dorazio 2008): specify vague priors, typically on "natural" scale
- Estimates then (very much) resemble MLEs

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

Graphical illustration of 4 Bayesian analyses of tadpole Ex.



vogelwart

Bayesian computation

- So why has not everyone always been a Bayesian ?

  --> Bayes rule was hard to apply in practice

- Denominator: n-dimensional integral for a model with n parameters

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

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta$$

- Integrals impossible to compute for most realistic models
- For centuries, Bayesian analysis of complex models not possible

Bayesian computation

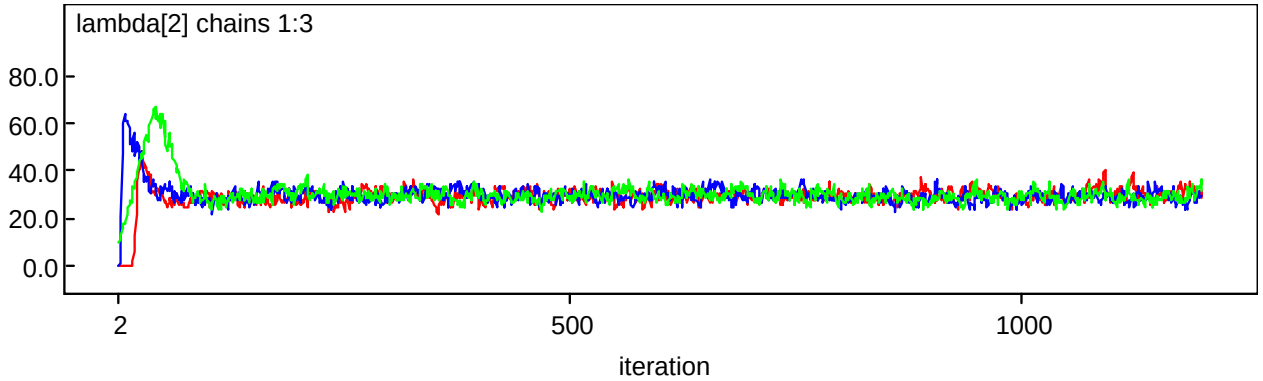
- Early 1990s: statisticians rediscover work from the 1950's in physics

  --> Use stochastic simulation to draw dependent samples from posterior distribution
- Don't actually evaluate integrals in Bayes rule
- Instead, approximate posterior to arbitrary degree of accuracy by drawing large sample
- Markov chain Monte Carlo (MCMC)
- Boost to Bayesian statistics in statistics community

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

Bayesian computation

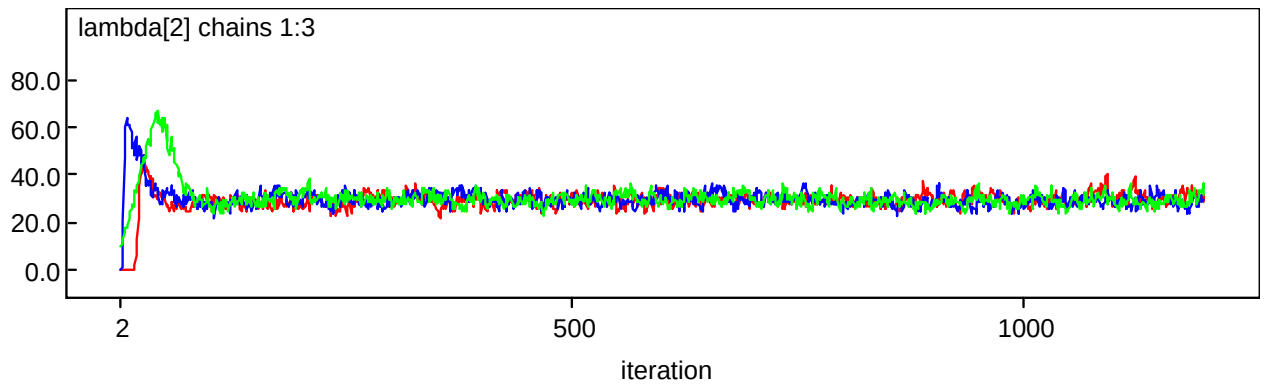
- MCMC: Stochastic algorithm to produce sequence of dependent random numbers (= Markov chain)
- Converge to equilibrium distribution (usually)
- Equilibrium distribution = desired posterior distribution (if algorithm constructed well)



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

Bayesian computation

- When is equilibrium attained ?
- Run multiple chains from arbitrary starting places (inits)
- Assume convergence when all cover same ground
- Discard initial 'burn-in' phase
- Summarize remainder (mean: point estimate; sd: analogue of SE)



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

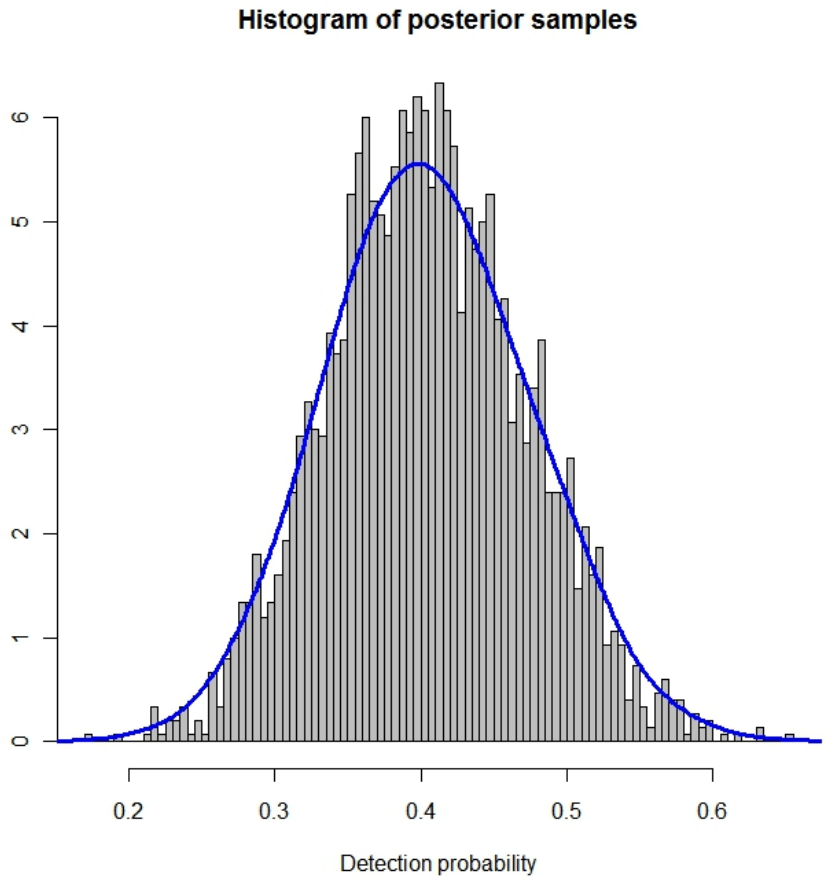
Bayesian computation

```
> p
   [1] 0.5265 0.4088 0.3885 0.3482 0.3850 0.3311
   [7] 0.4042 0.3593 0.3580 0.3880 0.3688 0.3793
  [13] 0.4935 0.2831 0.4827 0.4632 0.3765 0.4186
  [19] 0.4579 0.3605 0.4488 0.3914 0.3474 0.4444
    ...
[2983] 0.3866 0.3265 0.3121 0.2337 0.3255 0.3912
[2989] 0.3446 0.3584 0.3839 0.4920 0.4068 0.3202
[2995] 0.3844 0.5067 0.4212 0.5759 0.2485 0.2362
```

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

Bayesian computation

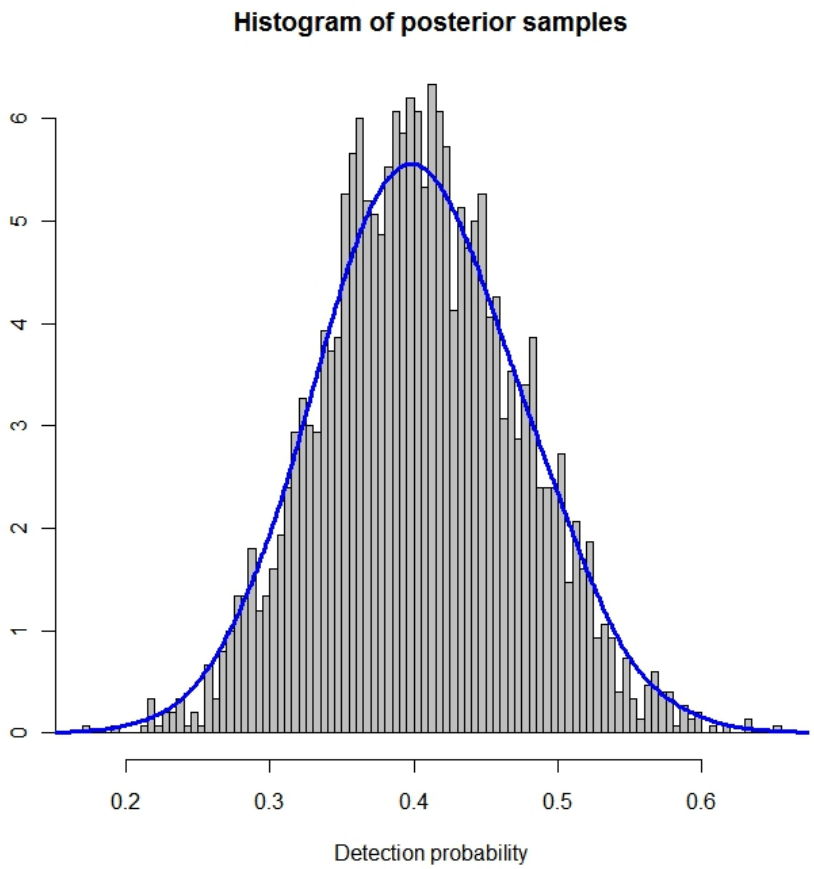
```
> p
   [1] 0.5265 0.4088 0.3885 0.3482
   [7] 0.4042 0.3593 0.3580 0.3880
  [13] 0.4935 0.2831 0.4827 0.4632
  [19] 0.4579 0.3605 0.4488 0.3914
    ...
[2983] 0.3866 0.3265 0.3121 0.2337
[2989] 0.3446 0.3584 0.3839 0.4920
[2995] 0.3844 0.5067 0.4212 0.5759
```



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

Bayesian computation

```
> p
   [1] 0.5265 0.4088 0.3885 0.3482
   [7] 0.4042 0.3593 0.3580 0.3880
  [13] 0.4935 0.2831 0.4827 0.4632
  [19] 0.4579 0.3605 0.4488 0.3914
    ...
[2983] 0.3866 0.3265 0.3121 0.2337
[2989] 0.3446 0.3584 0.3839 0.4920
[2995] 0.3844 0.5067 0.4212 0.5759
```



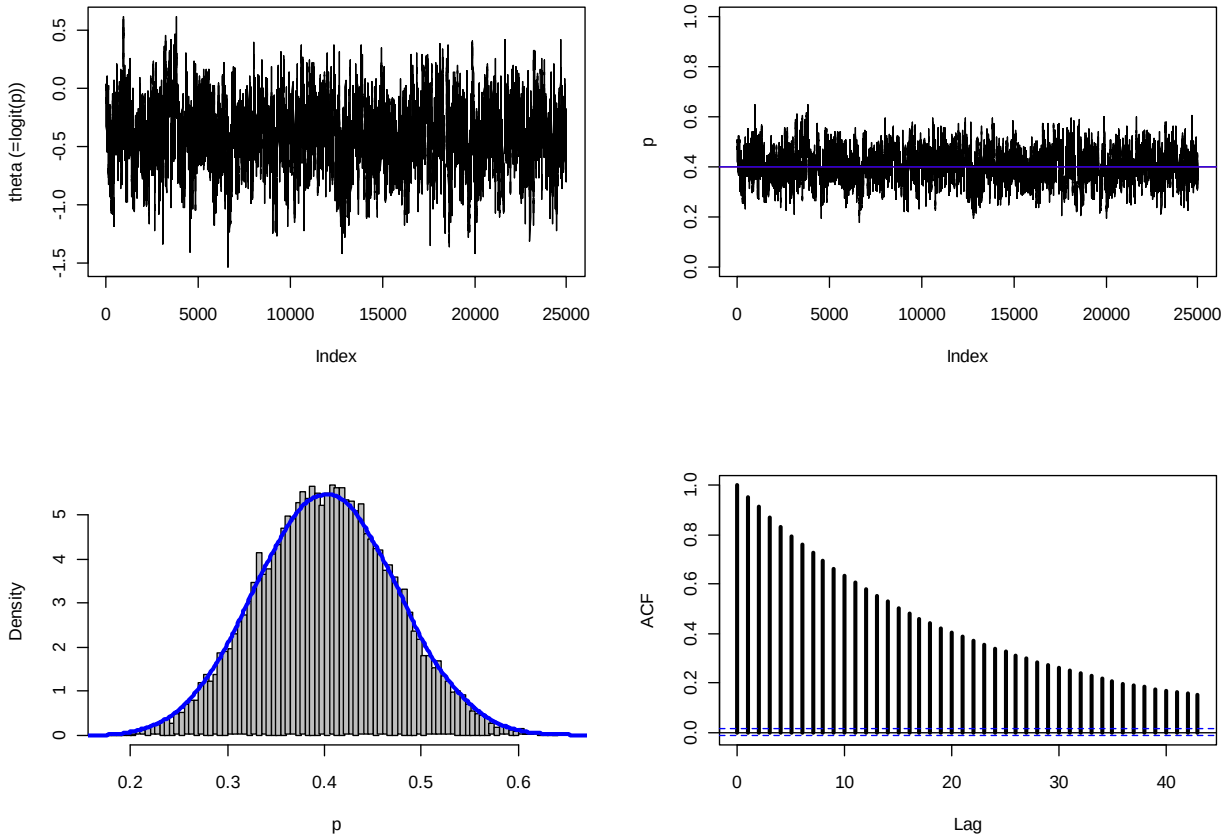

```
> mean(p)
[1] 0.4047
> sd(p)
[1] 0.0674
> quantile(p, probs = c(0.025, 0.975))
     2.5%     97.5%
   0.2771    0.5375
```

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

vogelwarte.ch

Bayesian computation

- Custom MCMC code for binomial proportion (tadpoles)



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

vogelwarte.ch

The BUGS project

- Boost in Bayesian statistics initially *not in ecology*
- To code MCMC algorithms, need to know something about statistics and especially about computing
- Change due to BUGS project:
  Bayesian inference **u**sing **G**ibbs **s**ampling
- Gibbs sampling: variant of MCMC
- Statisticians/Epidemiologists in Cambridge/UK
- Lunn et al. (2009), *Statistics in Medicine,* 3049–3067

The BUGS project

- BUGS: Flexible, generic software, does:
  1. Simple and intuitive model description language (BUGS language)
  2. Automatic development of MCMC algorithms (algorithmic black box)
  3. Run algorithm: produce posterior samples

- Three variants:
  - **WinBUGS:** www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml
  - **OpenBUGS:** www.openbugs.info/w/
  - **JAGS:** mcmc-jags.sourceforge.net/

The BUGS language

- Simple and intuitive model description language
- Implicit description of likelihood of model by nested sequence of simple *probability statements* and *deterministic relationships* between quantities
- *Unexpected side-effect:* BUGS language great to *really* understand GLMs, random-effects/mixed models
- BUGS is **not** a black box in terms of the model fitted !
- Rather:

  One of the most transparent ways of building a model is by describing it in the BUGS language.

Hierarchical models (HMs)

- BUGS particularly good (natural) for hierarchical models
- HM: Nested sequence of observed and unobserved r.v.s:

$$x \sim f(\omega)$$
$$y \sim g(x, \theta)$$

- Factorization of joint distribution [x,y] to marginal ([x]) * conditional distribution ([y|x])
- Flexible modeling of hidden structure and correlations
- Latent effects, random effects, **mixed models** …
- Can describe a large class of models as HM
- E.g., site-occupancy model:

$$z_i \sim Bern(\psi)$$
$$y_{ij} \sim Bern(z_i \times p_{ij})$$

Why we have become Bayesians

Why we have become Bayesians

… and why you might want to become one, too !

(Quote from Bill Link)

Why we have become Bayesians

3 types of advantages of Bayesian analysis by MCMC in BUGS:

(1) Bayesian paradigm:

- 'Natural' use of probability
- Formal introduction of prior information possible

Why we have become Bayesians

3 types of advantages of Bayesian analysis by MCMC in BUGS:

(1) Bayesian paradigm:

- 'Natural' use of probability
- Formal introduction of prior information possible

(2) Bayesian computation (MCMC):

- Easy to fit HMs
- Trivial to compute functions of parameters
  (with exact uncertainty intervals: error propagation)

Why we have become Bayesians

3 types of advantages of Bayesian analysis by MCMC in BUGS:

(1) Bayesian paradigm:

- 'Natural' use of probability
- Formal introduction of prior information possible

(2) Bayesian computation (MCMC):

- Easy to fit HMs
- Trivial to compute functions of parameters
  (with exact uncertainty intervals: error propagation)

(3) BUGS language and software (WinBUGS, OpenBUGS, JAGS):

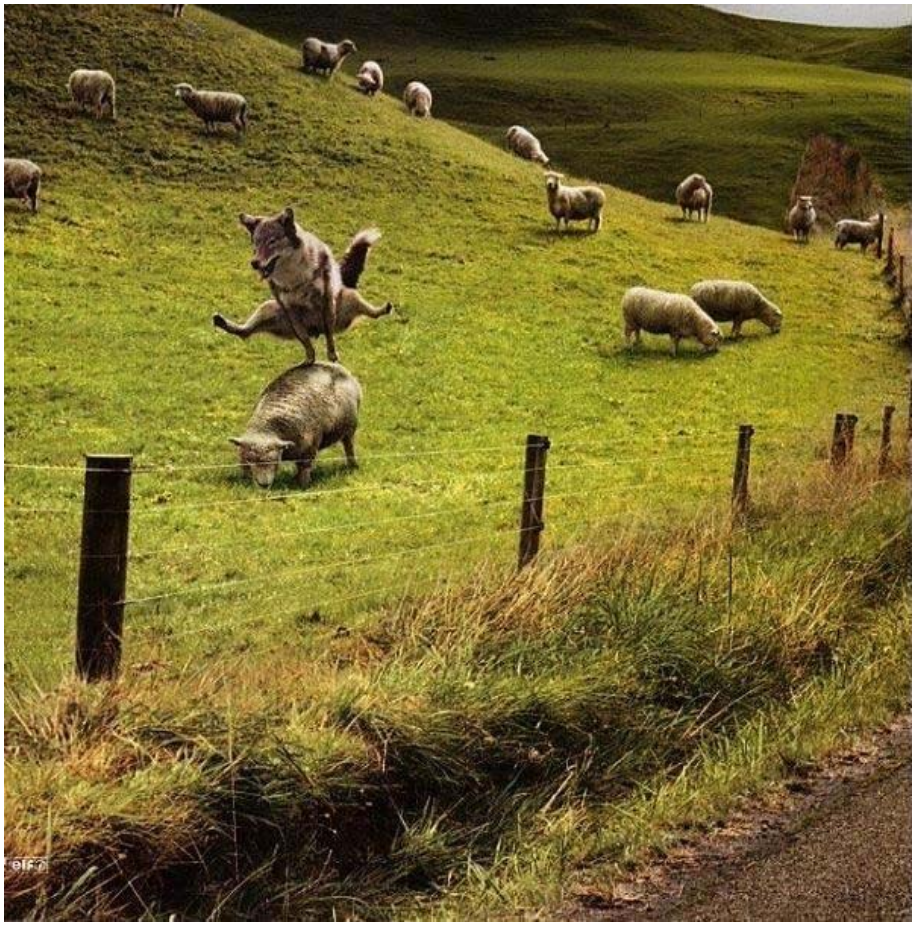
- Implementation of complex, custom models
  within reach of ecologists
- Enforces understanding of model
- **WinBUGS frees the modeler in you !**

Why we are not real Bayesians

- Seldom use informative priors
- Plus, some inconveniences of Bayesian analysis in BUGS:
  - Take long time to run (often less for ML)
  - Model selection is a pain (cf. AIC with ML)
  - Sensitivity of results to prior choice (not with ML)
  - Harder to explain
  - BUGS so flexible that may fit nonsensical models
- Hence, happy to use maximum likelihood as well

Conclusion on the Bayesian / frequentist choice

- Be eclectic !
- Choose what is most useful for *you*
- Usually will not use BUGS for trivial problems
- BUGS is fantastic for more complex models (except for large data sets !)
- BUGS language is great to actually understand your models
- Stay tuned: in the future, there will (hopefully !) be better MCMC and even likelihood software for complex models, e.g., HMs



Should you learn to write your own MCMC samplers ?

- Claims made by one anonymous statistician:

  "should discontinue use of BUGS software"

  "only writing own sampler allows you to understand what's going on"

  "writing own samplers is really easy"

- Two answers:
  - emphatic **no** !
  - cautious **yes**

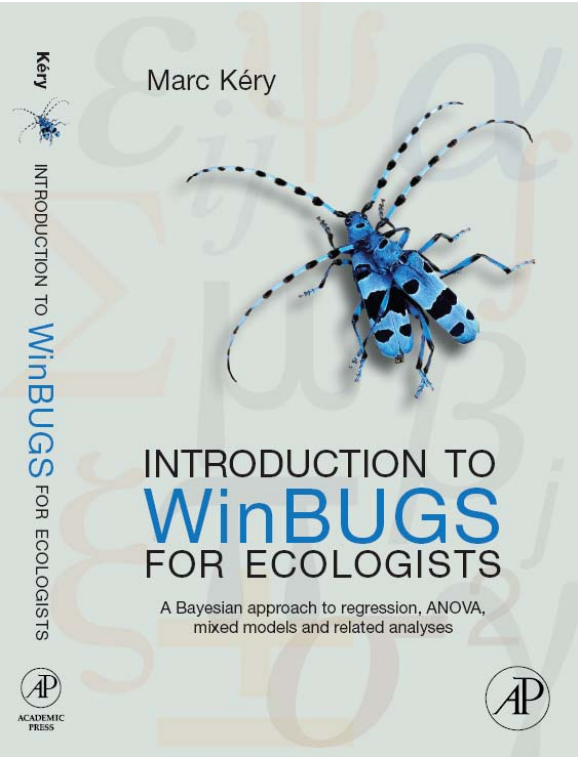
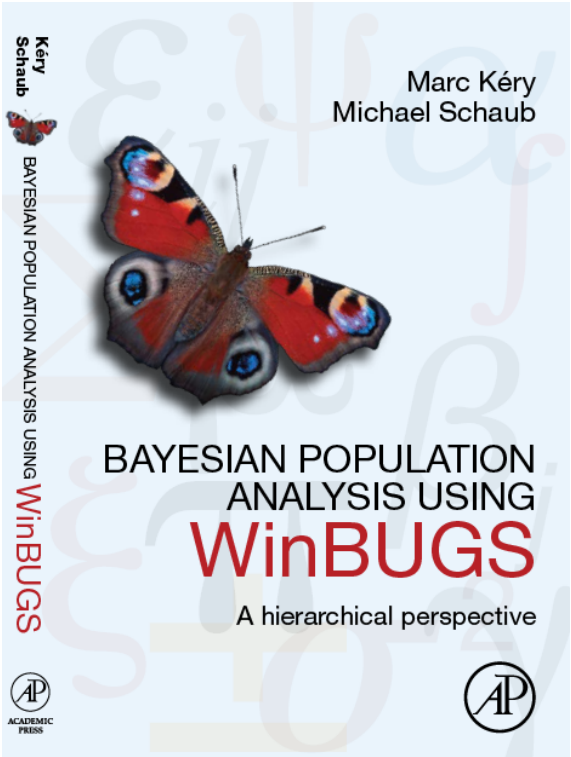
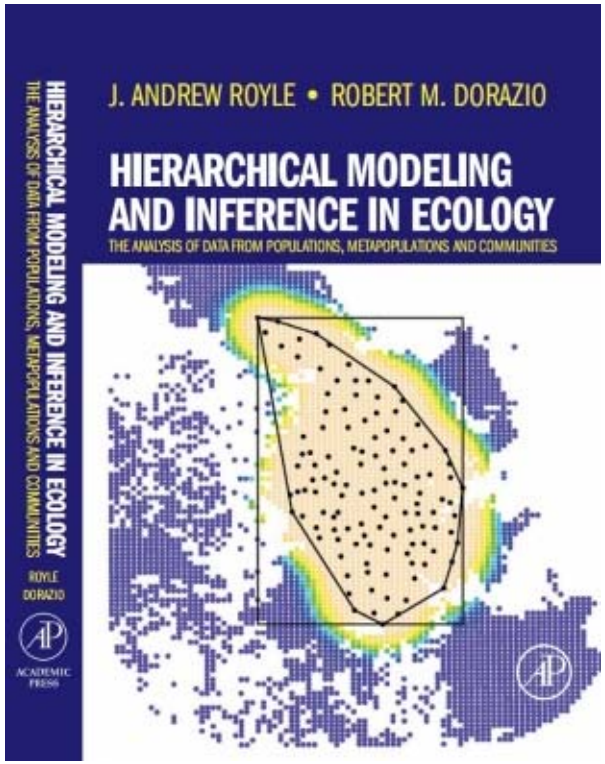
NO: you should NOT learn to write your own MCMC samplers

- CLEARLY, writing own samplers is NOT always easy, because otherwise no statistician would EVER use BUGS
- ... is not easy for everybody
- ... is error-prone
- ... is time-consuming
- ... nobody codes up his own Newton-Raphson algorithm for GLM or Laplace approximation for mixed models
- **... is a pure waste of time !**
- ... takes away time that could be used for exploring more models and better understand them, write more or better papers etc. ....
- BUGS *lingua franca* of Bayesian modeling: BUGS code usable and adaptable for many more people than MCMC code

YES: you should learn to write your own MCMC samplers

- BUGS can be painfully slow
- even hacked MCMC algorithms can be much faster
- writing MCMC *can* give you much more insights and prevent errors of interpretation
- more things can go wrong in BUGS than in simple GLM Newton-Raphson or mixed-model-Laplace approximation in **R**
- Hence, if you are savvy statistically and programming-wise, **yes, by all means do learn to write your own samplers !**
- (but may still use BUGS for exploration and dissemination)

BUGS frees the (hierarchical) modeler in you

- Can build statistical model in (almost) exactly the way you imagine data-generating process, i.e. as an HM
- Invites a principled and mechanistic approach to statistical modeling, novel to most ecologists, i.e. HM
- Can allow ecologists to go in creative statistical modeling where they have never even dreamt to go, i.e., by HM

Want to learn WinBUGS/JAGS (and HMs) ?