Introduction to Bayesian inference

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

Marc Kéry

Swiss Ornithological Institute

Online workshop for UZH
12–14 April 2021

Outline of talk

- Intro: What’s the fuss ?
- Role of models in science
- Statistical models
- Analysis of statistical models:
  - Frequentist analysis by maximum likelihood
  - Bayesian analysis
- Simulation-based Bayesian inference via specialised RNGs: MCMC
- JAGS and Nimble, Stan, greta *et al.*
- Concluding remarks on Bayesian/frequentist choice
- BUGS frees the (hierarchical) modeler in you !

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

vogelwarte.ch

What's the fuss ?

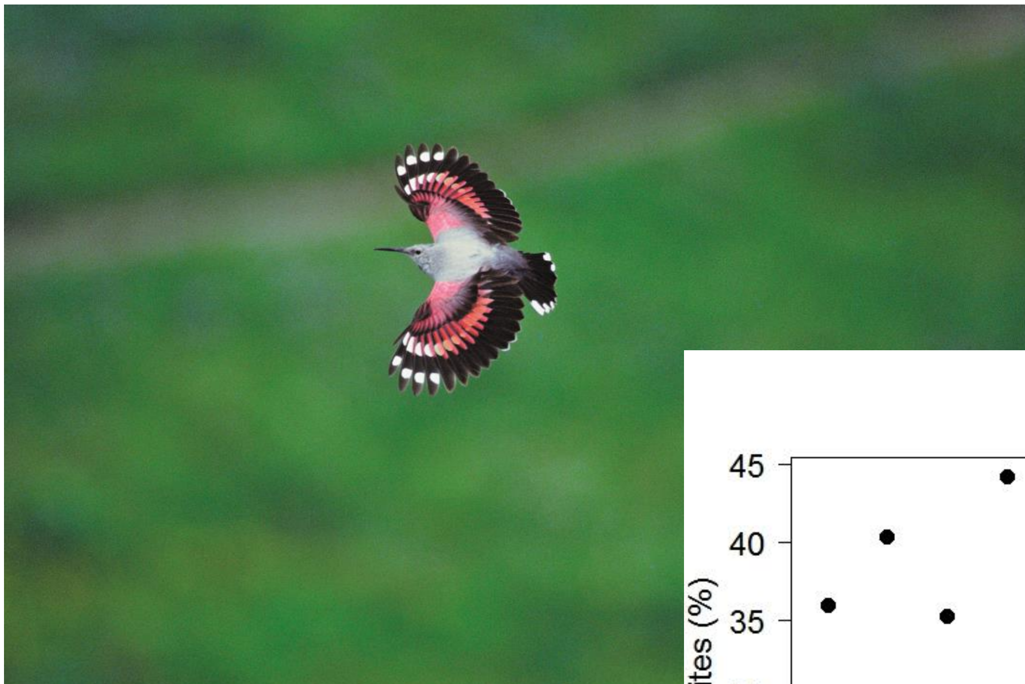
- A simple example

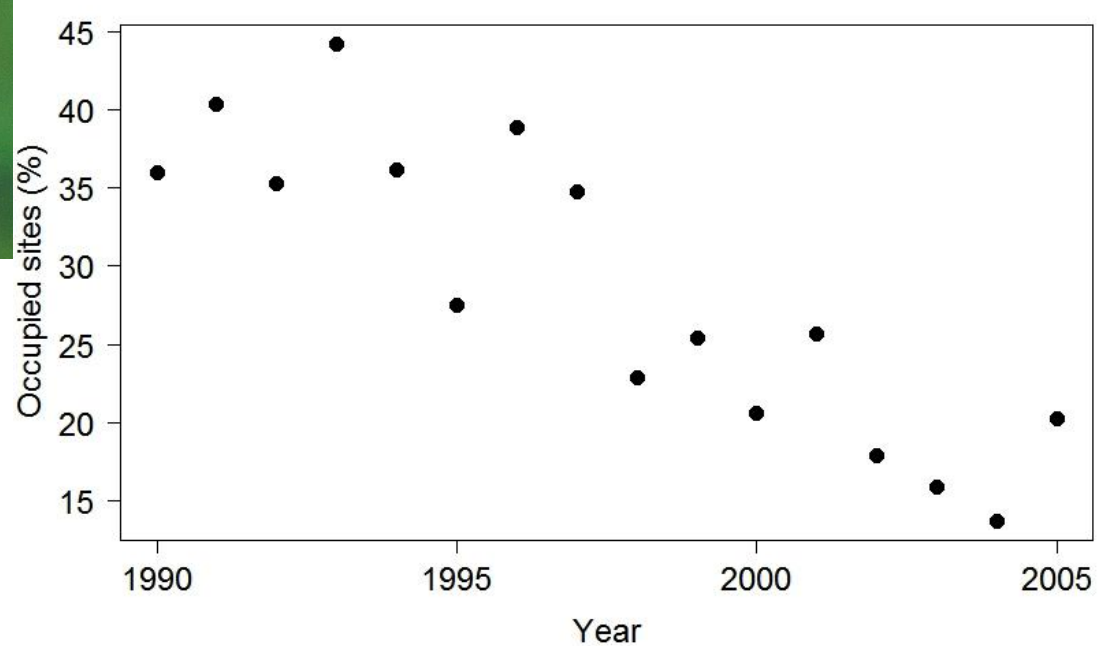
$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

vogelwarte.ch

What's the fuss ?

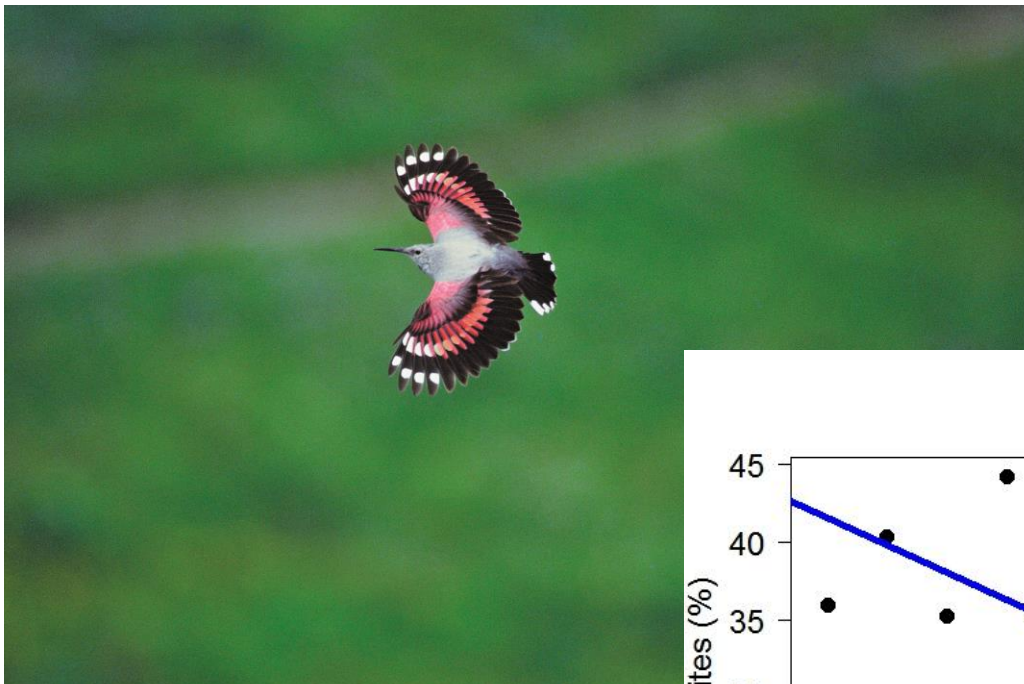
- A simple example





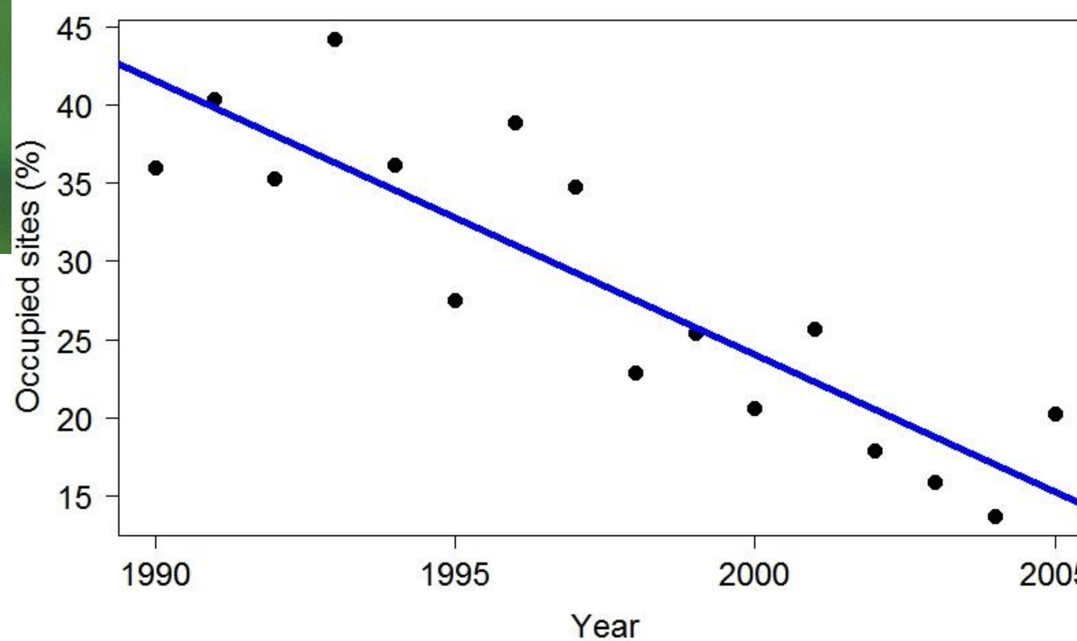
What's the fuss ?

- A simple example



Trend estimate

b = -1.754

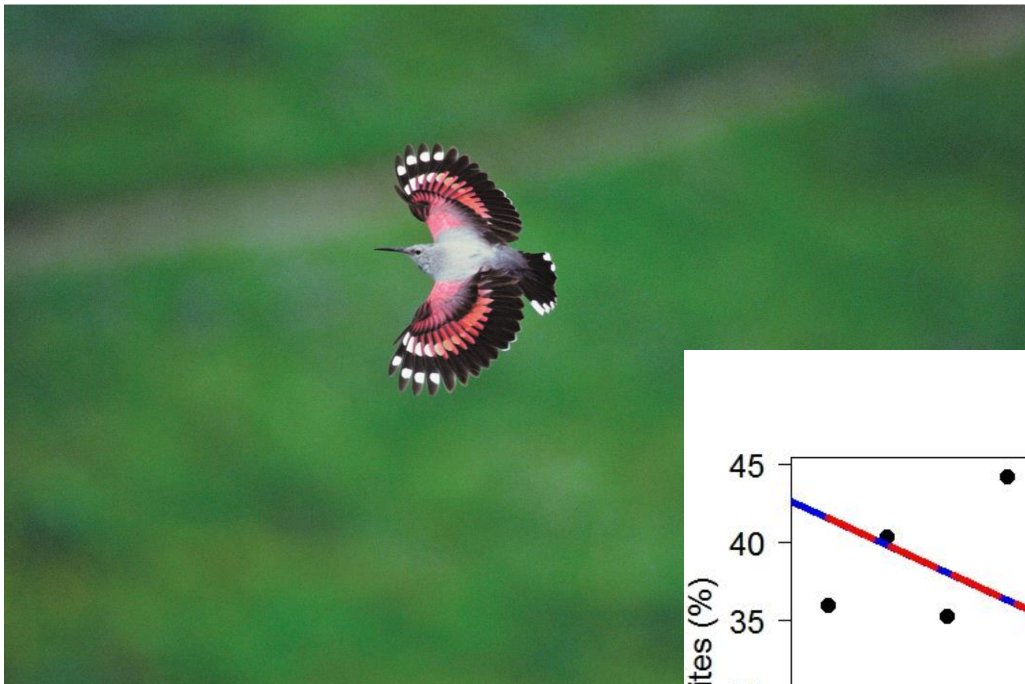


$$y = a + b * X + \varepsilon$$

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

What's the fuss ?

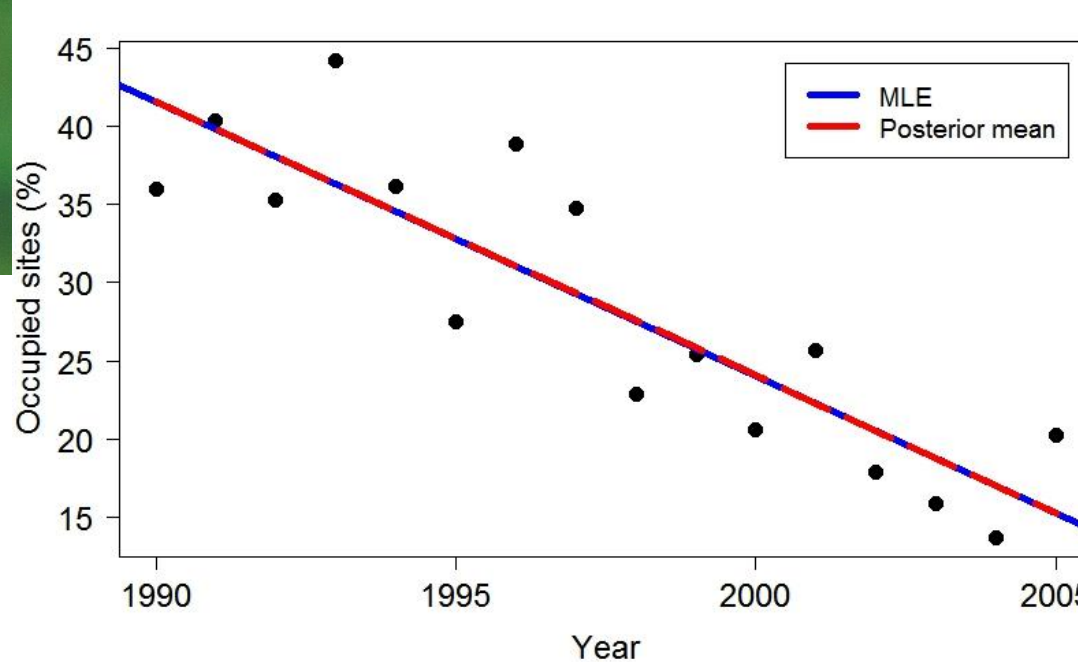
- A simple example



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 their statistical analysis !
- There are no "Bayesian models" or "frequentist models"
- First and foremost, must choose and understand a model
- Then, may choose to analyse that model (e.g., a linear regression) in a Bayesian or non-Bayesian way
- Typically, Bayesian and frequentist analyses yield numerically very similar estimates

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

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

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

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,
To explain ...
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/machine learning ?]*
- Explicit models are better than implicit models
  (e.g., assumptions more transparent, can test them, know what you're doing ..)

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.

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

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

Statistics

- **Statistics: Science of uncertainty ...**
- Or: **Science of learning from data/observations**
- virtually NOTHING in science -- and in life --
  is perfectly predictable, i.e., totally certain
- rather, virtually EVERYTHING in science/life is stochastic
- hence, great importance of statistics in science/life:
  grammar of science; meta-science
- Statisticians:
  "custodians of the scientific method" (Hooke, 1980)
- contrast this with the popular meaning of "statistics" as
  a mere tabulation of numbers !
  e.g. sports statistics

Statistical models

- describe processes underlying observed data
- treat some observed response as outcome from a random variable (r.v.), use probability to describe variation
- r.v.: stats jargon for “something that varies”
- r.v. not fully predictable, only in an average sense
- description of r.v. by probability density function (pdf, for continuous r.v.’s) or probability mass function (pmf, for discrete r.v.’s)
- pdf gives probability density (and pmf gives probability) of every possible observation (outcome) of the random variable
- statistical model *is* a pdf (or pmf)
- This is the way in which statisticians think about statistical models

Statistical models

- describe processes underlying observed data
- treat some observed response as outcome from a random variable (r.v.), use probability to describe variation
- r.v.: stats jargon for "something that varies"
- r.v. not fully predictable, only in an average sense
- description of r.v. by probability density function (pdf, for continuous r.v.'s) or probability mass function (pmf, for discrete r.v.'s)
- pdf gives probability density (and pmf gives probability) of every possible observation (outcome) of the random variable
- statistical model *is* a pdf (or pmf)
- This is the way in which statisticians think about statistical models **--- and in which we biologists should, too !**

Statistical models

- **Trivial example (continuous rv):** model for body mass y
- Body mass y varies, is a random variable
- Use normal probability density function (pdf) for process description:

$$p(y \mid \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2})$$

- Other notation: y ~ Normal(μ, σ^2)
- or (in `R`): `lm(y ~ 1)`
- or: `glm(y ~ 1, family = "gaussian")`

Statistical models

- **Less trivial example (cont. rv):** mass y as a function of height x
- Use normal pdf, with μ replaced by α, β and x:

$$p(y \mid \alpha, \beta, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(y-(\alpha+\beta * x))^2}{2\sigma^2}\right)$$

- Other notation: `y ~ Normal(α + β*x, σ²)`
- or: `y = α + β*x` + ε, with ε `~ Normal(0,σ²)`
- or (in `R`): `lm(y ~ x)`
- or: `glm(y ~ x, family = "gaussian")`

Statistical models

- **Trivial example (discrete rv):**
  number of species detections (y) during N visits
  to an occupied site
- Use binomial probability mass function (pmf):

$$p(y \mid N, p) = \frac{N!}{y!(N-y)!} p^{y} (1-p)^{(N-y)}$$

- Other notation: `y ~ Binomial(N, p)`
- or (in `R`): `glm(y ~ 1, family = "binomial")`

Statistical models

- Statistical model describes both the systematic pattern in a random variable (= response), perhaps as function of covariates ...
- .... as well as the random (=unexplained) variability around the mean
- Response = systematic part + random part

$$y = \mu + \varepsilon$$

- other pairs of terms: deterministic+ stochastic, mean + dispersion structure of model
- 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: $y \sim \text{Binomial}(p, N) = N * \text{Bernoulli}(p)$

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

Statistical models

- Parametric statistical model: description of the stochastic processes thought to have produced response y
- response y is random variable
- Often models with combinations of multiple stochastic sub-processes
- Linked random variables: hierarchical models (HMs) = mixed models etc.
- HMs tremendously rich and powerful manner of building statistical models
- Components of HMs: random variables

Statistical models

Hierarchical models as a combination of >=2 r.v.'s, or GLMs, ordered according to their conditional probability structure:

- **Normal/Normal HM:**

  Latent random variable: $\alpha \sim \text{Normal}(\mu, \tau^2)$

  Observed random variable: $y \sim \text{Normal}(\alpha, \sigma^2)$

- **Bernoulli/Bernoulli HM:**

  Latent random variable: $z \sim \text{Bernoulli}(\psi)$

  Observed random variable: $y \sim \text{Bernoulli}(z * p)$

Statistical models

The model is the fundamental thing to understand in statistics
.... and a fundamental thing in science, too.

And Bayes *vs.* non-Bayes comes only afterwards.

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) ?
- “to guess”: find good value and assess uncertainty

--> Statisticians devise many procedures for guessing, e.g.,
- method of moments
- least-squares
- maximum likelihood (ML), 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

(One) Frequentist way of guessing at θ: maximum likelihood

- Parametric model describes data-generating probabilistic mechanism: probability function, pdf or pmf 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; defines only variability
- E.g., binomial pmf:

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

Frequentist analysis of a model

Maximum likelihood

- **Idea:** good choice of θ is that which maximises function value of pdf/pmf for my data set
- **Likelihood function:** read pdf/pmf "in reverse", i.e., as a function of θ

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

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

$$L(\theta\,|\,y)=\frac{50!}{20!\left(50-20\right)!}\theta^{20}\left(1-\theta\right)^{50-20}$$

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

Frequentist analysis of a model

How to find the MLEs ?

- Analytically (sometimes)
- Numerically (most of the times): ″trial and error″:

  (0) ″Brute force″: simplest trial and error

  (1) Function minimisation

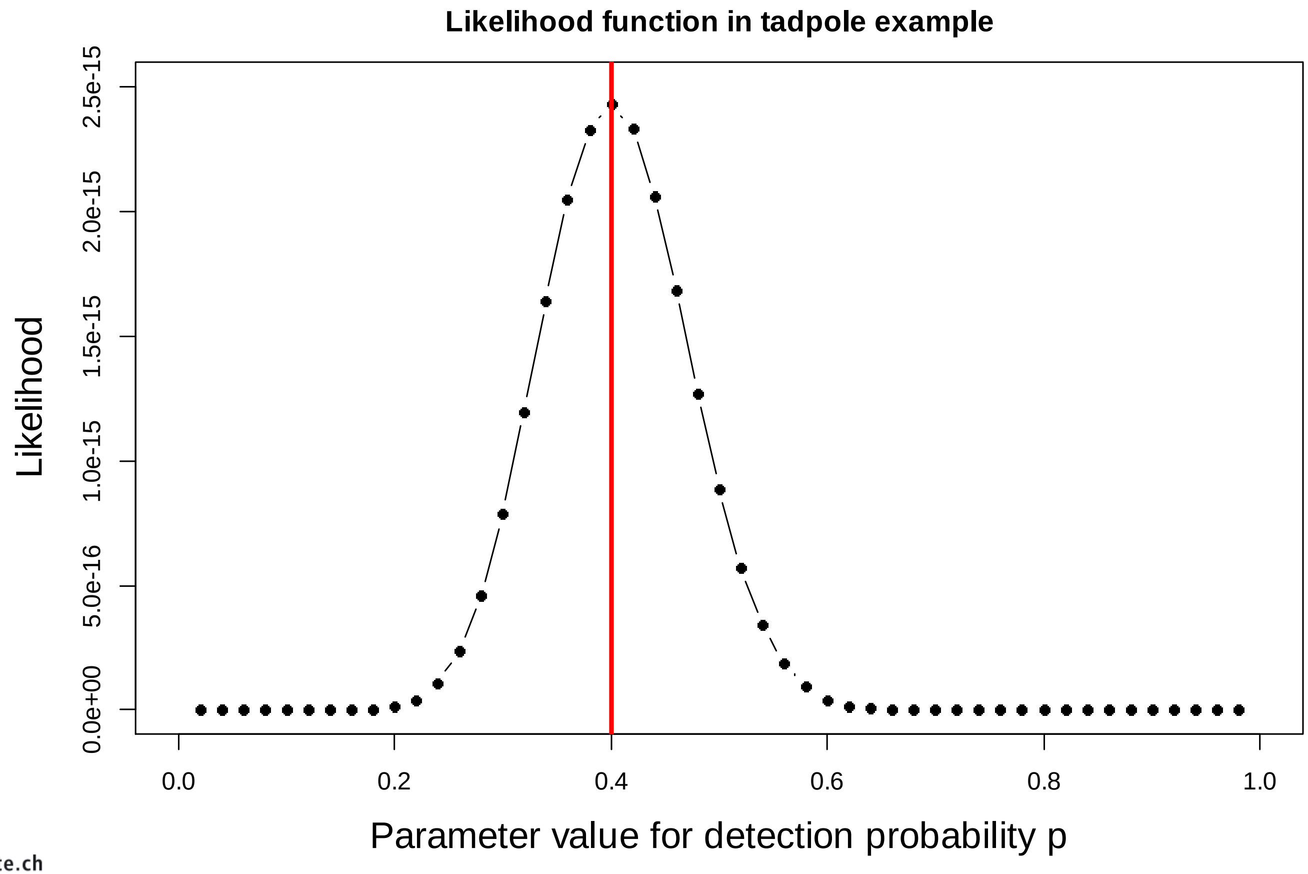
  (2) Using statistical functions in R

  [(3) Bayesian version; see later ...]

Frequentist analysis of a model

Maximum likelihood

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



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
- Much of statistical modeling in ecology is based on MLE

Frequentist analysis of a model

<u>BUT:</u>

- 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 (e.g., for small sample size, MLE are biased !)
- 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 "replicate populations of Panda bears" mean ?

Nice explanation of likelihood inference

See Mike Meredith's web site for a nice example of MLE in the context of an occupancy model:

www.mikemeredith.net/blog/201502/MLE_with_NelderMead.htm

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(θ|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(θ|y)
	"*Prob. of parameter, given data*"

- Data, once collected, are fixed (or, known perfectly)
- **Note:** probability statement about the parameter
- **Degree-of-belief concept of probability:**
  Use probability distribution to express imperfect knowledge (about θ)
- Hence, parameters treated **as if** they were random variables
- How should p(θ|y) be computed ?

Bayesian analysis of a model

- Bayes rule

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

- 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) ?
- Update p(b) to p(b|F)

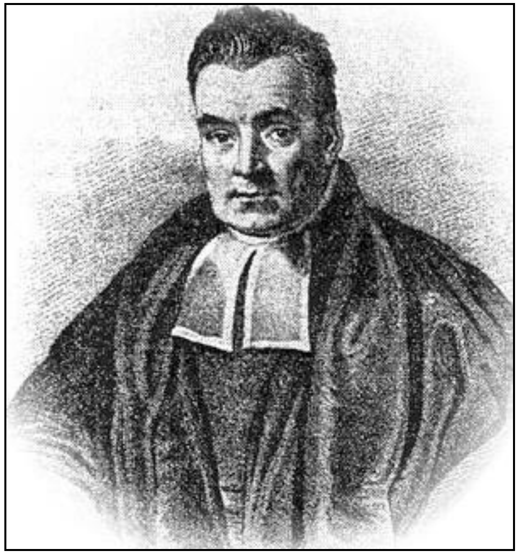
Bayesian analysis of a model

- Bayes rule

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

- 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 quantities:
  Can say "*... I am 95% certain that prob of detection > 0.2*" !

Bayesian analysis of a model

Formal steps underlying every Bayesian analysis

- Use probability as a universal measure of uncertainty about unknown quantities; here: θ
- Treat all statistical inference (parameter estimation, testing, missing values, ...) as a mere 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 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

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

Bayesian analysis of a model

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

- "Human" concept of probability ("*I am 95% certain that ...*")
- $p(\theta|y) \propto p(y|\theta) \times p(\theta)$
- can say, "Posterior = Likelihood x prior"
- Like human learning:
  - Conclusion is combination of experience and new information (e.g., problem of bird identification, such as "Andean Condor in the Alps")
  - 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

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

Bayesian analysis of a model

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

- Every scientific position/opinion (embodied in prior) can be modified by new evidence/data !
- Unlike religion (in the strict and in the wide sense),
  where no amount of evidence/data
  can ever overthrow The Prior Belief
- Avoid 0/1 priors in science:
  this would be the end of learning !

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

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

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

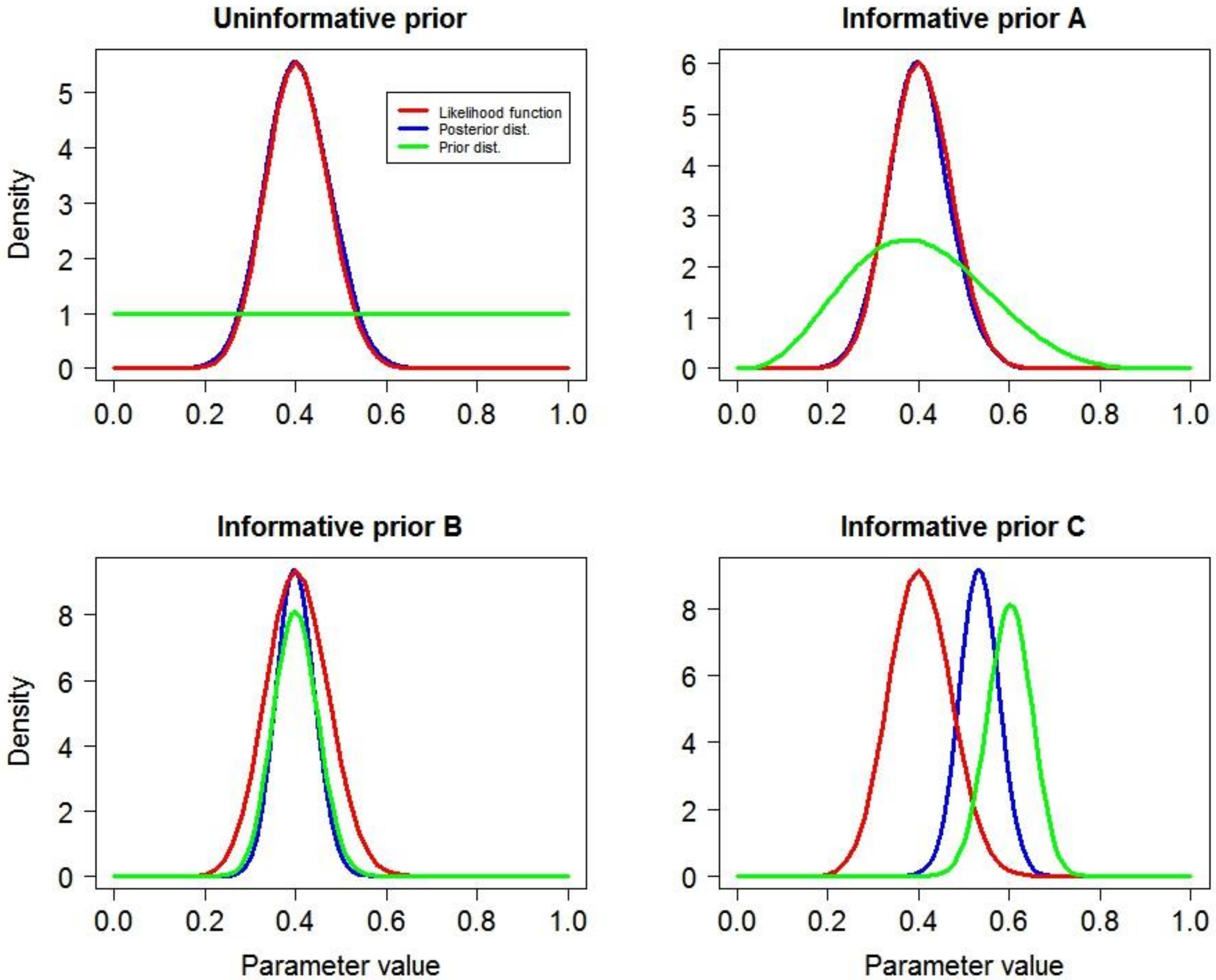
Bayesian analysis of a model

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 default 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.



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

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

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; only evaluate numerator (likelihood x prior)
- Approximate posterior to arbitrary degree of accuracy by drawing large sample
- **Markov chain Monte Carlo (MCMC) / Markov chain simulation, e.g.**
  - Metropolis(-Hastings) algorithm
  - Gibbs sampling
- Huge boost to Bayesian statistics in statistics community

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

Algorithm of Metropolis et al. (1953)

- Start with arbitrary value: θ^0
- Repeat large number of times (`for t in 1:T`):

(1) Propose (try) new value θ^* for parameter θ:
Draw θ^* from "rule", e.g. `Normal(`θ^{t-1}`, `$\sigma_{\mathtt{proposal}}$`)`

(2) Compare posterior densities for θ^* and θ^{t-1} by ratio R

$$R = \frac{p(y|\theta^*)\ p(\theta^*)\ /\ \cancel{p(y)}}{p(y|\theta^{t-1})\ p(\theta^{t-1})\ /\ \cancel{p(y)}}$$

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

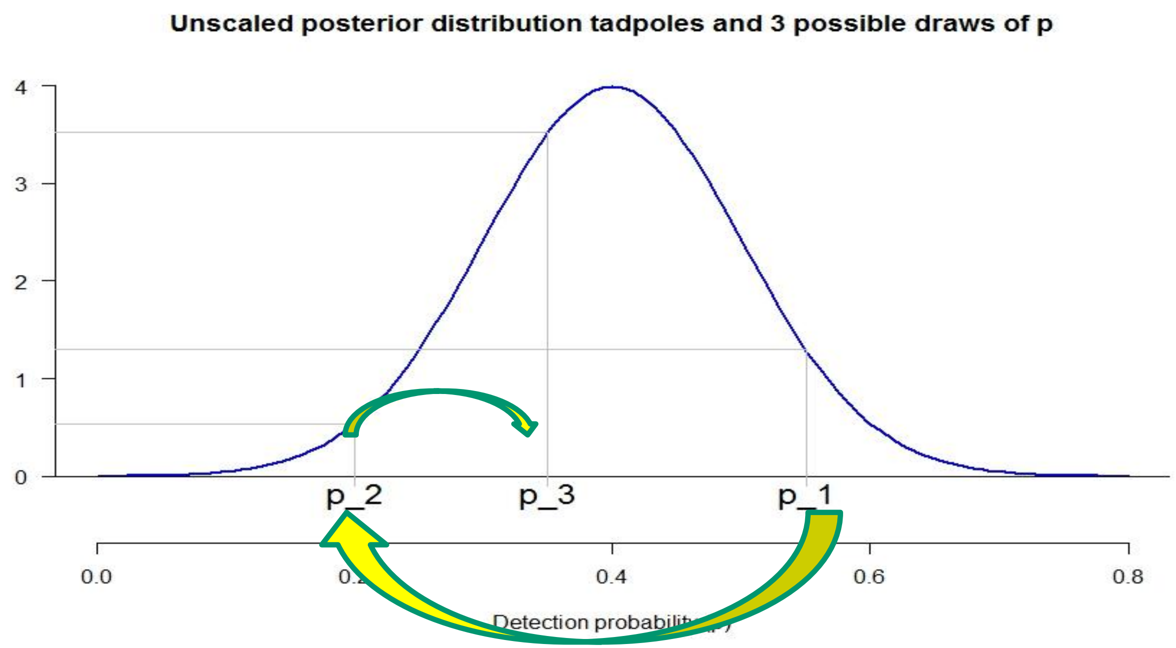
(3) If R >= 1, set θ^t <- θ^* (**accept** new value)

If R < 1, set θ^t <- θ^* with prob. R (**accept** new value)
else θ^t <- θ^{t-1} (**reject** new value, keep previous)

=> Frequency of values proportional to p(θ|y) !

Algorithm of Metropolis et al. (1953)

- sample p(θ | y) !
- repeat for multiple parameters (if $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_k\}$)
- MCMC: *jump "upwards" along posterior with greater prob.*



-> see `demoMCC` ex.

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

Gibbs sampling algorithm (Geman & Geman 1984)

- want $p(\theta|y)$ for $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_k\}$
- define *full conditional distributions* $p(\theta_1| \theta_2,\theta_3, \ldots \theta_k, y)$
- Set $\boldsymbol{\theta} = \{\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}\}$ at arbitrary initial values
- Repeat large number of times (`for t in 1:T`):

  (1) Draw $\theta_1^{(t)}$ from $p(\theta_1| \theta_2^{(t-1)},\theta_3^{(t-1)},\ldots, \theta_k^{(t-1)}, y)$

  (2) Draw $\theta_2^{(t)}$ from $p(\theta_2| \theta_1^{(t-1)}, \theta_3^{(t-1)},\ldots, \theta_k^{(t-1)}, y)$

  ....

  (3) Draw $\theta_k^{(t)}$ from $p(\theta_k| \theta_1^{(t-1)}, \theta_2^{(t-1)},\ldots, \theta_{k-1}^{(t-1)}, y)$

- again, sample $p(\theta|y)$!

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)}$$

vogelwarte.ch

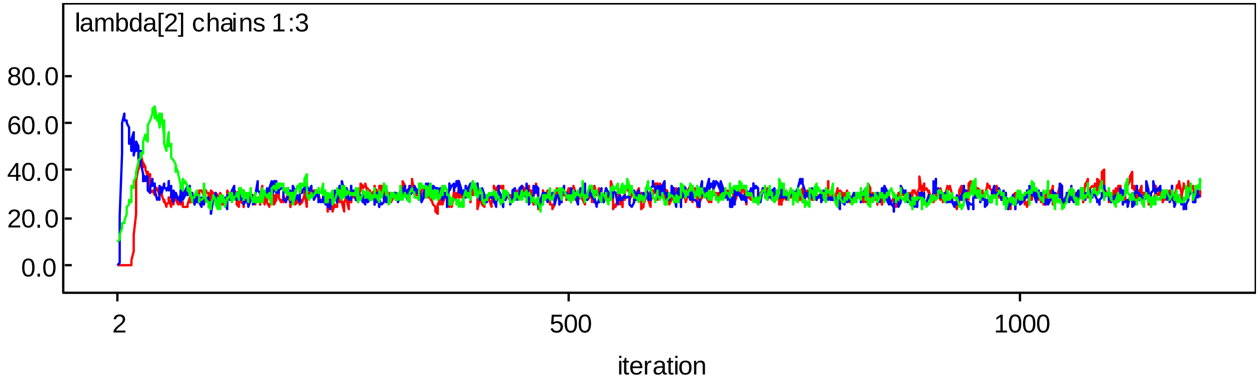
Markov chain Monte Carlo (MCMC)

- Metropolis-(Hastings) algorithm, Gibbs sampler, and MANY others !
- Often combinations (hybrids) of basic algorithms, e.g. Metropolis-within-Gibbs
- Purpose in life of many in statistics/computation: to devise more efficient algorithms
- MCMC can be great fun (see later)
- Great if you know how to construct algorithms
- However, except for 10-20% of ecologists, waste of time
- much better to use MCMC engine such as BUGS/JAGS
- However, necessary to understand principles

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

MCMC

- MCMC: Stochastic algorithm produces sequence of dependent random numbers (= Markov chain)
- **RNG for arbitrary and often unknown (posterior) distributions ! -> R example (for independent sample)**
- MCMC produces long streams of numbers
- Converge to equilibrium **distribution** (usually)
- Equilibrium distribution = desired posterior distribution (if algorithm constructed well)

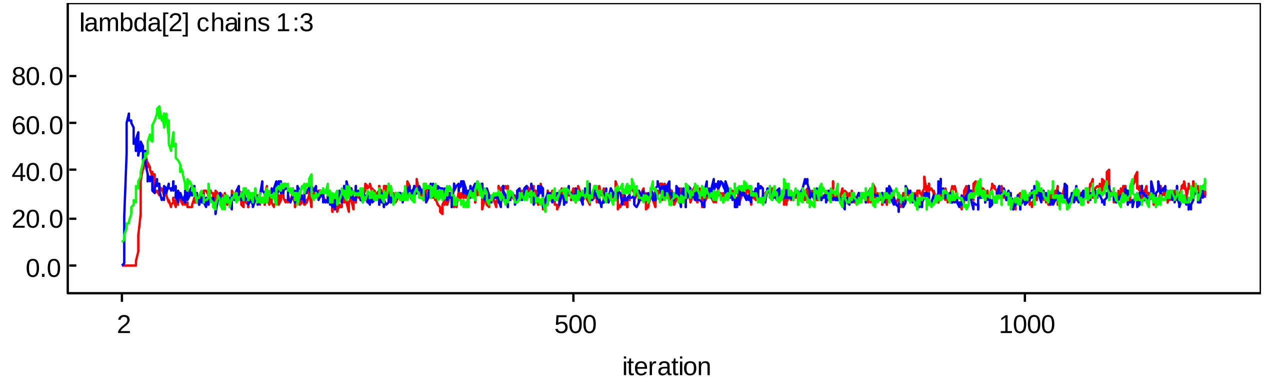


$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

vogelwarte.ch

MCMC

- 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)



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

vogelwarte.ch

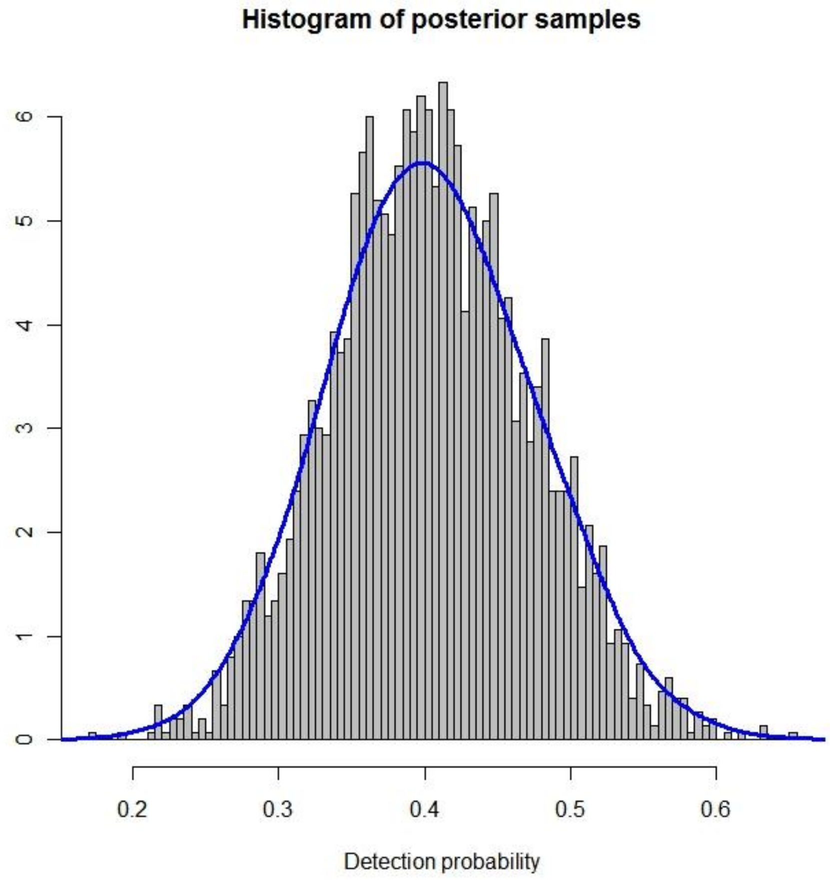
MCMC for tadpole example

```
> 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
```

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

MCMC for tadpole example

```
> 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
```

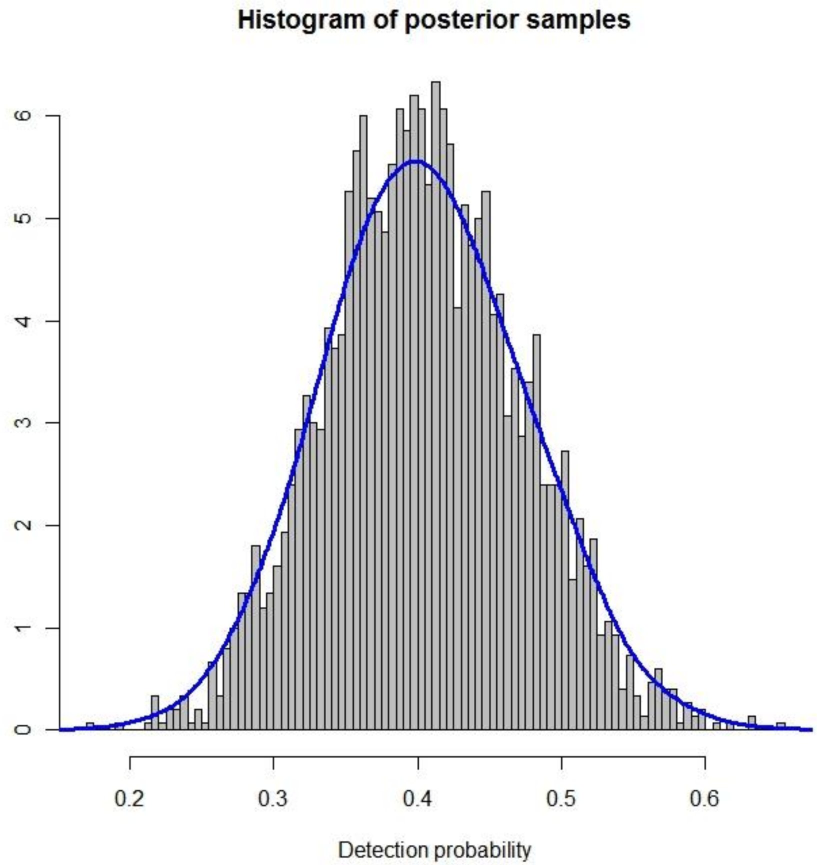


$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

vogelwarte.ch

MCMC for tadpole example

```
> 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
```



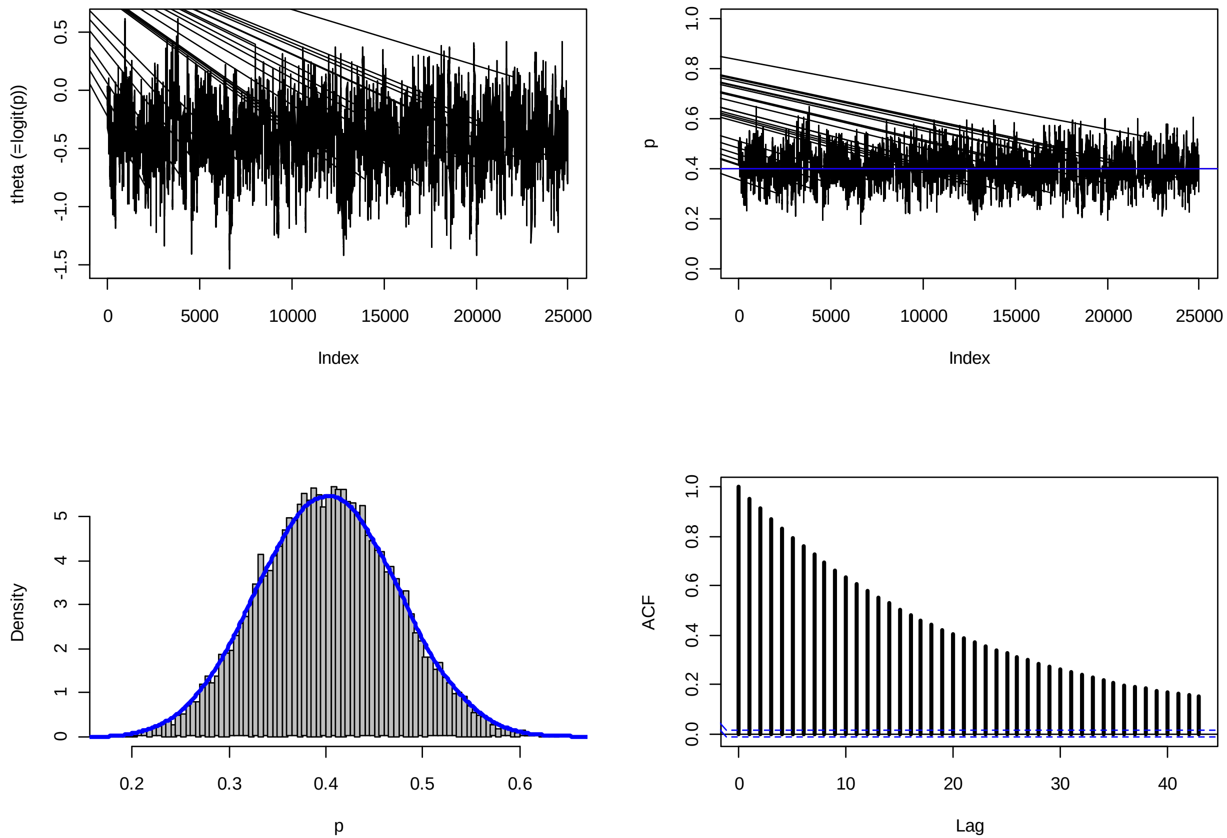

```
> mean(p)
[1] 0.4047
> sd(p)
[1] 0.0674
> quantile(p, probs = c(0.025, 0.975))
     2.5%     97.5%
   0.2771    0.5375
```

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

vogelwarte.ch

MCMC for tadpole example

- Custom MCMC code for binomial proportion (tadpoles)

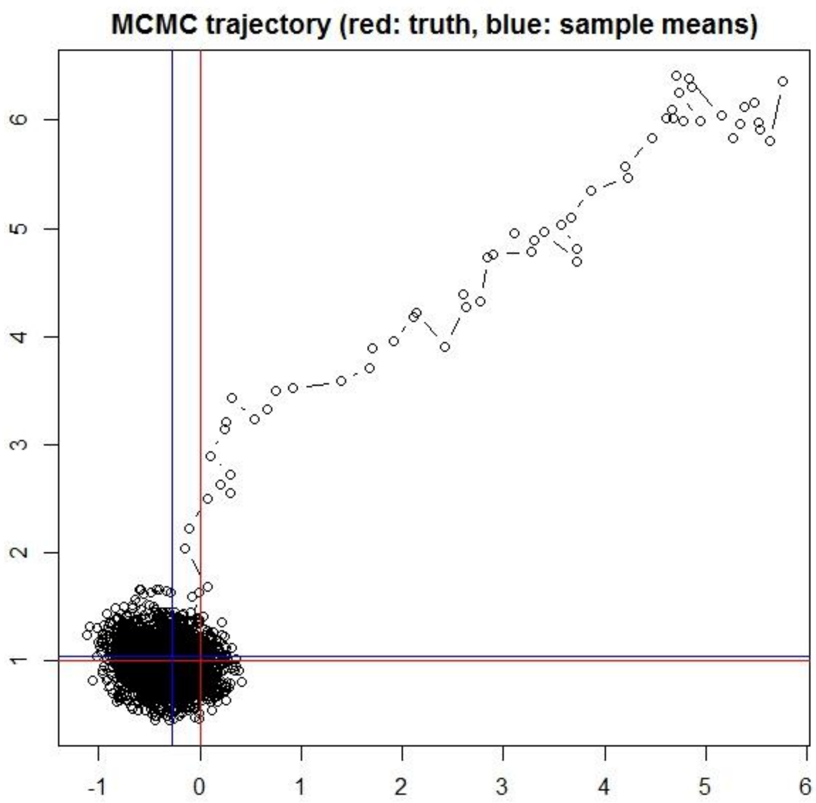


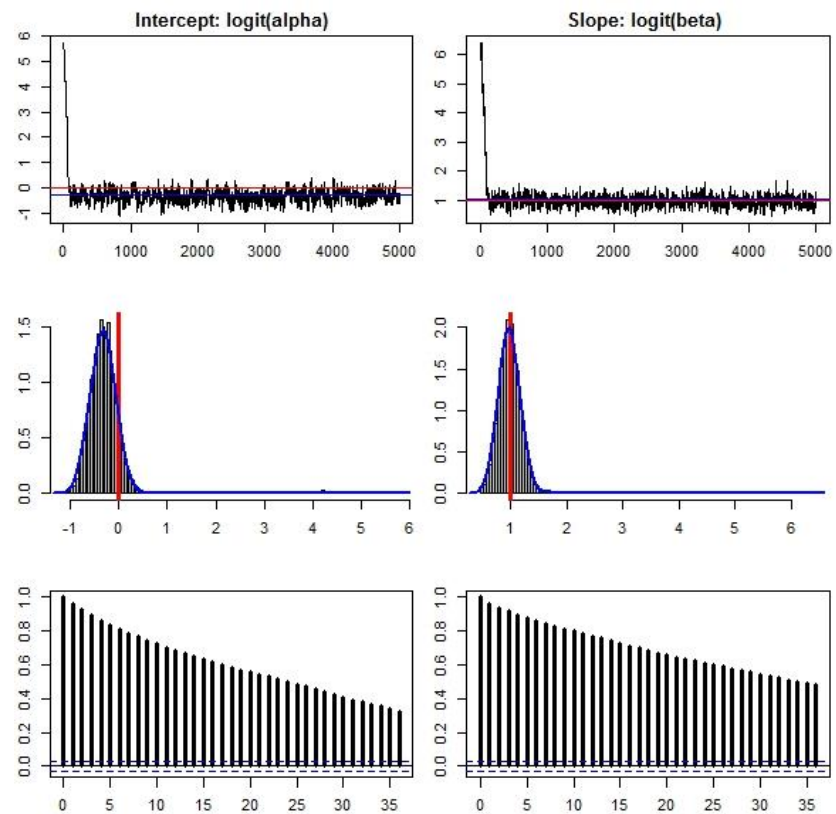
$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

vogelwarte.ch

MCMC for logistic regression example

- See cool animation (**-> R example with logistic reg.**) !



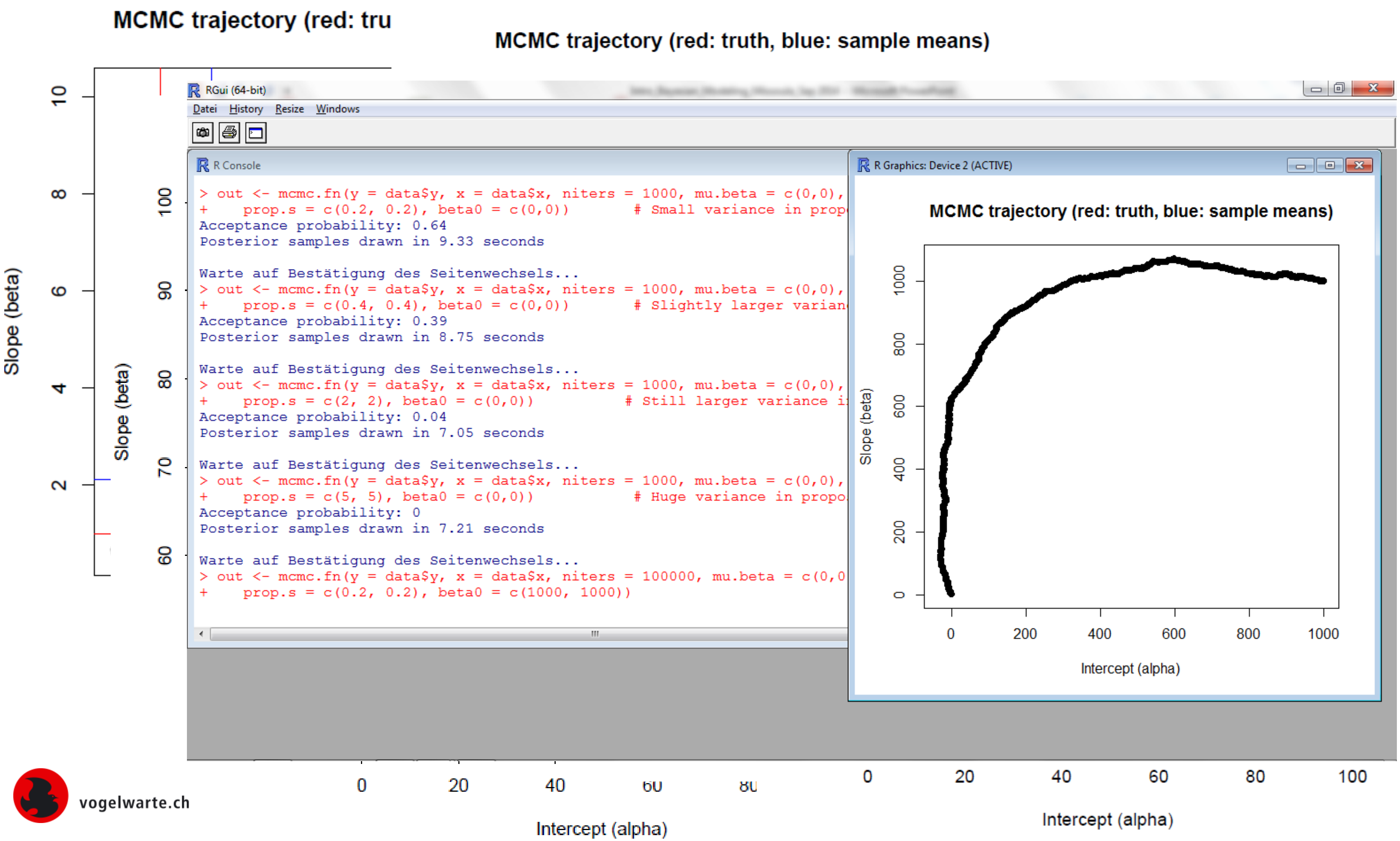


$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

vogelwarte.ch

MCMC for logistic regression example

- MCMC astonishing and crazily powerful family of algorithms !



vogelwarte.ch

Really nice explanation of Bayesian inference

See Mike Meredith's web site for a nice example of various flavours of Bayesian inference in the context of an occupancy model:

Gibbs sampler:

www.mikemeredith.net/blog/201502/Gibbs_sampler.htm

Metropolis-Hastings:

www.mikemeredith.net/blog/201503/RandomWalk_MCMC.htm

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 (see also later comments)
- Change due to BUGS project:
  Bayesian inference **u**sing **G**ibbs **s**ampling
- BUGS does Gibbs sampling and other variants of MCMC
- Statisticians/Epidemiologists in Cambridge/UK
- Lunn et al. (2009), *Statistics in Medicine,* 3049–3067

vogelwarte.ch

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

The BUGS project

- BUGS: Flexible, generic Bayesian modeling software; does:
  1. Simple and intuitive model description language (BUGS programming 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/ **(Andrew Thomas)**
  - **JAGS:** mcmc-jags.sourceforge.net/ **(Martyn Plummer)**
  - **(also NIMBLE, Stan, multiBUGS)**

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

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.

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

BUGS natural for hierarchical models (HMs)

- 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

<u>3 types of advantages of Bayesian analysis by MCMC in BUGS:</u>

<u>(1) Bayesian paradigm:</u>

- ‘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 (*“super-powerful glmer”*)
- Enforces understanding of model
- **BUGS software 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 (much) less for ML)
  - Model selection is a pain (cf. AIC with ML)
  - Sensitivity of results to prior choice (not with ML)
  - BUGS so flexible that may fit nonsensical models
  - ... that may fit models with unidentifiable params
- Hence, happy to use maximum likelihood as well

Conclusion on the Bayesian/frequentist choice

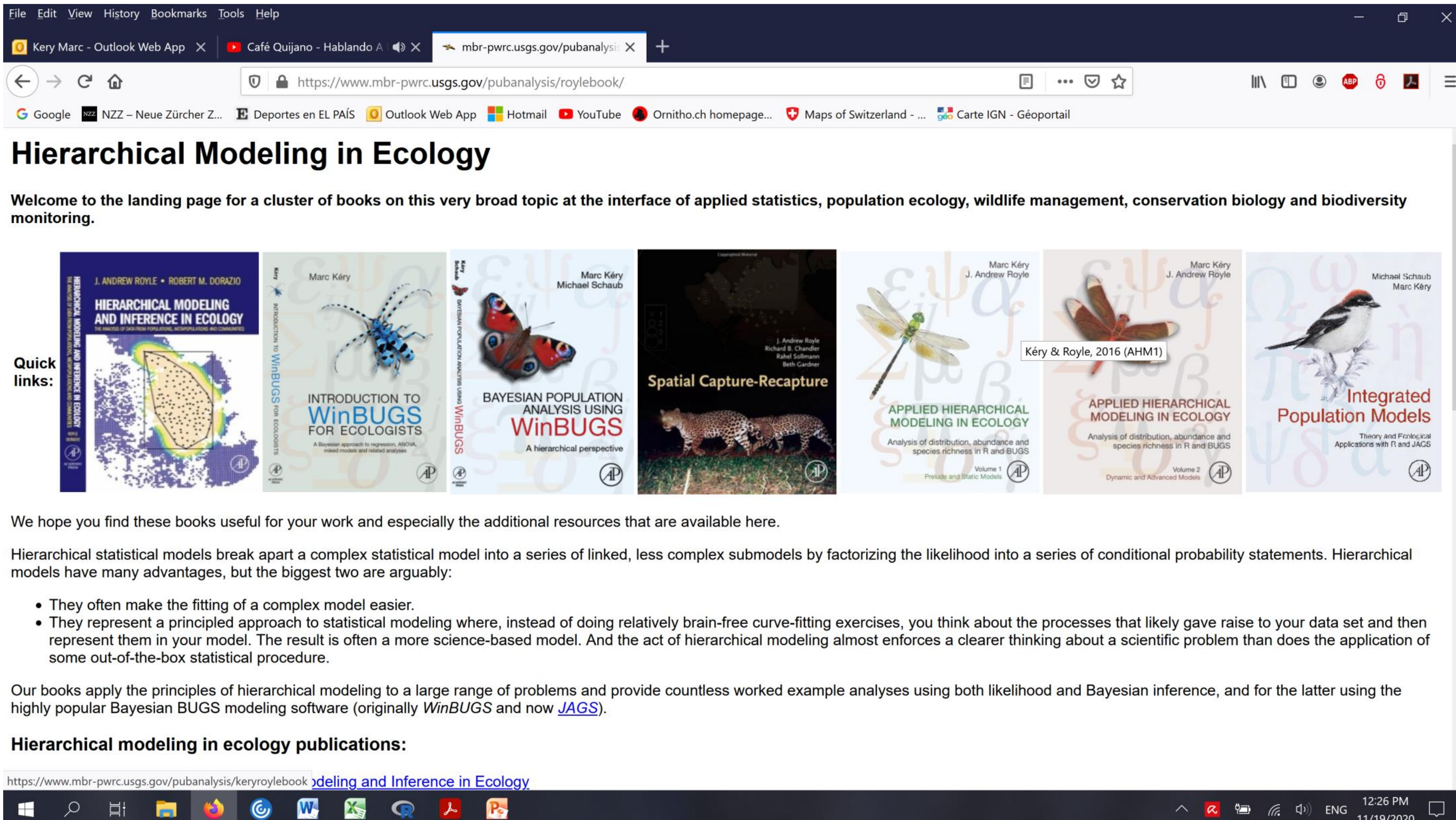
- Be eclectic !
- Choose what is most useful for *you*
- Usually will not use BUGS for trivial problems
- BUGS is fantastic for more complex models
  (except for large data sets !)
- BUGS language is great to actually understand a model
- Stay tuned: in the future, there will (hopefully !)
  be better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

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 ?

-> **www.hierarchicalmodels.com**



Hierarchical Modeling in Ecology

Welcome to the landing page for a cluster of books on this very broad topic at the interface of applied statistics, population ecology, wildlife management, conservation biology and biodiversity monitoring.

Quick links:

We hope you find these books useful for your work and especially the additional resources that are available here.

Hierarchical statistical models break apart a complex statistical model into a series of linked, less complex submodels by factorizing the likelihood into a series of conditional probability statements. Hierarchical models have many advantages, but the biggest two are arguably:

- They often make the fitting of a complex model easier.
- They represent a principled approach to statistical modeling where, instead of doing relatively brain-free curve-fitting exercises, you think about the processes that likely gave raise to your data set and then represent them in your model. The result is often a more science-based model. And the act of hierarchical modeling almost enforces a clearer thinking about a scientific problem than does the application of some out-of-the-box statistical procedure.

Our books apply the principles of hierarchical modeling to a large range of problems and provide countless worked example analyses using both likelihood and Bayesian inference, and for the latter using the highly popular Bayesian BUGS modeling software (originally *WinBUGS* and now *JAGS*).

Hierarchical modeling in ecology publications:

https://www.mbr-pwrc.usgs.gov/pubanalysis/keryroylebook odeling and Inference in Ecology

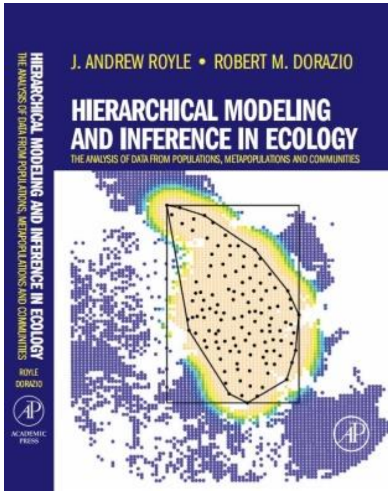
Summary

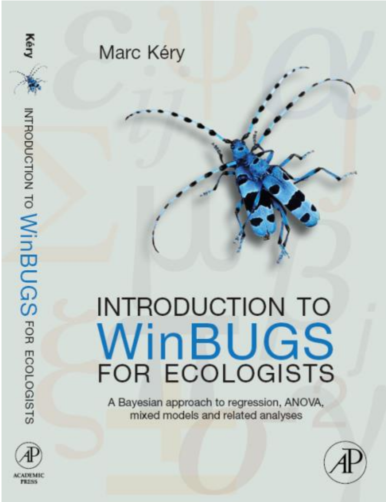
- Intro: What's the fuss ?
- Role of models in science
- Statistical models
- Analysis of statistical models:
  - frequentist analysis (maximum likelihood)
  - Bayesian analysis
- Bayesian computation via specialised RNGs: MCMC
- BUGS/JAGS
- Concluding remarks on Bayesian/frequentist choice
- BUGS frees the (hierarchical) modeler in you !

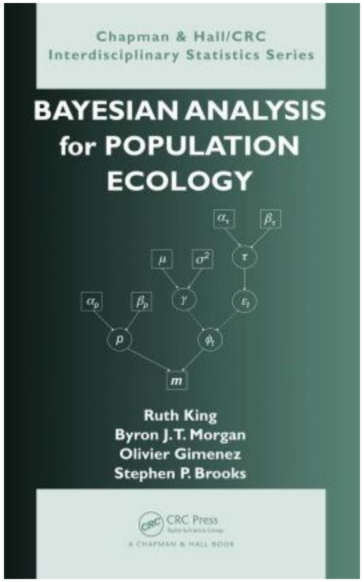
$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

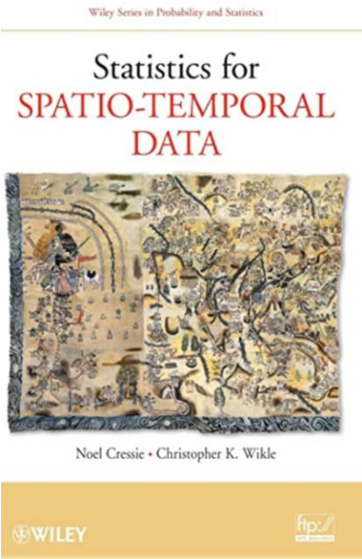
vogelwarte.ch

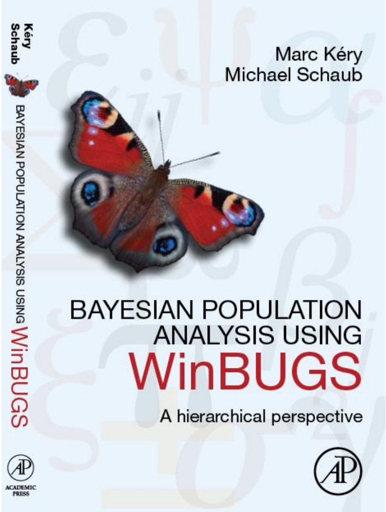
More books on HMs

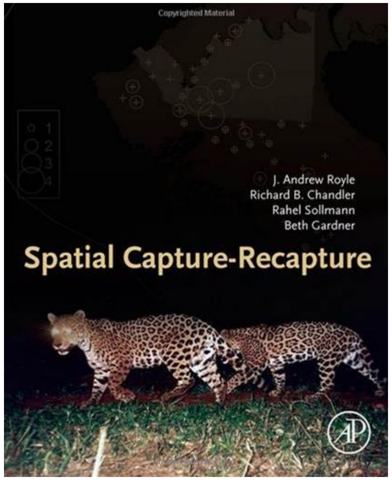


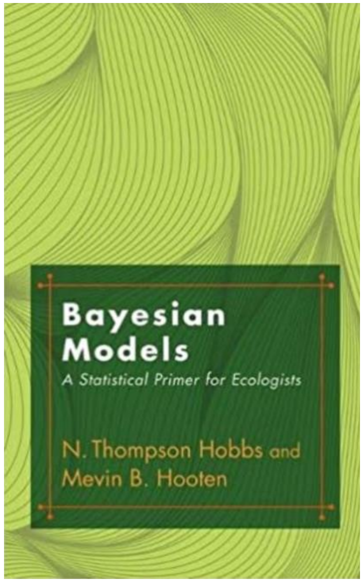


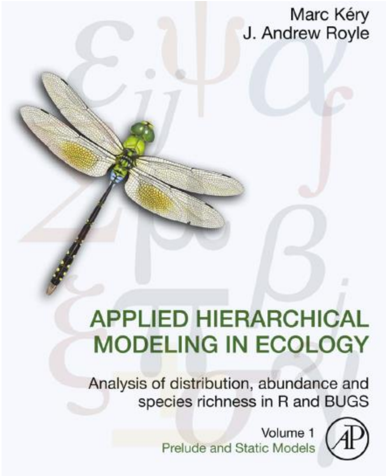




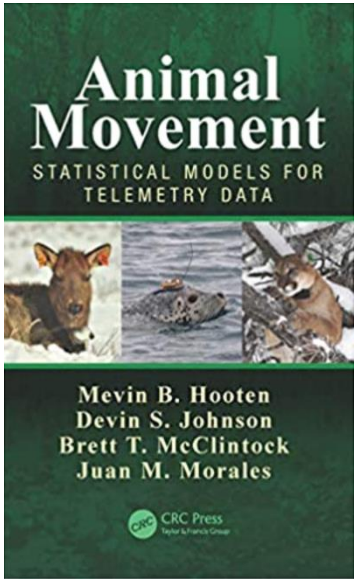


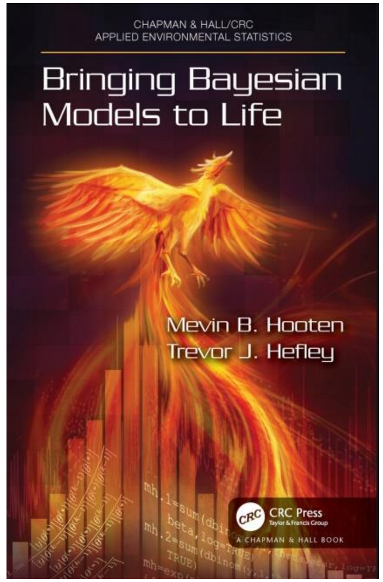


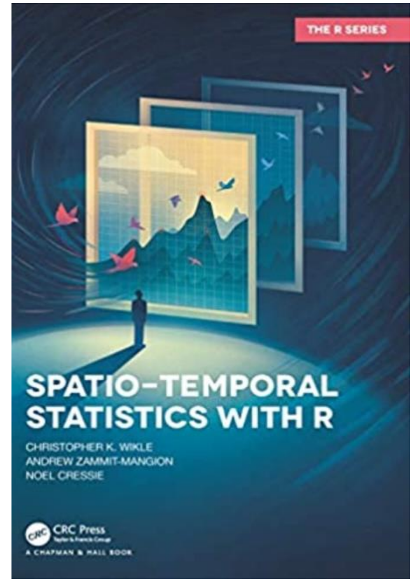


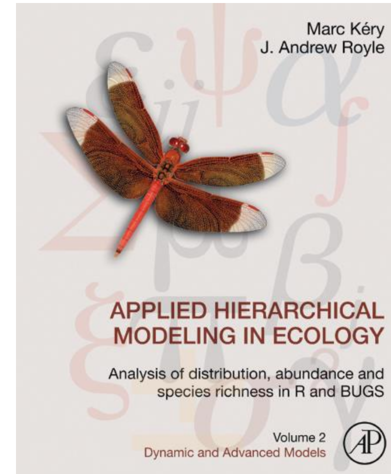


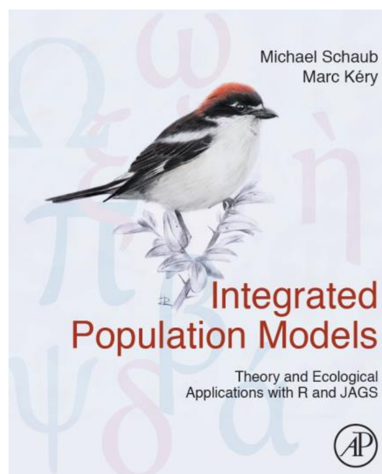
More books on HMs











vogelwarte.ch