Model analysis:

Revisiting modes of inference

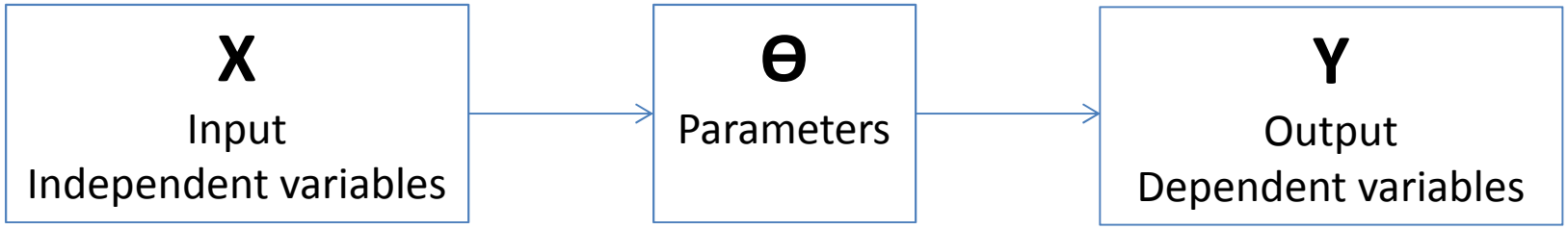
Outline

- Analysis of models
  - Frequentist
  - Bayesian
- Bayesian computation - WinBUGS and JAGS
- Discussion on Bayesian/frequentist choice

A model is a model no matter how it is analyzed!

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

Analysis of a statistical model



- Data viewed as result of random process(es)
- Parameters (θ) are unknown variables (of interest)
- How should we guess at value(s) of θ ?
  ...at missing covariates (x) ? ... at missing response (y) ?

--> Statisticians devise many procedures for guessing

- ***Frequentis approaches: maximum likelihood***
- ***Bayesian analysis***

Frequentist analysis of a model

- (One) Frequentist way of guessing at θ: maximum likelihood
- 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 θ
- Probability defined as long-run frequency in hypothetical replicate data sets
- E.g., Poisson distribution ($y \sim Pois(\Theta)$) with PMF:

$$p(y|\Theta) = \frac{\Theta^y}{y!} e^{-\Theta}$$

Frequentist analysis of a model

- Maximum likelihood
- **Idea:** good choice of θ is that which maximizes function value of sampling distribution for the data set

- **Likelihood function:** reading the sampling distribution "in reverse" as a function of θ -> $p(y|\Theta)$ = $L(\Theta|y)$

$$L(\Theta|\boldsymbol{y}) = \prod \frac{\Theta^{y}}{\boldsymbol{y}!} e^{-\Theta}$$

where $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$

- Call the value of θ that maximizes L the Maximum Likelihood estimate (MLE)

Frequentist analysis of a model

- Say we have three data points (2,3,4) and we want to estimate the mean using a Poisson distribution:

$$L(\theta|\boldsymbol{y}) = \prod \frac{\theta^{\boldsymbol{y}}}{\boldsymbol{y}!} e^{-\theta}$$

- Plugging in the numbers, we get:

$$L(\theta|\{2,3,4\}) = \frac{\theta^{\mathbf{2}}}{\mathbf{2}!} e^{-\theta} \cdot \frac{\theta^{\mathbf{3}}}{\mathbf{3}!} e^{-\theta} \cdot \frac{\theta^{\mathbf{4}}}{\mathbf{4}!} e^{-\theta}$$

Frequentist analysis of a model

Maximum likelihood

- How to find the MLE ?

  - Analytically (sometimes)

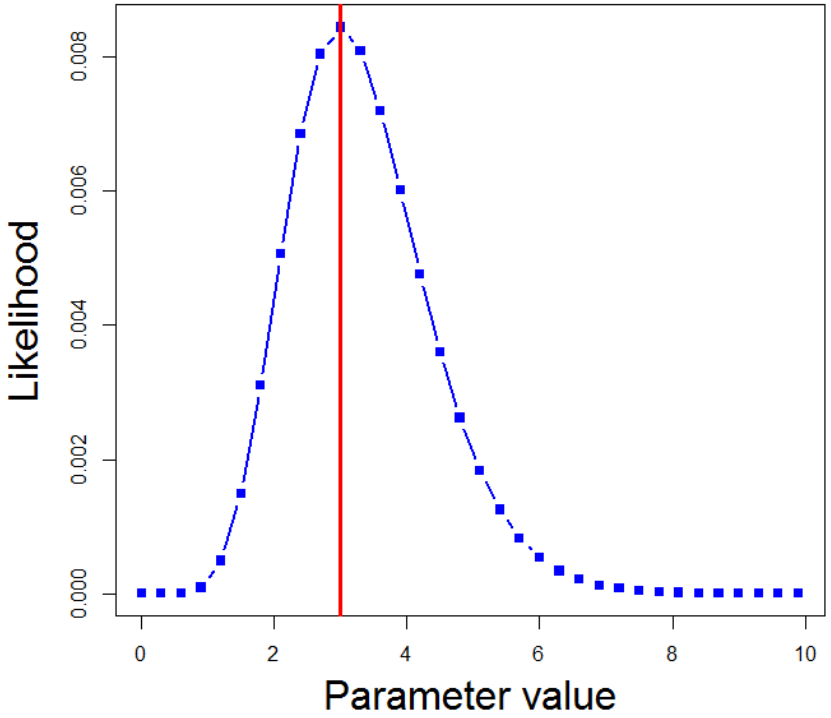
    Take the derivative of the likelihood and find the maximum value

  - 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



$$L(\theta|\{2,3,4\}) = \frac{\theta^2}{2!}e^{-\theta} \cdot \frac{\theta^3}{3!}e^{-\theta} \cdot \frac{\theta^4}{4!}e^{-\theta}$$

Frequentist analysis of a model

Some characteristics of maximum likelihood

- "Automatic inference": simply define likelihood function and then find parameter values that maximize it

- Produces "good estimates", e.g., asymptotically unbiased, consistent, transformation invariant

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
- "Indirect" probability statements about data, rather than about params: $p(y|\theta)$ ->

  If we repeated the procedure on multiple samples, the CI (which would differ for each sample) would encompass the true population parameter 95% of the time.

- **95% CI does not contain θ with P=0.95**. Impossible 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

Bayesian analysis of a model

- 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., dependent and independent data)

Bayesian analysis of a model

Recipe of every Bayesian analysis:

1. What is known?	The data (y = {2,3,4})
2. What is unknown?	Mean abundance (θ)
3. What to do?	Calculate p(θ|y)

- "*Probability 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(θ|y) be computed?

Bayesian analysis of a model

Bayes rule

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

- Can be deduced from p(A,B) = p(B | A) * p(A)

  (joint prob. = conditional prob. *marginal unconditional prob.)

Bayesian analysis of a model

Bayes theorem:

- Basic tool of Bayesian analysis
- Provides the means by which we can learn from data
- Given a prior state of knowledge, it tells us how to update this belief based on observations

Bayesian analysis of a model

Bayes rule for statistical inference:

$$P\ (\Theta|y) = \frac{P(y|\Theta)P(\Theta)}{P(y)} = \frac{P(\Theta, y)}{P(y)}$$

Posterior distribution:	$P(\Theta\|y)$
Likelihood function:	$P(y\|\Theta)$
Prior distribution:	$P(\Theta)$
Prob. of data:	$P(y) = \int P(y\|\Theta)P(\Theta)d\Theta$

- **NOTE:** Use probability to express imperfect knowledge
- Direct probability statements about unknown quantities

Bayesian analysis of a model

Heuristic appeal of Bayes rule as model for inference

- "Human" concept of probability ("*I am 95% certain that...*")
- Like human learning:

  - Conclusion is combination of experience and new information

  - New information changes ("updates") my previous state of knowledge to my current state of knowledge

  - Synthesizes *all* existing knowledge

Bayesian analysis of a model

Why would you want to use Bayesian methods?

- Sometimes finding the joint likelihood is really really hard
  - Interacting parameters
  - Lots of random effects in a hierarchical structure
  - Integrating hundreds (thousands!) of likelihoods
- Absence of asymptotics – unbiased regardless of sample size
- Ease of error computation – directly compute random sample (as opposed to delta method for MLE)
- Intuitive interpretation of parameters – directly calculate prob that parameter has a certain value (e.g., I am 99% sure that…)

Bayesian analysis of a model

Advantage of prior distribution:

- Bayesian inference allows formal incorporation of external knowledge into estimation via prior distribution
- Helps deal with small sample sizes (ecology of rare species)
- Advantage of ‘informative priors’:
    - Don’t feign to be stupid
    - More precise estimates
    - Can estimate additional parameters

Bayesian analysis of a model

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

  Some examples:

  - Beta prior and binomial likelihood lead to beta posterior
  - Normal prior and normal likelihood lead to normal posterior
  - Gamma prior and Poisson likelihood lead to gamma posterior

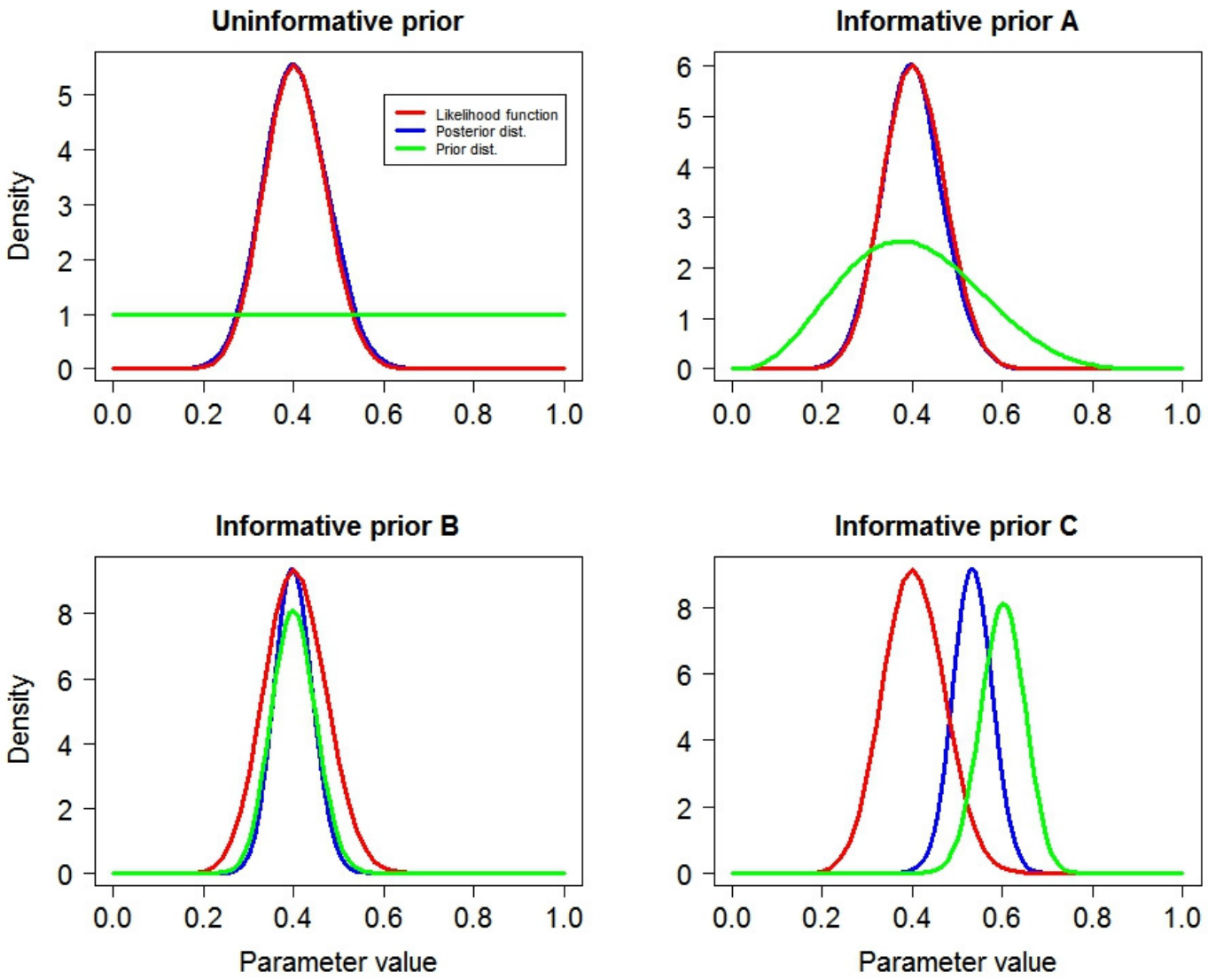
- Otherwise it is very difficult to find an analytical solution for the posterior.

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
- We will (almost exclusively) specify vague priors, typically on “natural“ scale
- Estimates (very much) resemble MLEs for vague priors; i.e., mode of the posterior = MLE

Bayesian analysis of a model



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

$$P\ (\Theta|y) = \frac{P(y|\Theta)P(\Theta)}{P(y)}$$

$$P(y) = \int P(y|\Theta)P(\Theta)d\Theta$$

- 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

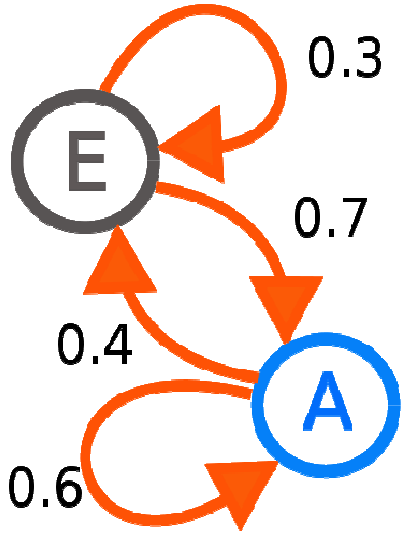
--> Use stochastic simulation to draw dependent samples from posterior distribution

- 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

Bayesian computation

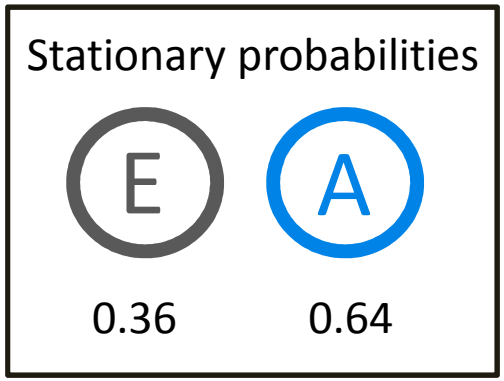
Markov chain

- A mathematical system that undergoes transitions from one state to another on a state space.
- A random process usually characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it.



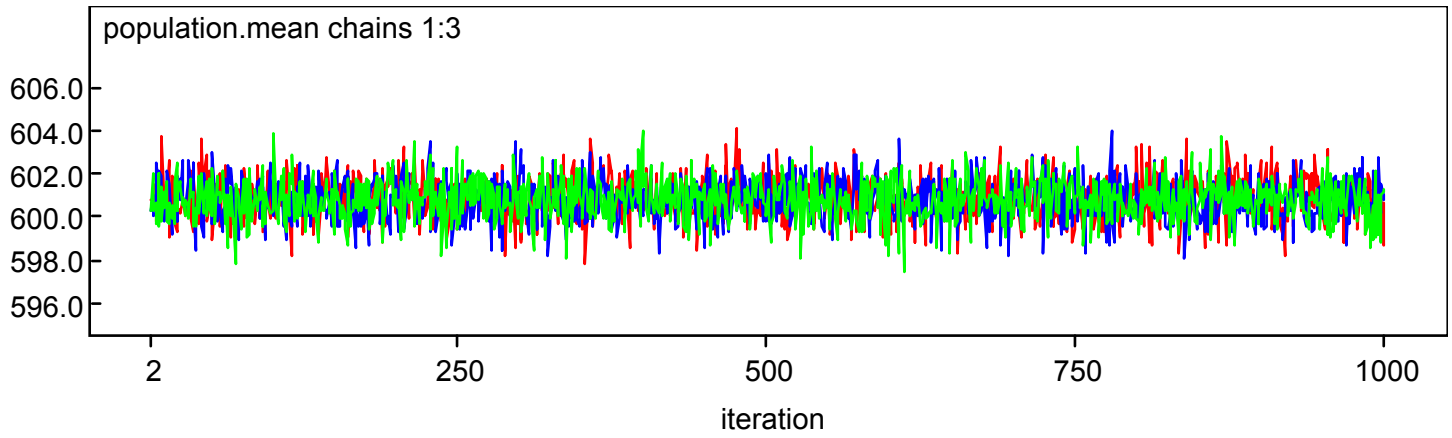
Markov chain Monte Carlo (**MCMC**)

- Class of algorithms for sampling from probability distributions based on a Markov chain
- The desired distribution is the chains equilibrium or stationary distribution. The state of the chain after a large number of steps is then used as a sample of the desired distribution.



Bayesian computation

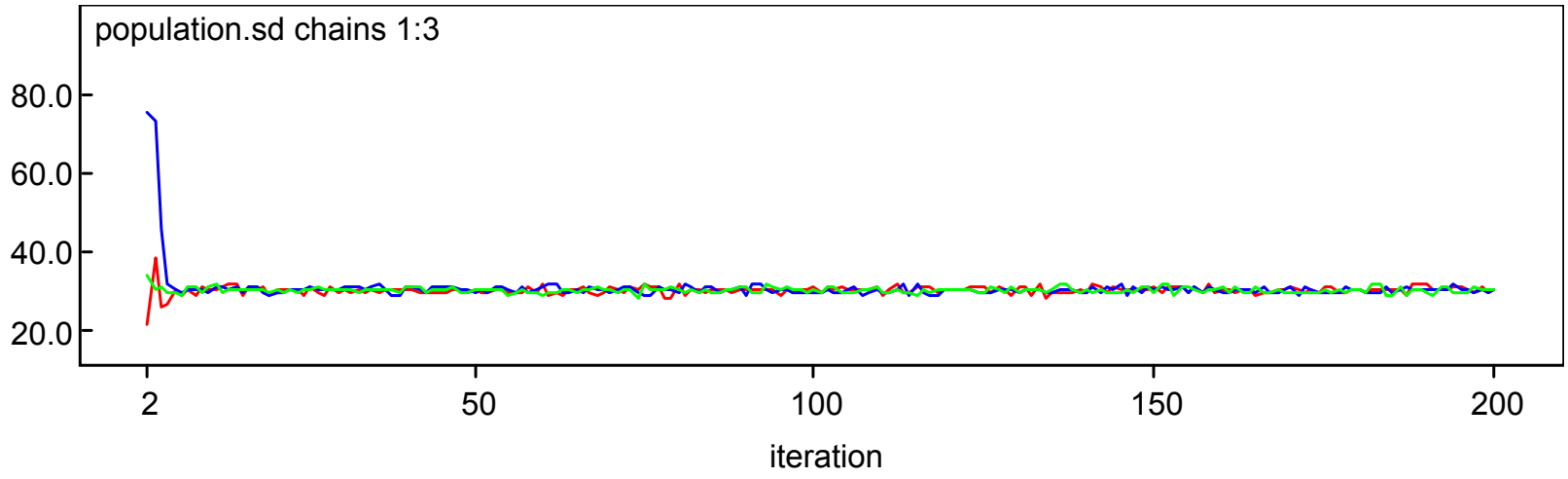
- MCMC: Stochastic algorithm to produce sequence of dependent random numbers (= Markov chain)
- Converge to equilibrium distribution (usually)
- Equilibrium distribution = desired posterior distribution (if algorithm constructed well)



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)
- Inspect the chains and the Rhat values



Bayesian computation

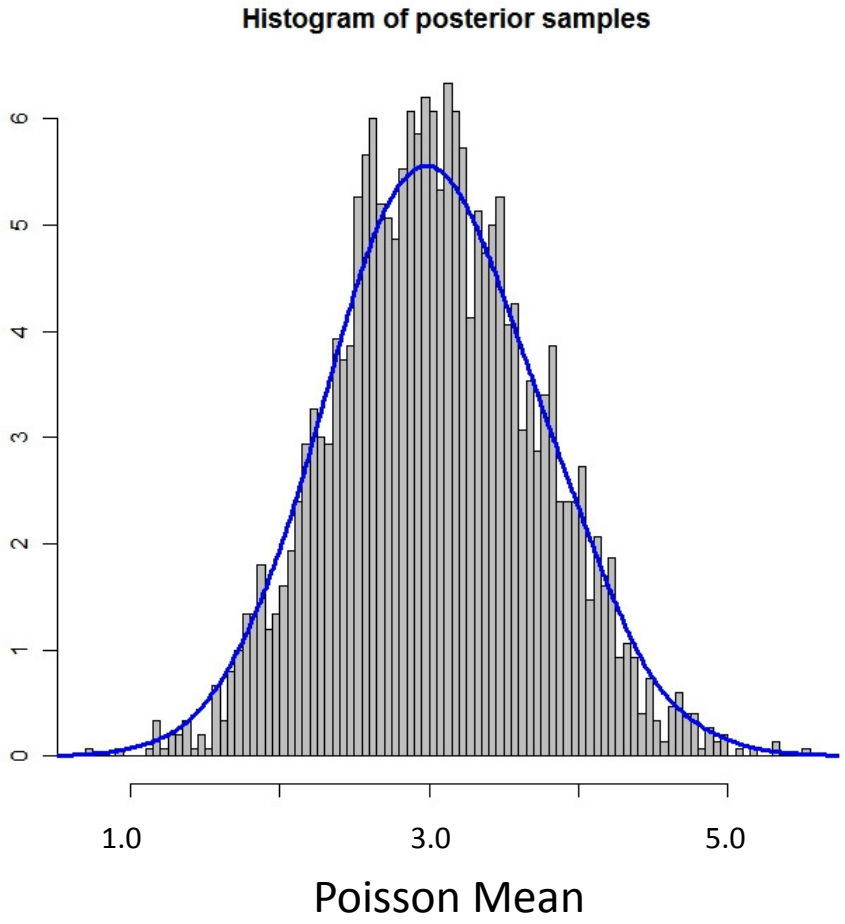
Poisson example

```
.P
[1] 2.5265 3.4088 3.3885 4.3482 3.3850 5.3311
[7] 2.5042 3.3593 5.0580 3.3880 3.3688 2.3793
[13] 2.9935 4.2831 3.4827 3.4632 1.9765 2.7186
[19] 4.1579 3.8605 3.4488 3.3914 2.9474 3.1444
...
[2983] 4.3866 3.3265 3.3121 5.2337 4.3255 2.7912
[2989] 5.0446 3.3584 3.3839 4.4920 4.4068 5.6202
[2995] 3.3844 3.5067 3.4212 4.5759 3.2485 2.2362
```

Bayesian computation

Poisson example

```
.P
[1] 2.5265 3.4088 3.3885 4.
[7] 2.5042 3.3593 5.0580 3.
[13] 2.9935 4.2831 3.4827 3
[19] 4.1579 3.8605 3.4488 3
...
[2983] 4.3866 3.3265 3.3121
[2989] 3.0446 3.3584 3.3839
[2995] 3.3844 3.5067 3.4212
```

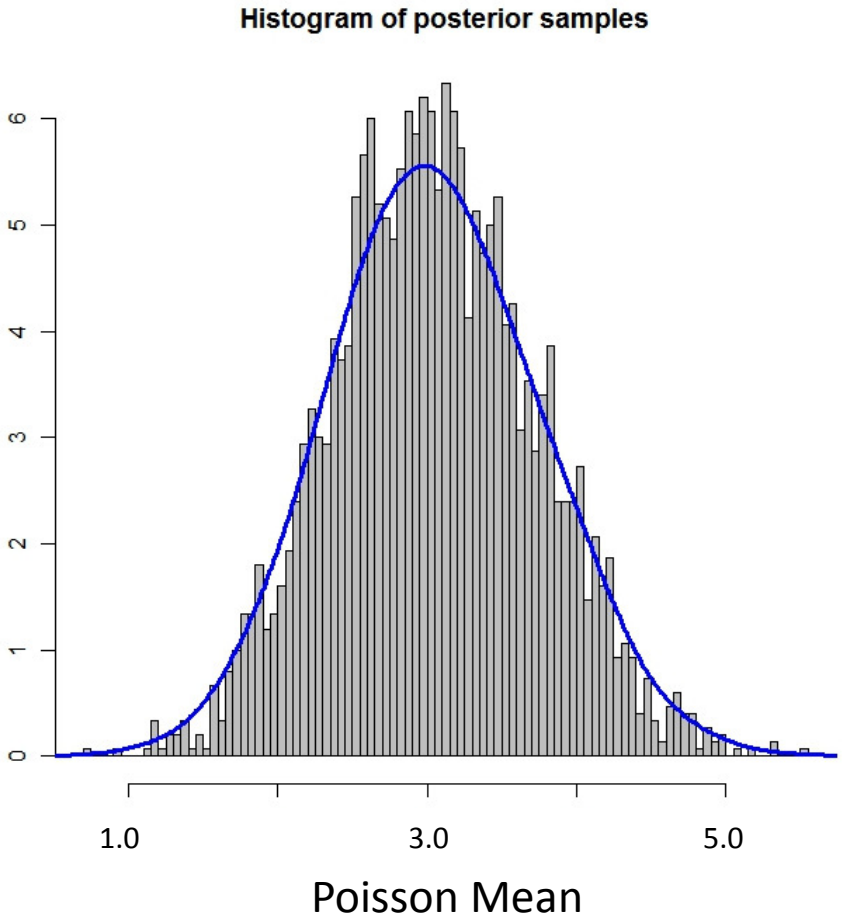


Bayesian computation

Poisson example

```
.P
[1] 2.5265 3.4088 3.3885 4.
[7] 2.5042 3.3593 5.0580 3.
[13] 2.9935 4.2831 3.4827 3
[19] 4.1579 3.8605 3.4488 3
...
[2983] 4.3866 3.3265 3.3121
[2989] 3.0446 3.3584 3.3839
[2995] 3.3844 3.5067 3.4212
```

```
> mean(lambda)
[1] 3.1231
> sd(p)
[1] 1.3637
> quantile(p, probs =
c(0.025,0.975))
2.5% 97.5%
1.7146 4.7385
```



Bayesian computation

The BUGS project

- Boost in Bayesian statistics initially *not in biology*
- 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, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS—A Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10: 325–337.

Bayesian computation

Gibbs Sampler

- A MCMC algorithm to obtain a sequence of observations which are approximated from a multivariate probability distribution (i.e. from the joint probability distribution of two or more random variables)
- Special case of the Metropolis–Hastings algorithm
- The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution.

Bayesian computation

Gibbs Sampler

Suppose we want to obtain *k* samples of $X = \left(x_1, x_2, ..., x_n \right)$

from a joint distribution: $p\left(x_1, x_2, ..., x_n \right)$

Denote the i th sample by $X^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)} \right)$

1. Begin with some initial value for each variable: $X^{(0)}$
2. For each i sample from the distribution, sample from the conditional distribution:

$$p\left(x_j \mid x_1^{(i)}, ..., x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, ..., x_n^{(i-1)} \right)$$

3. The samples approximate the joint distribution of all variables and the marginal distribution of any subset of variables can be approximated by examining the samples for that subset of variables.

Bayesian computation

The awesome part:

You don’t really need a deep understanding of this to use BUGS/JAGS efficiently!

Bayesian computation

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.mrcbsu.cam.ac.uk/bugs/winbugs/contents.shtml
- **OpenBUGS:** www.openbugs.info/w/
- **JAGS:** mcmc-jags.sourceforge.net/

Bayesian computation

The BUGS 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.*

Bayesian computation

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

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

- 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

Why I’ve become a Bayesian

And why you might want to as well…

Why I've become a Bayesian

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 I've become a Bayesian

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 I've become a Bayesian

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 biologists
- Enforces understanding of model

Why I’m not a real Bayesian

- 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

Thoughts on analysis paradigms

- 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 very large data sets!)
- BUGS language is great to actually understand your models
- Stay tuned: in the future, there will (hopefully) be better MCMC (STAN, ADMB?) and even likelihood software for complex models, e.g., HMs