Introduction to statistical inference

$$L(\theta \mid y) = \prod_i p(y_i \mid \theta)$$

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

Marc Kéry

Workshop @ MSU

22–26 July 2024

Statistical modeling in a nutshell

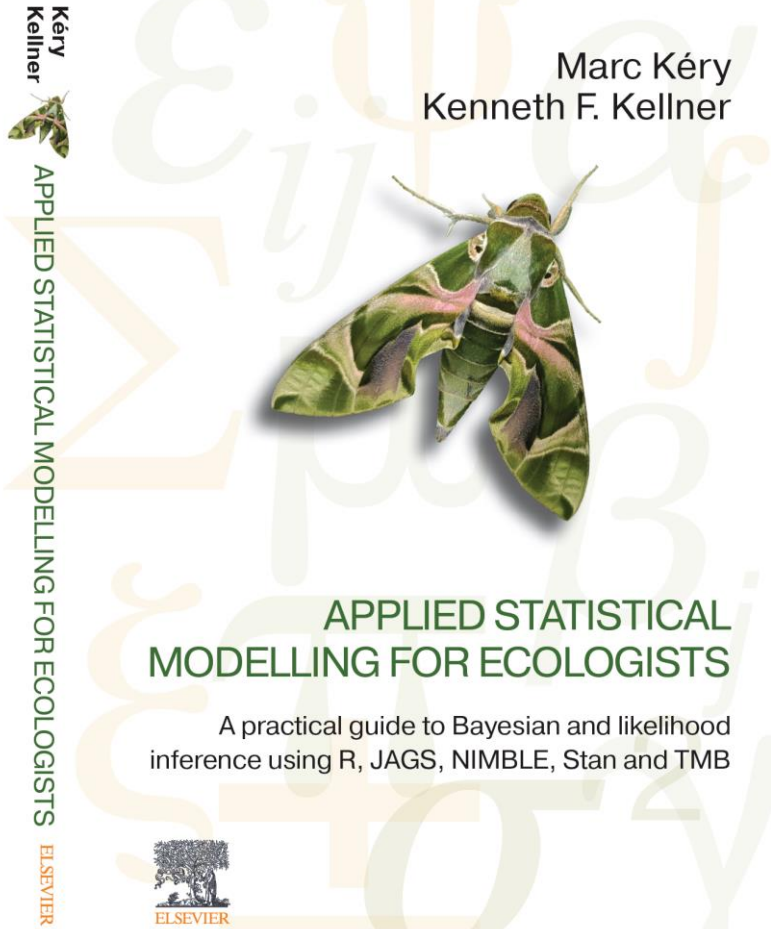
- Stat. model describes chance process that could have produced our data
- View our data set as just one out of a myriad of other possible data sets that same process could have produced
- With model plus data can make statistical inferences, i.e., infer features of hypothetical data-generating chance process
- Statistical modeling: building of model and its analysis using e.g. maximum likelihood or Bayesian posterior inference
- Topics of course and ASM book: how to build and analyse statistical models

Outline

- (1) Probability as the basis for statistical inference
- (2) Random variables and probability distributions
- (3) Statistical models and their usages
- (4) The likelihood function
- (5) Classical inference by maximum likelihood
  - Example of one- and two-parameter models
  - Computation of standard errors and confidence intervals
- (6) Bayesian inference using posterior distributions
  - Probability as a general measure of knowledge
  - Posterior distributions and prior distributions
- (7) Bayesian computation
  - Monte-Carlo integration
  - Markov chain Monte Carlo (MCMC)

- ***Thou shalt understand both ML and Bayes !***

For (even) more details, see Chapter 2 in the book



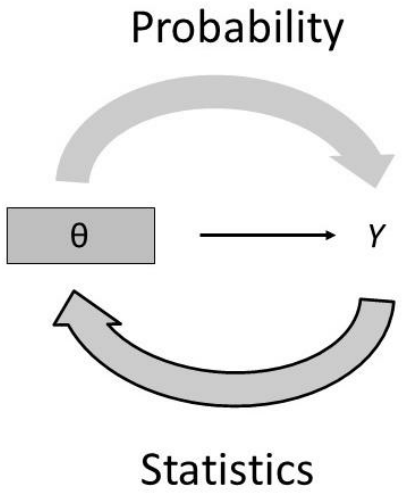
Probability as the basis for statistical inference

Probability as the basis for statistical inference

- World & life is uncertain
- Very few things perfectly known or completely invariable such that we can predict and understand them perfectly
- Need to draw conclusions, make decisions, or learn from observations in the face of resulting uncertainty
- Probability: branch of mathematics dealing with chance processes and their outcomes
- Extension of logic from certain events to all events in life
- Basis for statistical modeling and inference

Duality of probability and statistical inference

- Based on probability, but “goes the other way round” (Link & Barker, 2009)



- Probability takes model as a given, studies features of the outcome
- Statistical inference takes data, assumes model and infers features of the data-generating process (i.e., the model)

Random experiments

- "Experiment": Observing/measuring something unpredictable,
  not necessarily manipulative
- Single observation of chance process: *outcome*
- All possible outcomes of an experiment: *sample space S*
- Examples loved by statisticians:
  - Coin toss: $S = \{H, T\}$
  - Roll of a die: $S = \{1, 2, 3, 4, 5, 6\}$
- Ecological examples:

  (1) Choose an animal or plant population, select an individual and measure wing length, mass, color morph, number of parasites, sex, age …

  Individuals typically considered sample of some biological population about which we infer things.

Random experiments, ctd.

- Ecological examples *ctd.*:

  (2) Choose a sample of survey sites or local populations and assess species presence/absence, abundance, proportion with some disease, or other assessment at level of "site".

  Sites again assumed to be part of statistical population of sites (e.g., "all sites in region, country") about which we want to infer things

- Sample spaces (*S*) denote all possible measurements of a rand. experiment:

  (1) Wing length measurement between 80 and 120 cm, body mass of 500 and 1500 g, red/brown/grey/black morph, non-negative integers, male/female, juv/ad., ….

  (2) Presence/absence, non-negative integers, proportion between 0 and 1, …

Random experiments, ctd.

- Outcome of random experiment will vary when we repeat it, when multiple persons conduct it
- Due to spatial or temporal variability in the statistical population, measurement error, ….
- Hypothetical replicate outcomes of random experiments subject to sampling variation
- Quantified by sampling variance = squared SE, or by CI
- Hard concept to grasp –

  How can we infer the variability of a random experiment when all we have is a single number ???

Sample space, outcomes and events

- Outcome: one particular observation in a random experiment
- Sample space (*S*): all possible outcomes
- Event (*E*): Subspace of sample space *S*
- Examples of events:
  - Two heads in two coin tosses
  - A number > 4 in the roll of a die
  - Wing length greater than 10 cm
  - Mass between 600 and 610
  - Color morph “red”
  - More than 10 ectoparasites

Sample space, outcomes and events, ctd.

- More events:
  - Site is occupied (species present)
  - Number of black and grey vipers at a site

- Must define “event” such that is directly relevant to research question or management task.

Probability functions

- Use probability to describe effects of chance on the outcomes obtained, or the events observed, in a random experiment
- Probability function for an event *A*, *P*(*A*):

  (1) *P*(*A*) is a number between 0 and 1,
  denoting impossible or certain event *A*

  (2) Prob of entire sample space (*P*(*S*)) = 1
  (outcome must be in sample space)

  (3) Prob. of union of mutually exclusive events (either or both *A* and *B*) is the sum of both individual events: $P(A \cup B) = P(A) + P(B)$

- Statistics built on these 3 axioms of probability !
- Hold regardless of frequentist or Bayesian view of probability
- Host of other probability rules can be deduced from them

Random variables and probability distributions

Random variables (RVs)

- Random variable (RV): real-valued function defined on sample space of a random experiment
- Also, "uncertainty quantity" (Lindley, 2002)
- Quantitative description of the effects of chance on the feature of the experiment that is relevant to us
  - e.g., proportion grey and black vipers
  - number of grid cells occupied vs. unoccupied
- Key concept in probability and statistical modeling
- Data set and anything else affected by random variation is considered a RV
- May also be parameters:
  - random effects
  - priors in Bayesian analysis
- RVs described by probability distributions
- RVs/prob. distributions essential building blocks of statistical models

RVs and their realized value

- RV *Y* (upper case) describes an abstract stochastic process, produces outcome or realization *y* (lower case)
- Can be confusing at first, owing to different usage of term
  - e.g., body mass of peregrine may be RV *Y*
  - when we measure mass of first bird, get measurement *y*, which is our RV (= realized value of RV *Y*)
- When we take multiple measurements (e.g., repeated on same bird or multiple birds), distinguish multiple RVs, Y_1, Y_2, …., Y_n, each with associated realized RV y_1, y_2, …, y_n .
- Can distinguish RV *X* (sex or age of bird) from RV *Y* (body mass), such that each bird may have realized RVs such as (male, 589 g), (female, 968 g), ….
- In an analysis, have at least as many RVs as data points (*n*)

Discrete and continuous RVs

- Two broad classes:
  - Discrete RVs: ~ nominal, ordinal scales of measurement
  - Continuous RVs: ~ metric, interval scales
- Examples discrete RVs:
  - Labels or names, e.g. sex, color, geographic stratum, population, demographic states (dead/alive …)
  - Counts, e.g., #female nestlings in a nest of six,
    #birds counted at a point count station
- Examples continuous RVs:
  - Measurements of lengths, durations, mass, proportions, …
  - (in practice, also discrete due to finite measurement accuracy, but this is ignored)

Description of RVs by probability distributions

- A statistical model contains a probabilistic description of all its RVs (data, latent variables)
- Probability distributions describe how total probability of 1 is distributed among all possible realizations of a RV
- Probability mass function (PMF) for discrete RVs
- Probability density function (PDF) for continuous RVs
- Cumulative distribution function (CDF) for both

Probability distributions for discrete RVs

- Probability mass function (PMF) gives *probability* of every possible outcome of a discrete RV *Y*

$$f(Y) = P(Y = y)$$

- Function value always between 0 and 1
- Sum is 1
- Examples: Bernoulli/binomial, Poisson; see below

Probability distributions for continuous RVs

- Probability density function (PDF) gives *probability density* of every possible outcome of a continuous RV *Y*
- Density is non-negative value that is limit of area of rectangle with base (y – δ, y + δ) as δ -> 0.
- Prob. (but not density !) of any given value of *Y* is 0, and density (but not prob.) of some values may be >1.
- Can get probability of a range of values (y_1, y_2) by integrating density

$$P(y_1 < y < y_2) = \int_{y_1}^{y_2} f(Y)dY$$

- Integral over entire support of PDF is 1
- Examples: normal, *t*, see below

Named distributions

- Many chance processes so common and so well-studied that they have been given a name
- e.g., Bernoulli, binomial, Poisson, normal/Gaussian or uniform distributions
- Many, many, MANY more ….
  (e.g., Rayleigh, Cauchy distributions ….. Pareto … von Mises …)
- But can do surprising things when you know just a few of them !

  …. e.g., build all of linear models, generalized linear models, mixed models
- Have each a small number of constants (typically 1–3), which provide specific form: parameters

Ex. discrete distribution: Poisson PMF

$$f(y \mid \lambda) = P(Y = y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- "**probability** of observing value *y* of the RV *Y*, given parameter *λ*"
- Single parameter *λ*, which is mean and variance
- To better understand: Fill in values and observe what comes out
- E.g., what is *P*(*Y* = 1 | λ = 2) ?

$$\lambda^y e^{-\lambda} / y! = 2^1 e^{-2} / 1! \approx 0.27$$

- E.g., what is *P*(*Y* = 0 | λ = 2) ?

$$2^0 e^{-2} / 0! \approx 0.14$$

- In R:

```
pmf <- function(y,lambda){lambda^y*exp(-lambda)/factorial(y)}
pmf(1,2)  ;  pmf(0, 2)
dpois(1, 2)  ;  dpois(0, 2)
```

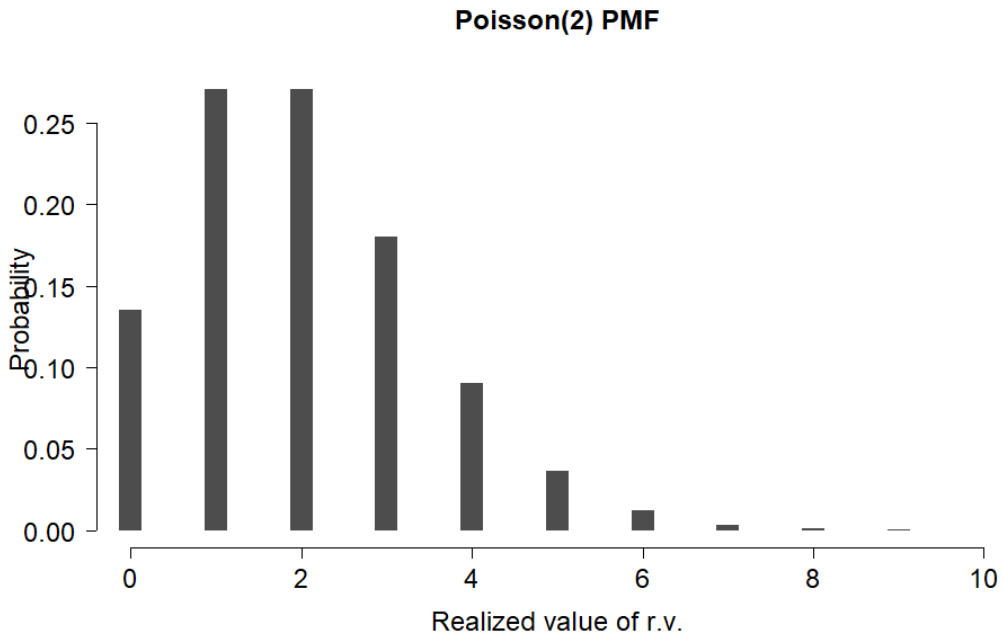
Poisson PMF, ctd.

- E.g., what is $P(Y = 2 \text{ or } 3 \mid \lambda = 2)$?

$$2^2 e^{-2} / 2! + 2^3 e^{-2} / 3! \approx 0.27 + 0.18 = 0.45$$

- The full PMF (for λ = 2):

```
plot(dpois(0:10, 2), type = 'h', lend = 'butt', lwd = 10)
```



Poisson PMF, ctd.

- Many short-hands for probability distributions, e.g.,
  - $y \sim f(Y \mid \lambda)$
  - $y \sim f_Y(Y \mid \lambda)$
  - $y \sim \text{Poisson}(\lambda)$
  - $y \sim \text{Pois}(\lambda)$
  - $y \sim \text{Poisson}(Y \mid \lambda)$
  - $y \sim \text{Pois}(Y \mid \lambda)$
  - `y ~ dpois(lambda)` BUGS language (see later)
  - `dpois(1, 2)` PMF in R, see below
  - `glm(y~1, family=poisson)` to fit associated GLM in R

- … but they all just mean this:

$$f(y \mid \lambda) = P(Y = y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Poisson distribution in R: demo in R

```
# Look up help text for the Poisson
?dpois                  # Also '?Poisson'

# Poisson probability mass function (PMF) evaluated for y = 0
dpois(0, lambda = 2)    # Probability of getting a value 0
dpois(0, lambda = 2, log = TRUE) # Same on log scale

# Get same density 'by hand' to emphasize dpois() is just shorthand !
lam <- 2; y <- 0
(lam^y)*exp(-lam) / factorial(y) # Probability of getting a value 0

# Cumulative distribution function (CDF), or just 'distr. function'
ppois(3, lambda = 2)    # Probability of getting a value of 0, 1, 2 or 3
sum(dpois(0:3, 2))      # Same, summing up probs 'by hand'

# Quantile function: value of y for which CDF(y) has a certain value
qpois(0.85, lambda = 2)

# Random number generator (RNG) function
set.seed(2016)            # Set seed if want same numbers as we have
rpois(n = 10, lambda = 2)# 10 Poisson(2) random numbers
```


Ex. continuous distribution: normal PDF

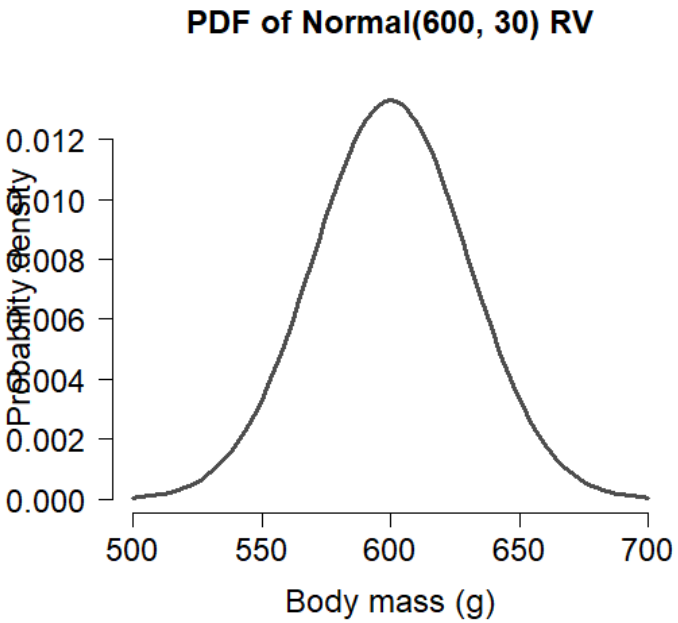
$$f(y \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

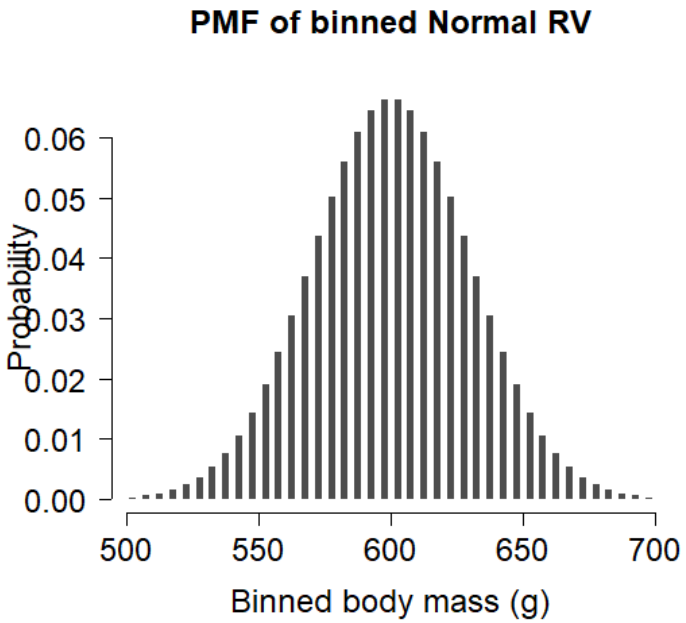
- "**probability density** of value y, given parameters μ and σ^2"
- Parameters μ (mean) and σ^2 (variance) .. or SD σ … or precision $1 / \sigma^2$
- To understand: Fill in values and observe what comes out
- E.g., what is probability density of 590,
  i.e., $f(Y = 590 \mid \mu = 600, \sigma^2 = 30^2)$?

$$(1/(30)\sqrt{2\pi}) \times \exp\left(-\frac{(590-600)^2}{2\sigma^2}\right)$$

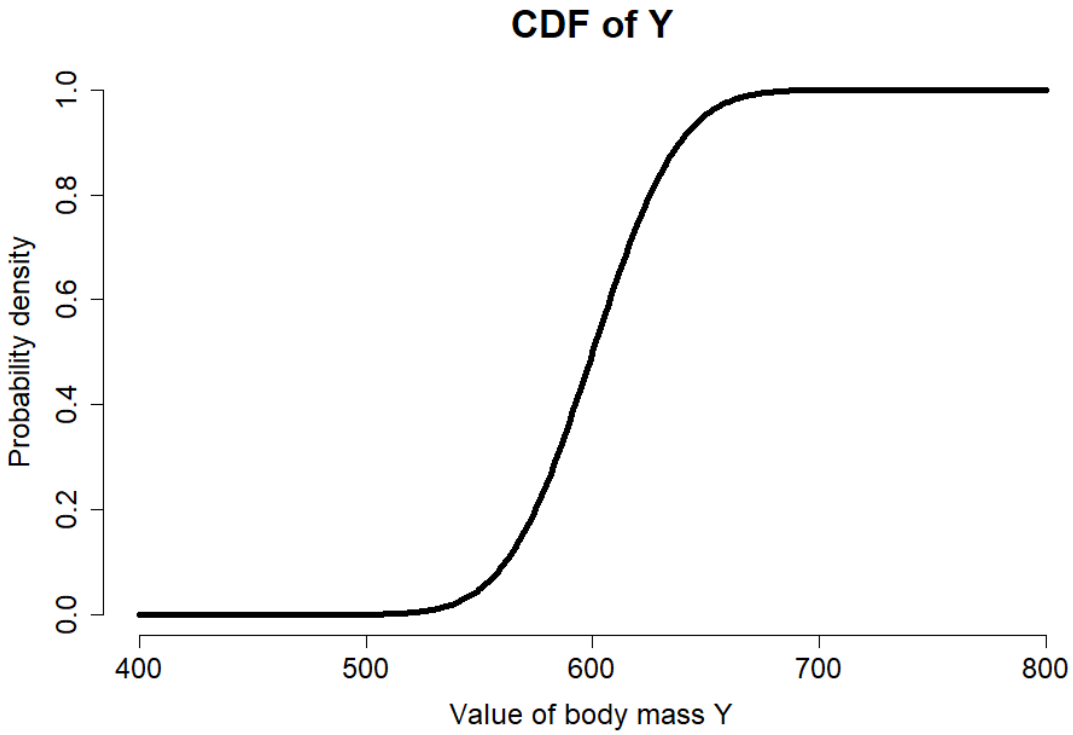
Normal PDF, ctd.

- Probability of any given value *y* is equal to 0, although probability density is non-negative !
- To get a probability from a density, must integrate
- Ex. of normal density and PMF of same with 5g bins





Normal cumulative distribution function (CDF)



- Note: derivative of CDF is the PDF

Normal distribution in R: demo in R

```
?dnorm  # Also '?Normal': Look up the help text for the Normal
# Gaussian, or normal, probability density function (PDF) for y = 650
dnorm(650, mean = 600, sd = 30)
dnorm(650, mean = 600, sd = 30, log = TRUE)     # Same on log scale

# Get the density 'by hand' to emphasize dnorm() is just a shortcut
mu <- 600; sig <- 30; y <- 650
1/(sig*sqrt(2*pi)) * exp(-(y - mu)^2 / (2 * sig^2))

# Cumulative distribution function (CDF), or just 'distr. function'
pnorm(600, mean = 600, sd = 30) # Prob. of value between -Inf and 600
# Next is same, but integrating the PDF 'by hand'
f <- function(x, mean, sd) dnorm(x, mean = mean, sd = sd)
integrate(f, lower = -Inf, upper = 600, mean = 600, sd = 30)
plot(pnorm(seq(400, 800, by = 0.01), 600, 30), type = 'l', lwd = 5, ylim = c(0, 1),
xlab = 'Value of Y', ylab = 'Prob. density', main = "CDF of Y", frame = FALSE)

# Quantile function: value of y for which CDF(y) has a certain value
qnorm(0.95, mean = 600, sd = 30)

# Random number generator (RNG) function
set.seed(2016)
rnorm(n = 10, mean = 600, sd = 30)# 10 Normal(600, 30) random numbers
```


Normal distribution in R: demo 2 in R

```
# Compute probabilities for classes of binned Normal rv
limits <- seq(500, 700, by = 5)        # 5g classes
midpts <- seq(500 + (5/2), 700 - (5/2), by = 5)    # ...

limits <- seq(500, 700, by = 0.1)      # 0.1g classes
midpts <- seq(500+(0.1/2), 700-(0.1/2), by = 0.1) # ...

cumProb <- pnorm(limits, mean = 600, sd = 30)
probs <- diff(cumProb)

# Plot probability density function (PDF) of Normal(600, 30)
par(mfrow = c(1, 2), mar = c(6,6,5,2), cex.lab = 1.5, cex.axis = 1.5,
cex.main = 1.5)                  # Fig. 2-3
curve(dnorm(x, mean = 600, sd = 30), 500, 700, xlab = 'Body mass (g)',
ylab = 'Probability density', main = 'PDF of Normal(600, 30) r.v.', type =
'l', lwd = 3, col = 'gray30', frame = FALSE, las = 1)

# Plot probabilities for 5g-binned mass random variable
plot(midpts, probs, xlab = 'Binned body mass (g)', ylab = 'Probability',
main = 'PMF of binned Normal r.v.', type = 'h', lend = 'butt', lwd = 10,
col = 'gray30', frame = FALSE, las = 1)
```

Modeling a parameter

- PDF/PMF (and CDF) provide complete description of RV
- Often, want to characterize distribution in simple ways
- Especially central tendency or location (expectation, mean) and dispersion or spread (variance)
- Mean often corresponds to a parameter in a distribution, e.g., in normal, Bernoulli/binomial, Poisson
- Can "model" these parameters to explore patterns in RVs related to covariates = explanatory variables
- Replace parameter by a function of covariates, often on a transformed scale (link function)
- e.g.,

$$\log(E(N_i)) = \log(\lambda_i) = \alpha + \beta x_i$$

$$\text{logit}(p_{i,t}) = \log(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_{i,t}^{(2)}$$

Modeling a parameter, ctd.

- Replace the original parameter (e.g., Poisson λ or Bernoulli *p*) with new parameters α and β, and α, β_1 and β_1, respectively
- Almost always linear models to describe effects of covariates
- NOT necessarily straight lines, but models where parameters appear linearly, i.e., as sums
- Can have very wiggly graphs of a function that is linear in parameters, e.g., polynomials
- Rarely can also directly model a dispersion parameter, e.g., sd or variance in a normal response

$$\log(\sigma_i^2) = \alpha + \beta x_i$$

Section summary

- RV concept hugely important in applied statistics …

  … yet, as ecologists we hardly ever learn about RVs

- RV: "uncertain quantities"
- Statistical distributions describe RVs, both in terms of mean (or average tendency) and in terms of variation
- Can do many calculations with distributions
- Distributions for RV are basic building blocks of our models, regardless of whether have a simple (e.g., LM, GLM) or more complex model (e.g., hierarchical or mixed model)
- Typically, “model” some parameters as a linear function of some covariates and new parameters (α, β above)
- In practice, much of what we do when we “model” is to specify linear models for how covariates affect our parameters, as in

  `glm(y ~ A * x)`

Statistical models and their usages

What is a model ?

- An abstraction or simplification of something complicated
- Any explanation is really a model (because it must always simplify) !
- Every model has a goal:
  - enforce clarity of thought
  - summarize
  - search for patterns
  - understand mechanisms
  - predict
- We are rarely explicit about the goals of our models … though should be !
- Because, for instance, it is impossible to say what is a good model when we haven't decided on the "what for" question
- Very relevant for both goodness-of-fit and much more so for model selection
- e.g., maps:
  geological – street – species distribution – bus lines – topographical ...

What is a statistical model ?

- A mathematical abstraction using concepts and language of probability
- Millar (2011, *Wiley*):
  "A parametric statistical model is a collection of joint density functions p(y; θ)" […for its random variables y and indexed by parameter θ].

- Lee et al. (2017, *CRC Press*):
  "The model … should specify how the data could have been generated probabilistically."

- Thus, use probability distributions to describe all RVs, based on what we just covered, and form the joint density of all data
  (and possibly of other RVs, such as random effects)

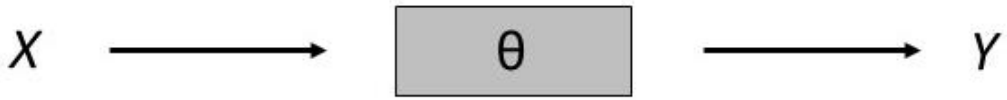
The model as a joint density for our data set

- For illustration assume we had a vector ***y*** with 10 counts, and want to construct a probabilistic model
- Simplest assumption: all observations of the same random variable *Y*
- *Y* independent and identically distributed (iid)
  - Identically distributed: e.g., all Poisson, with identical λ
  - Independence: "y_i contains no information about any other y"
- Joint density is product of densities of each component of the vector of counts **y**

$$f(\mathbf{y} \mid \lambda) = f(\{y_1, \ldots, y_{10}\} \mid \lambda) =$$

$$f(y_1 \mid \lambda) \cdot f(y_2 \mid \lambda) \ldots f(y_{10} \mid \lambda) = \prod_{i=1}^{10} f(y_i \mid \lambda)$$

Schematic of a model



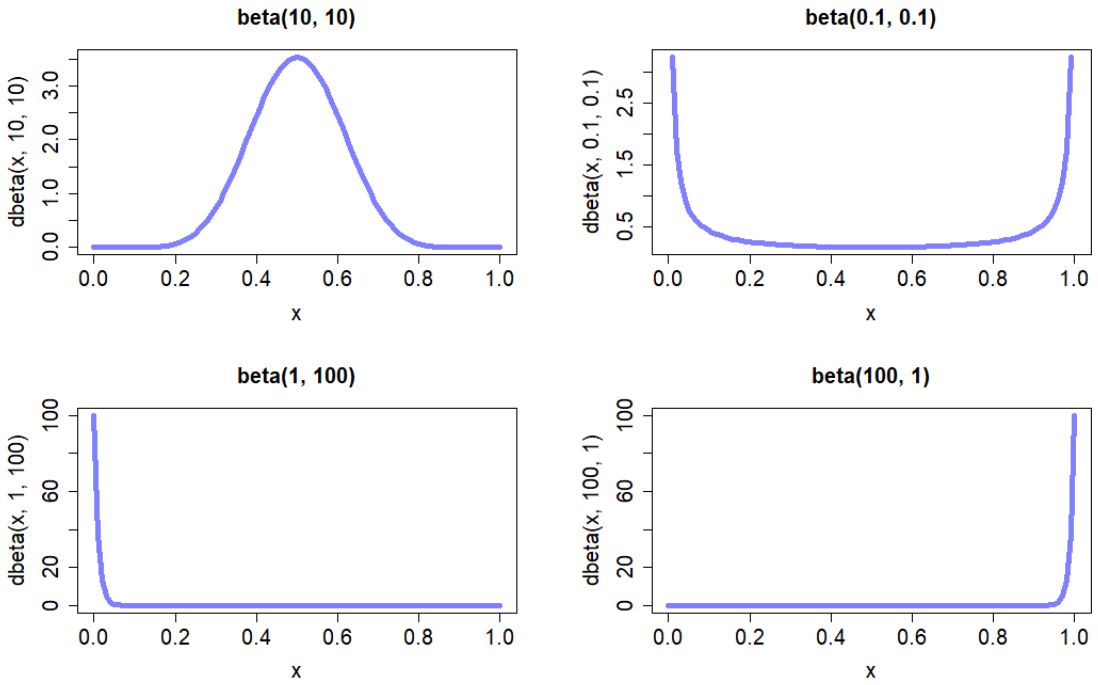
(From Breiman, 2001)

- Grey box: model (abstraction of data-generating process), governed by unknown constants **θ** (i.e., the parameter(s))
- *X*: Input (~ covariates)
- *Y*: Output (~ response, measurement, observed random variables)

- Three important forms of inferences:
  - Estimation of parameters θ (with uncertainty)
  - … of missing responses (= making predictions); with uncertainty
  - … of missing covariates (= covariate 'imputation'); with uncertainty

Parametric statistical modeling: rigid & flexible

- Rigid, since we claim that we know exactly the random variables and how they are connected (e.g., in a hierarchical model)
- But flexible, because we estimate the parameters such that they best fit the observed data, which gives considerable flexibility
- Depending on parameter value, a statistical distribution may look very different, see example of four betas



Confronting models with data for inference

- Using data assumed to be produced by processes represented by our model lets us estimate parameters and doing other statistical inferences
- Two main schools of statistical inference: frequentist/likelihood and Bayesian
- Apparent differences often dramatically exaggerated
- In fact, close relationships:
  - Both based on parametric statistical model (model-based inference)
  - Likelihood function central in both
  - Should understand them both

The likelihood function

Likelihood function

- Key concept in statistical inference
- Formal connection between data and parameters,
- between what we observe, and thus know and what we don't observe, but would like to know
- Millar (2011, *Wiley*):
  "the likelihood function is the (joint) density function evaluated at the observed data, and regarded as a function of **θ** … That is,

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$$

- Provides a measure of relative support for different values of **θ**, in terms of the probability (density) of the data set at hand **y**
- Conceptually simple, but in practice challenging for ecologists
- Algebraically same as joint density of the data, but used in the "opposite direction", as a function of given data

Likelihood function, ctd.

- Likelihood function is NOT a probability density function over parameter space: does not integrate to 1
- Likelihood of an iid sample of data is a product of small numbers

$$L(\theta \mid y) = \prod_i p(y_i \mid \theta)$$

- Working with likelihood directly may cause numerical under- or (rarely) overflow on computer
- Therefore, typically work with the log-likelihood; taking logs turns product into a sum of log-densities

$$\log L(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(f(y_i \mid \theta))$$

Likelihood function, ctd.

- Likelihood function in a hierarchical model is based on joint density of the data and latent variables, i.e., observed and unobserved RVs (= random effects)
- Random effects removed by integration (or summation), leading to marginal or integrated likelihood
- See Chapters 2 in Royle & Dorazio (2008), Royle et al. (2014) and Kéry & Royle (2016) and Chapters 7, 10, 14, 17, 19, 19B in Kéry & Kellner (2024)
- R package `unmarked` works with such integrated likelihoods, as do Programs MARK, PRESENCE (and `lme4`, `glmmTMB` and many others)

Likelihood function summary

- Likelihood function is the connection between observed data and unobserved parameters
- Both frequentist and Bayesian inference use the likelihood as a key ingredient, but in different ways

Classical inference by maximum likelihood

Maximum likelihood

- Most common method for statistical inference in statistics
- Key distinction between frequentist and Bayesian inference is different interpretation and use of ‘probability’
- Frequentist inference uses probability ***only*** as a measure of variability of data and things that we estimate from data (including parameter estimates or confidence intervals)
- Frequentists ***never*** use probability for direct statements about the uncertainty of their estimates as Bayesians do

- (In contrast, Bayesians use probability both as a measure of variability of observable things (i.e., data) and as a measure of uncertainty about unobservable things, such as parameters; see later)

Maximum likelihood, ctd.

- Frequentist, or classical, statistics views probability as the limit of a relative frequency of data (and functions of data) in hypothetical replicate data sets, as their number goes to infinity
- Estimator: method for obtaining from data an estimate of a parameter
- Random variable characteristic of data carry over to estimator
- Thus, a single estimate for a data set at hand viewed as one of an infinitely large population of such estimates that we *could* have gotten in hypothetical replicates of our data set, our study
- Distribution of these estimates is *sampling distribution* of the estimator
- Frequentist uncertainty assessments of estimates refer to the sampling distribution (e.g., sampling variance = squared standard error)

Maximum likelihood, ctd.

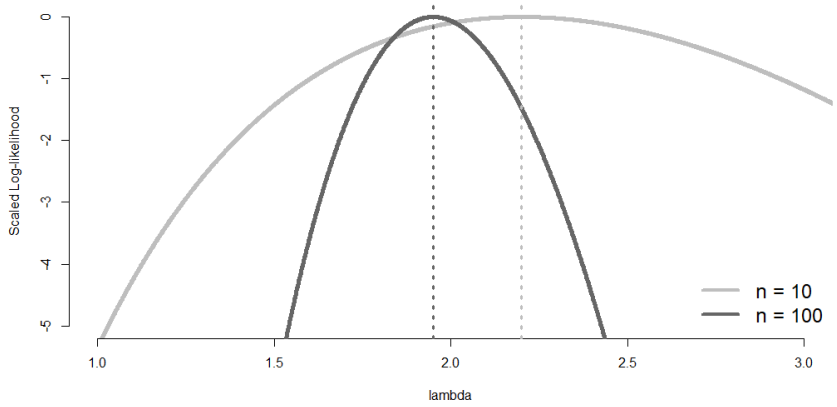
- Narrow sampling distribution means small CI, narrow CI, high precision of the estimate
- Statistical theory lets us make statement about sampling variance *from a single data set*
- (Alternative: resampling methods such as bootstrap evaluate sampling distribution empirically)

Maximum likelihood, ctd.

- Key of method of maximum likelihood: pick those parameter values that maximize likelihood function for observed data set
- Resulting maximum likelihood estimates (MLEs) make observed data the most likely
- Most widespread estimation method in statistics since its invention 100 years ago by Fisher (1922) and others
- Numerous other estimation methods produce MLEs for certain classes of models, e.g., least-squares for normal linear models, IWLS for GLMs
- MLEs can be obtained "automatically": form the likelihood and maximize it (typically numerically with optimization algorithm)
- MLEs have desirable properties, i.e., are "good", in large samples:
  - transformation invariance
  - efficiency
  - consistency
  - asymptotically unbiased and normally distributed around true value

MLEs, SEs, CIs in a single-parameter model

- **Demo in R** with two samples of different size of some counts
- Forming likelihood, log-likelihood and negative log-likelihood functions
- Get MLEs by different methods for finding extremum of function
- Observe how “concentration” of the likelihood function contains information about the precision of estimate
- Present three methods of obtaining uncertainty assessments for MLEs:
  - Profiling
  - Asymptotic normality of the MLEs, “inverting the Hessian”
  - Bootstrapping



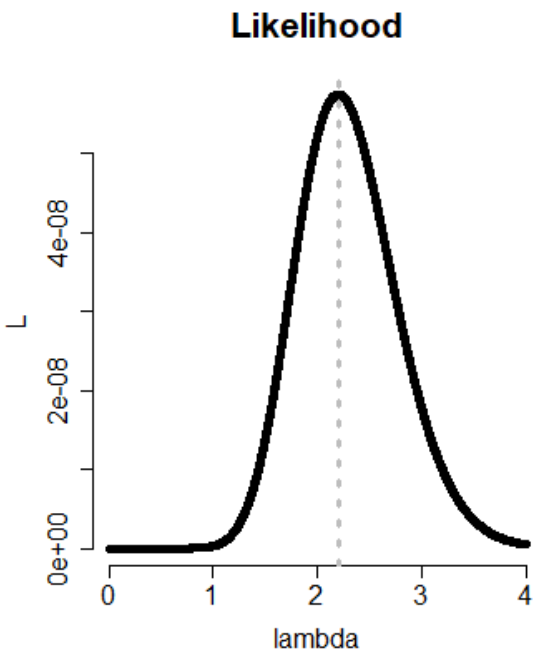
MLEs, SEs, CIs in a single-parameter model, ctd.

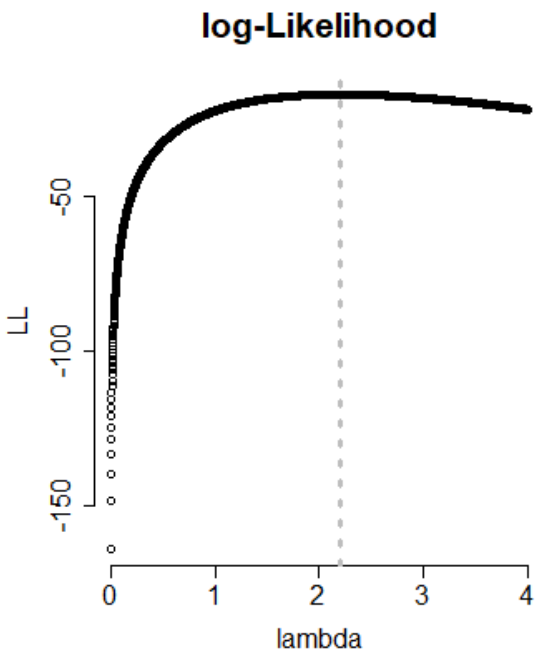
- Defining R functions for *L*, *LL* and *-LL*

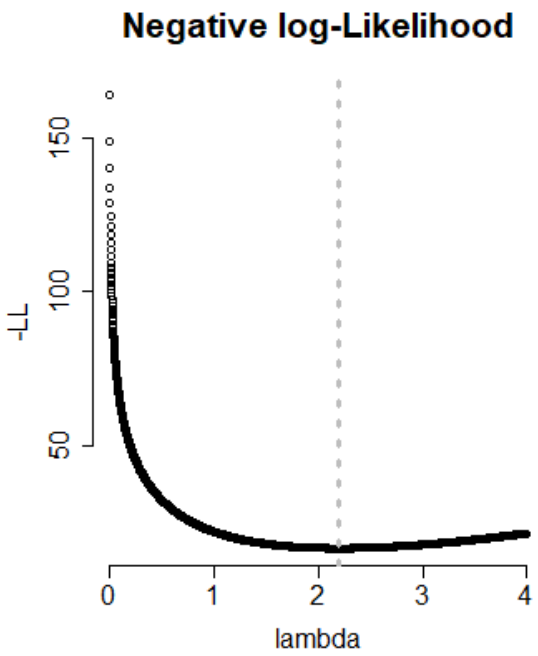
```
# Define likelihood function in R
L <- function(lambda, y) {
  Li <- dpois(y, lambda) # likelihood contribution of each data point i
  L <- prod(Li)      # Likelihood for entire data set is a product
  return(L)
}

# Define log-likelihood function in R
LL <- function(lambda, y) {
  LLi <- dpois(y, lambda, log = TRUE) # log-likelihood contribution of i
  LL <- sum(LLi)     # Log-likelihood for entire data set is a sum
  return(LL)
}

# Define negative log-likelihood function in R
NLL <- function(lambda, y) {
  LL <- dpois(y, lambda, log=TRUE) # log-likelihood contribution of i
  NLL <- -sum(LL)    # *neg* log-likelihood for entire data set is a sum
  return(NLL)
}
```


MLEs, SEs, CIs in a single-parameter model

- Equivalence of solutions for *L*, *LL* and *-LL*







MLEs, SEs, CIs in a single-parameter model, ctd.

- Function minimization using R function `optim()`:
  need R function, data set and inits

```
set.seed(2)
y1 <- rpois(10, lambda = 2)  # Small data set (n = 10)

# Define negative log-likelihood function in R (as before)
NLL <- function(lambda, y) {
  NLL <- -sum(dpois(y, lambda, log=TRUE))
  NLL
}

# Get MLE by minimisation of function for data set
inits <- c('lambda' = 1)
out <- optim(inits, NLL, y = y1) # Optimize function for y1 over lambda
out
(MLE <- out$par)                 # Grab the MLEs
```


MLEs, SEs, CIs in a single-parameter model, ctd.

- Output from `optim()`

```
> out
$par
lambda
   2.2

$value
[1] 16.67301

$counts
function gradient
      30       NA

$convergence
[1] 0

$message
NULL

> (MLE <- out$par)                    # Grab the MLEs
lambda
   2.2
```


MLEs, SEs, CIs in a single-parameter model, ctd.

- Getting asymptotic SEs and Wald-type CIs
- Theory says that sampling distribution of MLEs for large sample size (i.e., asymptotically) becomes normal:

$$\hat{\theta} \sim Normal(\theta, I(\hat{\theta})^{-1})$$

- *I* is called observed Fisher information: more information means more precise estimates, means likelihood function more peaked around MLEs
- Variance is inverse of the precision (i.e., precision^{-1})
- SE is square root of precision^{-1}

MLEs, SEs, CIs in a single-parameter model, ctd.

- Fisher information is negative of 2nd partial derivative of the log-likelihood (*LL*) function with respect to parameters, when evaluated at the MLE
- 2nd derivative of LL: Hessian matrix **H**
- Here, *I* is scalar, but for s-parameter model, will become the s x s Fisher information matrix

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}(\hat{\boldsymbol{\theta}}) = -l''(\hat{\boldsymbol{\theta}})$$

- To better understand Fisher information (*I*), note:
  - First derivative of *LL* gives slope of curve at the MLE
  - Second derivative of *LL* gives curvature, i.e., rate of change of slope
  - At the MLE, function is negative, since slope changes from positive to negative
  - Therefore, Fisher information is the negative of the function value of **H**

MLEs, SEs, CIs in a single-parameter model, ctd.

- Numerical estimate of Hessian can be obtained when we set `optim(…, hessian = TRUE)`
- Thus, we can get estimate of the variance-covariance (VC) matrix by taking inverse of the Hessian and negating the result

$$\mathbf{VC}(\hat{\boldsymbol{\theta}}) = [\mathbf{I}(\hat{\boldsymbol{\theta}})]^{-1} = -\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}$$

However, with `optim()` we always work with the negative *LL*

- Thus, get the VC matrix simply by inverting the Hessian given by `optim()`:

$$\mathbf{VC}(\hat{\boldsymbol{\theta}}) = \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}$$

- Therefore, may say *"obtain the variance by inverting the Hessian"*

MLEs, SEs, CIs in a single-parameter model, ctd.

- "Inverting the Hessian" is by far most common method of obtaining asymptotic SEs around MLEs
- e.g., in `unmarked` (and no doubt also in MARK or PRESENCE)
- Can obtain Wald-type confidence intervals by as MLE plus/minus *z* times asymptotic SE, e.g., for 95% CI (with *z* = 1.96)

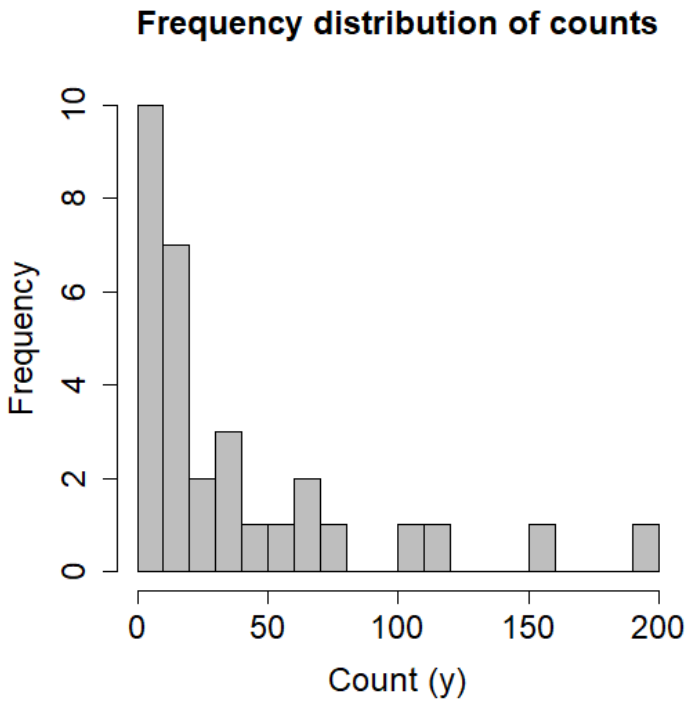
$$\hat{\theta} \pm 1.96 \times ASE(\hat{\theta})$$

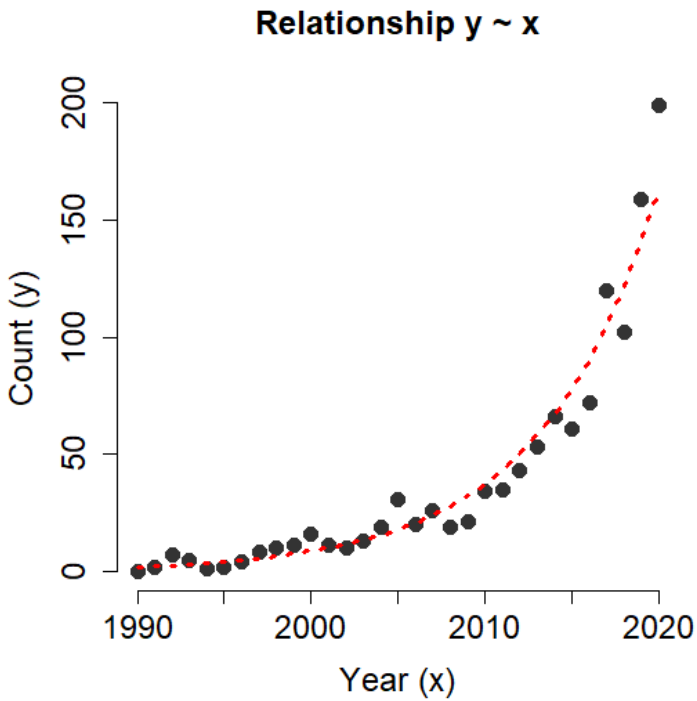
MLEs, SEs, CIs in a two-parameter model

- Counts of nesting pairs of Bee-eaters (*Merops apiaster*) in Switzerland, 1990–2020

MLEs, SEs, CIs in a two-parameter model

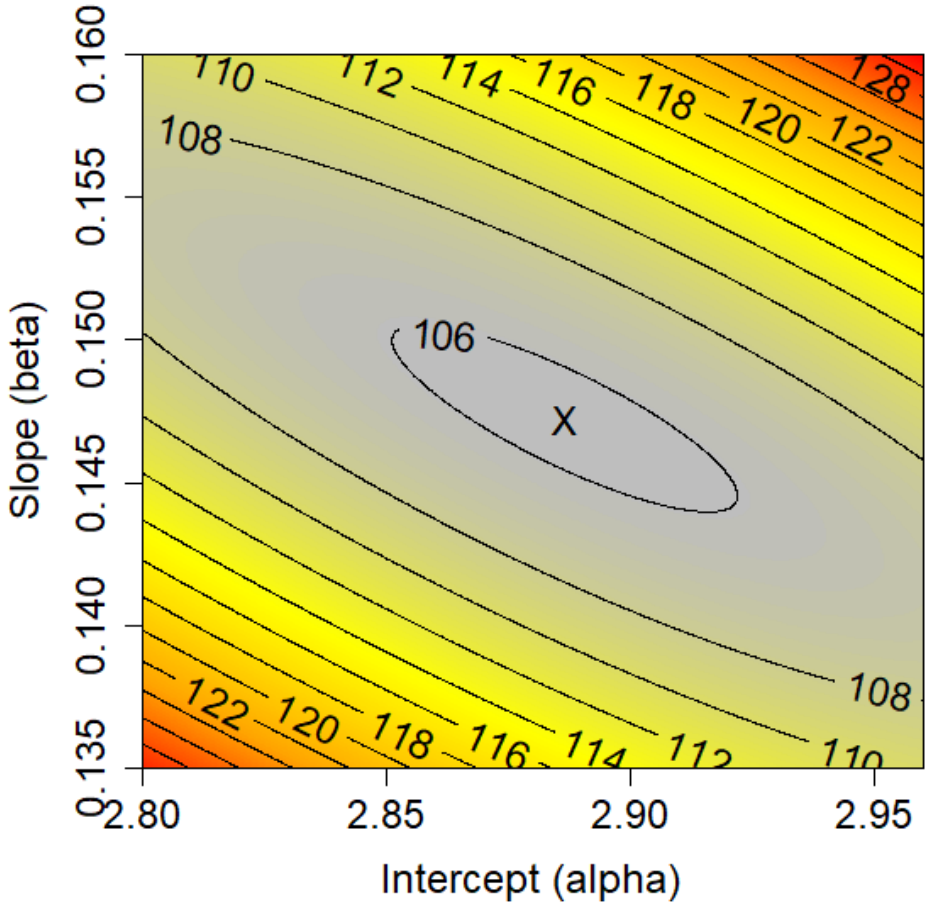
- Demo in R ….





MLEs, SEs, CIs in a two-parameter model

- A negative log-likelihood landscape with the minimum marked as an X



vogelwarte.ch

Bayesian inference using posterior distributions

The Bayesian way of learning from data

(From Breiman, 2001)

- Start exactly as before (i.e., for ML)
- Model is abstraction of data-generating process using probability, governed by unknown constants **θ** (i.e., the parameter(s))
- *X*: Input (~ covariates)
- *Y*: Output (~ response, data set), viewed as outcome from stochastic process (observed random variables)
- Can do estimation of **θ**, of missing covariates *X*, of missing responses *Y*
- Only from here on where Bayesian inference differs

The Bayesian way of learning from data, ctd.

- World of a Bayesian is divided into known and unknown things
- "Broader" use of probability in Bayes:
  - As a *measure of variability* of observable things ….
  - … and as a ***measure of uncertainty*** about unobserved things, including parameters
- Uncertainty is personal: what is perfectly known to someone may be uncertain to another person
- Degree-of-belief probability, degree of personal knowledge
- That uncertainty varies by person does not make it unscientific

The Bayesian way of learning from data, ctd.

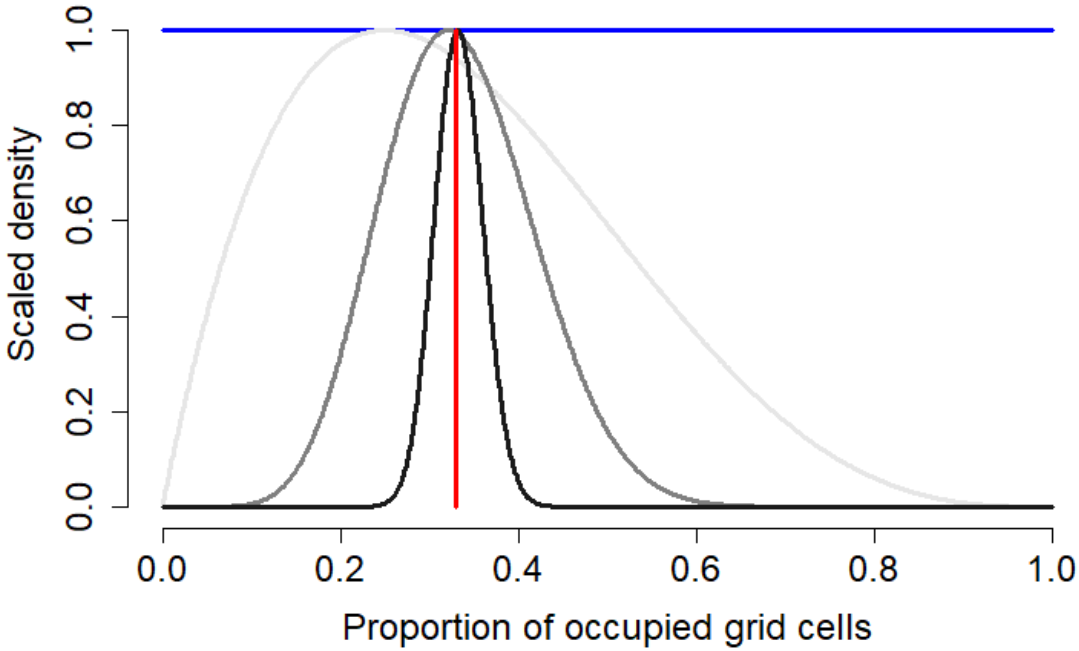
- Examples of uncertain things:
  - Whether a species still occurs at a site where it did 20 years ago
  - In a survival analysis, fate and mass of an individual in an occasion when not caught
  - In an SDM, presence/absence or abundance at a site that is not surveyed or number of occupied grid cells in a region
  - In a PVA, fate and size of a population some time into the future
  - But also things like "Does the Ivory-billed woodpecker (or Elvis) still exist?"
  - Or, "Who killed President Kennedy ?"
  - *"Is the world flat ?"*
  - And many, many more….
- Indeed, most things in the world are uncertain to us !
  (and this fact should really encourage us to get a good grasp of probability and stats ☺…)

The Bayesian way of learning from data, ctd.

- In a model, values of parameters are uncertain
- In Bayesian inference, use probability to quantify our knowledge about the values of parameters
- Put a probability distribution on a parameter
- Thus, parameter is treated mathematically as a random variable
- However, this does not mean that the value of a parameter cannot be a constant. It's still usually a constant, only we use the RV concept to express our imperfect knowledge
- *cf.* "parameters don't bounce up and down" & ex. of the 100th digit of *Pi* (Link & Barker 2009)

Use of probability as a measure of knowledge

- Can use probability to express every degree of knowledge between complete ignorance (blue) and certainty (red)
- Learning: move from blue towards red



Use of probability as a measure of knowledge

- More general use of probability (i.e., for variability & uncertainty) is defining feature of Bayesian inference
- Use coincides with the way in which humans use probability in everyday life
- But Bayesians do formal computations based on this concept of probability

Posterior distribution as target of inference

- Assume you have some data **y**, which is outcome from some random process with unknown parameter θ
- How should we make a good guess of θ, i.e., estimate it ?
- Have seen that frequentist form likelihood $p(y|\theta)$ and use that directly when going after MLEs
- A Bayesian analysis instead uses conditional probability to go after $p(\theta|y)$
- *y* are considered outcome from random process (i.e., RV), but in terms of our knowledge about them, are perfectly known once we have them
- $p(\theta|y)$ is expression about what we don't know, given what we know
- So, how should this conditional probability be computed ?

Posterior distribution as target of inference

- Step back and consider so-called Bayes' rule for two events A and B

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

- Mathematical fact, can be derived from definition of conditional probability
- Can be used on non-Bayesian probability calculations such as clinical testing or the estimation of site-specific occupancy or abundance with `occu()` and `pcount()` in `unmarked` (see later)
- General expression for how we update our knowledge in the light of new information (Blitzstein & Hwang, 2019, CRC Press)

Posterior distribution as target of inference

- Thomas Bayes (1702–1761) was one of the first to use conditional probability in this way for statistical inference, to estimate an unknown parameter θ from known data *y*

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

Components:

- Posterior distribution $p(\theta \mid y)$: probability (density) of the parameters given the information in the data and the priors
- Likelihood function $p(y \mid \theta)$: joint density of the data viewed as a function of the parameters
- Prior distribution $p(\theta)$
- Marginal distribution of the data: $p(y) = \int p(y \mid \theta)\, p(\theta) d\theta$
- Joint distribution of data and parameters $p(y, \theta)$

Posterior distribution as target of inference

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

- Use of probability as a measure of uncertainty about θ in two places
  - Prior: uncertainty assessment before we include the data
  - Posterior: …. after we use the information in the data
- In Bayes it's valid to make probability statements directly about parameters, e.g.

"We are 99% certain that the population is declining",

"I am 80% certain that the population will be extinct in 10 years",

"There is a 99.999% probability that the Ivory-billed woodpecker is extinct"

Posterior distribution as target of inference

Formal steps implicit in any Bayesian analysis:

- Use of probability as a measure of uncertainty about θ
- Treat all statistical inference (e.g., parameter estimation, testing, imputation of missing values, predictions) as a probability calculation by using Bayes' rule
- Start by expressing initial knowledge about θ in one probability distribution: the prior (must not use the data to choose prior !)
- Use Bayes' rule to update prior knowledge with information contained in the data and embodied by likelihood function p(y|θ)
- Out comes another probability distribution, the posterior distribution p(θ|y), which is our new state of knowledge

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

Posterior distribution as target of inference

Lot of heuristic appeal in Bayes' rule as a method for learning:

- Based on a "human" concept of probability as a measure of knowledge
- Bayes' rule can be described as $p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$
  i.e., the posterior is proportional to the product of likelihood and the prior.
- Exactly as we learn as humans: conclusions/decisions always a product of information of some observation at hand, plus contextual knowledge/experience.
- Bird identification good example: *"Condors in the Himalayas"*

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

Posterior distribution as target of inference

Lot of heuristic appeal in Bayes' rule as a method for learning:

- …..
- Every scientific position (represented by prior) can be modified by new information (represented by the likelihood)
- Do not use 0/1 priors ! End of learning, religion …
- Bayesian result is always a distribution: richer inference than simple point estimate with maximum likelihood
- Can characterize by measures of location and spread

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

Prior distributions

- Prior distribution are hallmark of Bayesian inference
- Allow us to incorporate other information into our inference (i.e., external to the data)
- Interestingly, Bayes is NOT the only way for using prior, or external, information, e.g., penalized likelihood (Moreno & Lele, 2010, Lele et al. 2012), regularization (ridge, lasso, …)
- Make some parameter values *a priori* more likely than others
- Bayes' rule has a "plug" with a label that says:
  "if you have some external information that you would like to incorporate into your estimation, put it in here"
- Looks like a good thing, since often may have external information

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

Prior distributions, ctd.

- Priors that do carry some information are called informative
- Can be formed based on gut-feeling of experts or taken from results that are “nearby” in space, time or taxonomy
- Bayes’ rule makes it very straightforward to incorporate such information, provided we can express it in a named probability distribution
- Advantage of informative priors:
  - They make sense (use all available information)
  - Make your estimates more precise
  - Sometimes enable estimation of additional parameters
  - (sounds like motivation for an integrated population model, IPM)
  - Also make MCMC algorithms more stable, better convergence

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

Prior distributions, ctd.

- Some might argue priors are a bad thing
- Historically, major reason for opposition to Bayes was belief that priors made a statistical analysis subjective in an “anything-goes” manner
- Appeared to contradict tenet that science must be objective
- We think that this is a red herring
- ***Any*** analysis contains literally dozens of subjective choices, e.g., where, when and what to study, what factor levels ? Which covariates ? Toss out outliers ? Log-transform covariates or not ? And many more.
- We must make dozens of decisions that are subjective
- Seek repeatability/transparency, rather than objectivity in the sense of absence of subjective decisions
- Thus, always must specify the priors to make a study transparent

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

Prior distributions, ctd.

- For better or worse, most people seek to minimize the effects of the priors by making them as vague as possible
- E.g., uniform or “flat” normal distributions
- Ideally do a prior sensitivity analysis: run analysis with 2-3 different sets of priors and see how much results (i.e, the posterior) changes
- Many people choose priors that are locally approx. uniform (locally means, where the likelihood has some support)
- Specify priors on a scale that matters, e.g., on probability scale in an occupancy model or on the bird-scale (rather than the log-bird scale) in an N-mixture model
- With vague priors, unless sample size tiny Bayesian point and uncertainty estimates are numerically VERY similar to MLEs, associated SEs/CIs

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

Bayesian computation using MCMC

Why did Bayesian statistics take off so late only ?

- Up to the 1980s, there was essentially no Bayesian inference in ecology; why ?
- Reason was mostly a practical one: denominator in Bayes' rule

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

- $p(y)$ is evaluated by integrating the numerator of the rule over every single parameter
- Mathematically intractable in almost all cases
- Therefore, could not evaluate posterior distributions in interesting models
- That is, even though some people always might have thought, Bayes was great, they just couldn't do it

Why did Bayesian statistics take off so late only ?

- 1980/90s: statisticians re-discover work conducted by physicists in the 1950s
- Can construct simulation algorithms that produce random numbers from “unknown” distributions, i.e., un-named distributions that can not be dealt with analytically
- Don’t need to evaluate the denominator in Bayes’ rule
- RNGs (random number generators) for the posterior distributions arising in Bayesian models !
- Approximate posterior distribution to arbitrary degree of accuracy by drawing larger samples
- Markov chain Monte Carlo (MCMC), e.g.
  - Metropolis(-Hastings) algorithm
  - Gibbs sampling
- Huge boost to Bayes in statistics community

Home-grown MCMC vs. Bayesian engines

- Several engines that let you fit models using Bayesian MCMC techniques:

  (1) BUGS, WinBUGS, OpenBUGS, multiBUGS

  (2) JAGS

  (3) Stan

  (4) Nimble

  (5) also many others, e.g. greta, …

- Should you learn how to write your own MCMC algorithms ?

  - YES: 10–20% of us

  - NO: all the rest

- Steep learning curve, debugging can take a lot of time, and “engines” just too powerful

- see books by Clark (2007), Hooten & Hefley (2019), Zhao (2024), also good chapter in Royle *et al.* (2014)

Getting a feel for MCMC

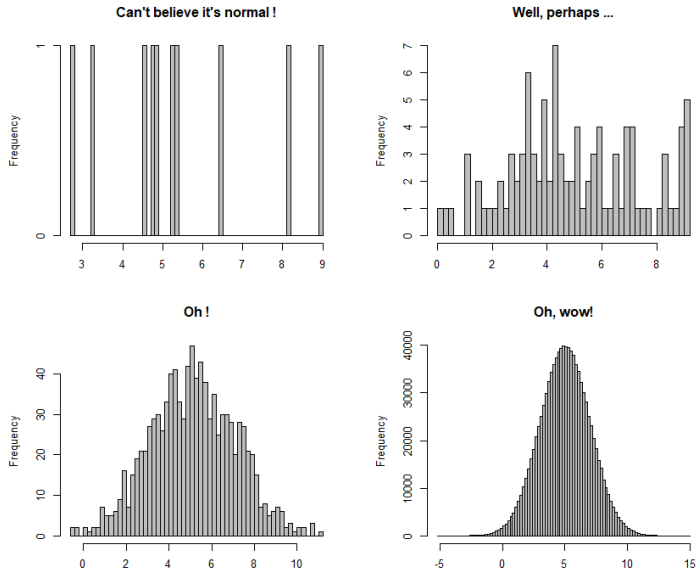
- But good to know how MCMC works
- Therefore, next give a taste of Monte Carlo integration and of one of simplest MCMC algorithm, for the bee-eater data
- MCMC algorithms are RNGs for the posterior distributions in our Bayesian models
- Can simply draw large number of samples and then summarize sample for inference about population quantities including parameters
- Larger sample yields better approximation than smaller sample
- Sample mean approximates posterior mean, sample SD or percentiles approximates SD or 95% CRI of posterior distribution

Getting a feel for MCMC, ctd.

- Technically, use simulation to solve an integral
- For instance, for the posterior mean, we must solve this integral

$$E(x) = \int p(x)xdx$$

- **Show an R demo** to illustrate how we can produce a sample from a (known) distribution and then use sample statistics to infer a parent distribution: Monte Carlo integration
- Larger samples are better, but even very small samples can be already pretty good



Example of Markov Chain Monte Carlo (MCMC)

- Develop very simple MCMC for Swiss bee-eaters (Poisson log-linear regression for counts)
- Two parameters: intercept (alpha) and slope (beta)
- Algorithm is essentially a glorified RNG for the joint posterior of alpha and beta
- Do not need to evaluate the integral in Bayes' rule !
- In this way, can fit even very complex models, e.g., in Royle & Dorazio (2008) and later ecology books on hierarchical models
- For many of these, inferences would have been impossible before the advent of MCMC

Example of Markov Chain Monte Carlo (MCMC)

- RNG for target posterior in a Bayesian model
- Simulation (= Monte Carlo) produces a Markov chain: one-step-back dependence, given current value, future values are independent from past values
- Serial dependence in output from a MCMC-RNG, by construction
- Does not affect ability to summarize samples to learn about the posterior distribution of parameters
- Only reduces effective sample size = information content

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}, $\sigma_{proposal}$)

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

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

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

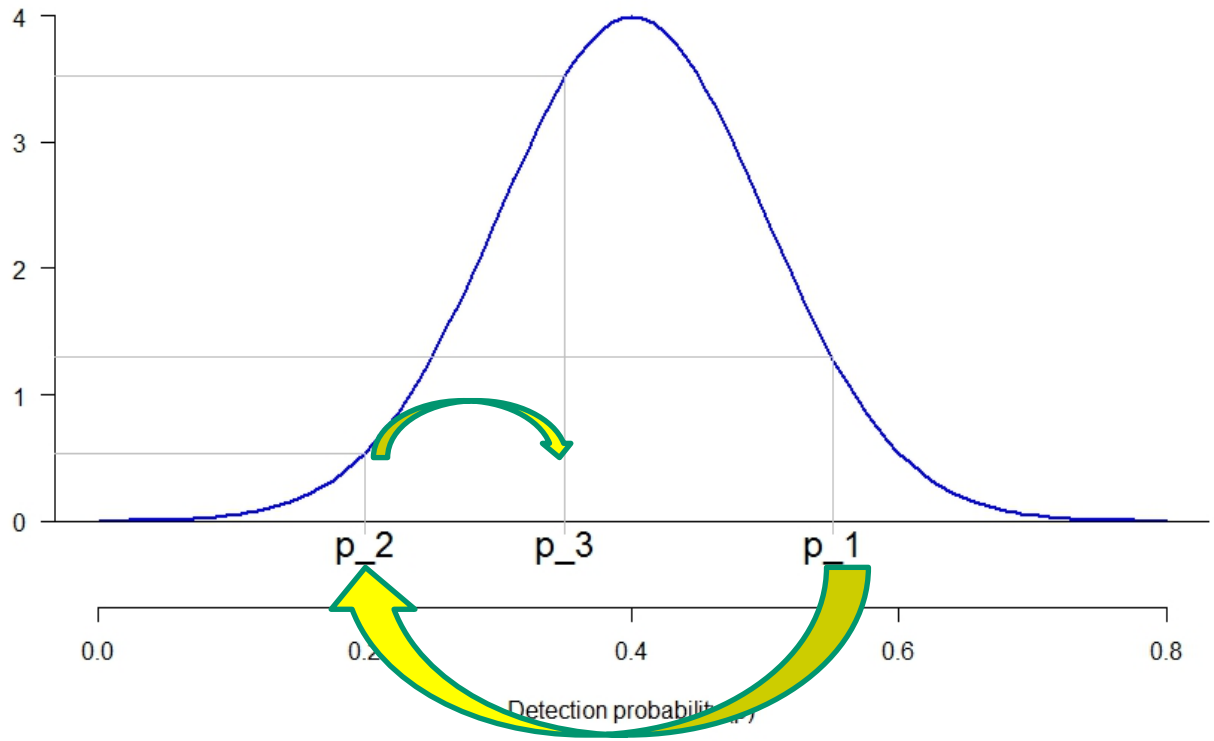
  (3) If R >= 1, set θ^t <- θ^* (**accept** new value)

  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)
- MCMC: *jump "upwards" along posterior with greater prob.*
- (example for a one-parameter binomial model)



Comments on simple (random-walk) MCMC

(1) The stochastic rule for proposing a new candidate is often a normal RNG. That is, we choose $\theta^* \sim Normal(\theta^{t-1}, \sigma_h)$. Note that this is not a distribution that is part of the model, but that the normal RNG is simply a convenient way of obtaining a random number. Proposal SD is best thought as a tuning parameter of the MCMC algorithm. Large values cause large "step length" of algorithm and vice versa

(2) Note how p(y) cancels from the expression for R:

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

Therefore, we don't need to evaluate the ugly integrals !

(3) If we repeat this algorithm a large number of times (e.g., 1000s or more), we will usually get random numbers from the desired posterior distributions.

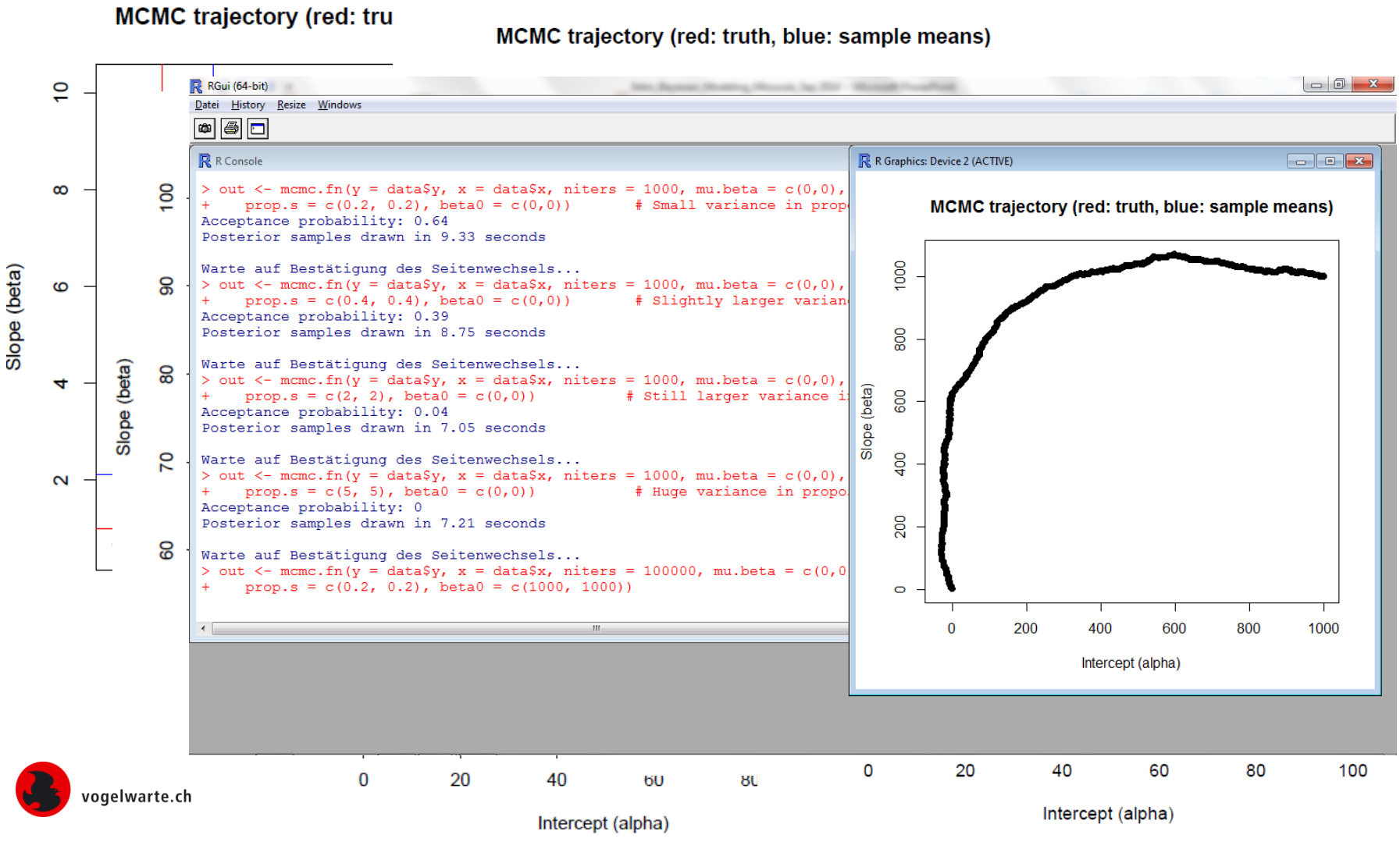
(4) Note we accept some values also with R<1, because we also want to characterize the tails of the posterior distributions.

Demo of simple (random-walk) MCMC

- See **R code with demo** of the algorithm for the Swiss bee-eater data
- Also see some scenarios with the R function `demoMCMC()`
- Also see Section 2.8.2 in the handout chapter 2 of the new ASM book
- ….and `'Logistic regression code with ML and Metropolis MCMC.R'`

The power of MCMC

- MCMC astonishing and crazily powerful family of algorithms !



vogelwarte.ch

Should you be a frequentist or a Bayesian ?

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, JAGS, NIMBLE):

- Implementation of complex, custom models within reach of ecologists
- 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 with MCMC:

  - Take long time to run (often (much) less for 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, very 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 even better MCMC and likelihood software for complex models

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