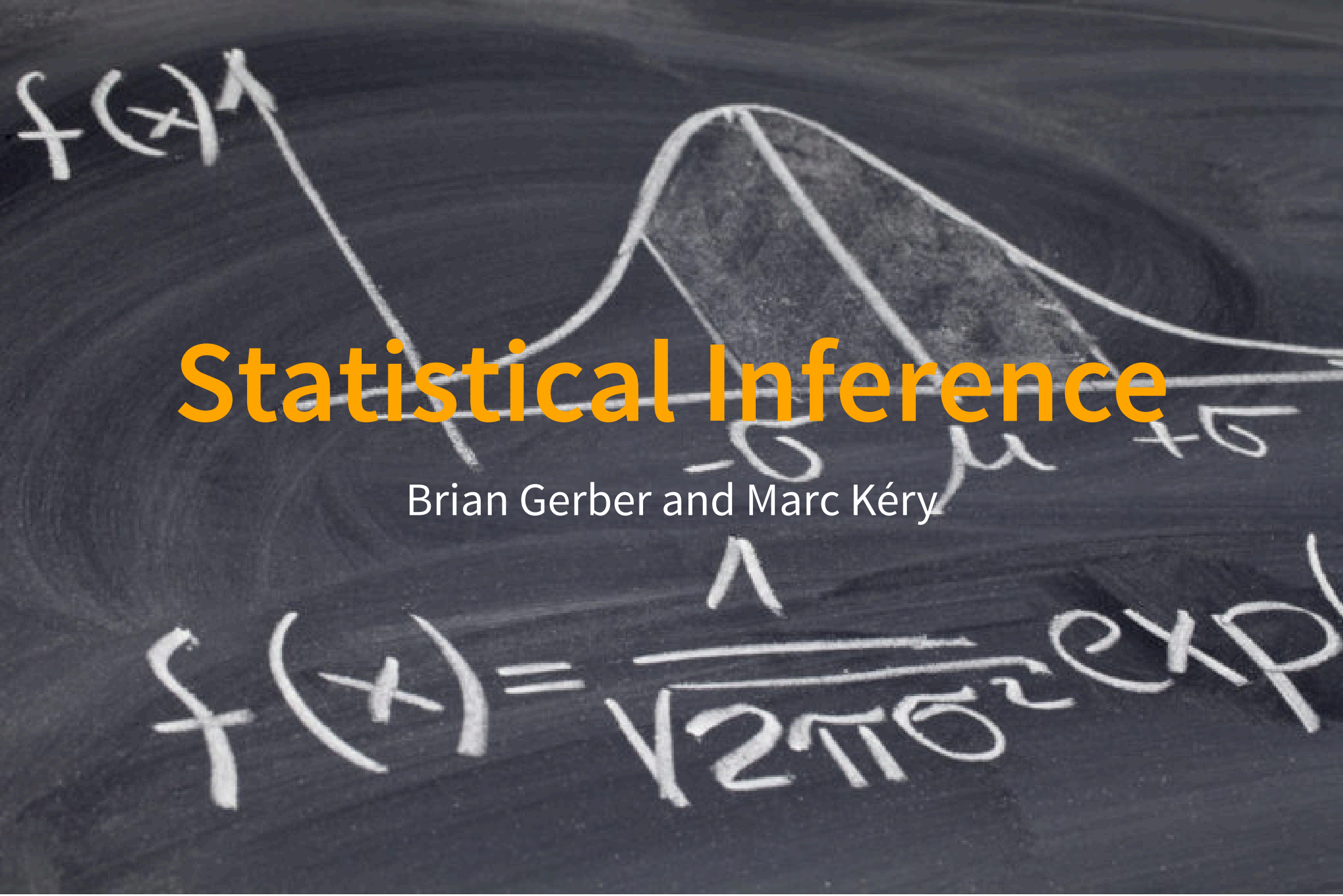


Statistical Inference

Brian Gerber and Marc Kéry

Outline

- Two types of Inference
- Probability
- Likelihood
- Bayesian

Kéry
Kellner

APPLIED STATISTICAL MODELLING FOR ECOLOGISTS

ELSEVIER

Marc Kéry
Kenneth F. Kellner

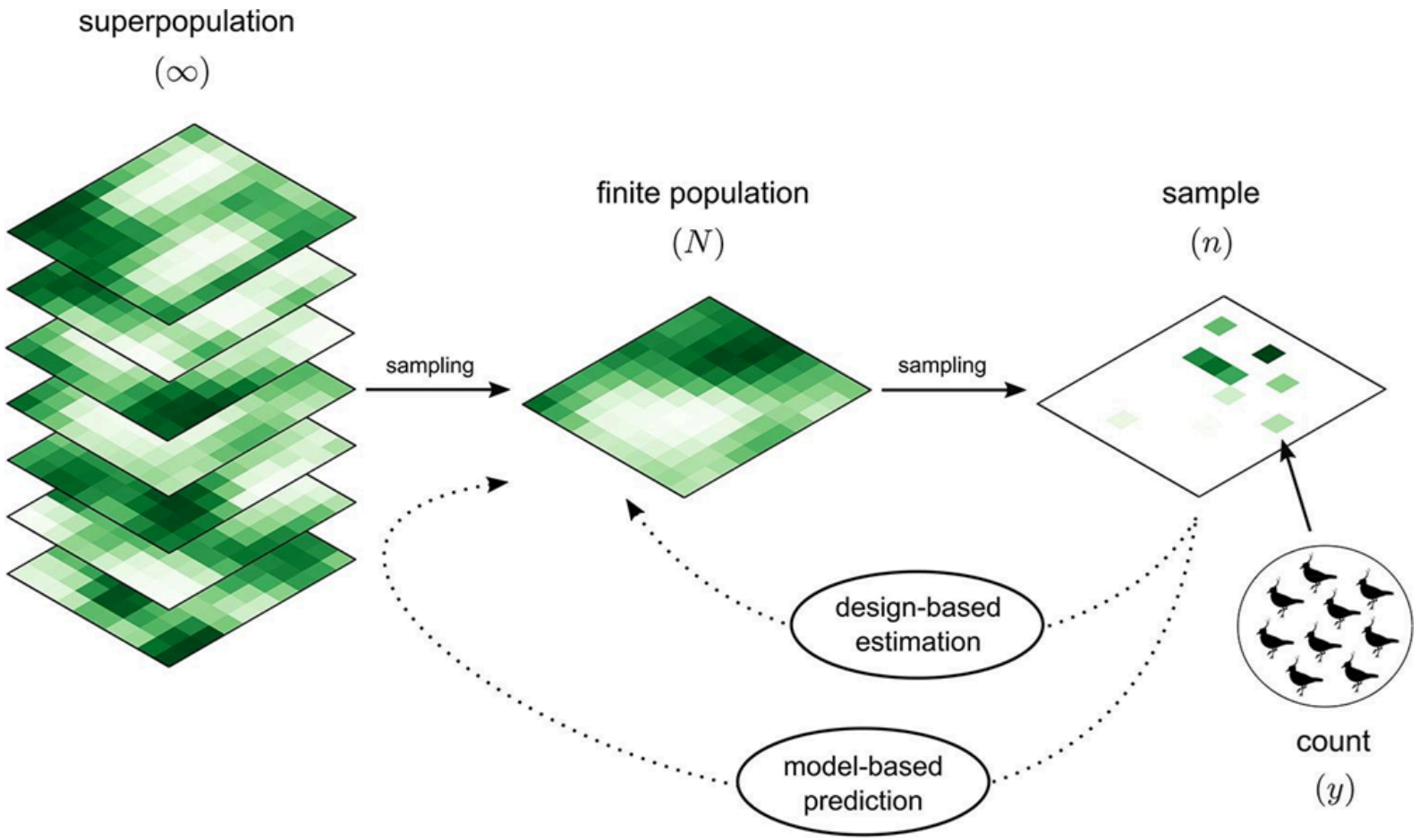
APPLIED STATISTICAL MODELLING FOR ECOLOGISTS

A practical guide to Bayesian and likelihood inference using R, JAGS, NIMBLE, Stan and TMB

ELSEVIER

Two types of Inference

Two types of Inference

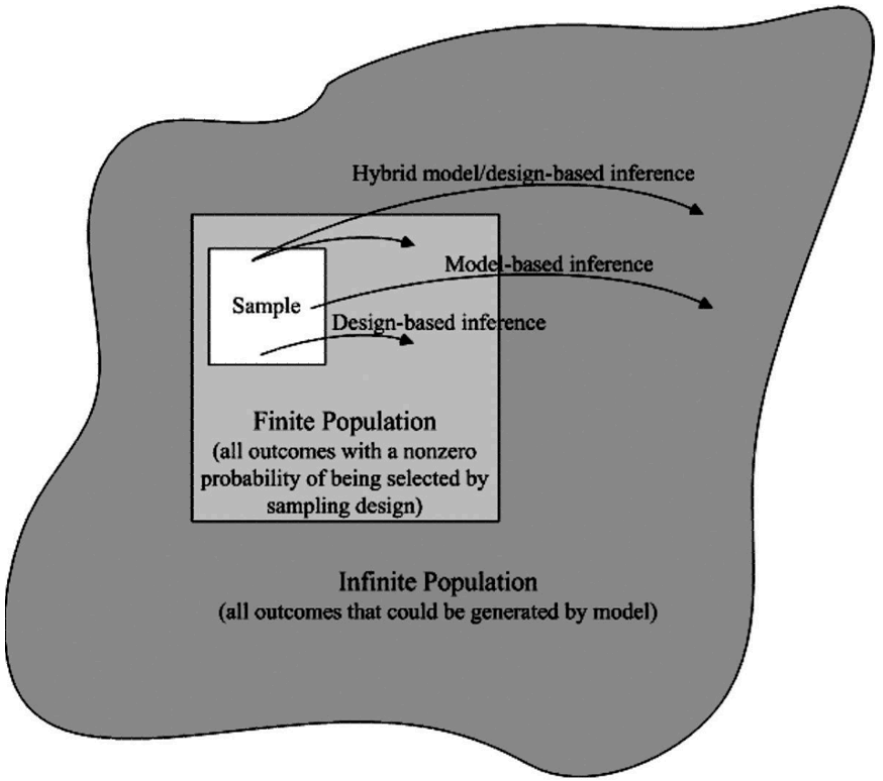


Aubry, P., & Francesiaz, C. (2022). On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. Ecological Indicators, 144, 109394.

Design-based

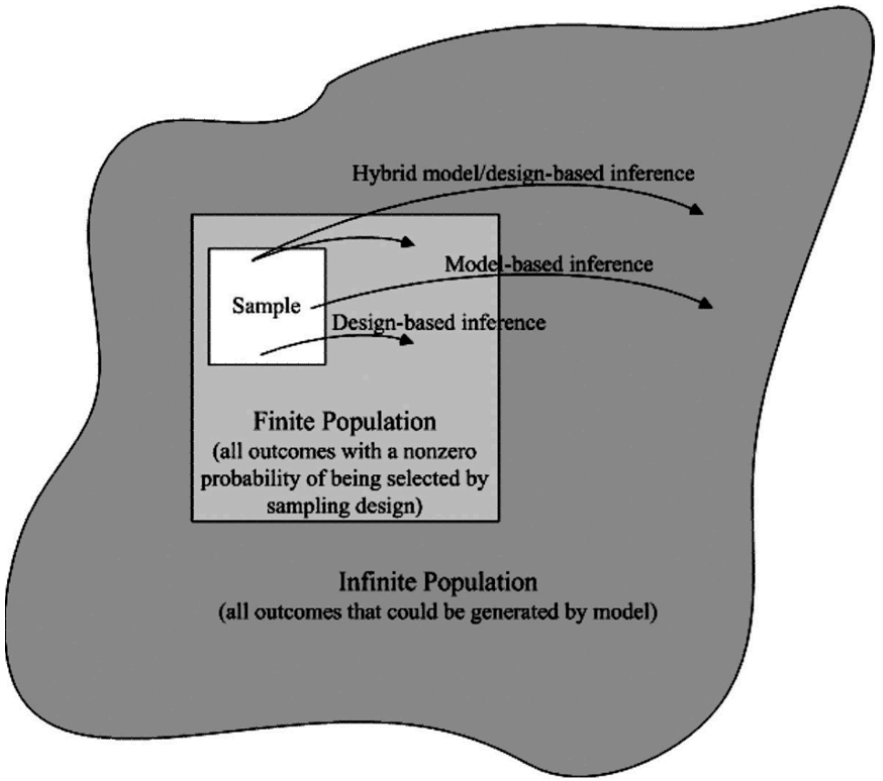
Model-based

- Strong theoretical basis
- Population quantities are a random realization of a stochastic generating process; ‘superpopulation’(ξ)
- Sampling frame and prob. sampling are not necessary



Model-based

- Unbiased estimators are not guaranteed
- Randomness is implicit in the process and assumed to follow stochastic distribution
- Some-to-very many assumptions
- Highly flexible / Basically Dark Magic!



Two types of Inference

- Know your population by defining a sampling frame
- Think about non random samples (not ignorable)
- Use statistical models for flexibility

Statistical Modeling (In a nutshell)

- Describes stochastic process that could produce the data
- Observed data is just one possible
- Model + data allows 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

What is a statistical model?

- An abstraction / simplification / explanation
- Every model has a goal
  - enforce clarity of thought
  - summarize
  - search for patterns
  - understand mechanisms
  - predict

What is a statistical model?

We are rarely explicit about the goals of our models…though should be!

Is my model good ?

Inference vs prediction

Parameteric Models

- 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."

Statistical Modeling Concerns

- Many hypothetical data-generating mechanisms of stochastic process
  - leads to “messiness” of model selection

- When sampling a large fraction of a finite population
  - e.g., sample a small finite population

$$\hat{p} \pm t_{\alpha/2} \sqrt{(\frac{\hat{p}(1-\hat{p})}{n} \frac{N-n}{n-1})}$$

Probability

Probability

Probability as the basis for statistical inference

- The world is uncertain
- Very few things perfectly known (i.e., deterministic)
- 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

David Spiegelhalter

ESSAY | 16 December 2024 | Correction 18 December 2024

Why probability probably doesn't exist (but it is useful to act like it does)

All of statistics and much of science depends on probability – an astonishing achievement, considering no one's really sure what it is.

https://www.nature.com/articles/d41586-024-04096-5

Objectives

- Connect random variables, probabilities, and parameters
- define prob. functions
  - discrete and continuous random variables
- use/plot prob. functions
- notation!

Probability/Statistics

Probability and statistics are the opposite sides of the same coin.

To understand statistics, we need to understand probability and probability functions.

The two key things to understand this connection is the random variable (RV) and parameters (e.g., $\theta, \sigma, \epsilon, \mu$).

Motivation

Why learn about RVs and probability math?

Foundations of:

- linear regression
- generalized linear models
- mixed models

Our Goal:

Not Random Variables

$$
\begin{aligned}
a =& 10 \\
b =& \log(a) \times 12 \\
c =& \frac{a}{b} \\
y =& \beta_0 + \beta_1 \times c + \epsilon
\end{aligned}
$$

All variables here are scalars. They are what they are and that is it. β variables and y are currently unknown, but still scalars.

Scalars are quantities that are fully described by a magnitude (or numerical value) alone.

Random Variables

$$y \sim f(y)$$

y is a random variable which may change values each observation; it changes based on a probability function, $f(y)$.

The tilde ($\sim$) denotes “has the probability distribution of”.

Which value (y) is observed is predictable. Need to know *parameters* (θ) of the probability function $f(y)$.

Random Variables

$$y \sim f(y)$$

Most often, $f(y|\theta)$, where '|' is read as 'given'.

Toss of a coin

Roll of a die

Weight of a captured elk

Count of plants in a sampled plot

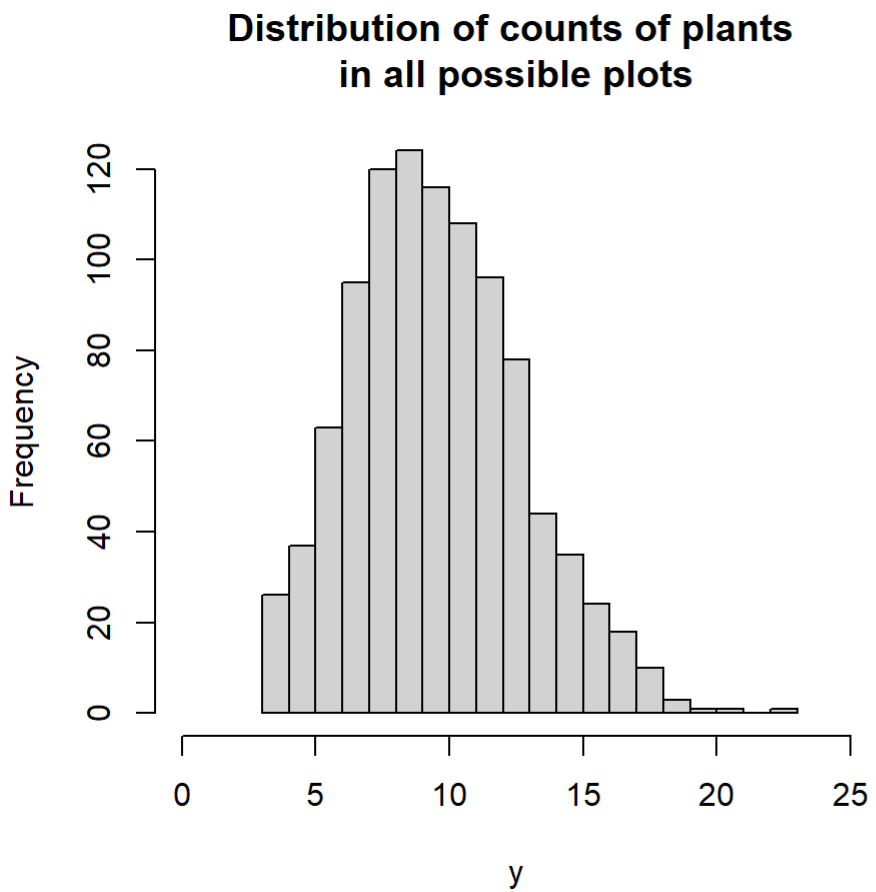
Random Variables

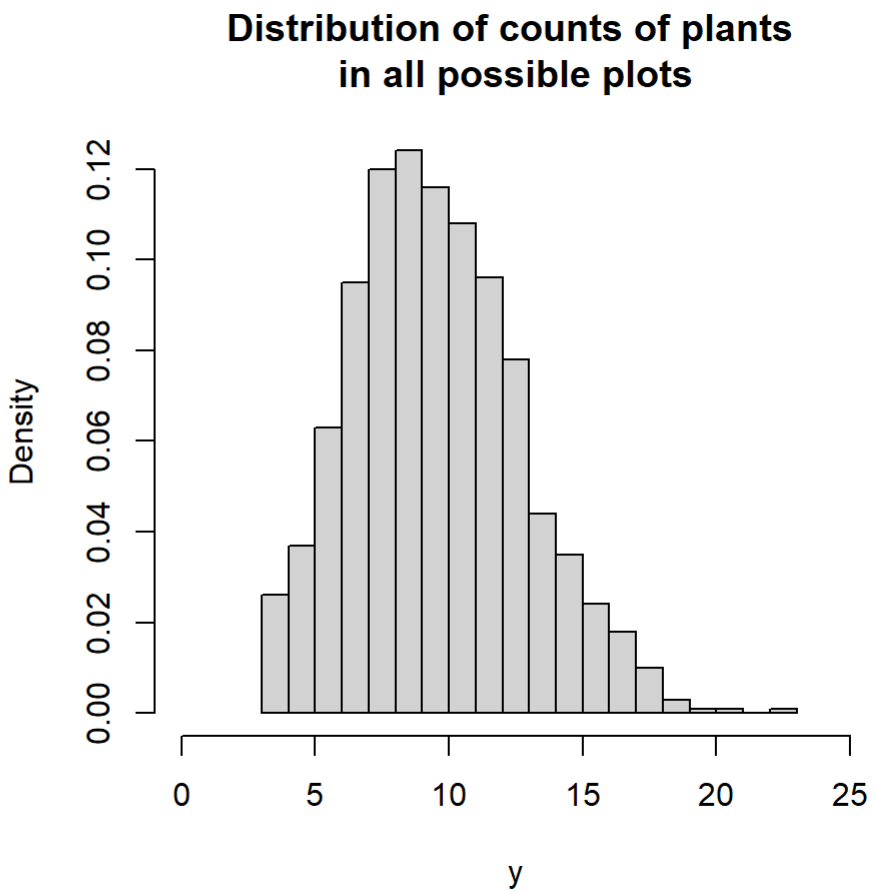
$$y \sim f(y)$$

The values observed can be understand based on the frequency within the population or presumed super-population. These frequencies can be described by probabilities.

Frequency / Probabilities

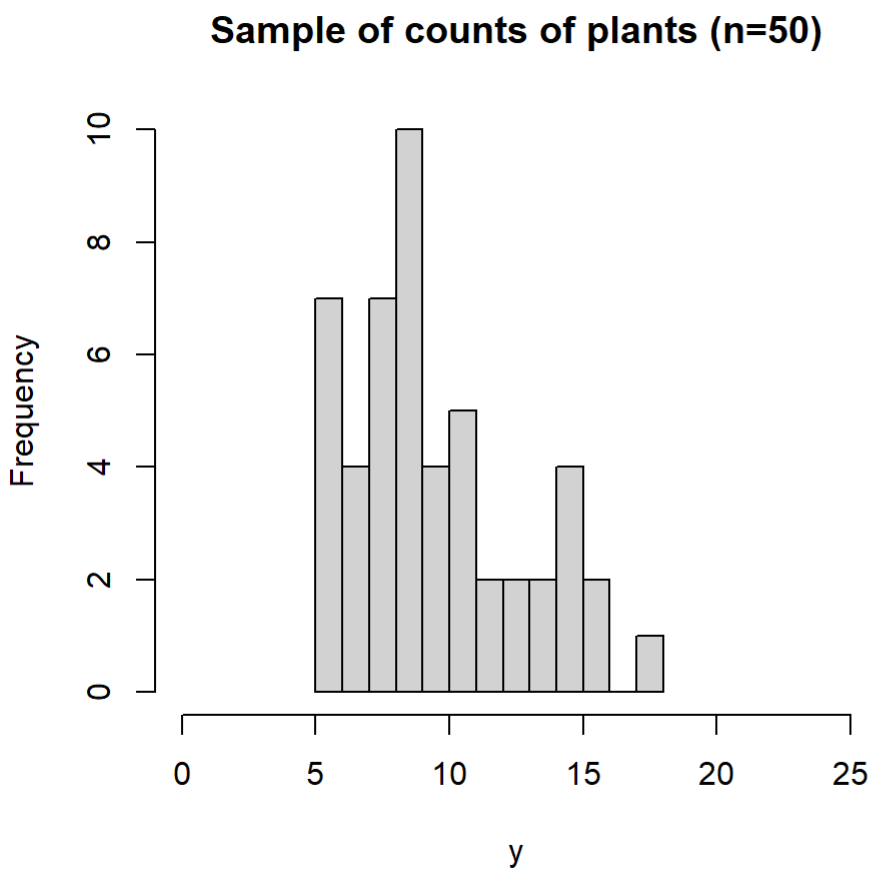
```
1 par(mfrow=c(1,2))
2 hist(y, breaks=20,xlim=c(0,25),main=main)
3 hist(y, breaks=20,xlim=c(0,25),freq = FALSE,main=main)
```

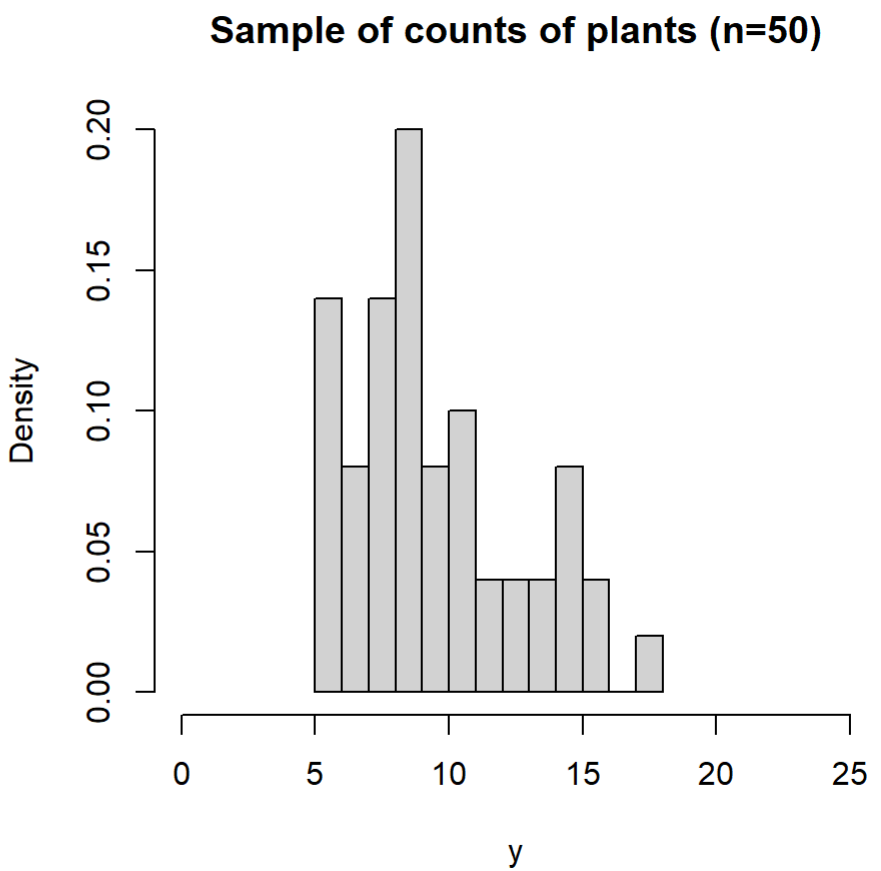




Frequency / Probabilities

We often only get to see ONE sample from this distribution.





Random Variables

We are often interested in the characteristics of the whole population of frequencies,

- central tendency (mean, mode, median)
- variability (var, sd)
- proportion of the population that meets some condition
  P($8 \leq y \leq 12$) =0.68

We infer what these are based on our sample (i.e., statistical inference).

Philosophy

Frequentist Paradigm:

Data (e.g., y) are random variables that can be described by probability distributions with unknown parameters that (e.g., θ) are *fixed* (scalars).

Bayesian Paradigm:

Data (e.g., y) are random variables (when observed, then fixed) that can be described by probability functions where the unknown parameters (e.g., θ) are also random variables that have probability functions that describe them.

Random Variables

$$
\begin{aligned}
y &= \text{event/outcome} \\
f(y|\boldsymbol{\theta}) = &[y|\boldsymbol{\theta}] = \text{ process governing the value of } y \\
\boldsymbol{\theta} &= \text{parameters}
\end{aligned}
$$

$f()$ or [ ] is conveying a function (math).

It is called a PDF when y is continuous and a PMF when y is discrete.

- PDF: **probability density function**
- PMF: **probability mass function**

Functions

We commonly use *deterministic* functions (indicated by non-italic letter); e.g., log(), exp(). Output is always the same with the same input.

$$x \overset{\mathrm{g}}{\Longrightarrow} \boxed{\text{DO STUFF}} \Longrightarrow \mathrm{g}(x)$$

$$x \overset{\mathrm{g}}{\Longrightarrow} \boxed{+7} \Longrightarrow \mathrm{g}(x)$$

$$\mathrm{g}(x) = x + 7$$

Random Variables

Probability: Interested in y, the data, and the probability function that "generates" the data.

$$y \leftarrow f(y|\boldsymbol{\theta})$$

Statistics: Interested in population characteristics of y; i.e., the parameters,

$$y \rightarrow f(y|\boldsymbol{\theta})$$

Probability Functions

Special functions with rules to guarantee our logic of probabilities are maintained.

Discrete RVs

y can only be a certain set of values.

1. $y \in \{0, 1\}$
    - 0 = dead, 1 = alive
2. $y \in \{0, 1, 2, \ldots, 15\}$
    - count of pups in a litter; max could by physiological constraint

PMF

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

Data has two outcomes (0 = dead, 1 = alive)

$y \in \{0, 1\}$

There are two probabilities

- $f(0) = P(Y = 0)$
- $f(1) = P(Y = 1)$

PMF

Axiom 1: The probability of an event is greater than or equal to zero and less than or equal to 1.

$$0 \leq f(y) \leq 1$$

Example,

- $f(0) = 0.1$
- $f(1) = 0.9$

PMF

Axiom 2: The sum of the probabilities of all possible values (sample space) is one.

$$\sum_{\forall i} f(y_i) = f(y_1) + f(y_2) + \ldots = P(\Omega) = 1$$

Example,

- $f(0) + f(1) = 0.1 + 0.9 = 1$

PMF

Still need to define $f()$, our PMF for $y \in \{0, 1\}$

The Bernoulli distribution

$$f(y|\theta) = [y|\theta] = \theta^y \times (1 - \theta)^{1-y}$$

θ = P(Y = 1) = 0.2

$$f(y|\theta) = [y|\theta] = = 0.2^1 \times (1 - 0.2)^{0-0}$$

$$f(y|\theta) = [y|\theta] = = 0.2 \times (0.8)^0 = 0.2$$

Bernoulli PMF

$$f(y|\theta) = [y|\theta] = \theta^y \times (1-\theta)^{1-y}$$

Sample space support (Ω):

- $y \in \{0, 1\}$

Parameter space support (Θ):

- $\theta \in [0, 1]$
- General: $\theta \in \Theta$

Bernoulli PMF (Code)

What would our data look like for 10 ducks that had a probability of survival (Y=1) of 0.20?

```
#Define inputs
  theta=0.2;  N=1

#Random sample - 1 duck
  rbinom(n=1,size=N,theta)
```

```
[1] 0
```

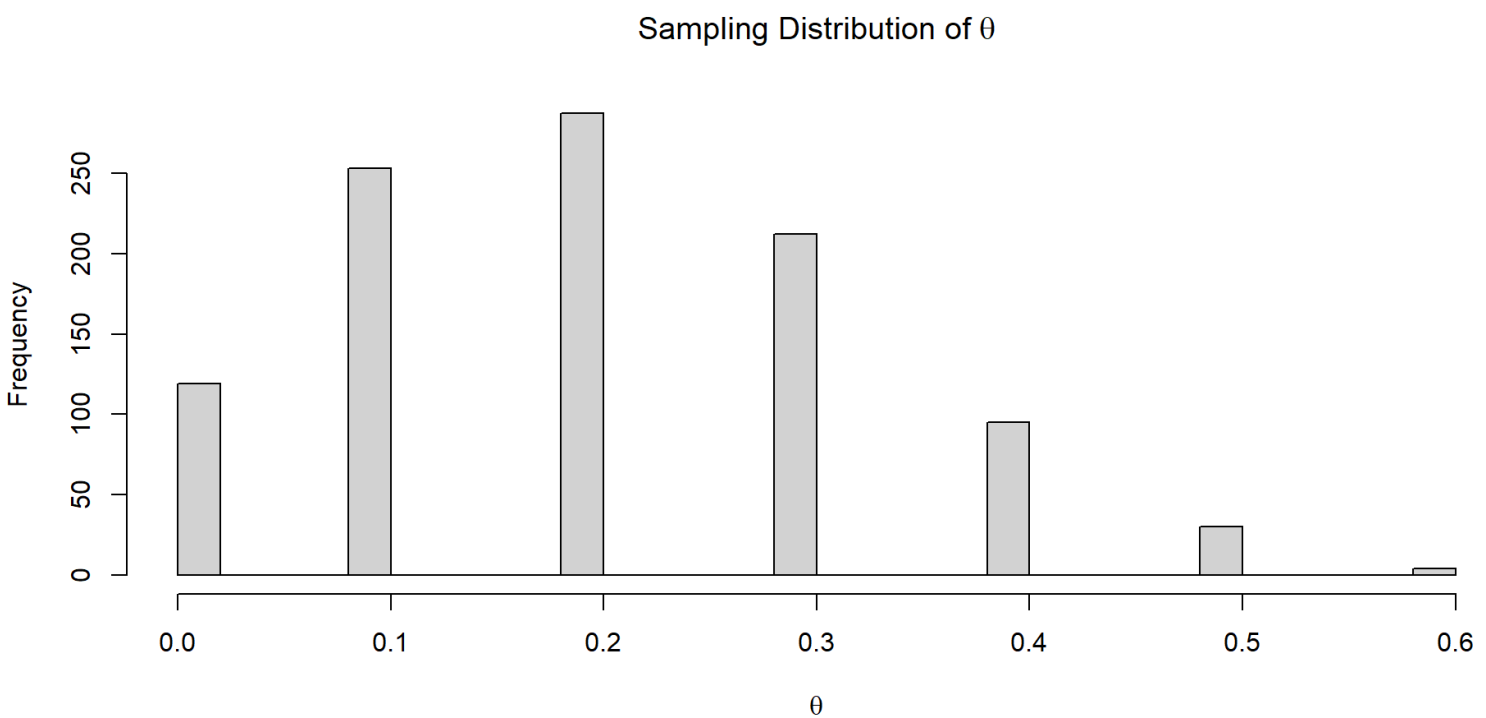
```
#Random sample - 10 ducks
  rbinom(n=10,size=N,theta)
```

```
 [1] 1 0 0 0 1 0 1 0 1 0
```

Why is this useful to us?

How about to evaluate the sample size of ducks needed to estimate θ?

```
y.mat = replicate(1000,rbinom(n = 10,size=N,theta))
theta.hat = apply(y.mat, 2, mean)
```



Binomial PMF

The Bernoulli is a special case of the Binomial Distribution.

[Math Processing Error]

N = total trials / tagged and released animals

y = number of successes / number of alive animals at the of the study.

Binomial PMF (Code)

```r
# 1 duck tagged/released and one simulation
  theta=0.2;  N=1
  rbinom(n=1,size=N,theta)
```

```
[1] 1
```

```r
# 1000 ducks tagged/released and one simulation
  theta=0.2;  N=1000
  rbinom(n=1,size=N,theta)
```

```
[1] 198
```

```r
# 1000 ducks tagged/released and 10 simulation
  theta=0.2;  N=1000
  rbinom(n=10,size=N,theta)
```

```
 [1] 180 190 198 192 169 192 192 217 206 216
```

```r
# 1 duck tagged for each of 1000 simulations
  theta=0.2;  N=1
  y = rbinom(n=1000,size=N,theta)
  y
```

```
   [1] 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0
  [38] 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
[112] 1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0
[149] 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0
[186] 1 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
[223] 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0 1 0 0
[260] 0 1 1 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
[297] 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0
[334] 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
[371] 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1
[408] 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0
[445] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0
[482] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
```

```
1 sum(y)
```

```
[1] 186
```

Support

Use a probability function that makes sense for your data/RV. In Bayesian infernece, we also pick prob. functions that make sense for parameters.

The sample space and parameter support can be found on Wikipedia for many probability functions.

Normal PDF

For example, the Normal/Gaussian distribution describes the sample space for all values on the real number line.

$$y \sim \text{Normal}(\mu, \sigma)$$
$$y \in (-\infty, \infty)$$
$$y \in \mathbb{R}$$

What is the parameter space for μ and σ?

Normal Distribution

We collect data on adult alligator lengths (in).

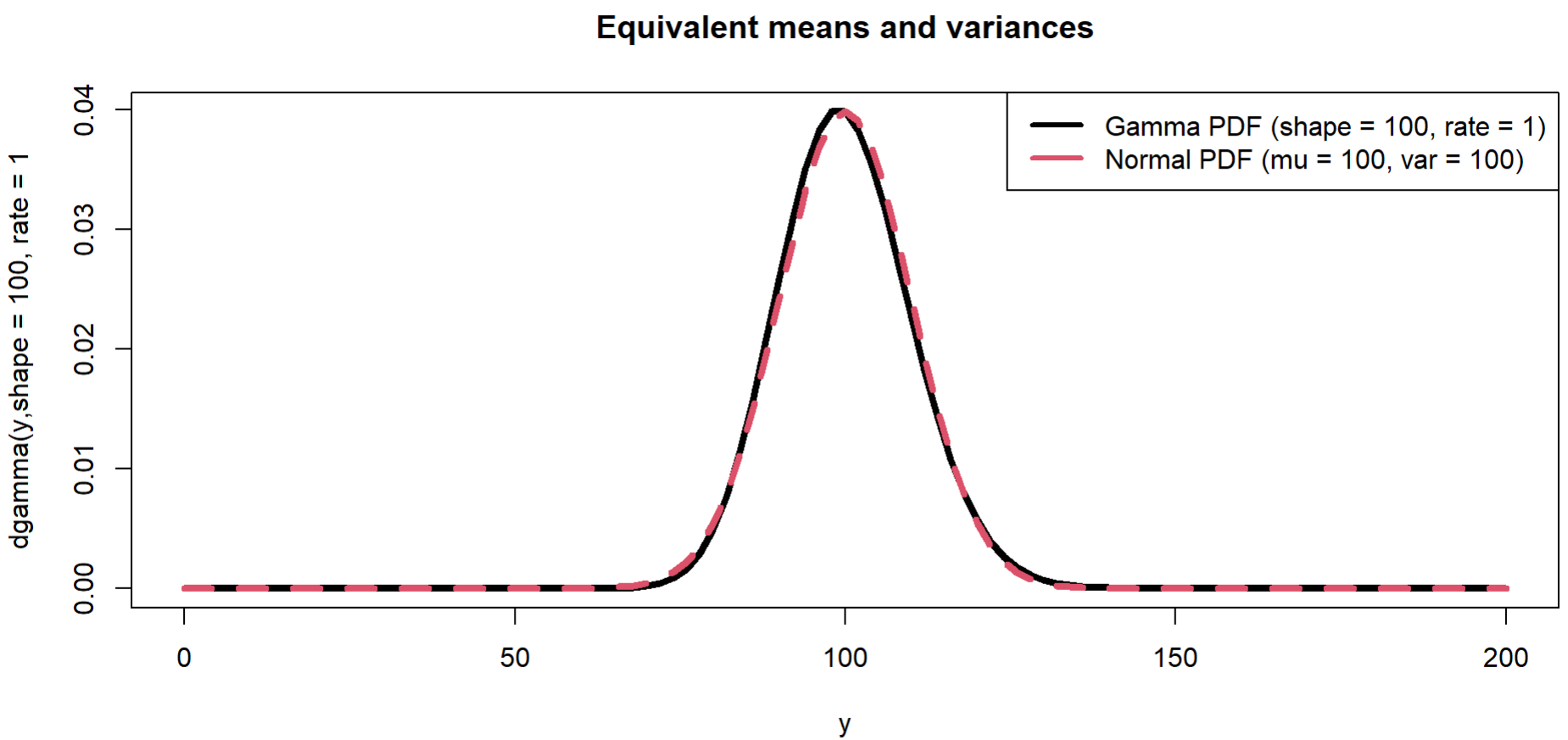
```
 [1]  90.30  83.02 103.67  85.17  99.20 106.74  90.76 105.28  99.41 101.72
```

Should we use the Normal Distribution
to estimate the mean?

Does the support of our data match
the support of the PDF?

What PDF does?

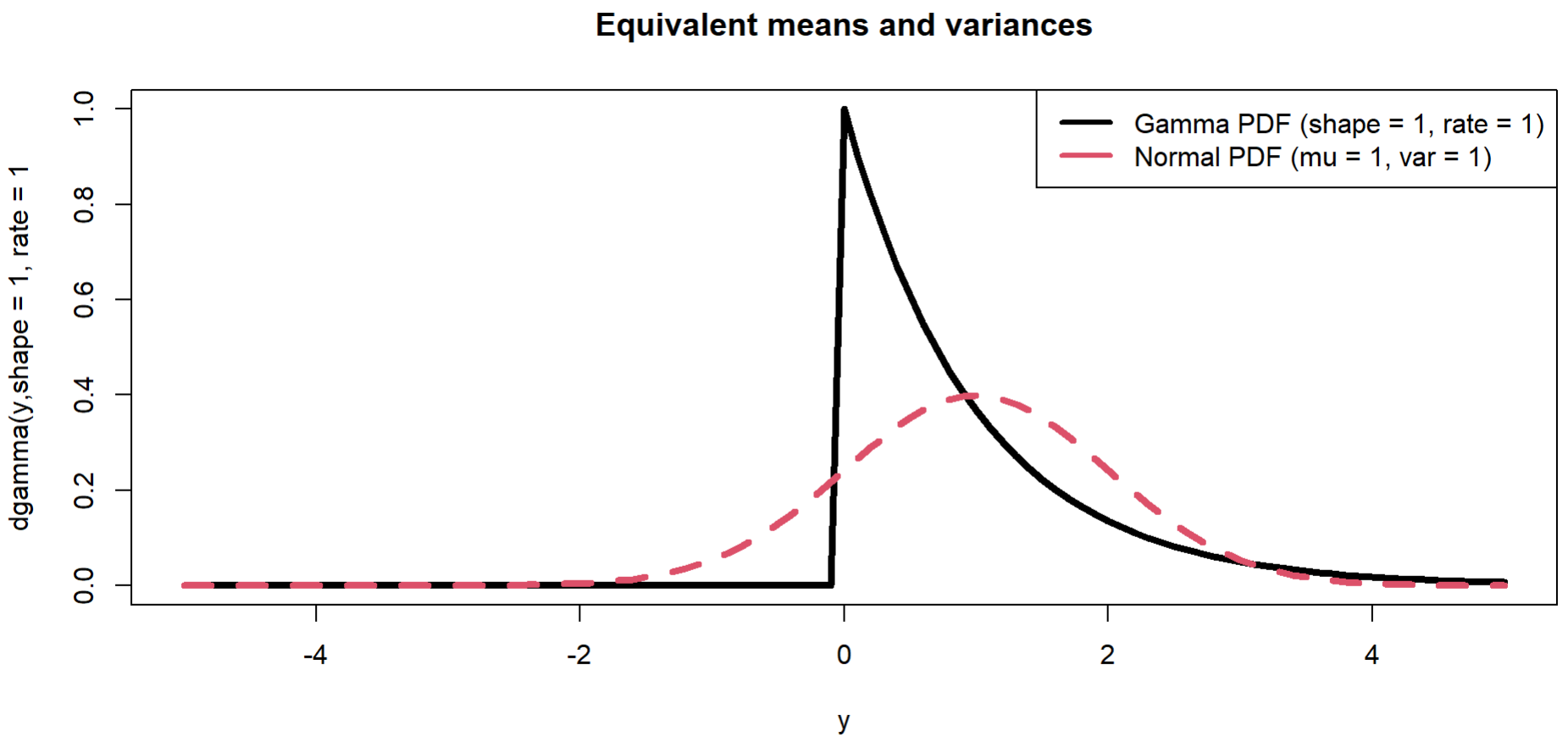
Normal Distribution



Are they exactly the same?

Normal Distribution

The issue is when the data are near 0, we might estimate non-sensical values (e.g. negative).



PDF

Continuous RVs

y are an uncountable set of values.

Provide ecological data examples that match the support?

1. Gamma: $y \in (0, \infty)$
2. Beta: $y \in (0, 1)$
3. Continuous Uniform: $y \in [a, b]$

PDF

PDFs of continious RVs follow the same rules as PMFs.

Confusing Differences

Axiom 1:

- $f(y) \geq 0$

PDFs output probability densities, not probabilities.

PDF

Axiom 2:

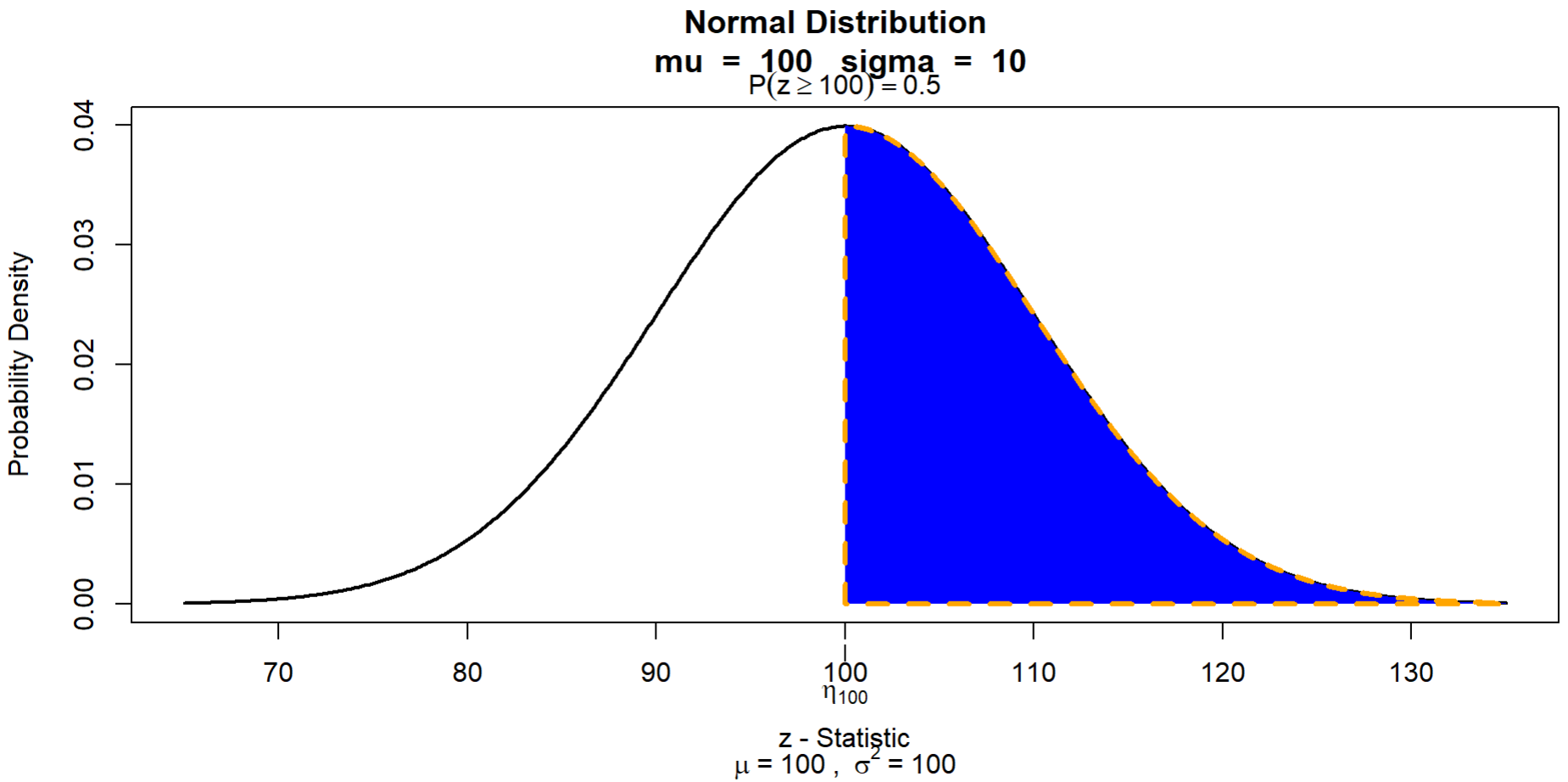
- Probabilities are the area b/w a lower and upper value of y; i.e, area under the curve

$$y \sim \text{Normal}(\mu, \sigma)$$

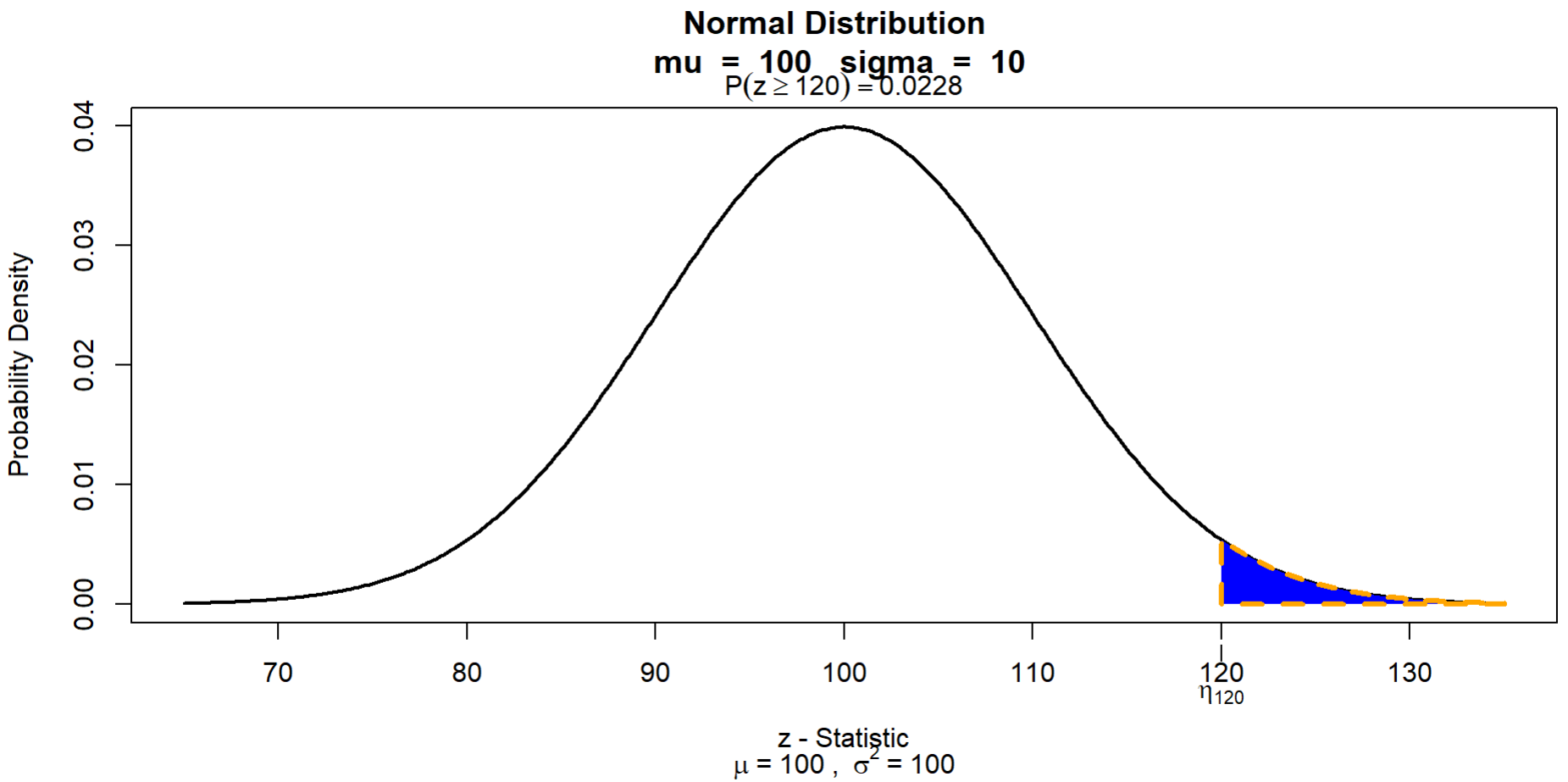
$$f(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

PDF

```
visualize.it(dist = 'norm', stat = c(100),
             list(mu = 100 , sd = 10), section = "upper")
```



PDF


Normal Distribution
mu = 100 sigma = 10
P(z ≥ 120) = 0.0228
Probability Density
0.00
0.01
0.02
0.03
0.04
70
80
90
100
110
120
130
η120
z - Statistic
μ = 100 , σ2 = 100

PDF

The math,

$$\int_{120}^{\infty} f(y|\mu,\sigma)dy = P(120 < Y < \infty)$$

Read this as “the integral of the PDF between 120 and infinity (on the left-hand side) is equal to the probability that the outcome of the random variable is between 120 and infinity (on the right-hand side)”.

The code

```
pnorm(120,mean=100,sd=10,lower.tail = FALSE)
```

```
[1] 0.02275013
```

Or, we could reverse the question.

```r
qnorm(0.02275,100,10,lower.tail = FALSE)
```

```
[1] 120
```

PDF

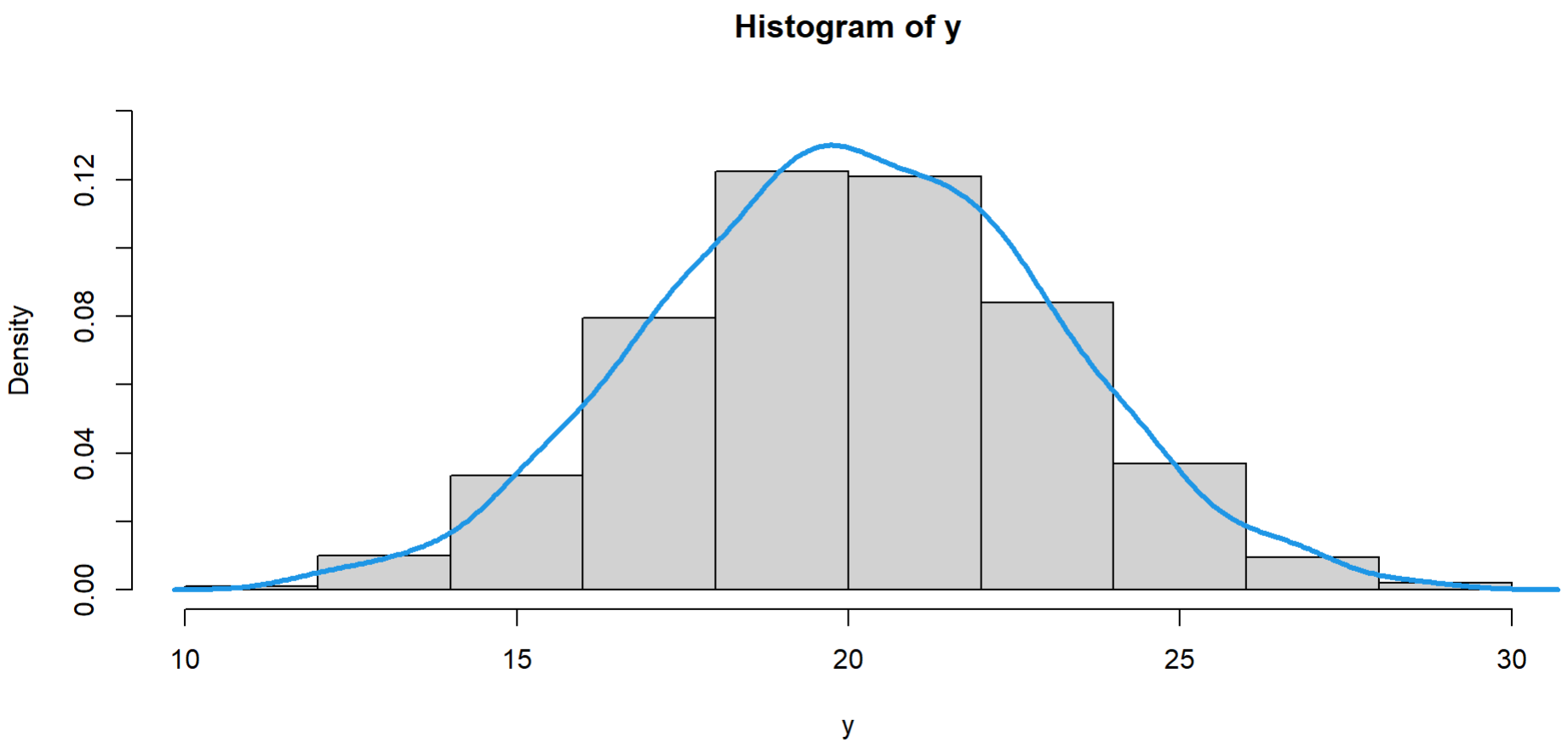
Axiom 3:

- $\int_{\text{lower support}}^{\text{upper suppport}} f(y)dy = 1$

The sum of the probability densities of all possible outcomes is equal to 1.

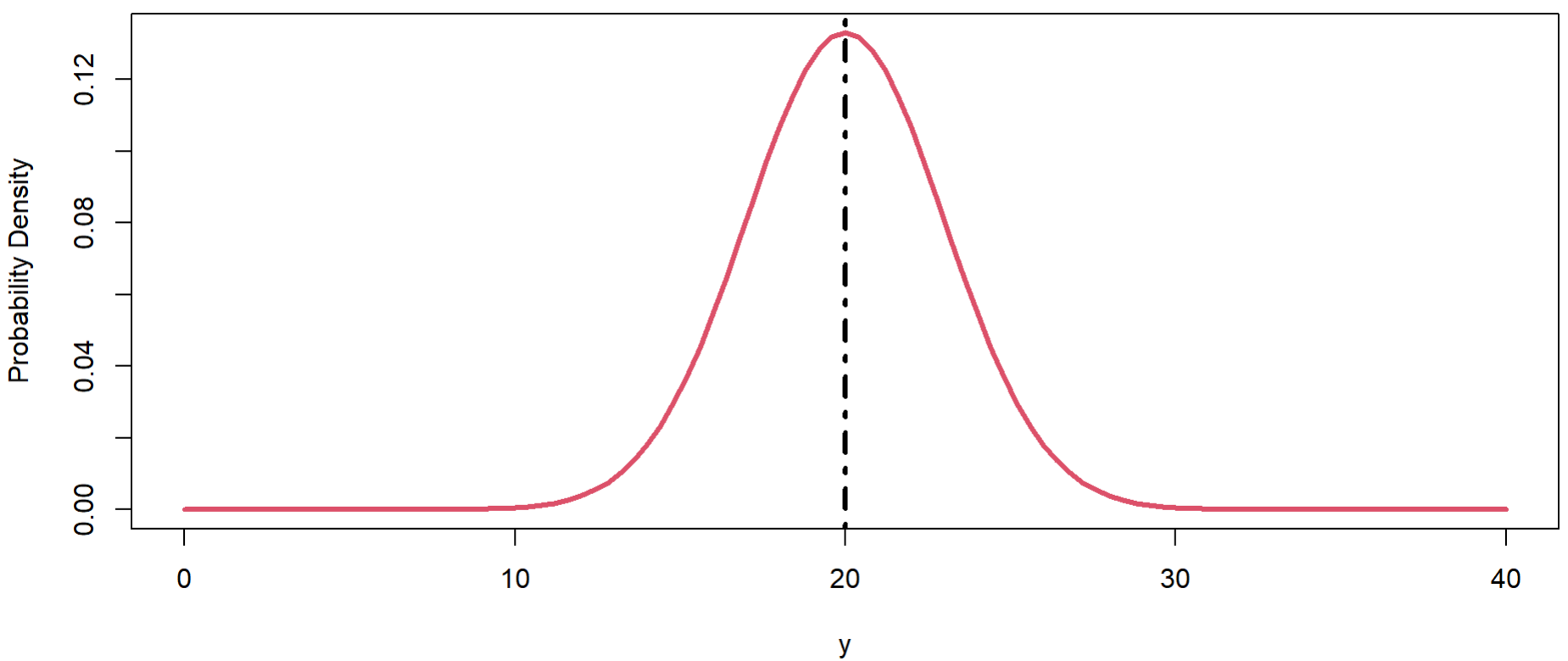
Normal Distribution (PDF Code)

```r
y = rnorm(1000, mean = 20, sd = 3)
hist(y,freq=FALSE,ylim=c(0,0.14))
lines(density(y),lwd=3,col=4)
```



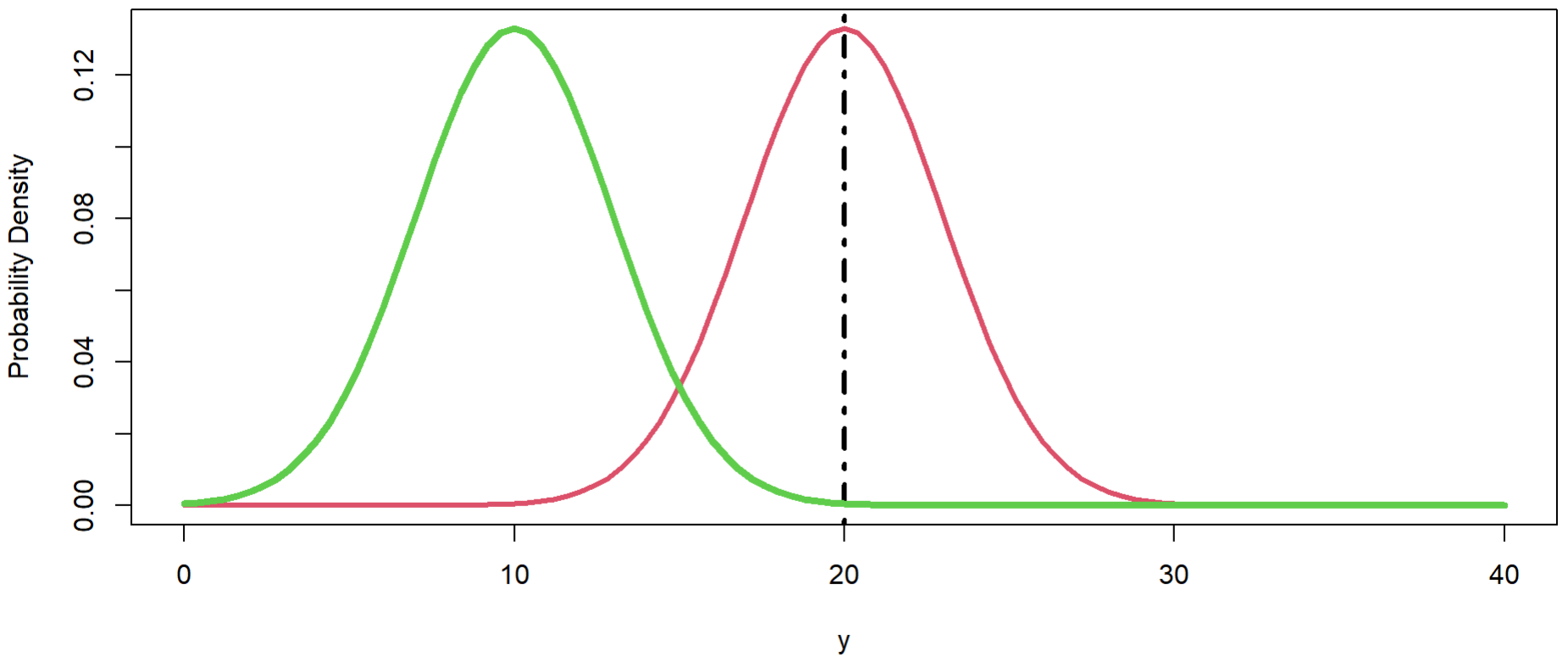
Normal Distribution (PDF Code)

```r
curve(dnorm(x, mean= 20, sd = 3),
      xlim=c(0,40),lwd=3,col=2,ylab="Probability Density",xlab="y")
abline(v=20, lwd=3, col=1, lty=4)
```



Normal Distribution (PDF Code)

```
curve(dnorm(x, mean = 10, sd = 3),xlim=c(0,40),lwd=4,col=3,add=TRUE)
```



Moments

Properties of all probability functions.

- 1^{st} moment is central tendency
- 2^{nd} moment is the dispersion
- ...

Normal Distribution: parameters (μ and σ) are 1^{st} and 2^{nd} moments

Moments

Gamma Distribution: parameters are not moments

Parameters

Shape = α, Rate = β

OR

Shape = κ, Scale = θ, where $\theta = \frac{1}{\beta}$

NOTE: probability functions can have Alternative Parameterizations, such they have different parameters.

Moments are functions of these parameters:

Gamma Distribution

Probability:

[Math Processing Error]

[Math Processing Error]

[Math Processing Error]

Sample/parameter Support:

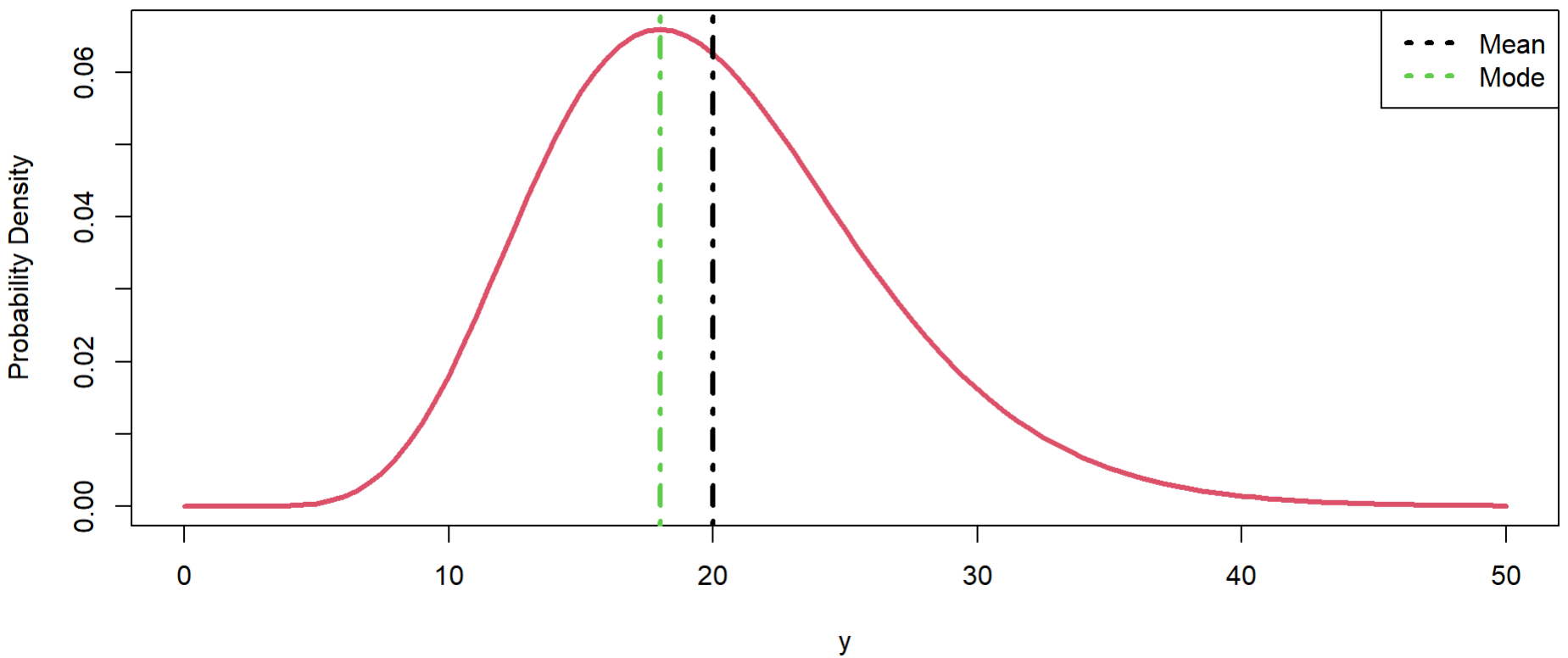
- $y \in (0, \infty)$
- $\kappa \in (0, \infty)$
- $\theta \in (0, \infty)$

```
[1] 20
```

```
[1] 18
```

```
[1] 6.324555
```

Gamma Distribution (PDF Code)



Gamma Distribution

What is the probability we would sample a value >40?

In this population, how common is a value >40?

[Math Processing Error]

```
1  pgamma(q=40, shape=10, scale=2,lower.tail=FALSE)
```

```
[1] 0.004995412
```

Gamma Distribution

What is the probability of observing $y < 20$

```
1 pgamma(q=20,shape=10, scale=2,lower.tail=TRUE)
```

```
[1] 0.5420703
```

Gamma Distribution

What is the probability of observing $20 < y < 40$

```
pgamma(q=40,shape=10, scale=2,lower.tail=TRUE)-
pgamma(q=20,shape=10, scale=2,lower.tail=TRUE)
```

```
[1] 0.4529343
```

Gamma Distribution

Reverse the question: What values of y and lower have a probability of 0.025

```
1  qgamma(p=0.025,shape=10, scale=2,lower.tail=TRUE)
```

```
[1] 9.590777
```

Gamma Distribution

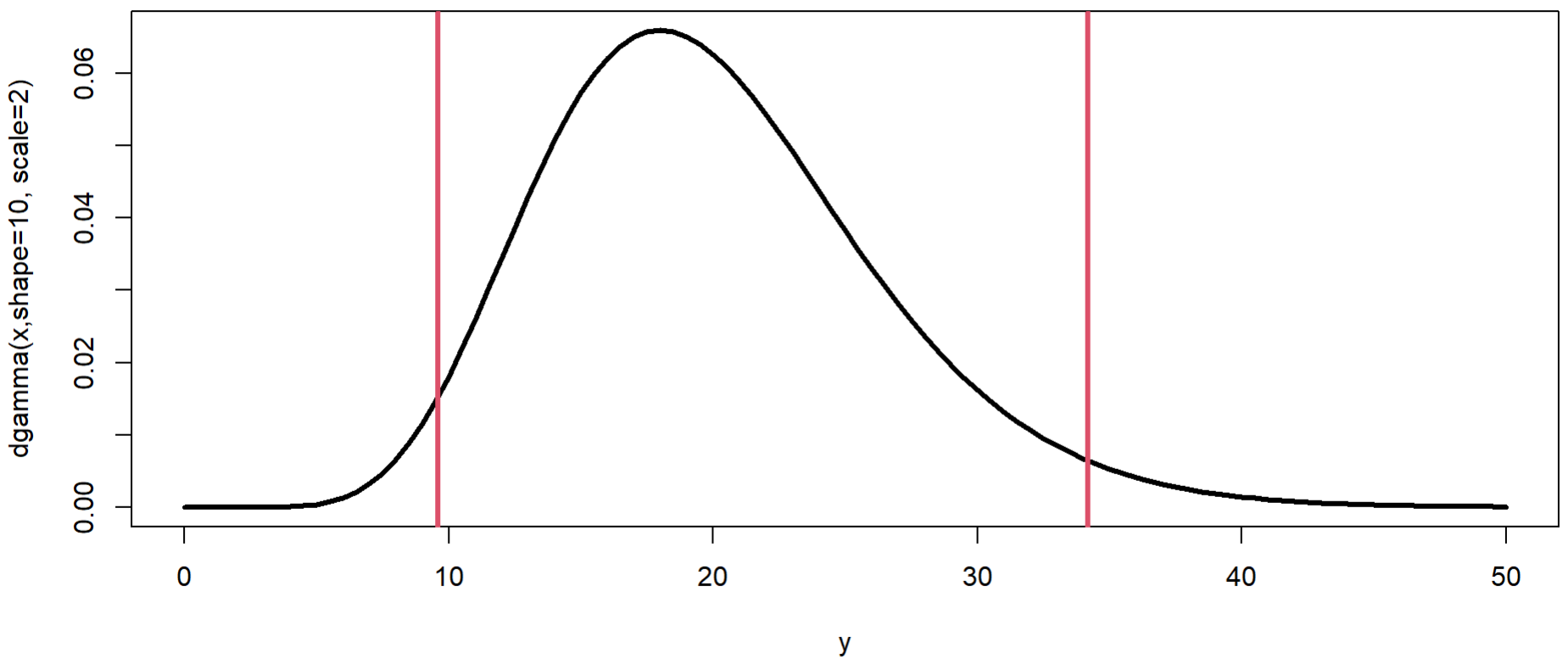
What values of y and higher have a probability of 0.025

```
1 qgamma(p=0.025,shape=10, scale=2,lower.tail=FALSE)
```

```
[1] 34.16961
```

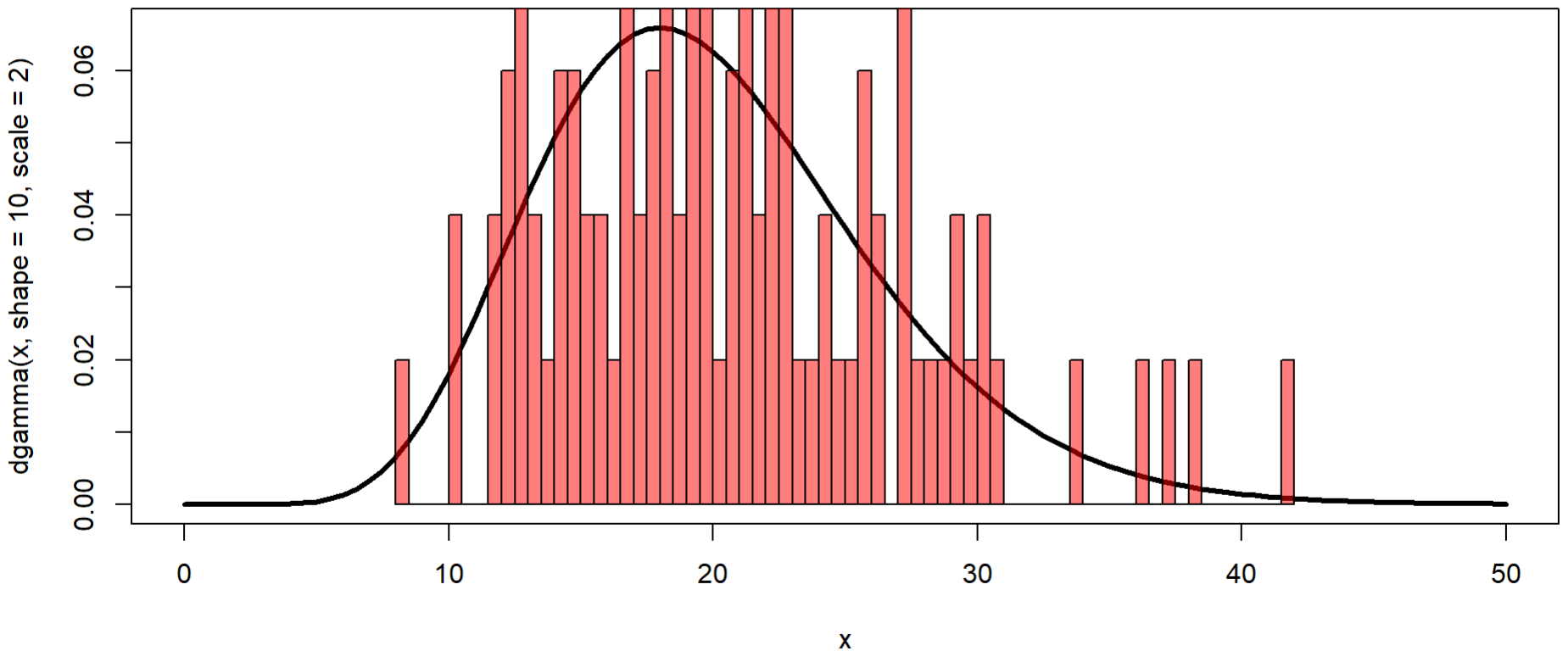
Gamma Distribution

```r
curve(dgamma(x,shape=10, scale=2),xlim=c(0,50),lwd=3,
      xlab="y", ylab="dgamma(x,shape=10, scale=2)")
abline(v=c(9.590777,34.16961),lwd=3,col=2)
```



We can consider samples from this population,

```
1  set.seed(154434)
2  y <- rgamma(100, shape=10, scale=2)
```



What do we know about this function?

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

- What is P(y = 1 | λ = 2) ?

```
1 dpois(1,lambda=2)
```

```
[1] 0.2706706
```

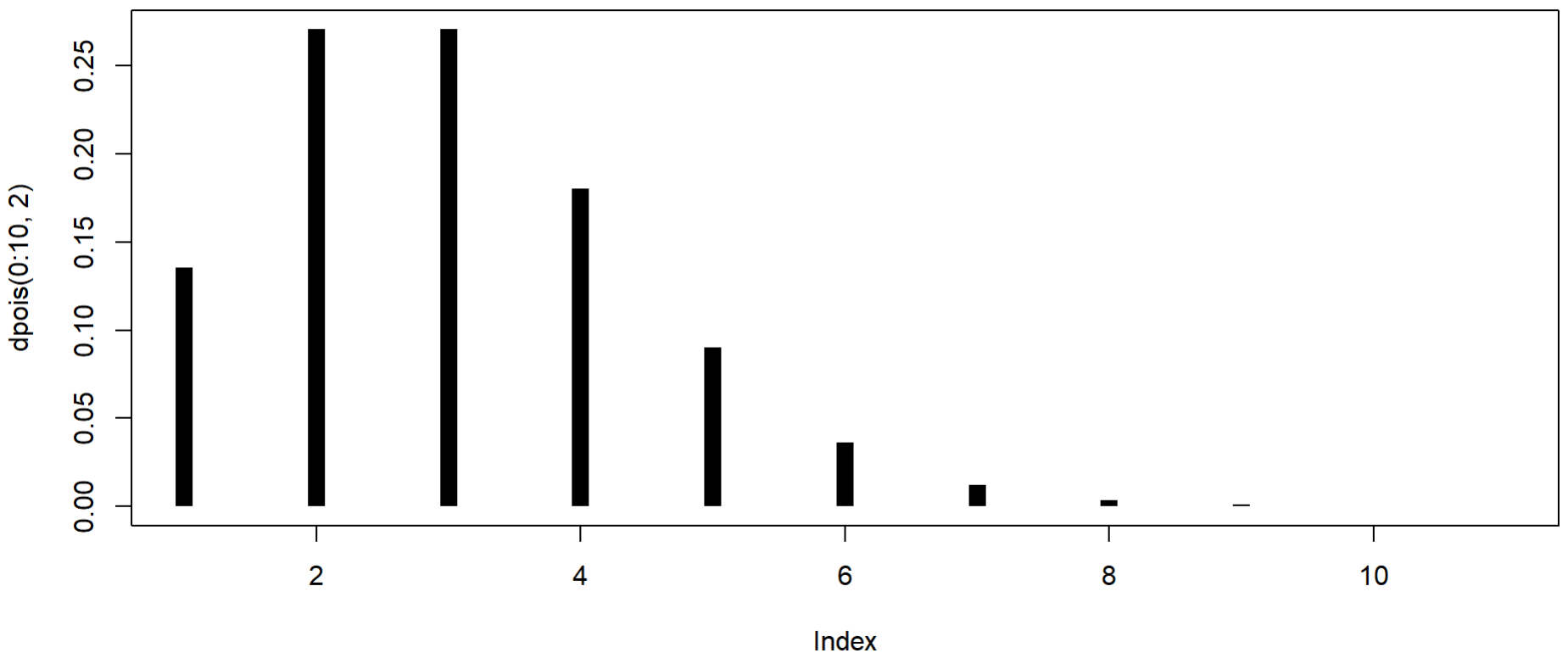
```
1 (2^1*exp(-2)) / (factorial(1))
```

```
[1] 0.2706706
```

Poisson

The full PMF (for $\lambda = 2$):

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



Poisson

- Many short-hands for probability distributions, e.g.,
  - $y \sim f(Y \mid \lambda)$
  - $y \sim f_Y(Y \mid \lambda)$
  - $y \sim$ Poisson(λ)
  - $y \sim$ Pois(λ)
  - $y \sim$ Poisson($Y \mid \lambda$)
  - $y \sim$ 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:

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

Named Probability Distributions

- Bernoulli, binomial, Poisson, normal/Gaussian or uniform distributions
  - Many, many, many more…. Rayleigh, Cauchy distributions, Pareto, von Mises)
- Good to really know at least a few!
  - e.g., build all of linear models, generalized linear models, mixed models

The others side of the coin

Statistics: Interested in estimating population-level characteristics; i.e., the parameters

[Math Processing Error]

REMEMBER

$f(y|\boldsymbol{\theta})$ is a probability statement about y, **NOT** $\boldsymbol{\theta}$.

Likelihood

Objectives

- likelihood principle
- likelihood connection to probability function
- optimization / parameter estimation

The others side of the coin

Statistics: Interested in estimating population-level characteristics; i.e., the parameters

[Math Processing Error]

Estimation

- likelihood
- Bayesian

Likelihood

Likelihood principle

All the evidence/information in a sample ($\mathbf{y}$, i.e., data) relevant to making inference on model parameters (θ) is contained in the *likelihood function*.

- "Conceptually simple, but in practice challenging for ecologists"

The pieces

- The sample data, $\mathbf{y}$
- A probability function for $\mathbf{y}$:
  - $f(\mathbf{y};\theta)$ or $[\mathbf{y}|\theta]$ or $P(\mathbf{y}|\theta)$
  - the unknown parameter(s) (θ) of the probability function

The Likelihood Function

[Math Processing Error]

The likelihood ($\mathcal{L}$) of the unknown parameters, given our data, can be calculated using our probability function.

The Likelihood Function

For example, for $y_1 \sim \text{Normal}(\mu, \sigma = 1)$

CODE:

```
# A data point
  y = c(10)

#the likelihood the mean is 8, given our data
  dnorm(y, mean = 8)
```

```
[1] 0.05399097
```

If we knew the mean is truly 8, it would also be the probability density of the observation y = 10. But, we don’t know what the mean truly is.

The Likelihood Function

For example, for $y_1 \sim \text{Normal}(\mu, \sigma = 1)$

The **key** is to understand that the likelihood values are relative, which means we need many guesses.

CODE:

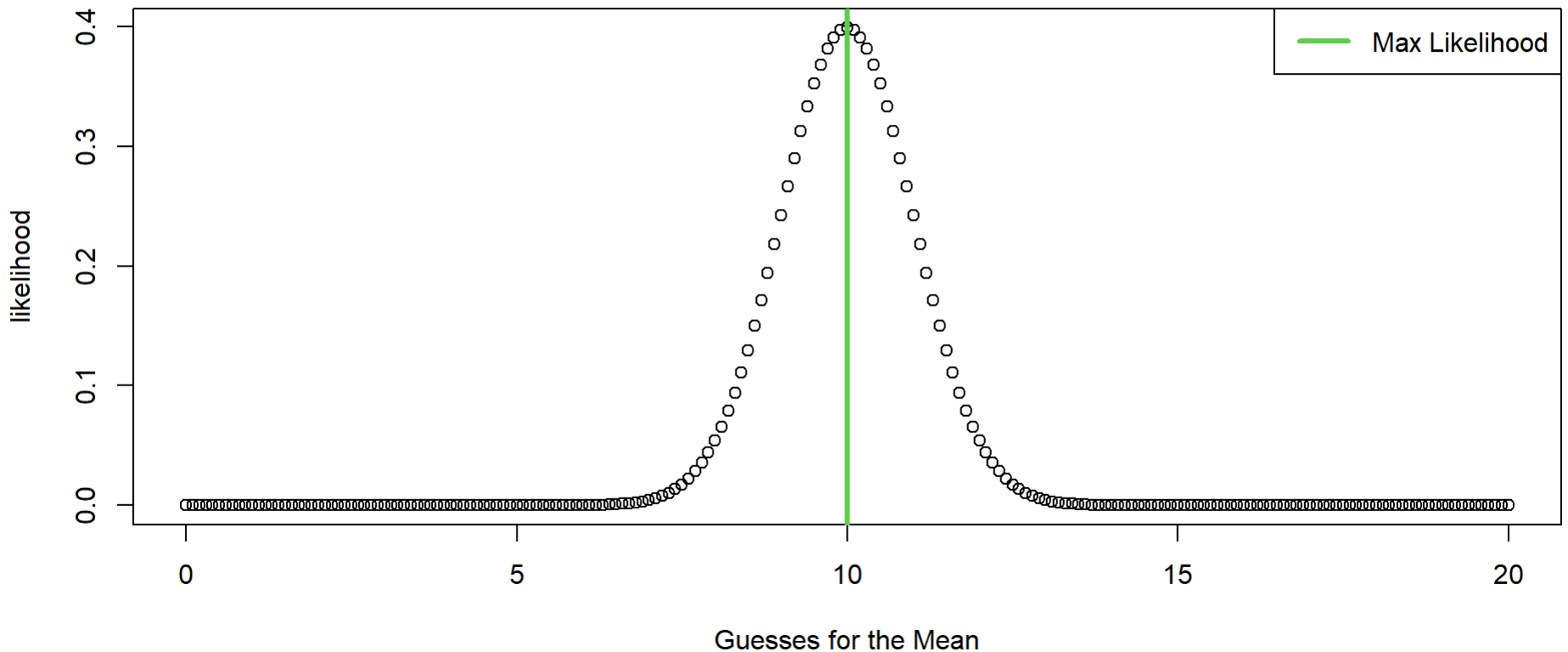
```
1 #the likelihood the mean is 9, given our data
2   dnorm(y, mean = 9)
```

```
[1] 0.2419707
```

Optimization

Grid Search

```
# many guesses of the mean
  means = seq(0, 20,by = 0.1)
# likelihood of each guess of the mean
  likelihood = dnorm(y, mean = means, sd = 1)
```


likelihood
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Guesses for the Mean
Max Likelihood

Maximum Likelihood Properties

- Central Tenet: evidence is relative.
- Parameters are not RVs. They are not defined by a PDF/PMF.
- MLEs are consistent. As sample size increases, they will converge to the true parameter value.
- MLEs are asymptotically unbiased. The $E[\hat{\theta}]$ converges to θ as the sample size gets larger.
- No guarantee that MLE is unbiased at small sample size.
- MLEs will have the minimum variance among all estimators, as the sample size gets larger.

MLE with n > 1

What is the mean height of King Penguins?

MLE with n > 1

We go and collect data,

[Math Processing Error]

Let’s decide to use the Normal Distribution as our PDF.

[Math Processing Error]

AND

[Math Processing Error]

AND

[Math Processing Error]

Need to connect data together

Or simply,

$$\mathbf{y} \overset{iid}{\sim} \text{Normal}(\mu, \sigma)$$

iid = independent and identically distributed

Need to connect data together

The joint probability of our data with shared parameters μ and σ,

[Math Processing Error]

$$P(Y_1 = 4.34, Y_2 = 3.53, Y_3 = 3.75 | \mu, \sigma)$$

Need to connect data together

IF each y_i is **independent**, the *likelihood* of our parameters is simply the multiplication of all three probability densities,

[Math Processing Error] [Math Processing Error]

Conditional Independence Assumption

We can do this because we are assuming knowing one observation does not tell us any new information about another observation.

$$P(y_2|y_1) = P(y_2)$$

Code

Translate the math to code…

```r
# penguin height data
  y = c(4.34, 3.53, 3.75)

# Joint likelihood of mu=3, sigma =1, given our data
  prod(dnorm(y, mean = 3,sd = 1))
```

```
[1] 0.01696987
```

Optimization Code (Grid Search)

Calculate likelihood of many guesses of μ and σ simultaneously,

```r
# The Guesses
  mu = seq(0,6,0.05)
  sigma = seq(0.01,2,0.05)
  try = expand.grid(mu,sigma)
  colnames(try) = c("mu","sigma")

# function
fun = function(a,b){
          prod(dnorm(y,mean = a, sd = b))
      }

# mapply the function with the inputs
  likelihood = mapply(a = try$mu, b = try$sigma, FUN=fun)
```

MLE Code

```
# maximum likelihood of parameters
  try[which.max(likelihood),]
```

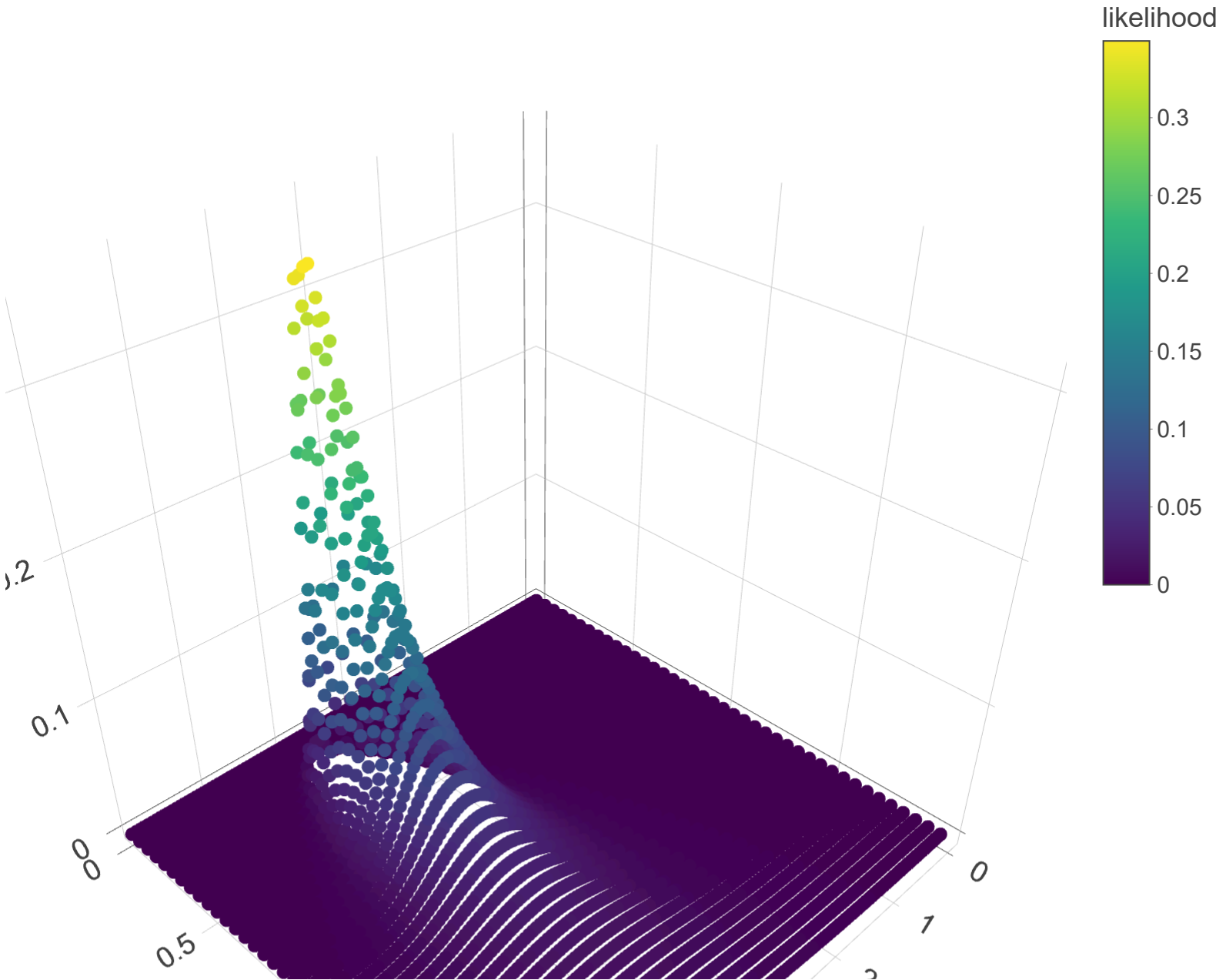
```
      mu sigma
925 3.85  0.36
```

Lets compare to,

```
sum(y)/length(y)
```

```
[1] 3.873333
```

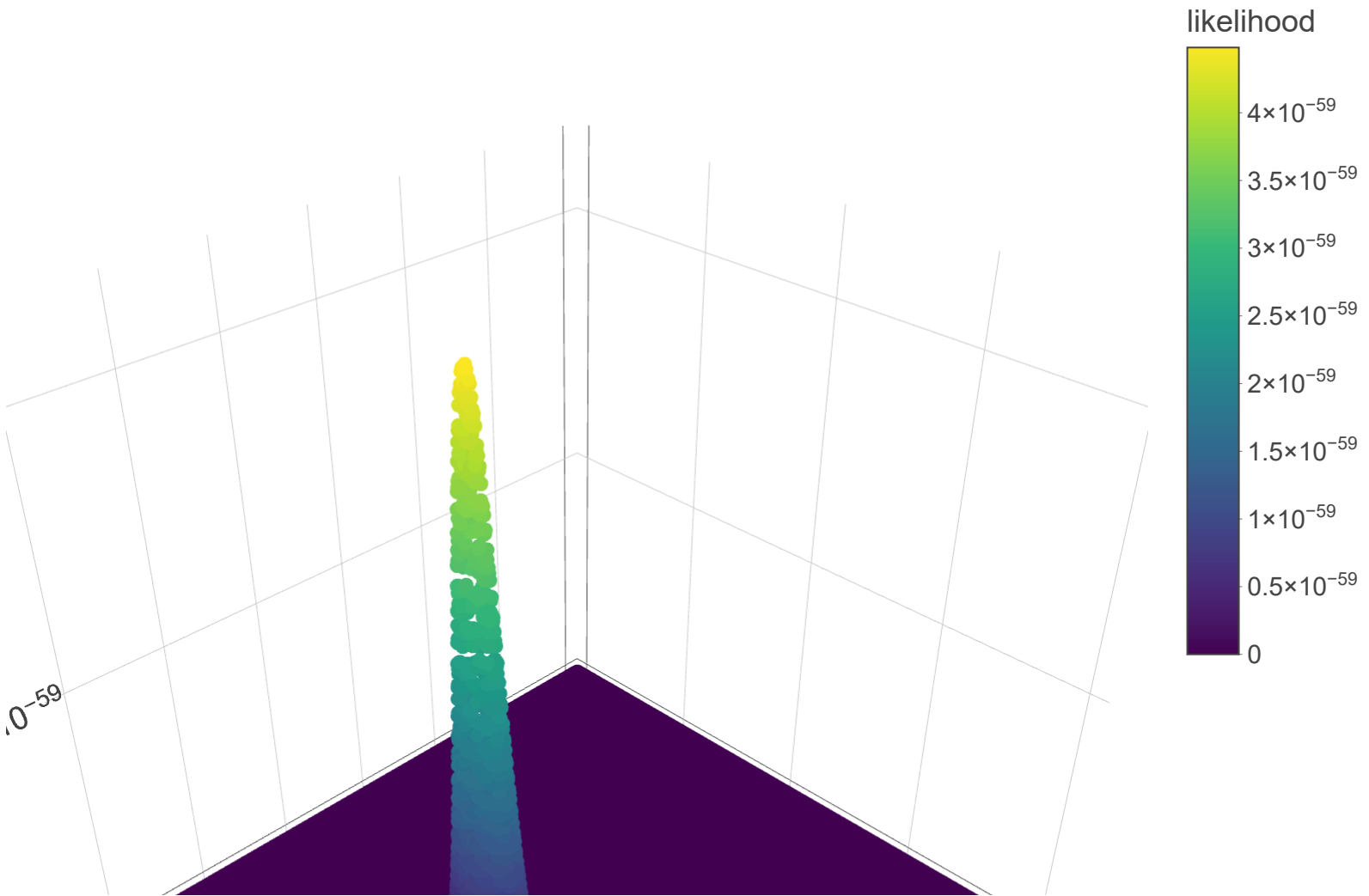
Likelihood plot (3D)

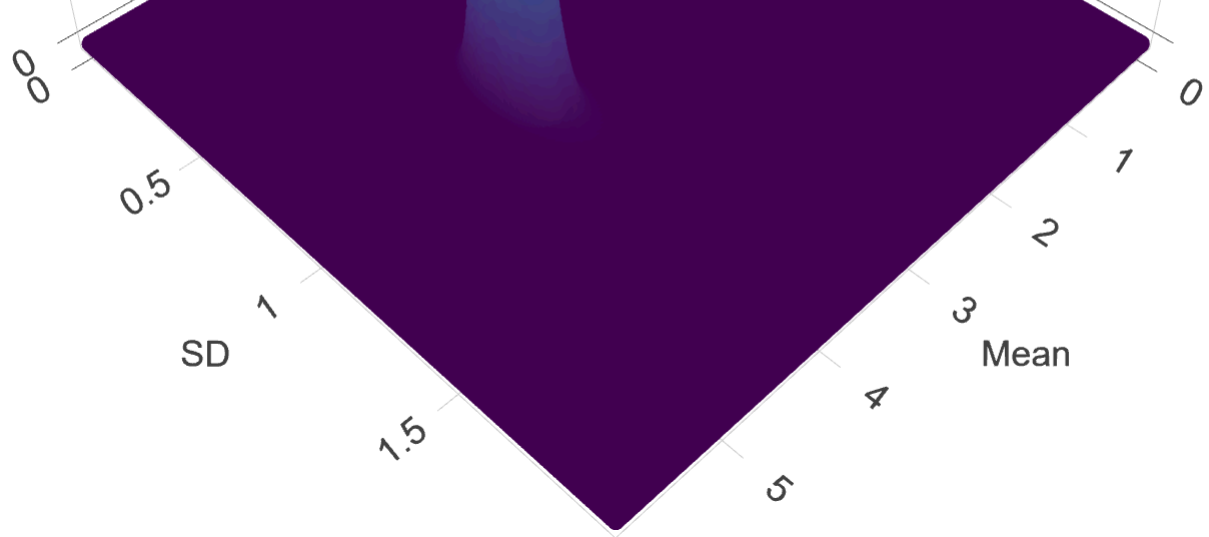


sigma
1
1.5
mu
3
4
5

Sample Size

What happens to the likelihood if we increase the sample size to N=100?




0
0.5
1
1.5
SD
0
1
2
3
4
5
Mean

Likelihood

[Math Processing Error]

Is the likelihood a probability?

Likelihood

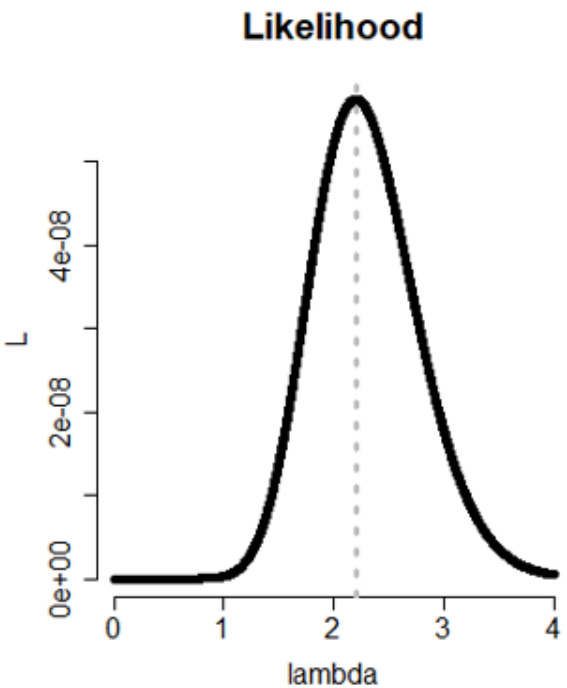
[Math Processing Error]

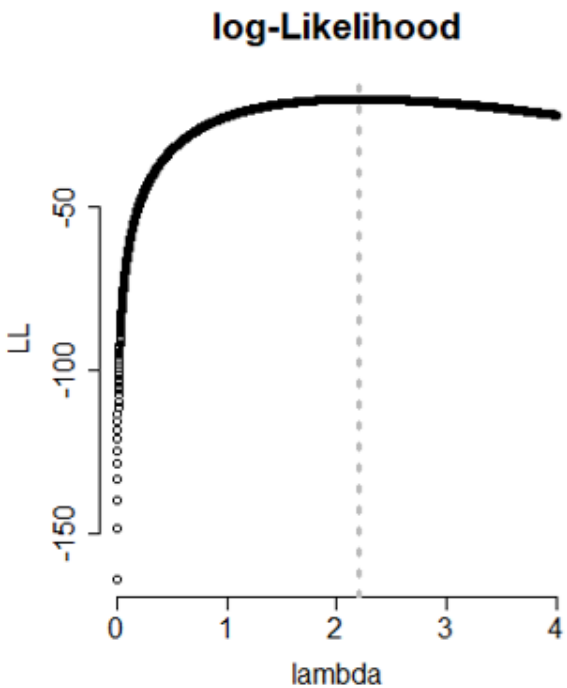
- Product of small numbers… makes computers sad!

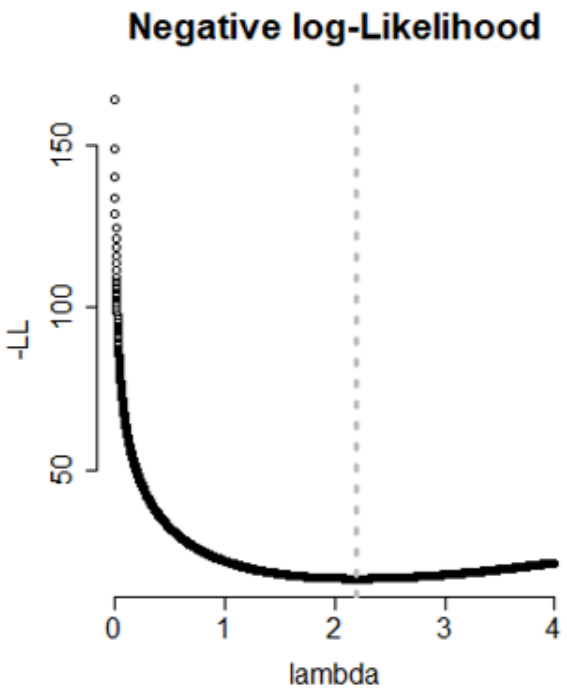
- Typically work with log-likelihood (sum of log densities)

[Math Processing Error]

Likelihood


Likelihood
L
0e+00
2e-08
4e-08
0
1
2
3
4
lambda


log-Likelihood
LL
-150
-100
-50
0
1
2
3
4
lambda


Negative log-Likelihood
-LL
50
100
150
0
1
2
3
4
lambda

Relativeness

log-likelihood

```r
fun.log = function(a,b){
              sum(dnorm(y,mean = a, sd = b, log=TRUE))
          }

log.likelihood = mapply(a = try$mu, b = try$sigma, FUN=fun.log)

# maximum log-likelihood of parameters
  try[which.max(log.likelihood),]
```

```
       mu sigma
55683 3.9  0.93
```

Optimization Code (Numerical)

Let's let the computer do some smarter guessing, i.e., optimization.

```r
# Note: optim function uses minimization, not maximization.
# WE want to find the minimum negative log-likelihood
# THUS, need to put negative in our function

neg.log.likelihood=function(par){
  -sum(dnorm(y,mean=par[1],sd=par[2],log=TRUE))
  }

#find the values that minimizes the function
#c(1,1) are the initial values for mu and sigma
fit <- optim(par=c(1,1), fn=neg.log.likelihood,
             method="L-BFGS-B",
             lower=c(0,0),upper=c(10,1)
             )

#Maximum likelihood estimates for mu and sigma
fit$par
```

Linear Regression

King Penguin Height Data (N=100)

```r
out = lm(y~1)
summary(out)
```

```
Call:
lm(formula = y ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00136 -0.61581  0.01208  0.67407  2.58369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9012     0.0932   41.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bayesian Inference

All things Bayesian

- Bayesian Inference
- Bayes Thereom
- Bayesian Components
- Conjugacy
- Hippo Case Study
- Bayesian Computation

Andrew Gelman's Blog

Statistical Modeling, Causal Inference, and Social Science

7 reasons to use Bayesian inference!

Posted on October 11, 2025 9:46 AM by Andrew

- Decision analysis
- Propagation of uncertainty
- Prior information
- Regularization
- Combining multiple sources of information
- Latent data and parameters.
  - When a model is full of parameters–perhaps even more parameters than data–you can't estimate them all.
- Enabling you to go further

Probability, Data, and Parameters

What do we want our model to tell us?

Do we want to make probability statements about our data?

Likelihood = P(data|parameters)

90% CI: the long-run proportion of corresponding CIs that will contain the true value 90% of the time.

Probability, Data, and Parameters

What do we want our model to tell us?

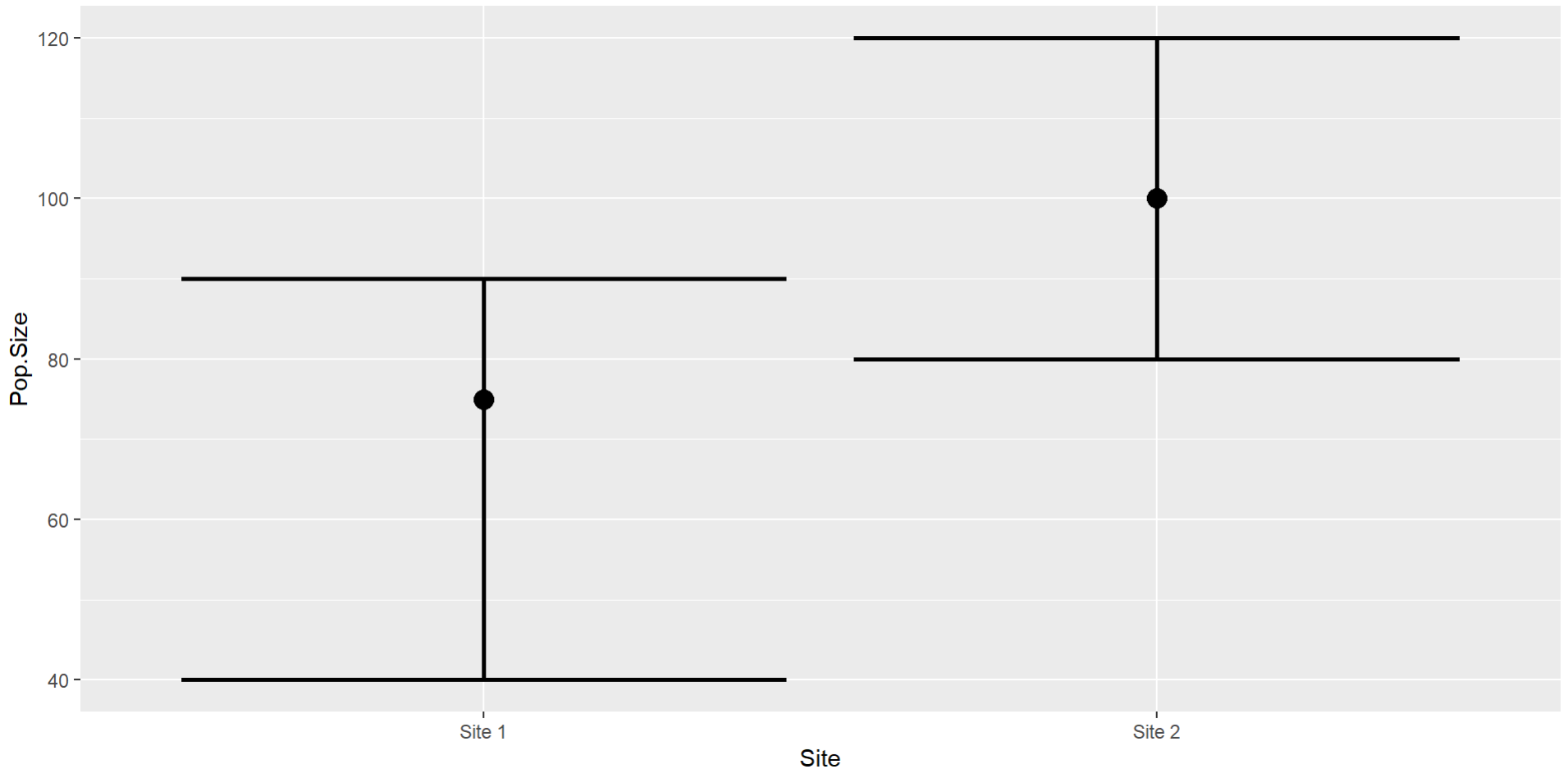
Do we want to make probability statements about our parameters?

Posterior = P(parameters|data)

Alternative Interval: 90% probability that the true value lies within the interval, given the evidence from the observed data.

Likelihood Inference

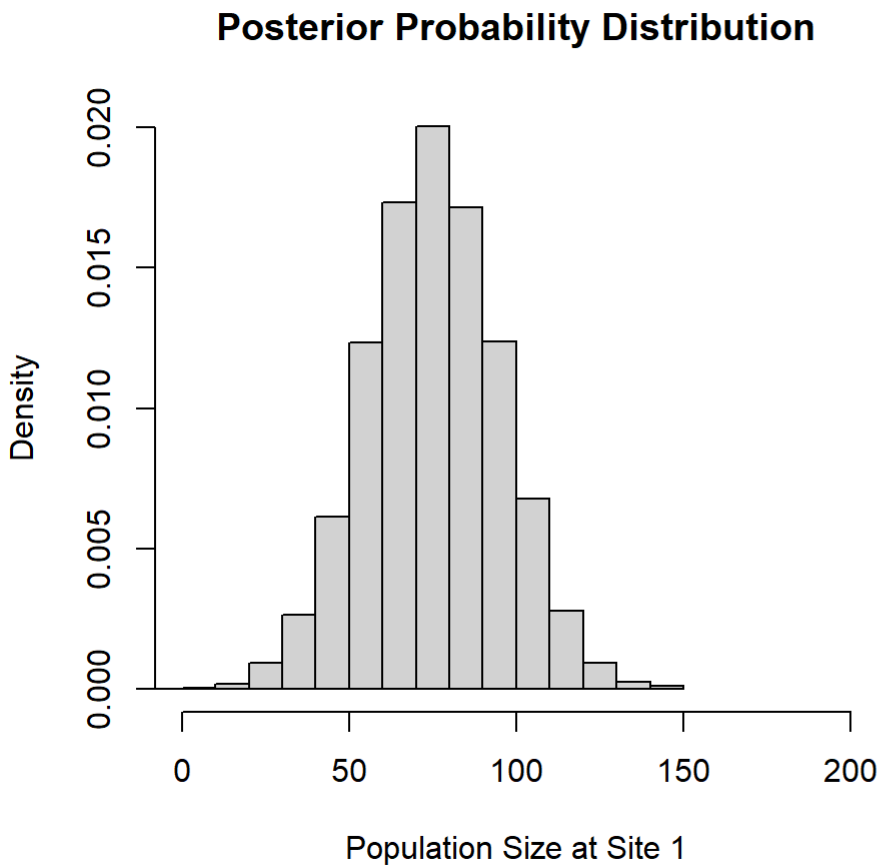
Estimate of the population size of hedgehogs at two sites.

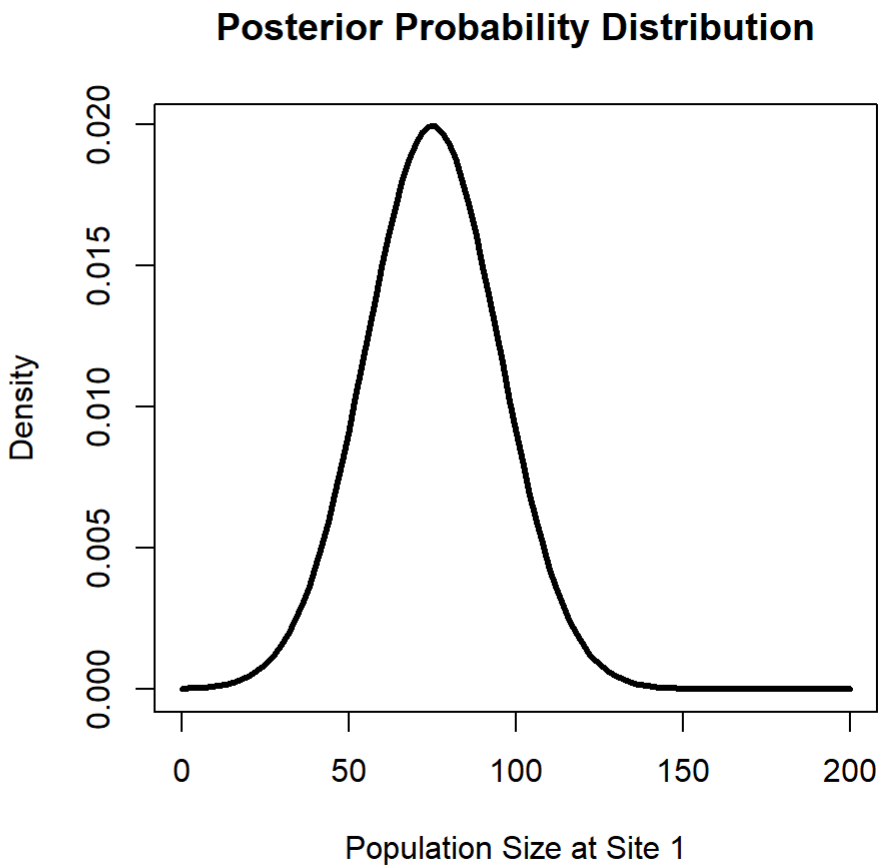


Bayesian Inference

Posterior Samples

```
[1] 102.67671  81.11546  81.75260  87.77246  73.99043  80.70631  76.26219
[8]  83.99927  64.74208  26.93133
```

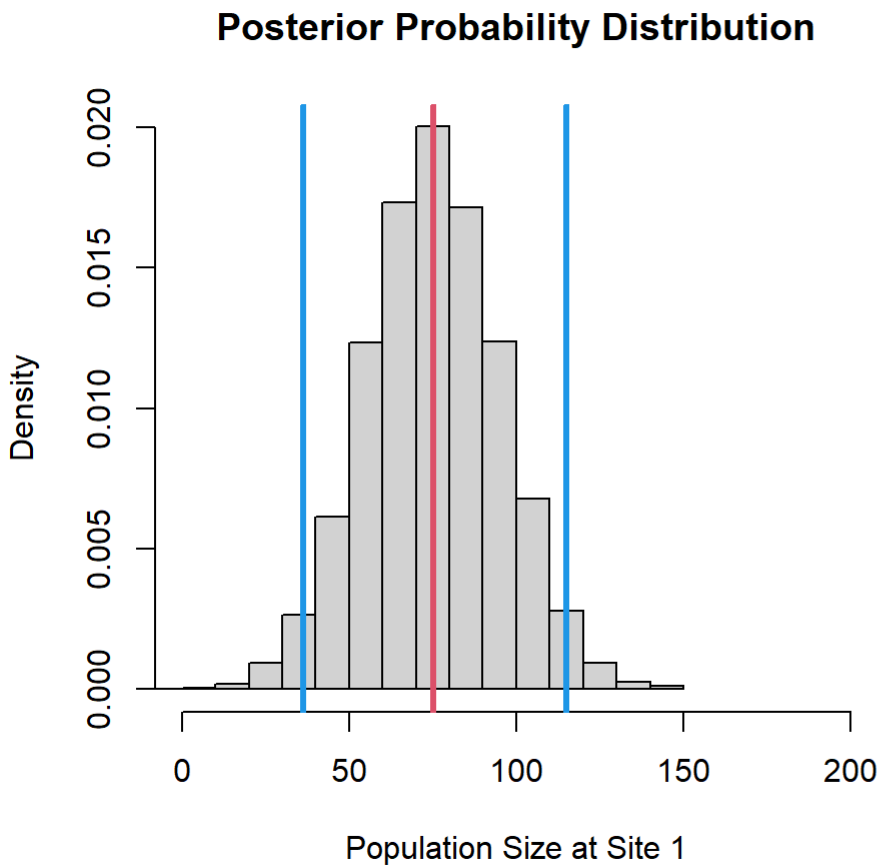

Posterior Probability Distribution
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
Population Size at Site 1

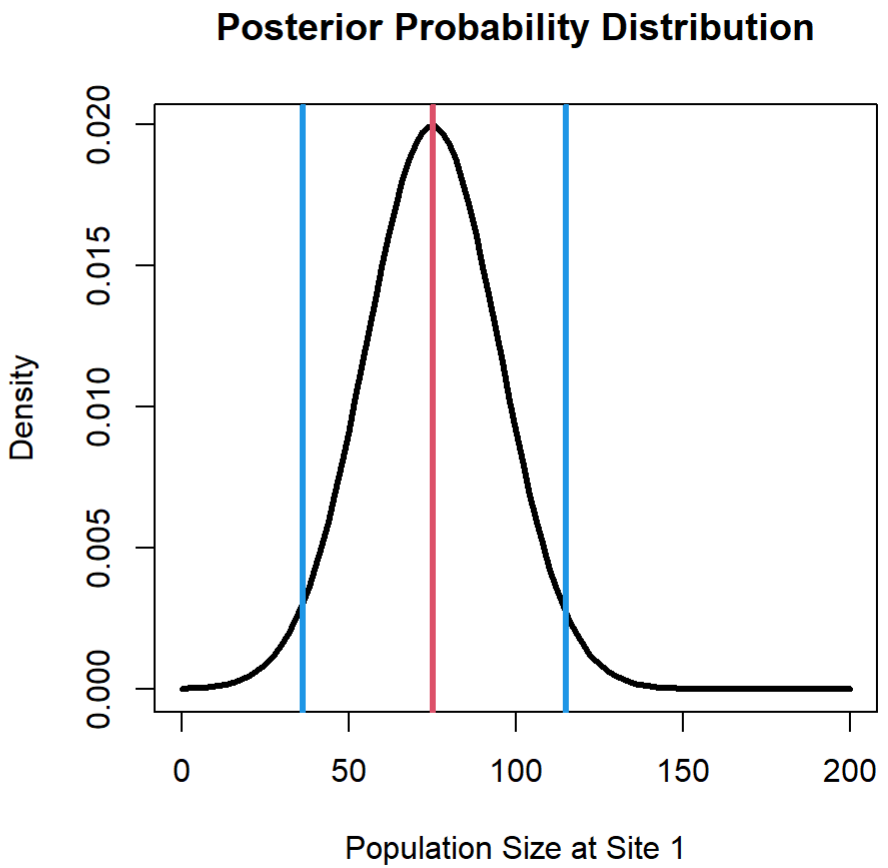

Posterior Probability Distribution
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
Population Size at Site 1

Bayesian Inference

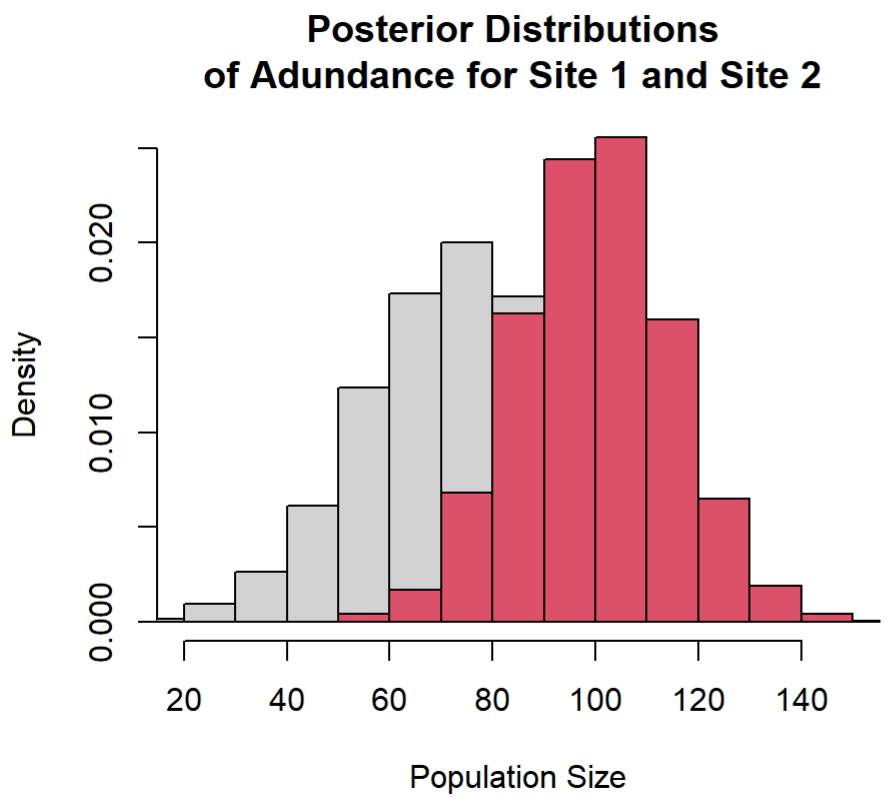
Posterior Samples

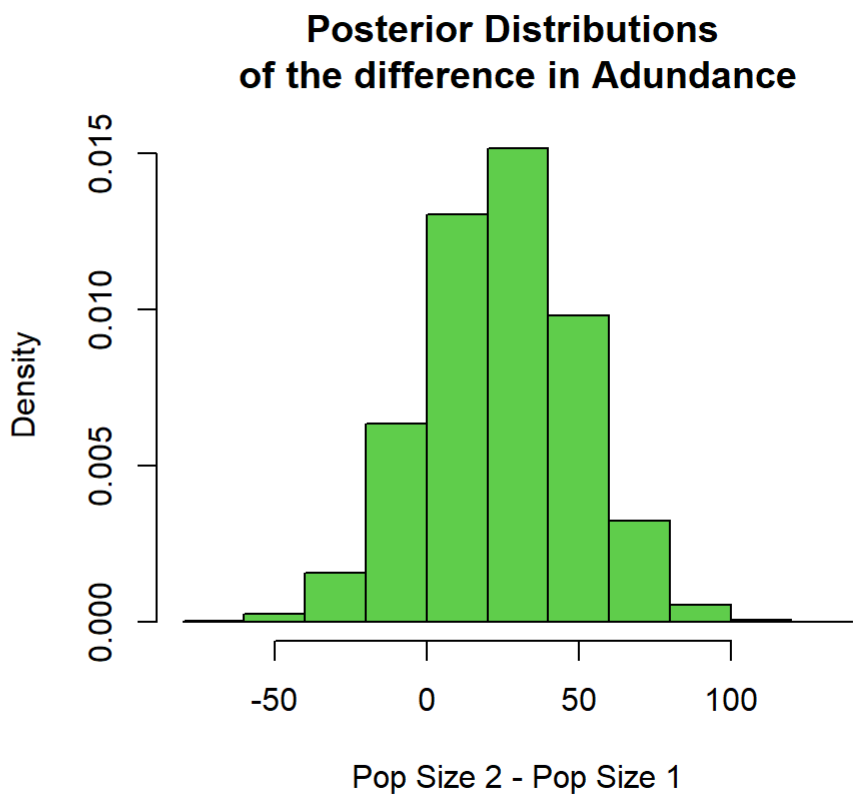
```
[1] 102.67671  81.11546  81.75260  87.77246  73.99043  80.70631  76.26219
[8]  83.99927  64.74208  26.93133
```


Posterior Probability Distribution
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
Population Size at Site 1


Posterior Probability Distribution
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
Population Size at Site 1

Bayesian Inference






```
1  diff=post2-post1
2  length(which(diff>0))/length(diff)
```

```
[1] 0.8362
```

Likelihood Inference (coeficient)

y is Body size of a beetle species

x is elevation

```
summary(glm(y~x))
```

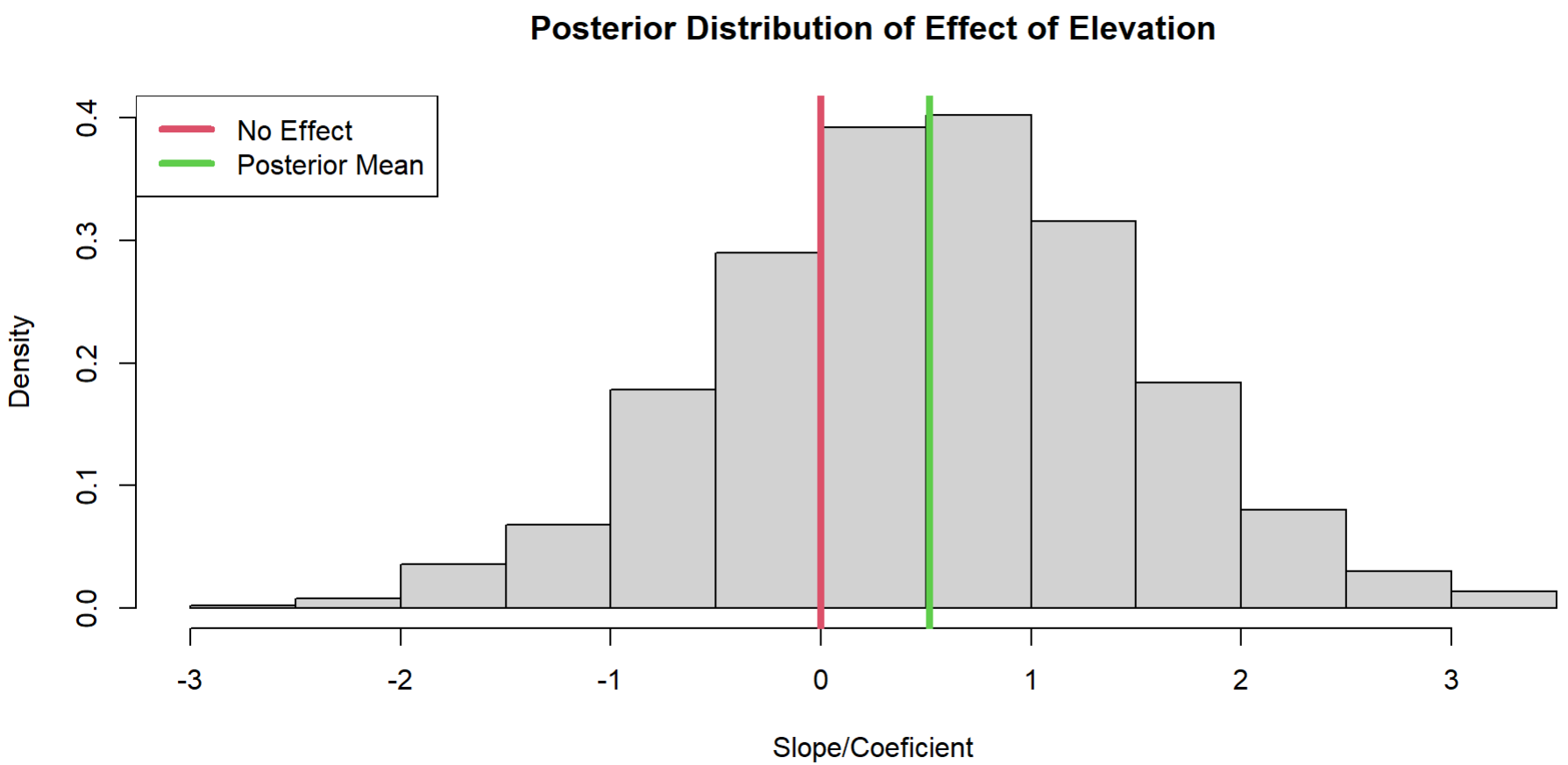
```
Call:
glm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9862     0.2435   4.049 0.000684 ***
x             0.5089     0.4022   1.265 0.221093
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.245548)

    Null deviance: 25.659  on 20  degrees of freedom
```

Bayesian Inference (coeficient)



Bayesian Inference (coeficient)

```r
#Posterior Mean
  mean(post)
```

```
[1] 0.5185186
```

```r
#Credible/Probability Intervals
  quantile(post,prob=c(0.025,0.975))
```

```
     2.5%     97.5% 
-1.461887  2.403232 
```

```r
# #Probabilty of a postive effect
 length(which(post>0))/length(post)
```

```
[1] 0.709
```

```r
# #Probabilty of a negative effect
 length(which(post<0))/length(post)
```

```
[1] 0.291
```

Bayesian Theorem

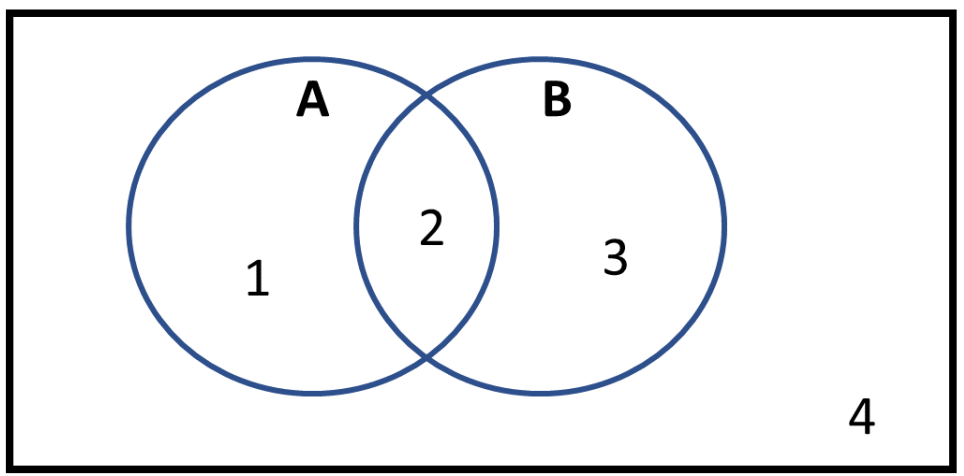
Bayes Theorem

Link

- Marginal Probability
  - $P(A)$
  - $P(B)$

▶ Click for Answer

Sampled N = 10 locations

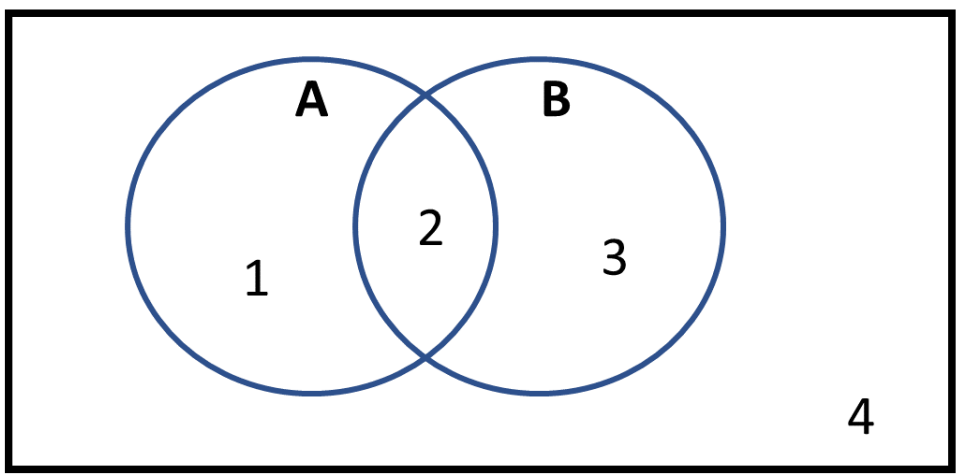


Bayes Theorem

- Joint Probability
  - $P(A \cap B)$

▶ Click for Answer

Sampled N = 10 locations

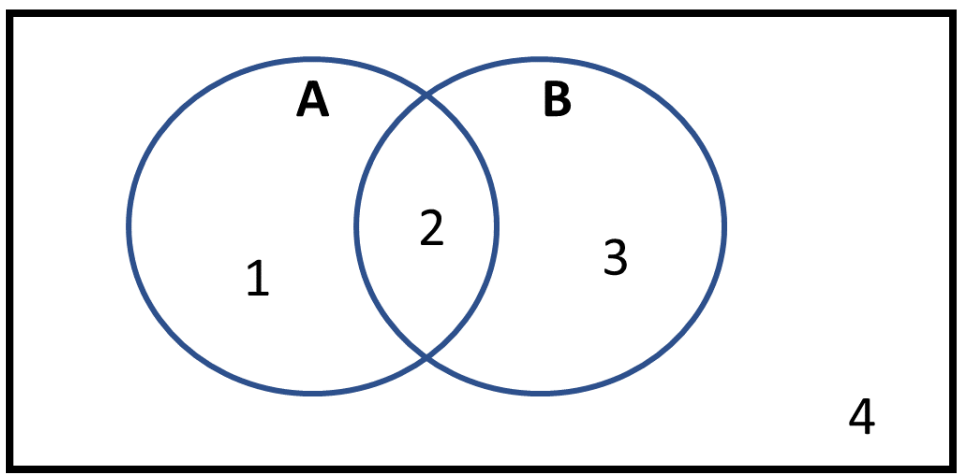


Bayes Theorem

- Conditional Probability
  - $P(A|B)$
  - $P(B|A)$

▶ Click for Answer

Sampled N = 10 locations

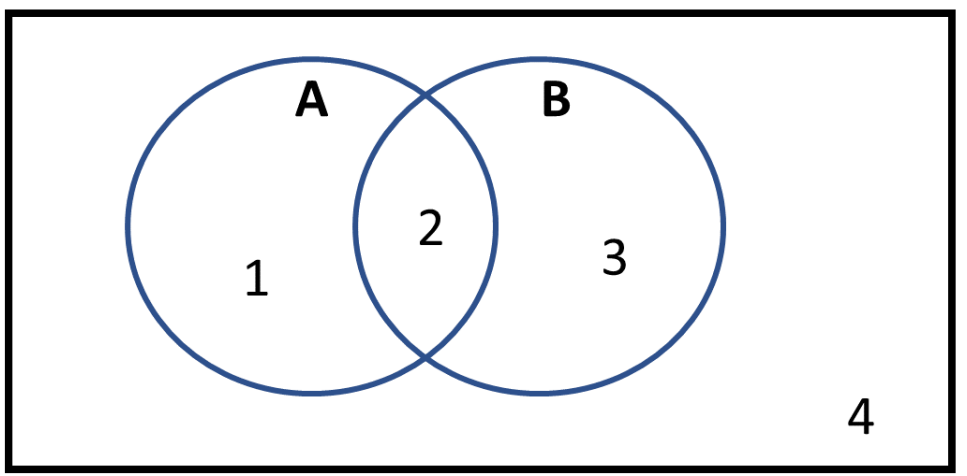


Bayes Theorem

- OR Probability
  - $P(A \cup B)$

▶ Click for Answer

Sampled N = 10 locations



Notice that...

$$P(A \cap B) = 0.2$$

$$P(A|B)P(B) = 0.4 \times 0.5 = 0.2$$

$$P(B|A)P(A) = 0.6666 \times 0.3 = 0.2$$

$$P(A|B)P(B) = P(A \cap B)$$

$$P(B|A)P(A) = P(A \cap B)$$

$$P(B|A)P(A) = P(A|B)P(B)$$

Bayes Theoreom

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

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

Bayesian Components

Bayes Components

param = parameters

$$P(\text{param}|\text{data}) = \frac{P(\text{data}|\text{param})P(\text{param})}{P(\text{data})}$$

Posterior Probability/Belief

Likelihood

Prior Probability

Evidence or Marginal Likelihood

Bayes Components

param = parameters

$$P(\text{param}|\text{data}) = \frac{P(\text{data}|\text{param})P(\text{param})}{\int_{\forall\ \text{Param}} P(\text{data}|\text{param})P(\text{param})}$$

Posterior Probability/Belief

Likelihood

Prior Probability

Evidence or Marginal Likelihood

Bayes Components

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$\text{Posterior} \propto \text{Likelihood}$$

Bayesian Steps

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 $p(y|\theta)$
- Out comes another probability distribution, the posterior distribution $p(\theta|y)$, which is our new state of knowledge

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

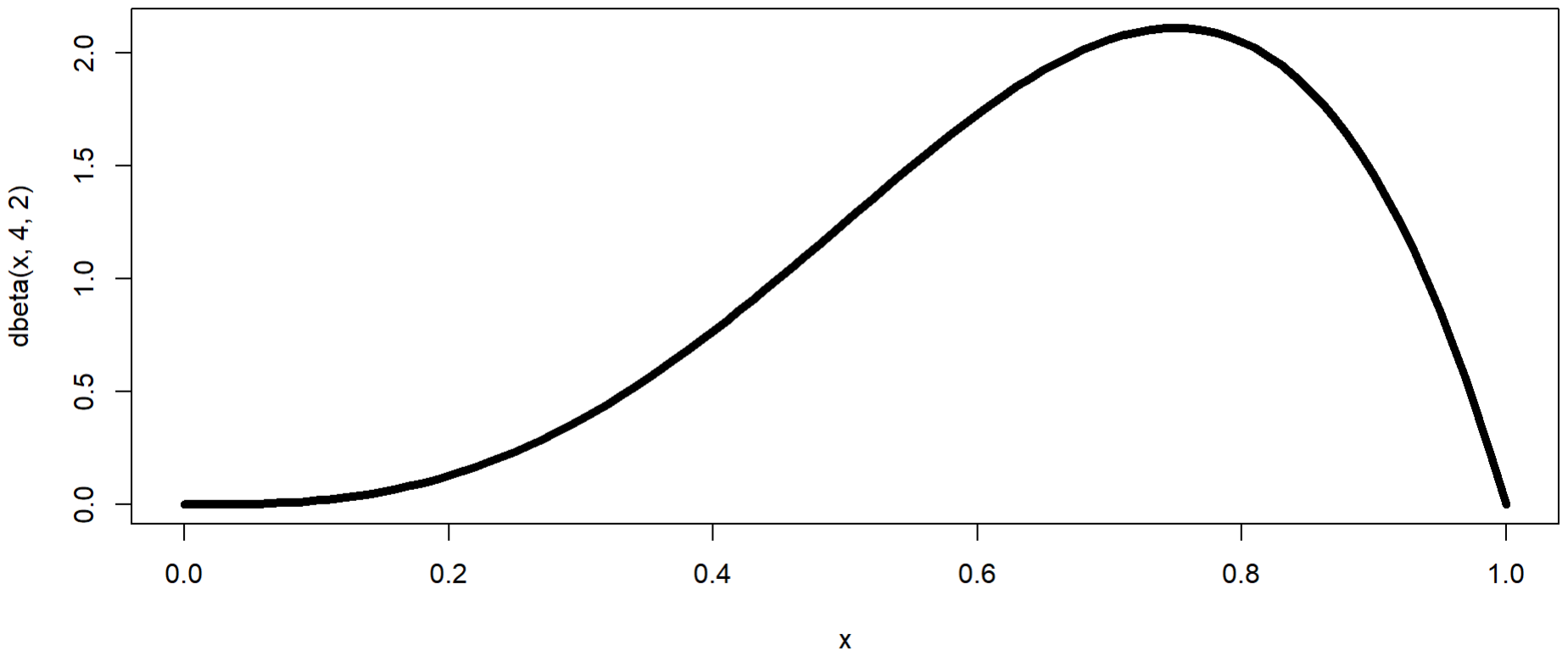
The Prior

All parameters in a Bayesian model require a prior specified; parameters are random variables.

$$y_i \sim \text{Binom}(N, \theta)$$

$$\theta \sim \text{Beta}(\alpha = 4, \beta = 2)$$

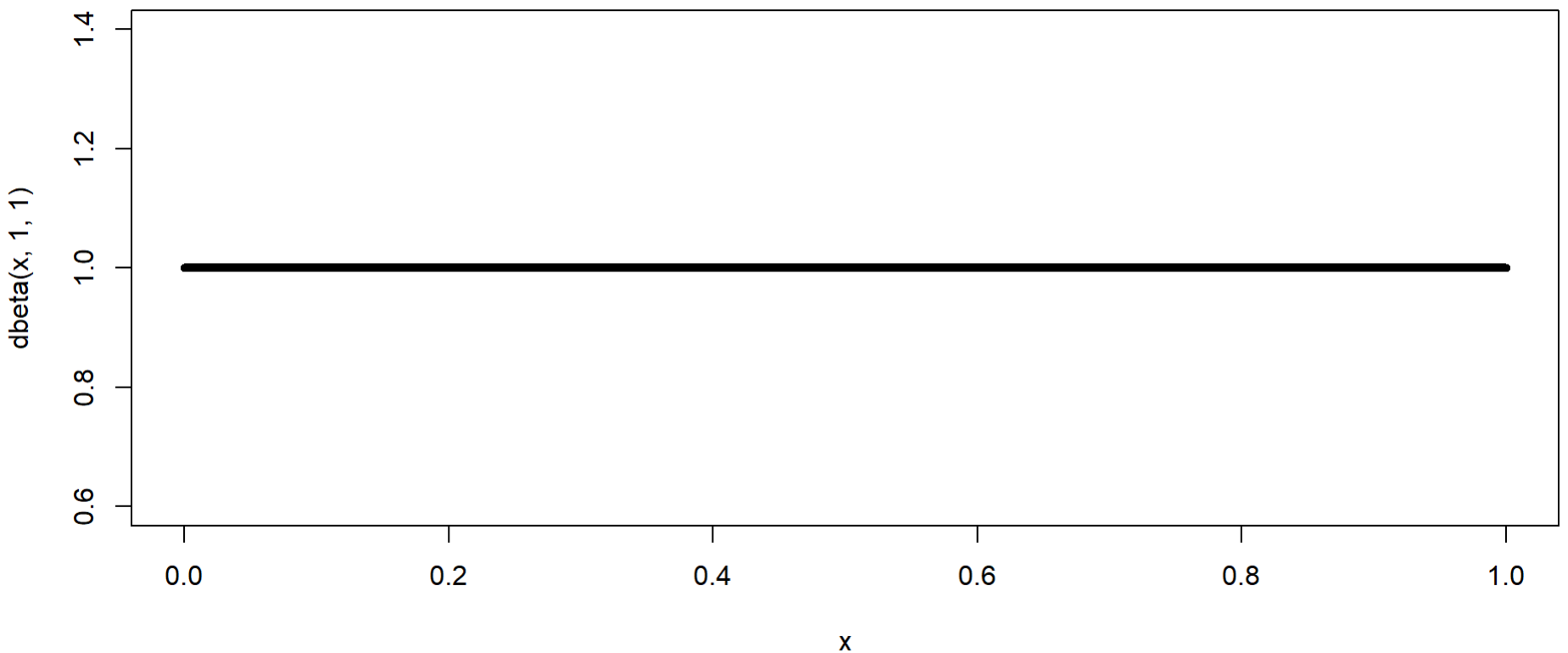
```
1  curve(dbeta(x, 4,2),xlim=c(0,1),lwd=5)
```


dbeta(x, 4, 2)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
x

The Prior

$$\theta \sim \text{Beta}(\alpha = 1, \beta = 1)$$

```
1  curve(dbeta(x, 1,1),xlim=c(0,1),lwd=5)
```



The Prior

- The prior describes what we know about the parameter before we collect any data

- Priors can contain a lot of information (informative priors) or very little (diffuse priors); No such thing as a ‘non-informative’ prior
- “Well-constructed” priors can also improve the behavior of our models (computational advantage)

The Prior

Use diffuse priors as a starting point

It’s fine to use diffuse priors as you develop your model but you should always prefer to use “appropriate, well-contructed informative priors” (Hobbs & Hooten, 2015)

The Prior

Use your “domain knowledge”

We can often come up with weakly informative priors just by knowing something about the range of plausible values of our parameters.

The Prior

Dive into the literature

Find published estimates and use moment matching and other methods to convert published estimates into prior distributions

The Prior

Gateway to regularization / model selection and optimal predictions

The Prior

Visualize your prior distribution

Be sure to look at the prior in terms of the parameters you want to make inferences on

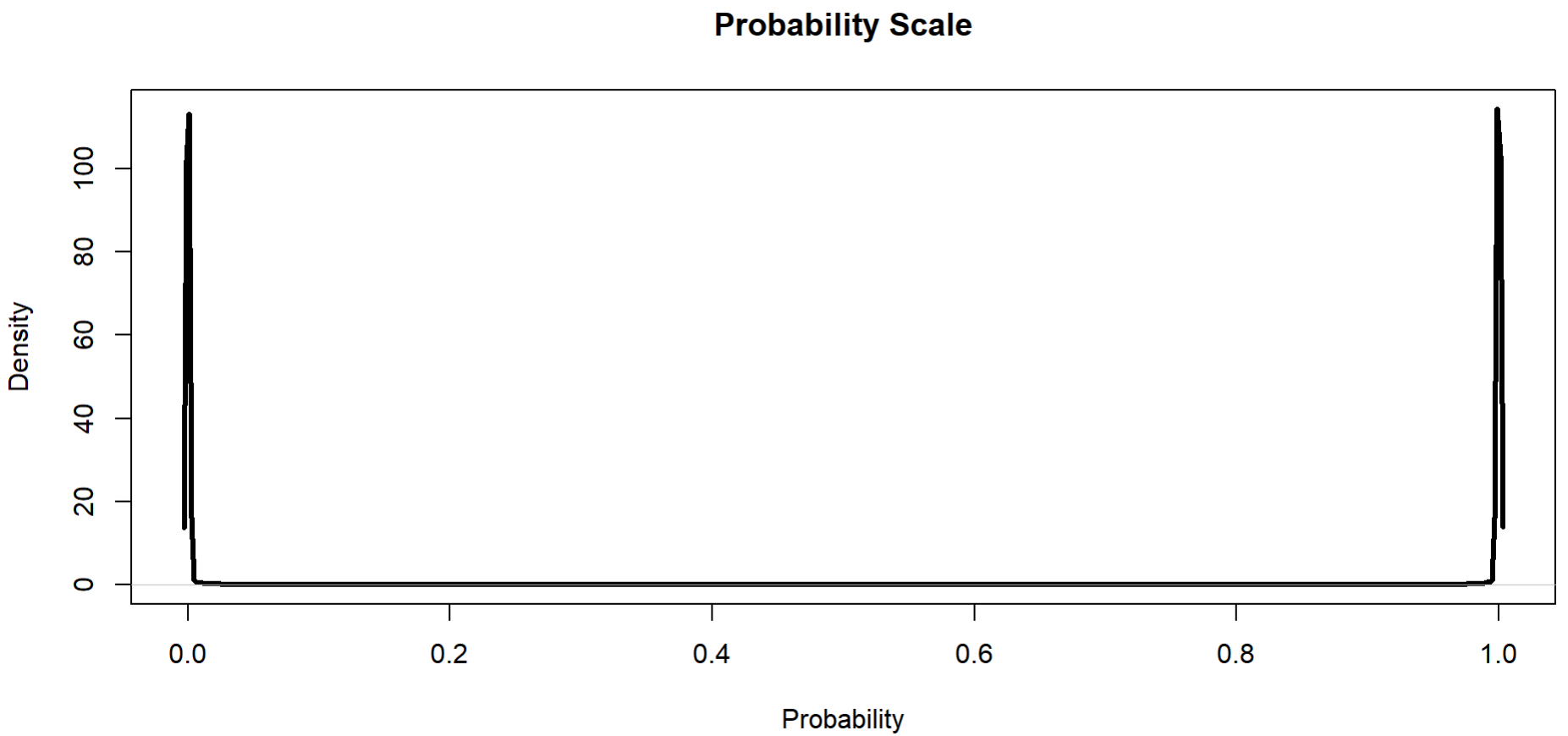
The Prior

Your prior is probabably not invariant to transformations

- $\beta \sim \text{Normal}(\mu = 0, \sigma = 100)$

The Prior

Your prior is probabably not invariant to transformations



The Prior

Do a sensitivity analysis

Does changing the prior change your posterior inference?

The Prior

Are priors bad?

Do they compromise scientific objecivity?

The Prior

- 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

Conjugacy

Bayesian Model

Model

$$\mathbf{y} \sim \text{Bernoulli}(p)$$

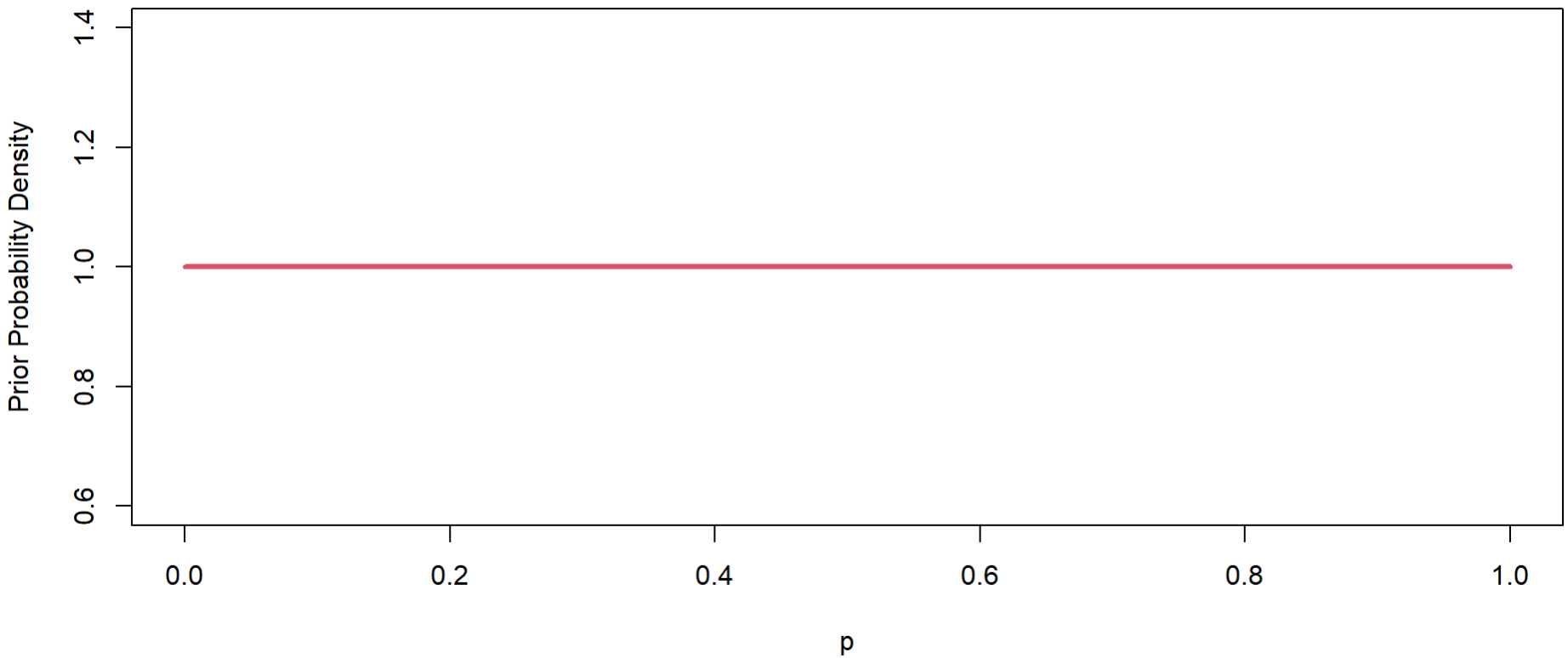
Prior

$$p \sim \text{Beta}(\alpha, \beta)$$

These are called Prior hyperparameters

$$\alpha = 1$$
$$\beta = 1$$

```r
curve(dbeta(x,1,1),xlim=c(0,1),lwd=3,col=2,xlab="p",
      ylab = "Prior Probability Density")
```



Conjugate Distribution

Likelihood (Joint Probability of y)

$$\mathscr{L}(p|y) = \prod_{i=1}^{n} P(y_i|p) = \prod_{i=1}^{N} (p^y (1-p)^{1-y_i})$$

Prior Distribution

$$P(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

Posterior Distribution of p

$$P(p|y) = \frac{\prod_{i=1}^{N}(p^{y}(1-p)^{1-y_i}) \times \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\int_p(\text{numerator})}$$

CONJUGACY!

$$P(p|y) \sim \text{Beta}(\alpha^*, \beta^*)$$

α^* and β^* are called Posterior hyperparameters

Case Study

Hippos

We do a small study on hippo survival and get these data…

7 Hippos Died

2 Hippos Lived

Hippos: Likelihood Model

[Math Processing Error]

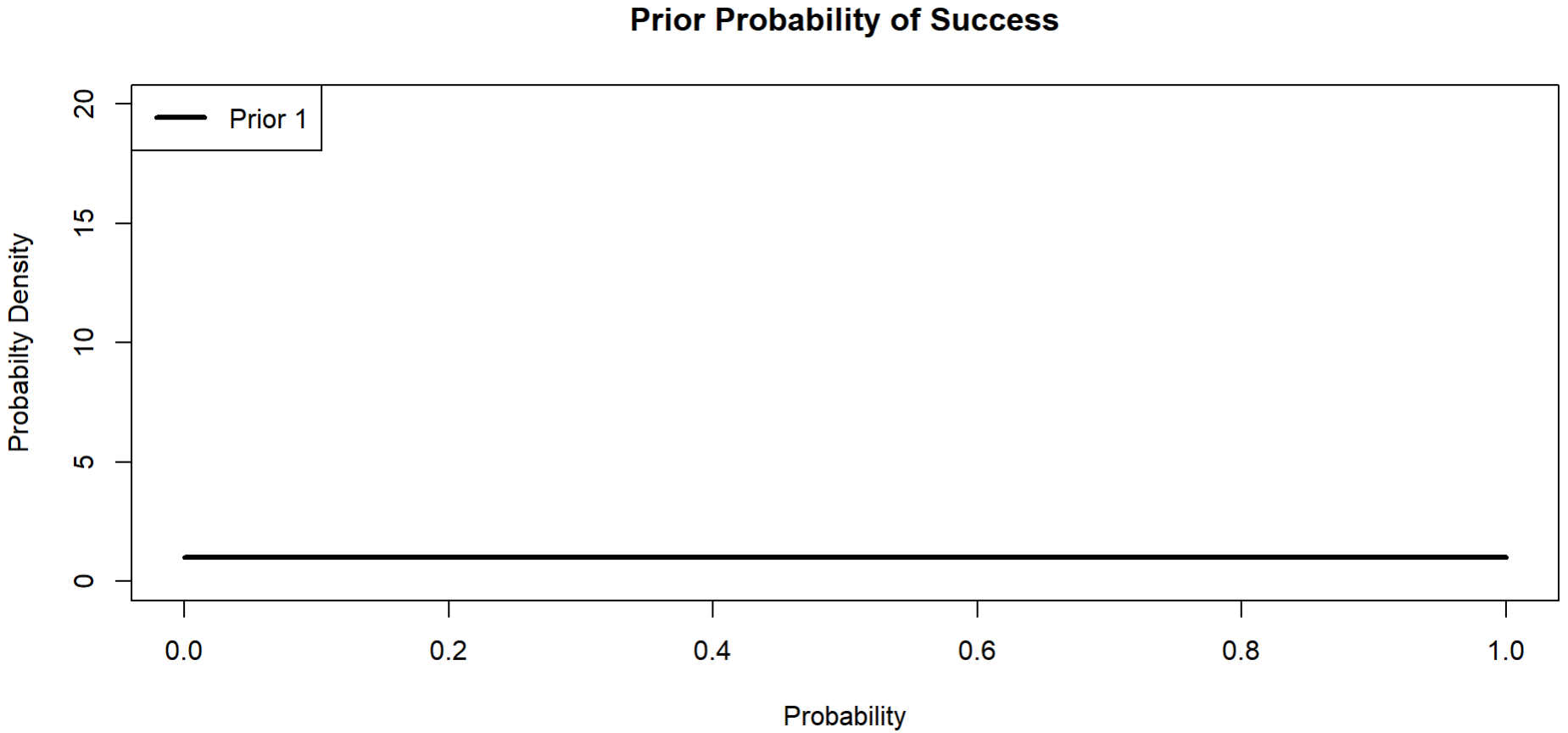
```
# Survival outcomes of three adult hippos
  y1=c(0,0,0,0,0,0,0,1,1)
  N1=length(y1)
  mle.p=mean(y1)
  mle.p
```

```
[1] 0.2222222
```

Hippos: Bayesian Model (Prior 1)

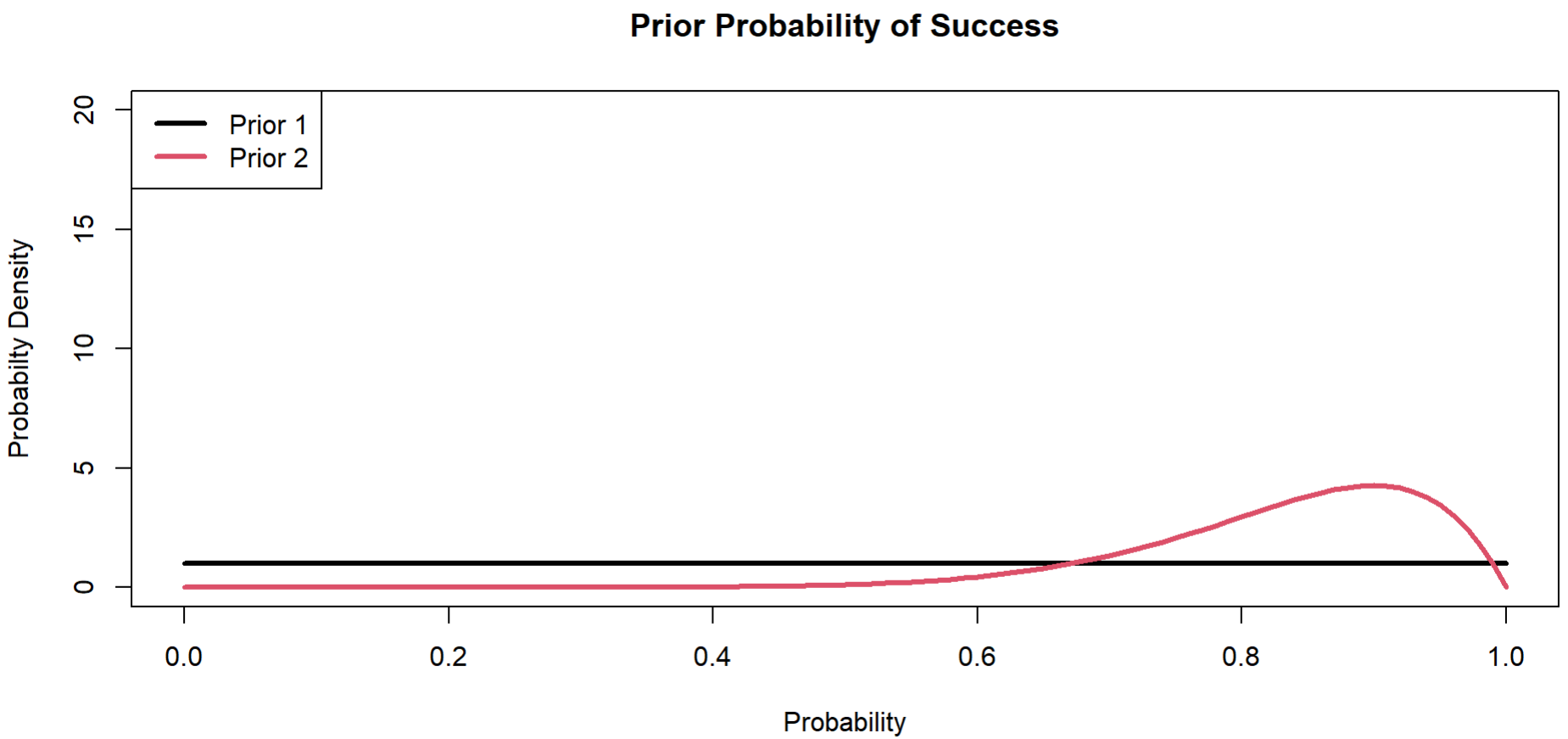
[Math Processing Error]

```
alpha.prior1=1
beta.prior1=1
```


Prior Probability of Success
Prior 1
Probabilty Density
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
Probability

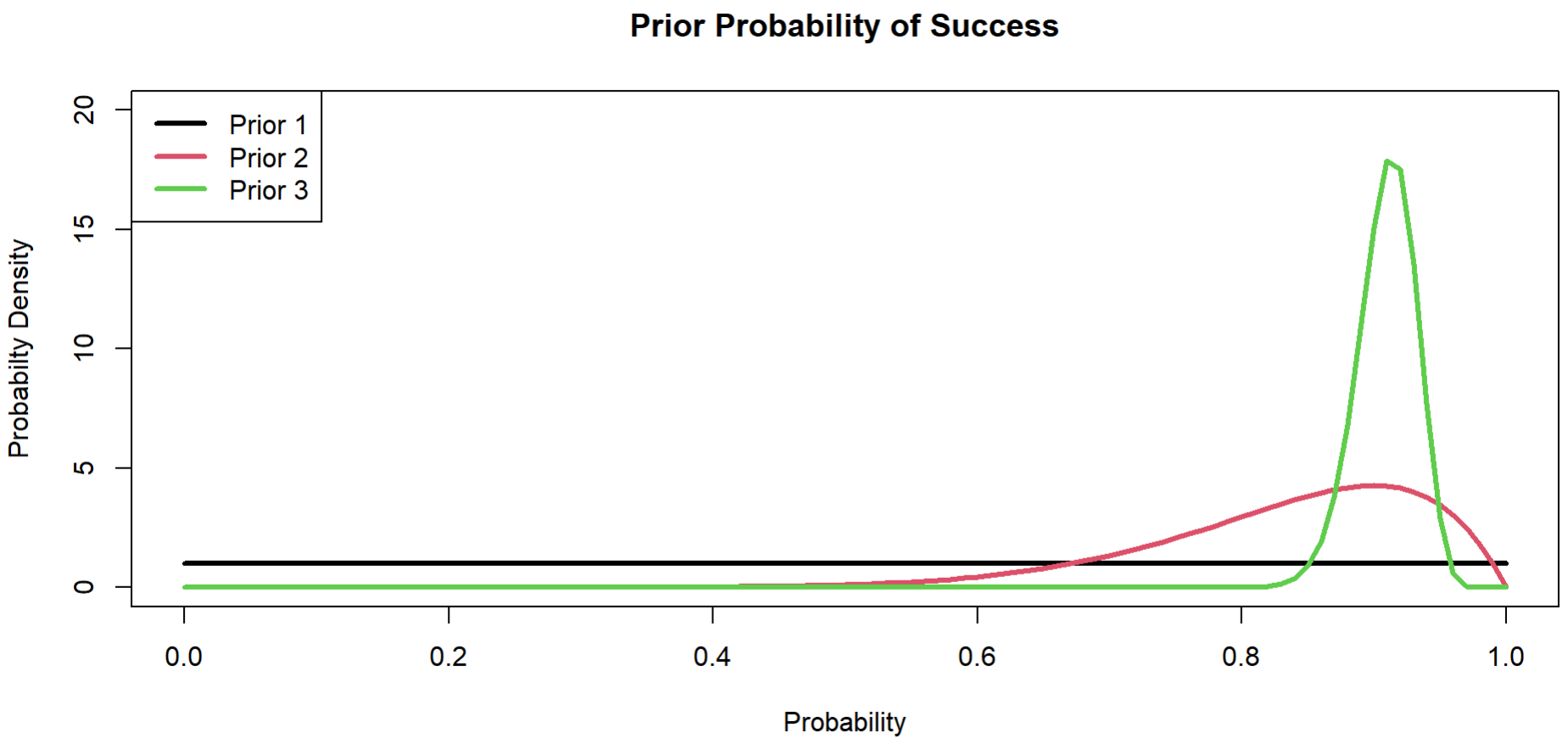
Hippos: Bayesian Model (Prior 2)

```
1    alpha.prior2=10
2    beta.prior2=2
```



Hippos: Bayesian Model (Prior 3)

```
1    alpha.prior3=150
2    beta.prior3=15
```



Hippos: Bayesian Model (Posteriors)

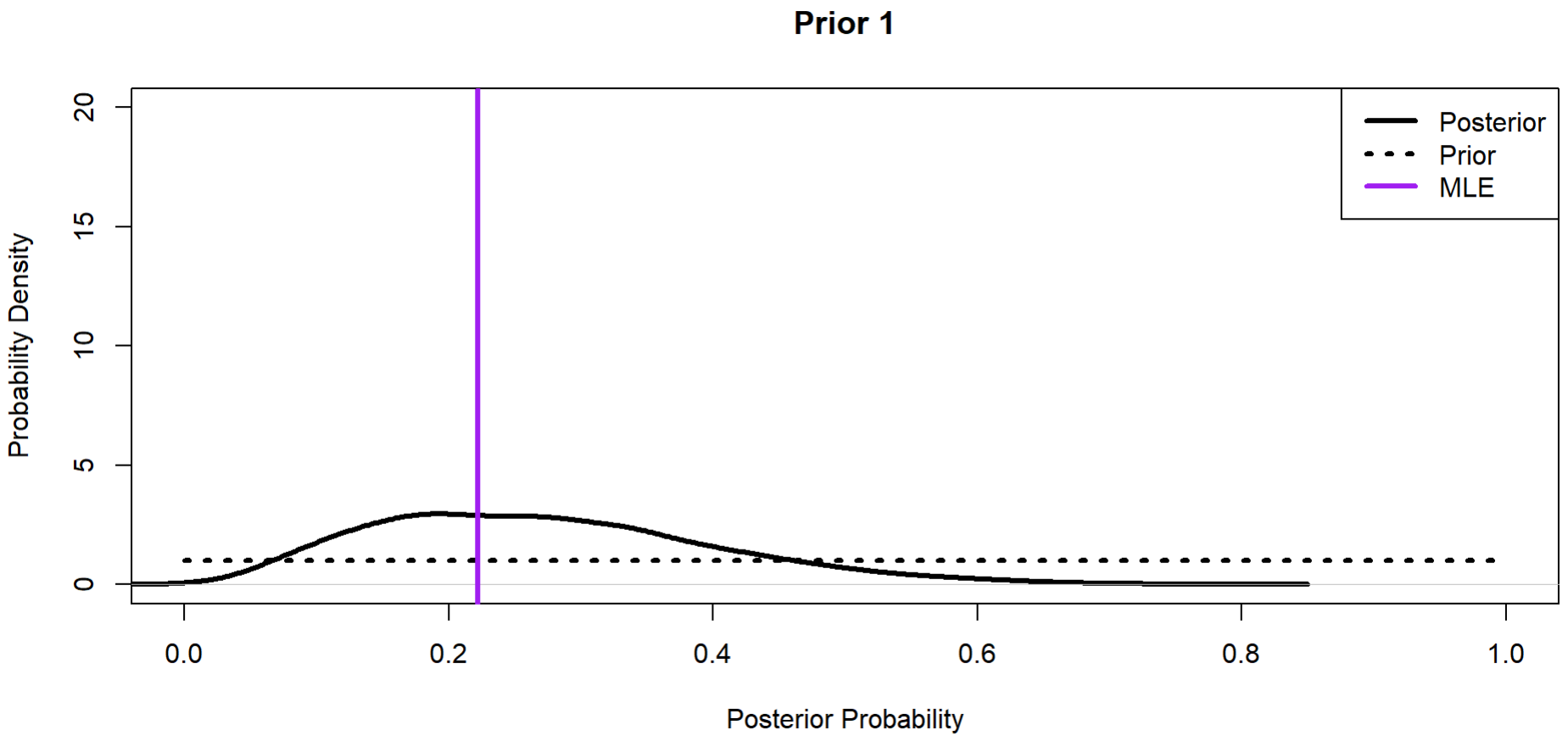
$$P(p|y) \sim \text{Beta}(\alpha^*, \beta^*)$$

$$\alpha^* = \alpha + \sum_{i=1}^{N} y_i$$

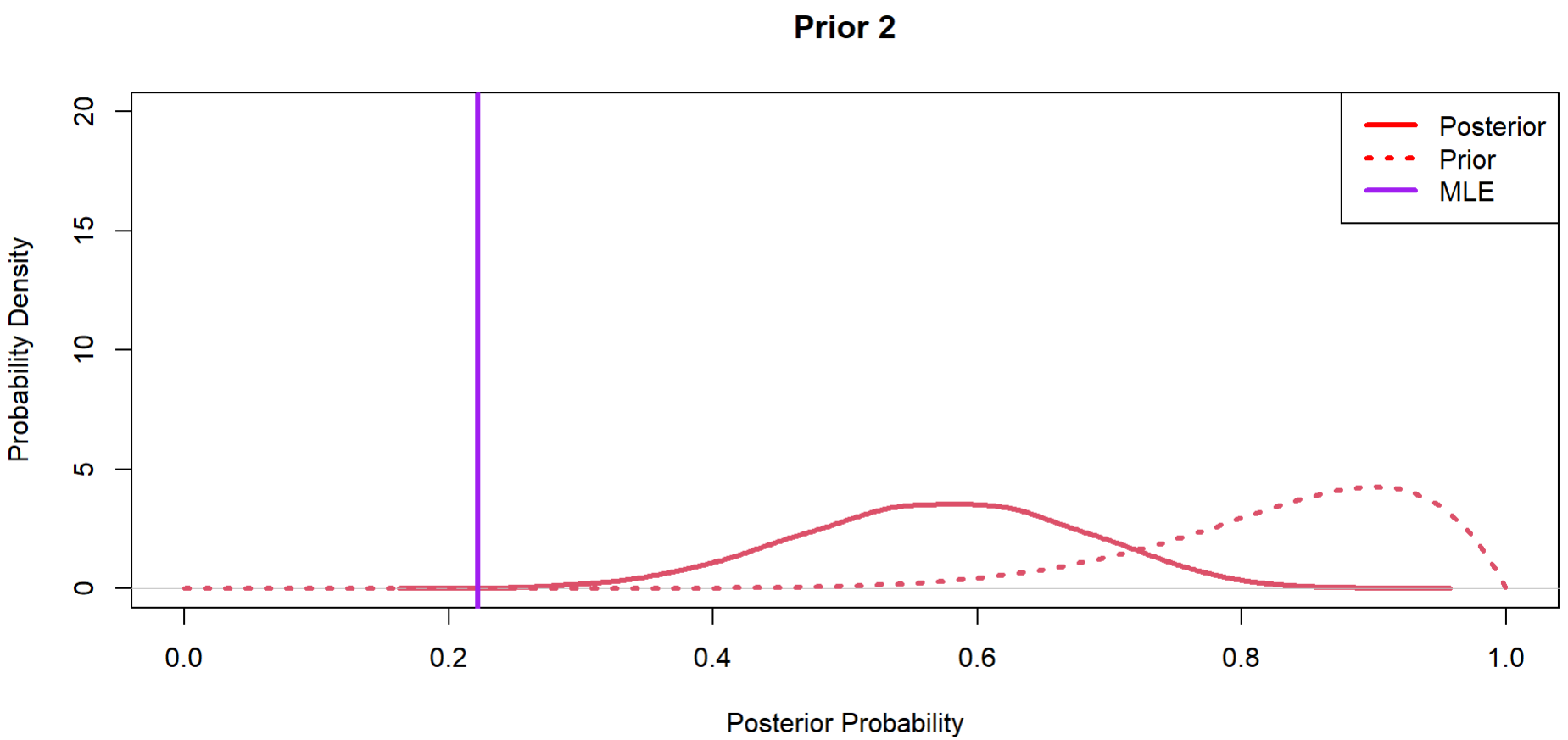
$$\beta^* = \beta + N - \sum_{i=1}^{N} y_i$$

```
# Note- the data are the same, but the prior is changing.
# Gibbs sampler
  post.1=rbeta(10000,alpha.prior1+sum(y1),beta.prior1+N1-sum(y1))
  post.2=rbeta(10000,alpha.prior2+sum(y1),beta.prior2+N1-sum(y1))
  post.3=rbeta(10000,alpha.prior3+sum(y1),beta.prior3+N1-sum(y1))
```

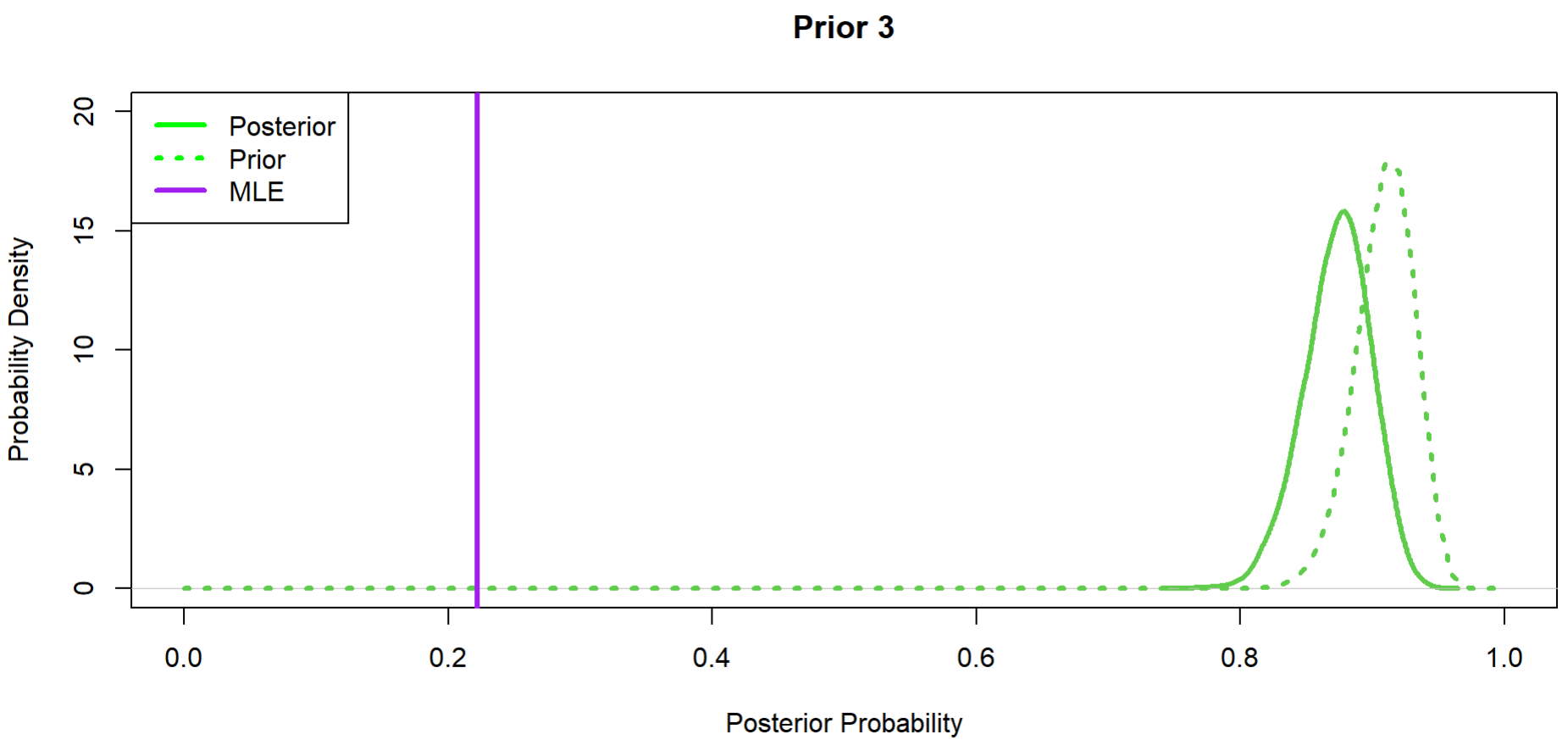
Hippos: Bayesian Model (Posteriors)



Hippos: Bayesian Model (Posteriors)



Hippos: Bayesian Model (Posteriors)

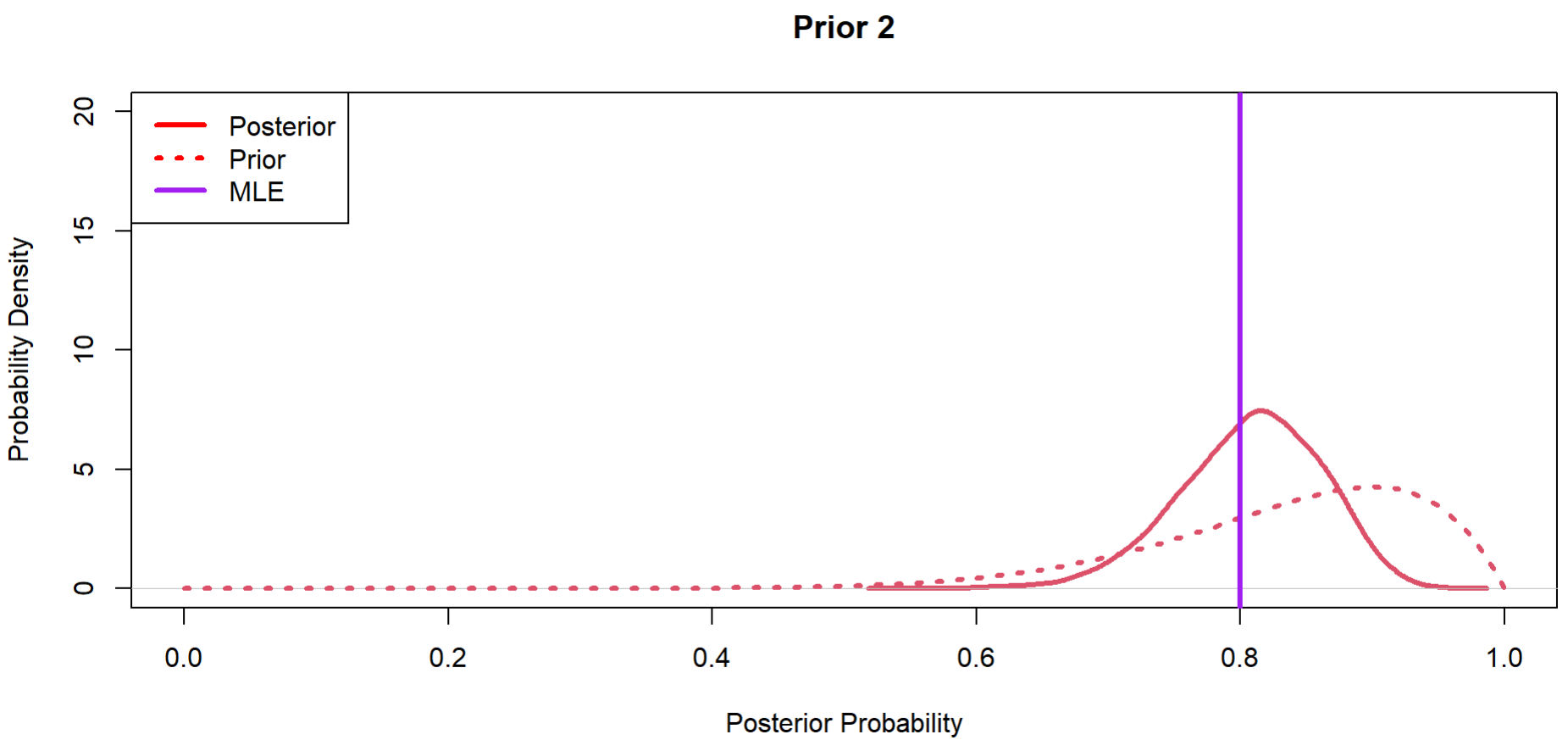


Hippos: More data! (Prior 1)

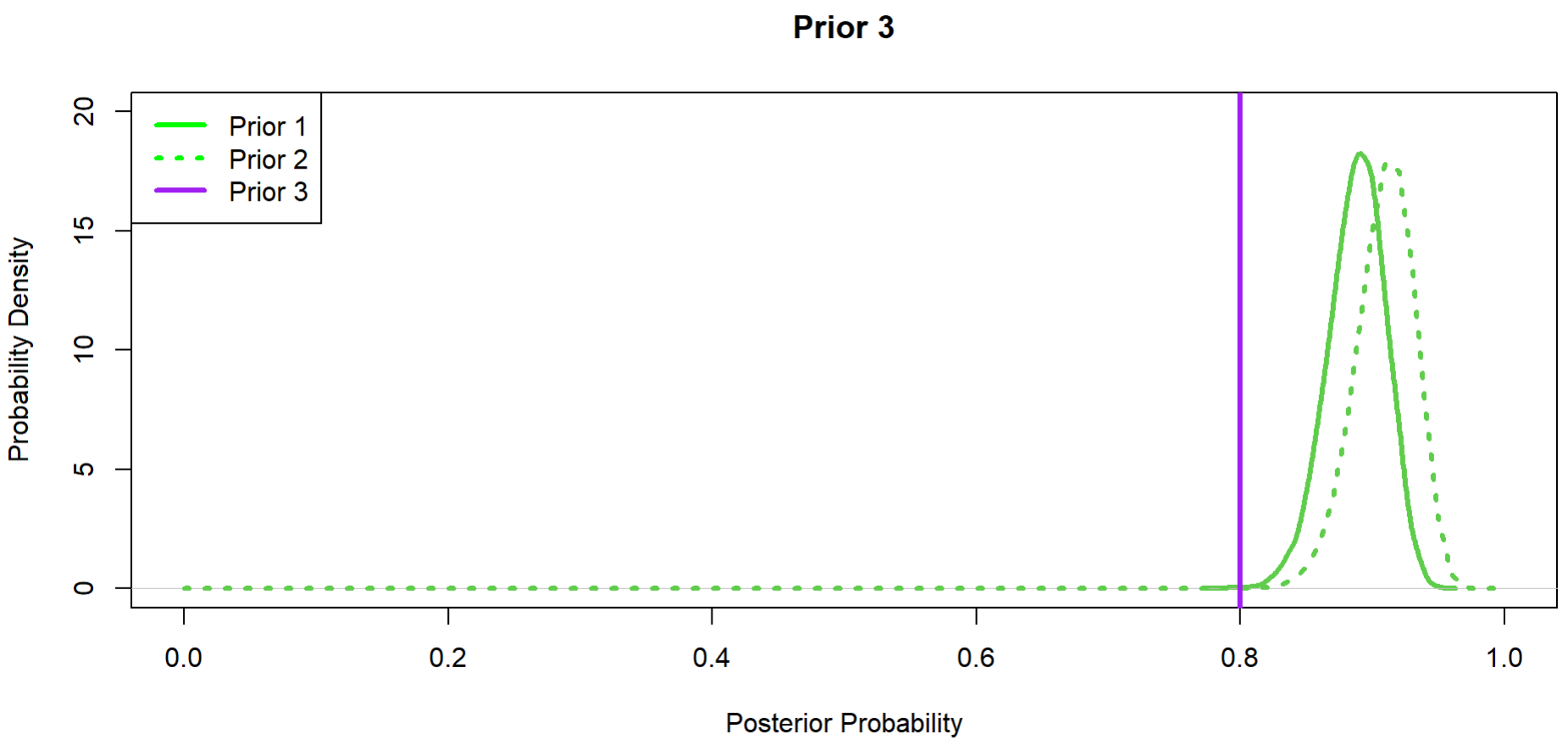
```r
y2=c(0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,0,
     1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
length(y2)
```

```
[1] 40
```

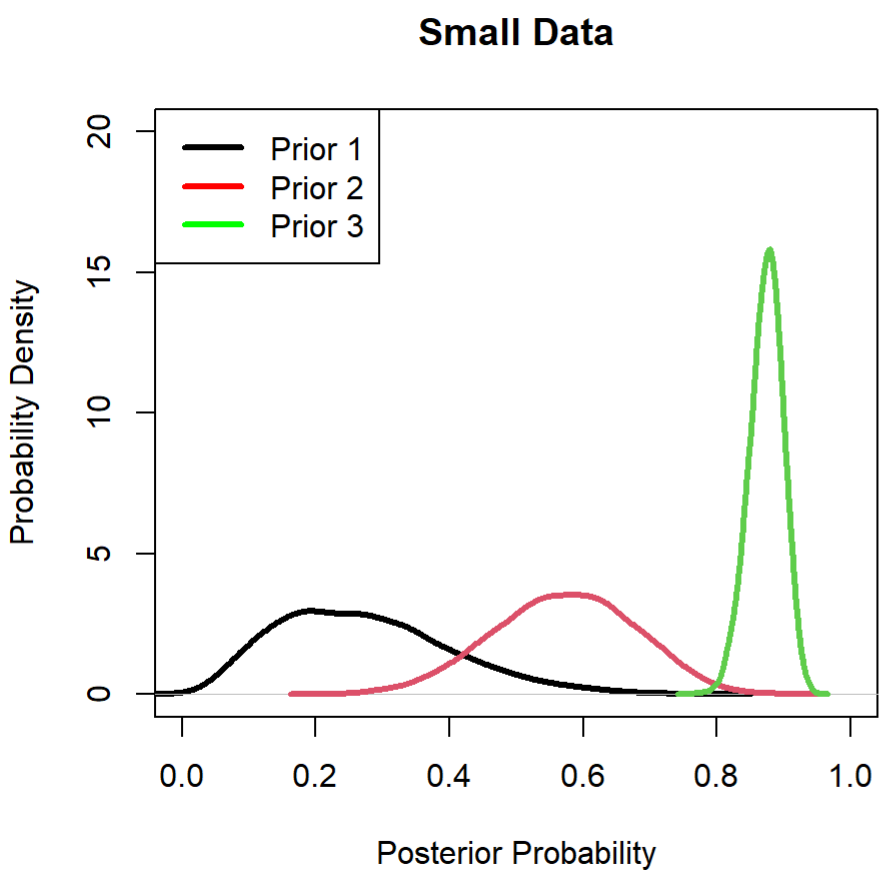
Hippos: More data! (Prior 2)

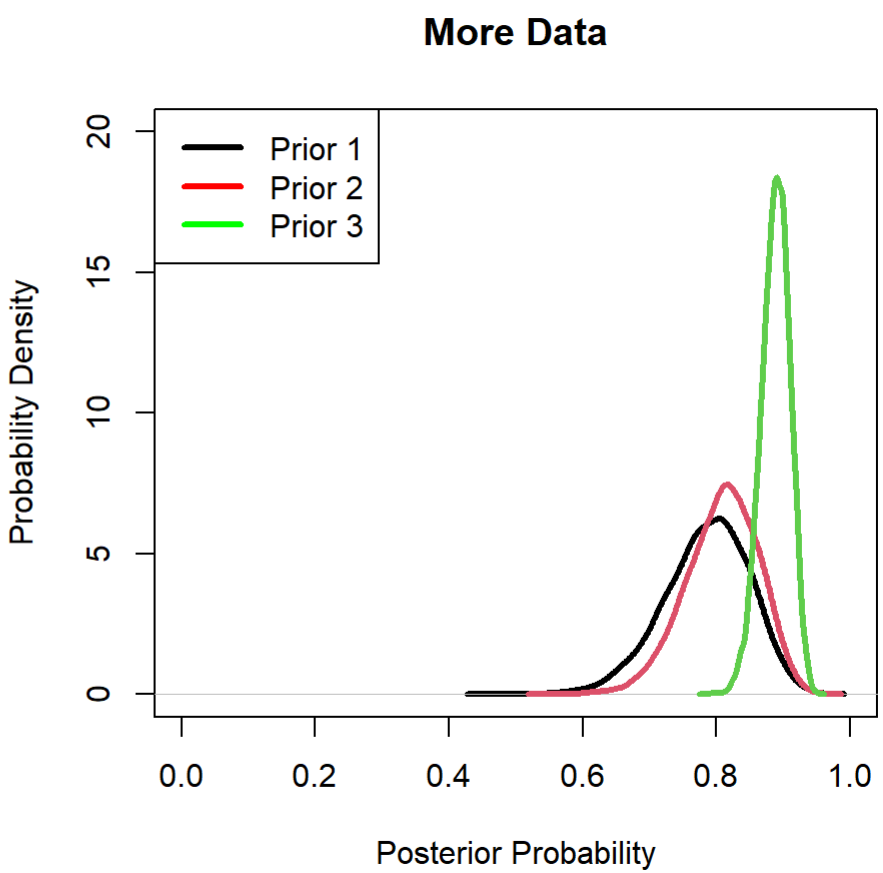


Hippos: More data! (Prior 3)



Hippos: Data/prior Comparison





Bayesian Computation

Markov Chain Monte Carlo

Often don’t have conjugate likelihood and priors, so we use MCMC algorithims to sample posteriors.

Class of algorithim

- Metropolis-Hastings
- Gibbs
- Reversible-Jump
- No U-turn Sampling

MCMC Sampling

- number of samples (iterations)
- thinning (which iterations to keep), every one (1), ever other one (2), every third one (3)
- burn-in (how many of the first samples to remove)
- chains (unique sets of samples; needed for convergence tests; requires different initial values)

Why did Bayesian statistics take off so late?

Many options now

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

- BUGS, WinBUGS, OpenBUGS, multiBUGS
- JAGS
- Stan
- Nimble
- also many others, e.g. greta…

Custom algrithims vs Software Engines

Should you learn how to write your own MCMC algorithms?

- Absolutely!
- Hell no!

What are the advantages/disadvantages?

Should you be a frequentist or a Bayesian?

Why we have become Bayesians…

Why we are not real Bayesians…

- Seldom use informative priors
- Plus, some inconveniences of Bayesian analysis with MCMC:
  - Take long time to run
  - Sensitivity of results to prior choice (not with ML)
  - BUGS so flexible that may fit nonsensical models

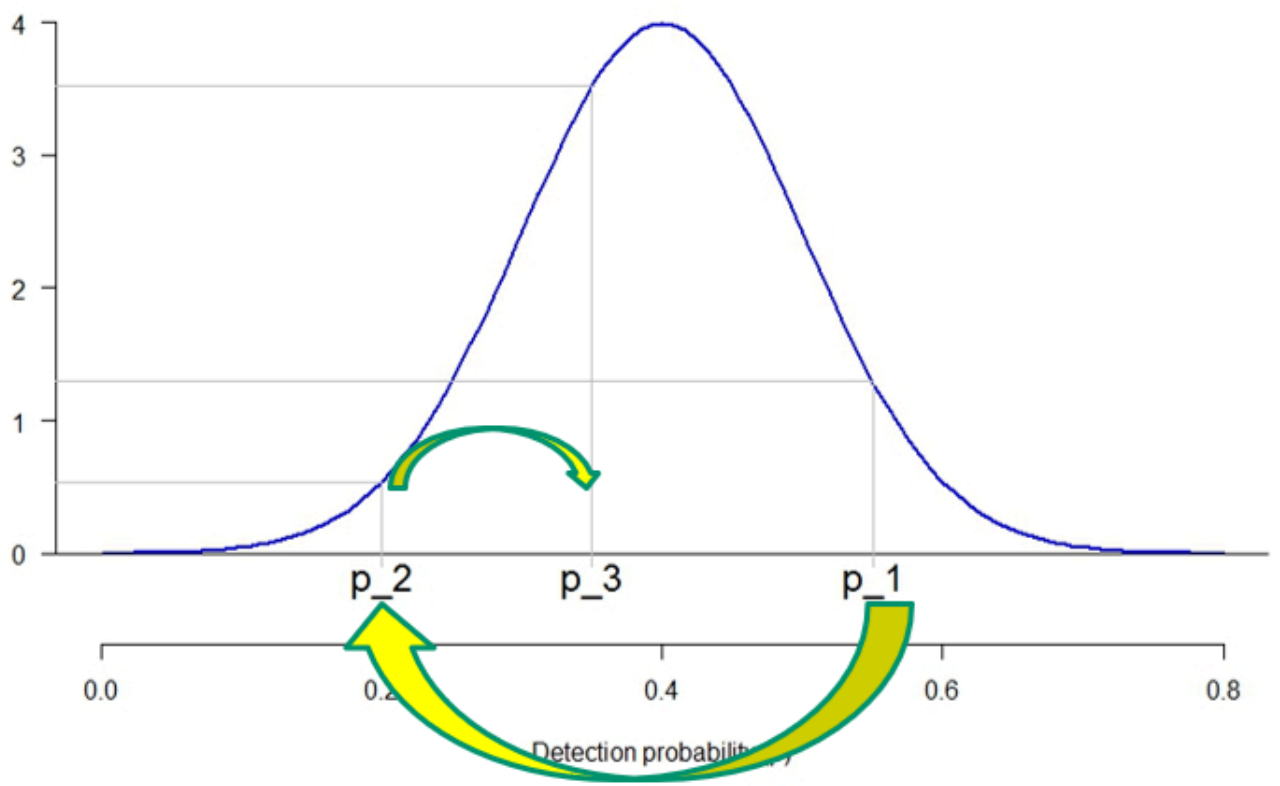
Very happy to use maximum likelihood as well

Conclusion on the Bayesian/frequentist choice

- Be eclectic!
- 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

Algorithm of Metropolis et al. (1953)

- sample p(θ | y)
- MCMC: *jump "upwards" along posterior with greater p*
- (example for a one-parameter binomial model)



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) !

THE END