

# Introduction to Bayesian Statistics

## *Part 2* Bayesian Principles

Benjamin Rosenbaum

iDiv 2025



# This lecture

Frequentist statistics and Null hypothesis significance testing

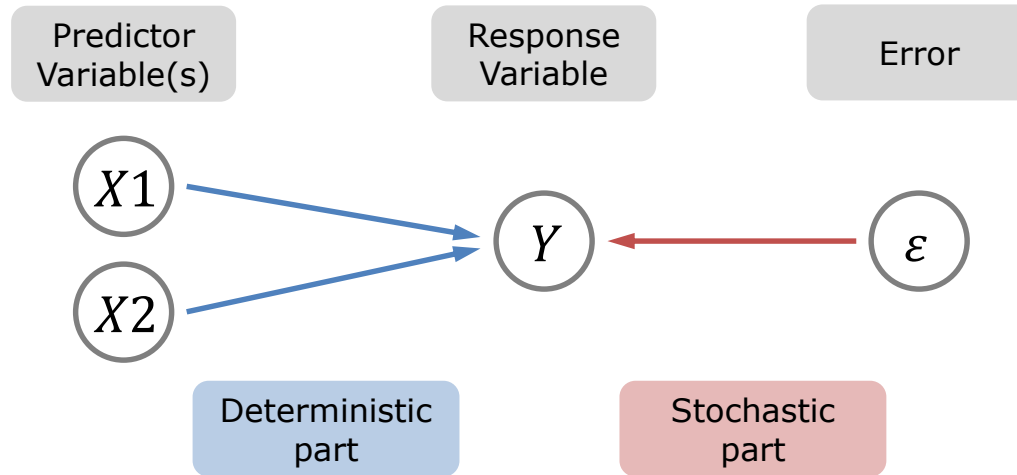
Bayes' rule

Markov Chain Monte Carlo sampling

MCMC software

Why Bayesian statistics?

# Statistical modeling



# Statistical modeling

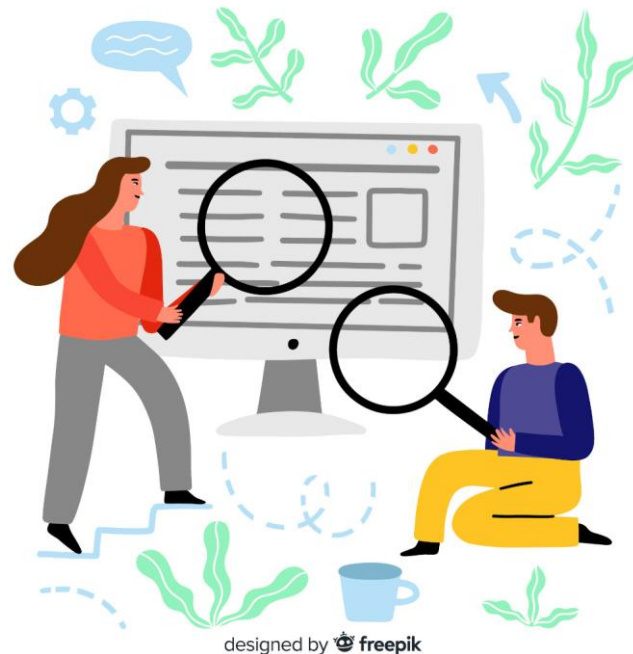
We are data detectives, trying to solve a mystery

In ecology, these mysteries can be extra tricky:

- Observational data instead of experiments
- Noisy data
- Many sources of variation

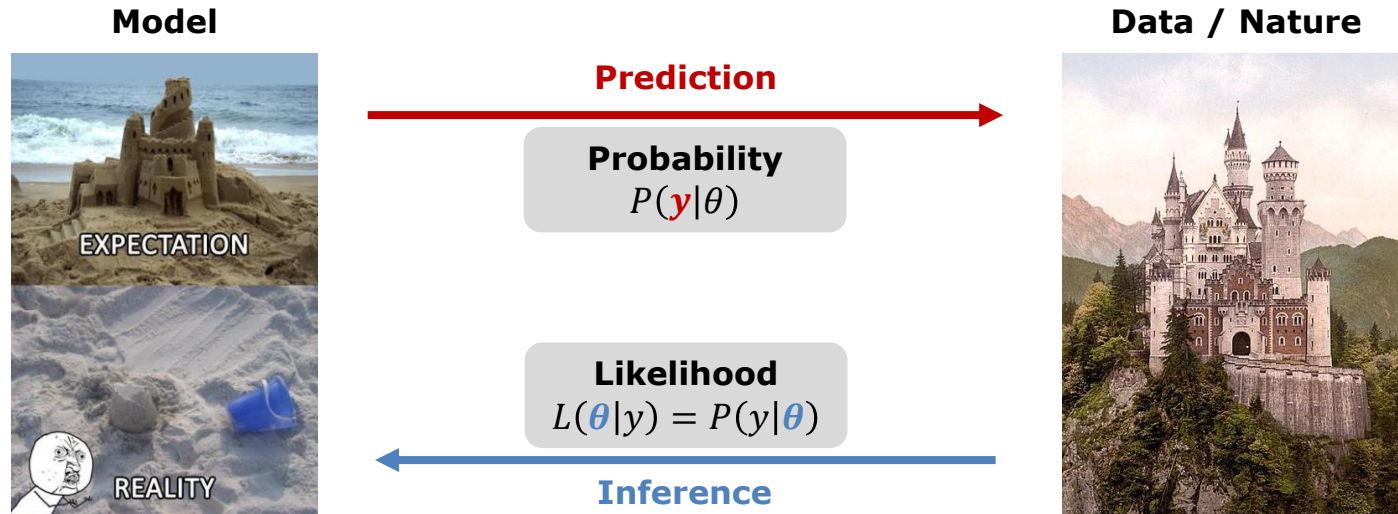
We're trying to unravel a signal from the noise  
(e.g. overall trend of biodiversity loss)

We want to make quantitative statements on research questions!



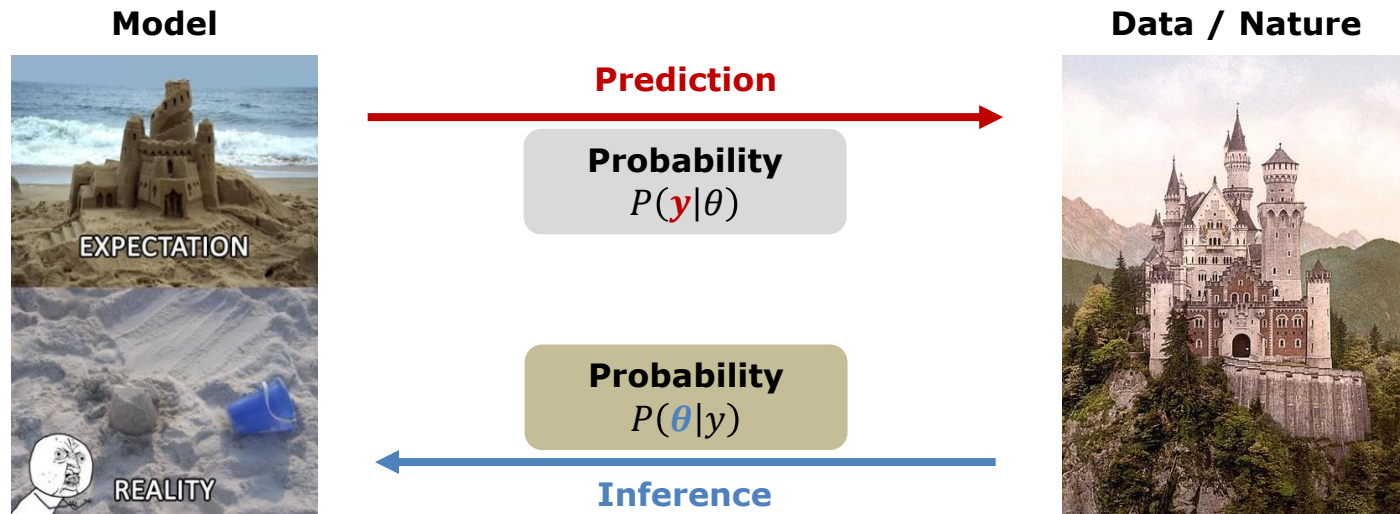
## *Some comments on frequentist stats*

# Statistical modeling



Maximum likelihood can't make probability statements about our research question!

# Statistical modeling

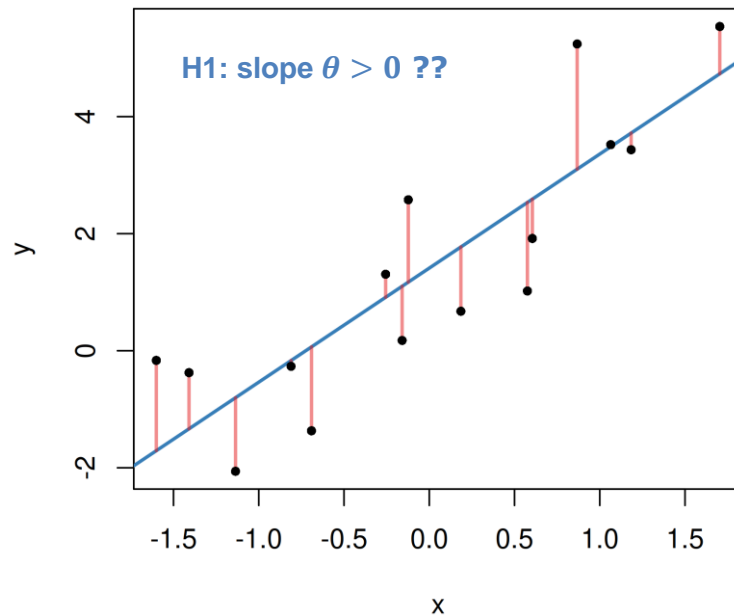


We want to assign probabilities to model parameters! E.g.,  $P(\theta > 0)$

# The frequentist „trick”: Null hypothesis significance testing

1. We want to test hypothesis **H1**  
(e.g., association is positive:  $\theta > 0$ )
2. We assume that the null hypothesis **H0** is true  
(e.g., association is zero:  $\theta = 0$ )
3. Use a transformation  $T$  for which data  $Y$  (residuals) have a well-known distribution  $P(T)$  under **H0**
4. If  $T(Y)$  deviates enough from the assumed distribution, reject the null hypothesis  
 $P(T > T(Y) \mid \theta = 0) < 0.05 \rightarrow \text{reject H0}$

Tests if the estimated association  $\theta^*$  (MLE)  
is just due to randomness of the data

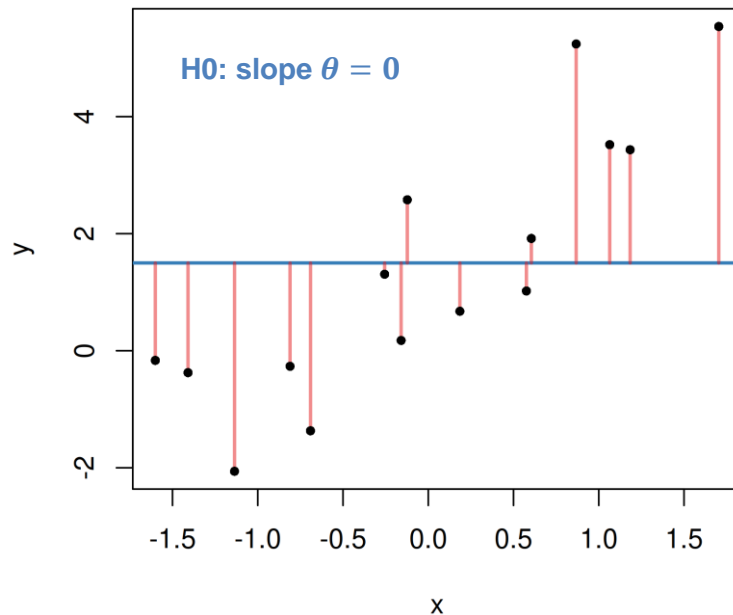




# The frequentist „trick”: Null hypothesis significance testing

1. We want to test hypothesis  $H_1$   
(e.g., association is positive:  $\theta > 0$ )
- 2. We assume that the null hypothesis  $H_0$  is true  
(e.g., association is zero:  $\theta = 0$ )**
3. Use a transformation  $T$  for which data  $Y$  (residuals) have a well-known distribution  $P(T)$  under  $H_0$
4. If  $T(Y)$  is improbable under assumed distribution, reject the null hypothesis  
 $P(T > T(Y) \mid \theta = 0) < 0.05 \rightarrow \text{reject } H_0$

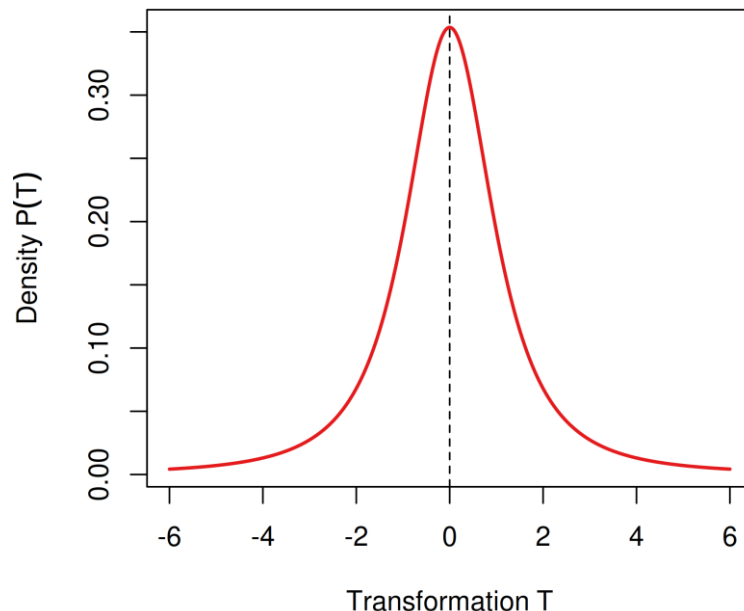
Tests if the estimated association  $\theta^*$  (MLE)  
is just due to randomness of the data



# The frequentist „trick”: Null hypothesis significance testing

1. We want to test hypothesis  $H_1$   
(e.g., association is positive:  $\theta > 0$ )
2. We assume that the null hypothesis  $H_0$  is true  
(e.g., association is zero:  $\theta = 0$ )
- 3. Use a transformation  $T$  for which data  $Y$  (residuals) have a well-known distribution  $P(T)$  under  $H_0$**
4. If  $T(Y)$  is improbable under assumed distribution, reject the null hypothesis  
 $P(T > T(Y) \mid \theta = 0) < 0.05 \rightarrow \text{reject } H_0$

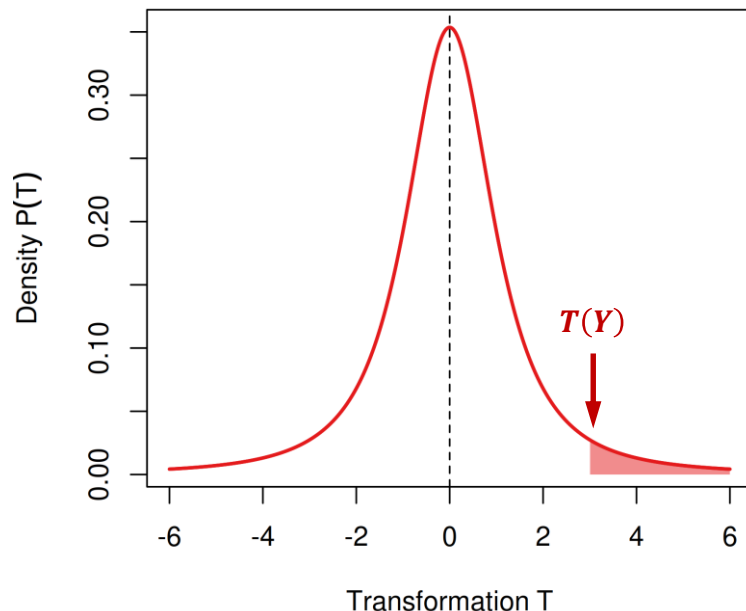
Tests if the estimated association  $\theta^*$  (MLE)  
is just due to randomness of the data



# The frequentist „trick”: Null hypothesis significance testing

1. We want to test hypothesis  $H_1$   
(e.g., association is positive:  $\theta > 0$ )
2. We assume that the null hypothesis  $H_0$  is true  
(e.g., association is zero:  $\theta = 0$ )
3. Use a transformation  $T$  for which data  $Y$  (residuals) have a well-known distribution  $P(T)$  under  $H_0$
4. **If  $T(Y)$  is improbable under assumed distribution, reject the null hypothesis**  
 **$P(T > T(Y) \mid \theta = 0) < 0.05 \rightarrow \text{reject } H_0$**

Tests if the estimated association  $\theta^*$  (MLE) is just due to randomness of the data

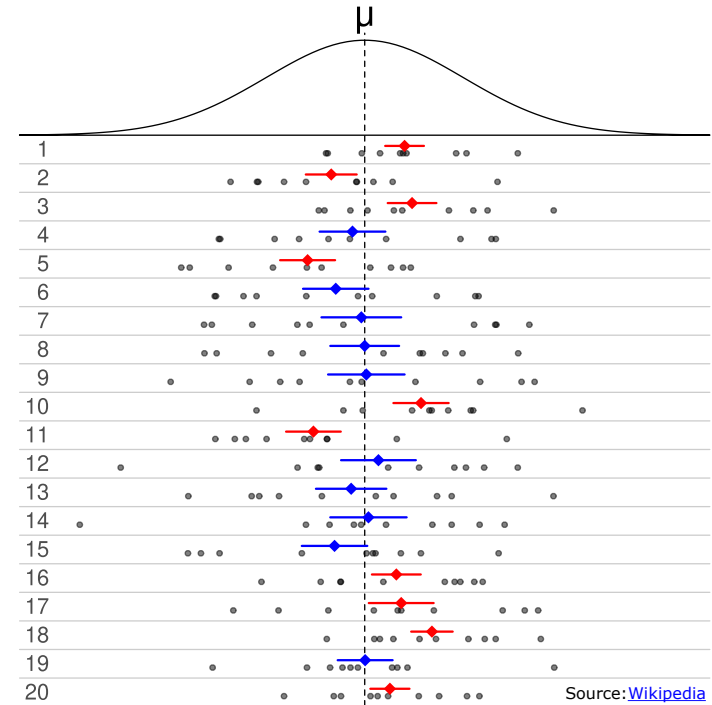


# Frequentist principles

→ Data are a random realization of an experiment.  
True (but unknown) parameter is fixed.

Some problems with NHST:

- P-value: Probability of the data under the null hypothesis
- Can't confirm hypotheses, just reject the null hypothesis
- Standard errors rely on assumptions & approximations
- Confidence intervals' interpretation tricky
- Limited to tests with known distributions (T-test, F-test, ...)



50%-confidence intervals in 20 repeated experiments:  
10 out of 20 contain the true value  $\mu$

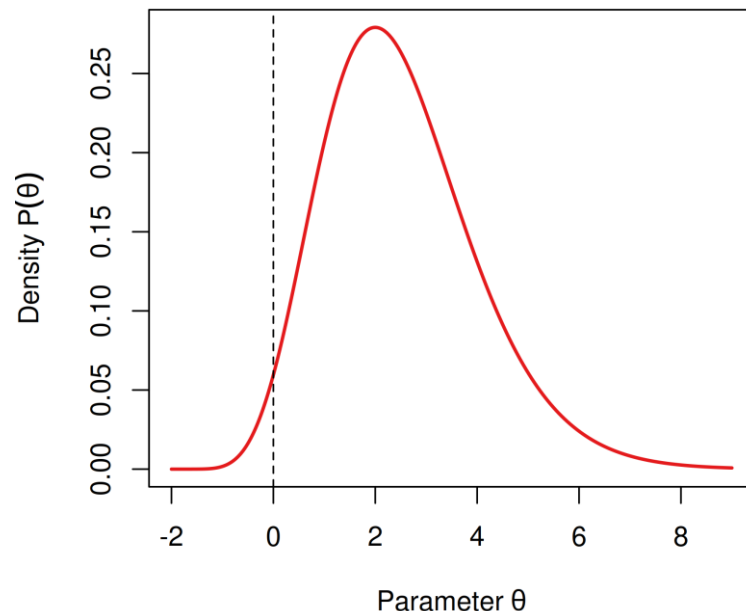
# Bayesian principles

To make quantitative statements about research questions, we need probability distribution for model parameters, after observing the data  $P(\theta|y)$

→ Data is fixed, parameters are random.

Some examples:

- $P(\theta > 0) = 0.99$ :  
"I am 99% sure the association is positive."
- 90%-quantile [0.5, 4.3]:  
"There is a 90% chance the slope is between 0.5 and 4.3."
- 2 population means  $\mu_1$  and  $\mu_2$ .  
 $P(\mu_1 - \mu_2)$  quantifies distribution of population-level difference.



# Bayesian models ???

There is no such thing as a „Bayesian model“!

*Maximum likelihood* → *Point estimates*

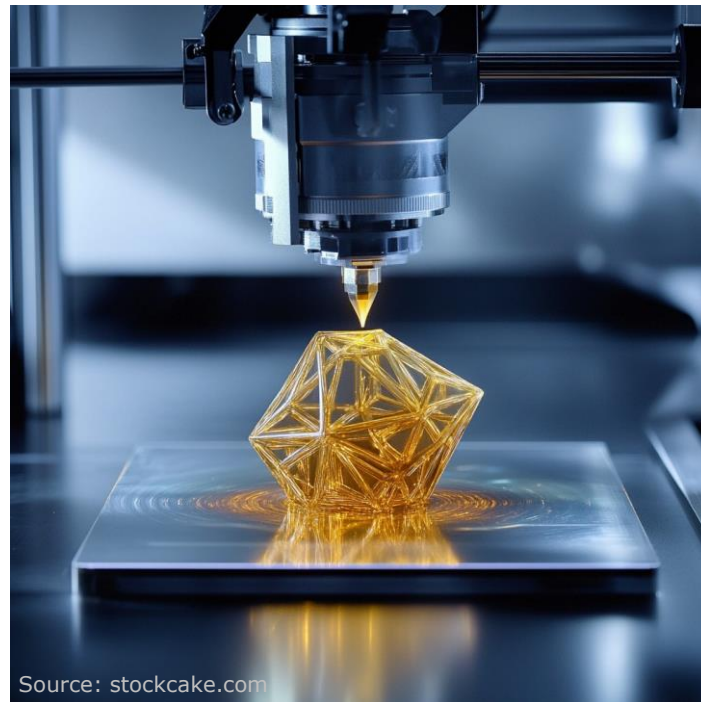
Frequentist NHST → P-values for Null hypothesis

Bayesian stats → True probability distribution  
for model parameters

“Full luxury Bayes” (Richard McElreath)

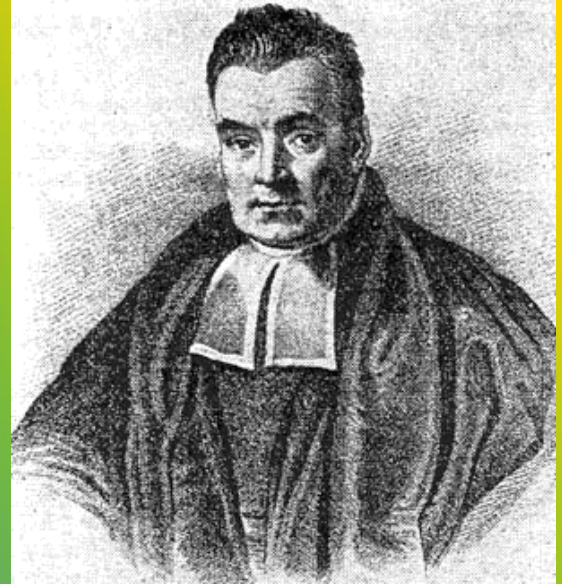
“Bayesian 3D printer” (me)

**The Bayesian 3D printer**



Source: stockcake.com

# *Bayes' rule*



Reverend Thomas Bayes (1701–1761)

# Bayes' rule

Conditional probability:

Prob. of event  $A$ , given that  $B$  occurred

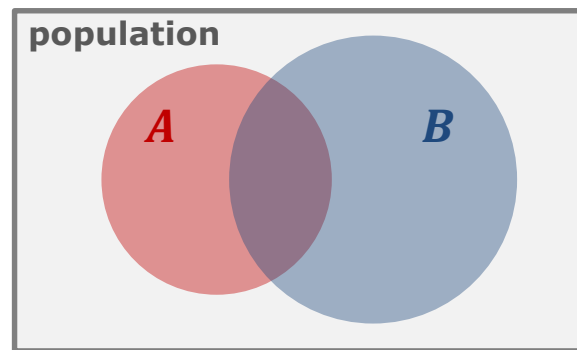
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(\text{Covid} \mid \text{positive}) = \frac{P(\text{positive}|\text{Covid}) \cdot P(\text{Covid})}{P(\text{positive})}$$

$P(\text{positive} \mid \text{Covid})$  test sensitivity

$P(\text{Covid})$  prevalence in the population

$P(\text{positive})$  positive test rate





# Bayes' rule

Probability distribution of model parameters  $\theta$   
after observing the data  $y$

The diagram illustrates Bayes' rule by showing the formula for the posterior distribution,  $p(\theta|y)$ , enclosed in a light beige rounded rectangle. Four arrows point to the components of the formula: 'Likelihood function  $L(\theta)$ ' points to  $p(y|\theta)$ , 'Prior distribution' points to  $p(\theta)$ , 'Posterior distribution' points to the entire formula, and 'Normalization constant Independent of  $\theta$ ' points to  $p(y)$  in the denominator.

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

Labels and arrows:

- Likelihood function  $L(\theta)$  (points to  $p(y|\theta)$ )
- Prior distribution (points to  $p(\theta)$ )
- Posterior distribution (points to the formula)
- Normalization constant Independent of  $\theta$  (points to  $p(y)$ )

# Bayes' rule

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

**$p(\theta|y)$  Posterior distribution**

Update prior information in light of new evidence (data)

**$p(\theta)$  Prior distribution**

Belief about model parameters before data is observed

**$p(y|\theta) = L(\theta)$  Likelihood function**

Data inform the parameters, but  $L$  is not a probability distribution for  $\theta$

**$p(y) = \int_{\theta} p(y|\theta) p(\theta)$  Normalization constant**

Ensures that posterior is a true probability distribution with  $\int_{\theta} p(\theta|y) = 1$

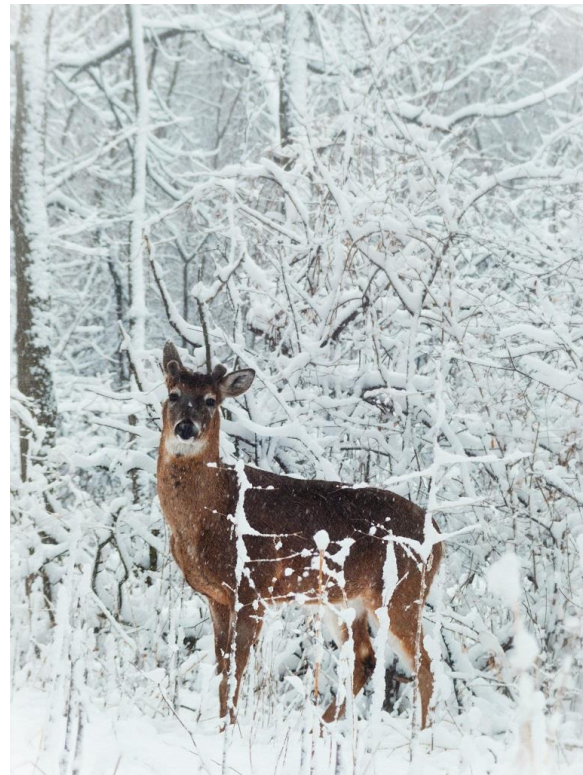
# Example

Survival rate of a deer population

1 datapoint: 7 out of 10 individuals survived

Deterministic part: average survival rate  $\theta$

Stochastic part:  $y \sim \text{Binomial}(N, \theta)$



# Example

Survival rate of a deer population

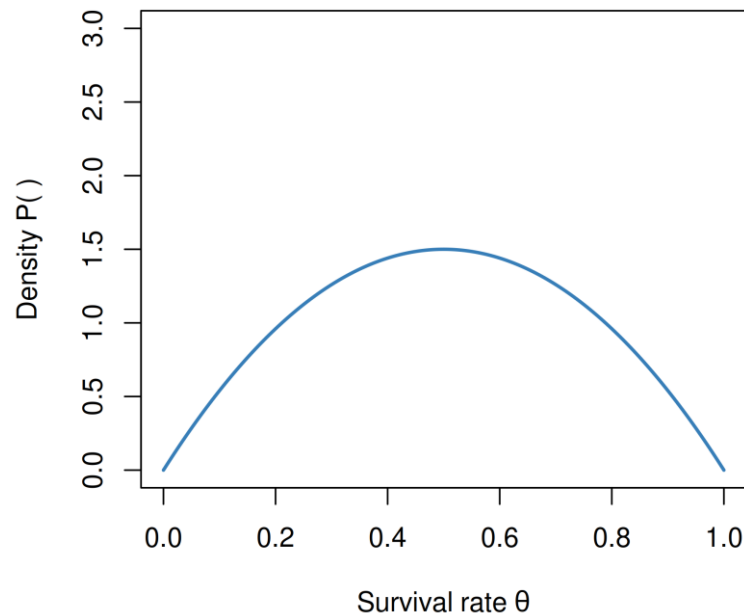
1 datapoint: 7 out of 10 individuals survived

Deterministic part: average survival rate  $\theta$

Stochastic part:  $y \sim \text{Binomial}(N, \theta)$

**Prior:**  $\theta$  almost 0 or almost 1 are improbable

$\theta \sim \text{beta}(2,2)$  or  $p(\theta) = \text{dbeta}(\theta | 2,2)$



# Example

Survival rate of a deer population

1 datapoint: 7 out of 10 individuals survived

Deterministic part: average survival rate  $\theta$

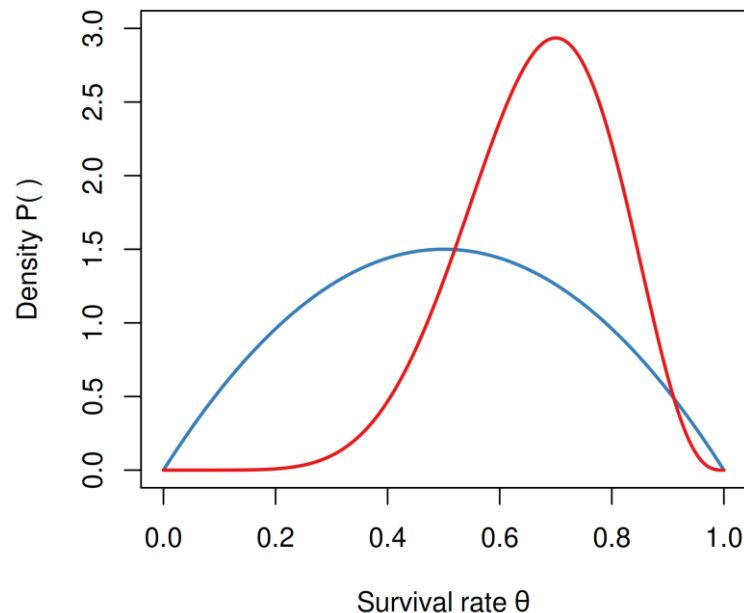
Stochastic part:  $y \sim \text{Binomial}(N, \theta)$

**Prior:**  $\theta$  almost 0 or almost 1 are improbable

$\theta \sim \text{beta}(2,2)$  or  $p(\theta) = \text{dbeta}(\theta | 2,2)$

**Likelihood:** defined by statistical model & data

$L(\theta) = p(y|\theta) = \text{dBinomial}(y = 7, N = 10 | \theta)$



# Example

Survival rate of a deer population

1 datapoint: 7 out of 10 individuals survived

Deterministic part: average survival rate  $\theta$

Stochastic part:  $y \sim \text{Binomial}(N, \theta)$

**Prior:**  $\theta$  almost 0 or almost 1 are improbable

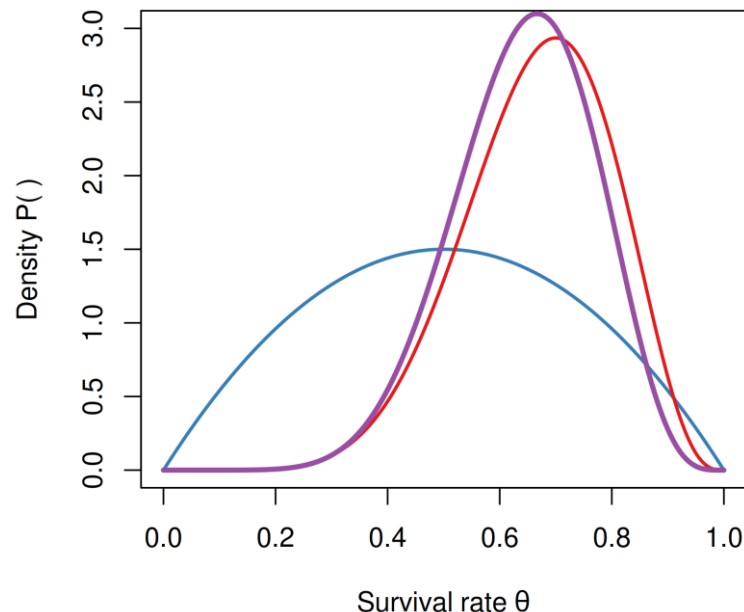
$\theta \sim \text{beta}(2,2)$  or  $p(\theta) = \text{dbeta}(\theta \mid 2,2)$

**Likelihood:** defined by statistical model & data

$L(\theta) = p(y|\theta) = \text{dBinomial}(y = 7, N = 10 \mid \theta)$

**Posterior:**

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



## Example: different prior

Survival rate of a deer population

1 datapoint: 7 out of 10 individuals survived

Deterministic part: average survival rate  $\theta$

Stochastic part:  $y \sim \text{Binomial}(N, \theta)$

**Prior:** uninformative

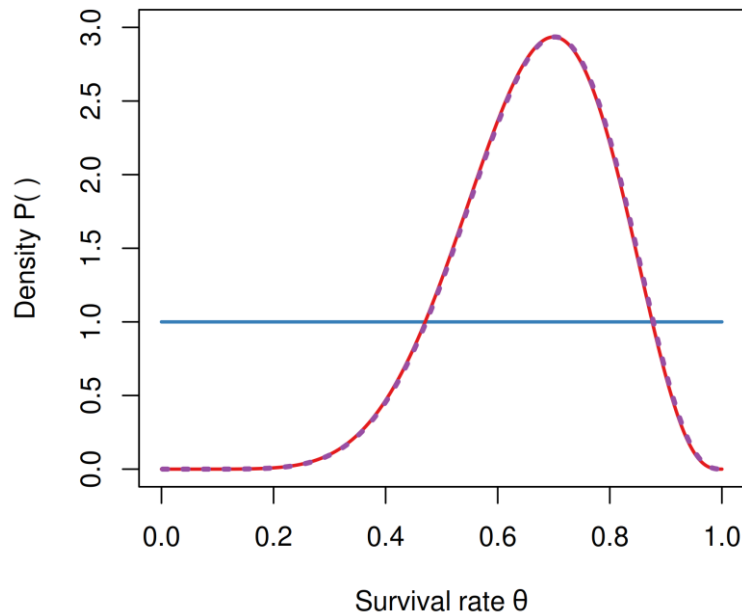
$\theta \sim \text{beta}(1,1) = \text{uniform}(1,1)$

**Likelihood:** defined by statistical model & data

$L(\theta) = p(y|\theta) = \text{dBinomial}(y = 7, N = 10 | \theta)$

**Posterior:**

$p(\theta|y) = \frac{L(\theta) \cdot p(\theta)}{c}$  proportional to likelihood here



# Calculation of the posterior?

To compute, e.g.  $\text{mean}(\theta)$ ,  $\text{sd}(\theta)$ ,  $P(\theta > 0.5)$  ...

we'd need to know  $p(\theta|y)$  for all  $\theta$  and also  $c = \int_{\theta} L(\theta)p(\theta)$

1) Analytical (mathematical formula)

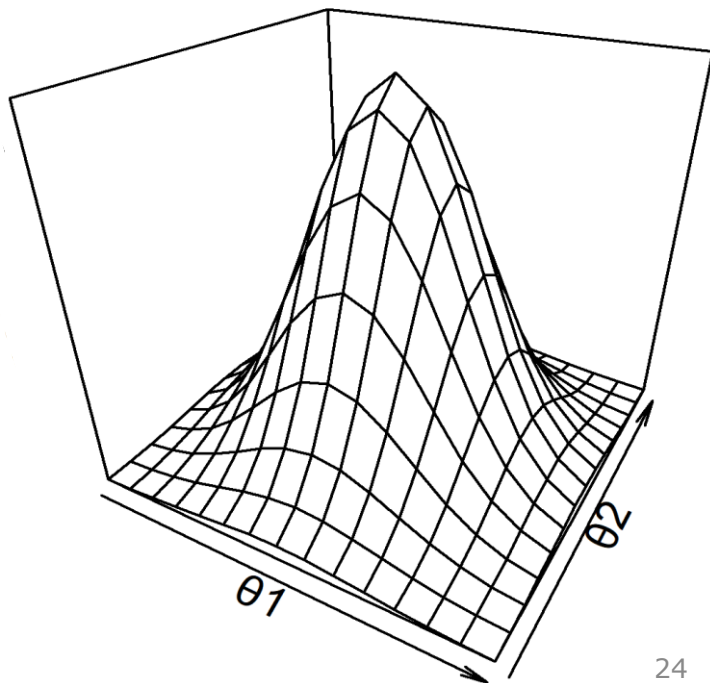
→ Much too complicated, often impossible

2) Numerical (e.g., grid)

→ Effort grows exponentially with  $\#\theta$

→ Computationally too expensive

„Curse of  
dimen-  
sionality“



**Oh no!** Same problem as before.



# *Markov Chain Monte Carlo (MCMC) sampling*

## New idea: sampling!

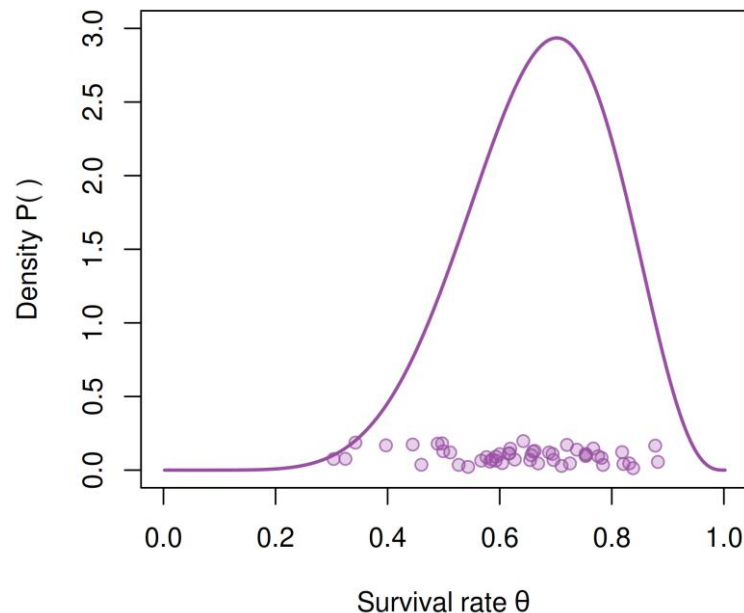
$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

Instead of calculating  $p(\theta|y)$ , draw random  $\theta$  samples.  
Many samples where  $p$  high, few samples where  $p$  low

- Sample density proportional to  $p(\theta|y)$
- We don't need the normalizing constant  $c = p(y)$

$$p(\theta|y) \sim p(y|\theta) \cdot p(\theta)$$

Posterior is proportional to likelihood x prior



# Markov Chain Monte Carlo

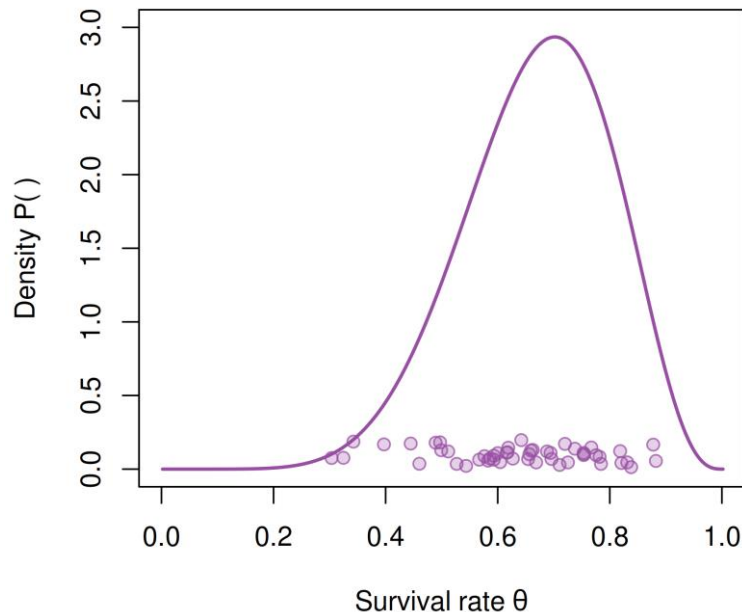
Start with initial  $\theta_1$

Compute  $f(\theta_1) = L(\theta_1) \cdot p(\theta_1)$

In each step  $i = 2, 3, 4, \dots$  :

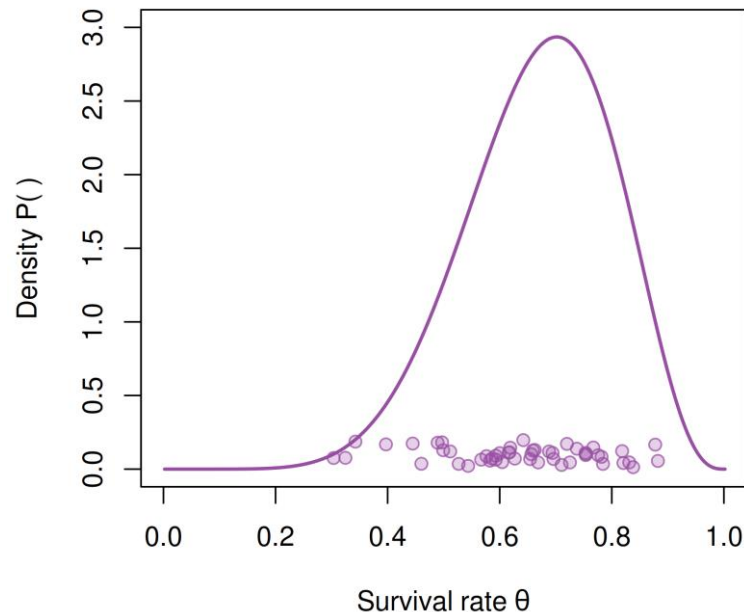
- Propose new  $\theta_{\text{new}}$ , e.g.  $\theta_{\text{new}} \sim \text{Normal}(\theta_{\text{old}}, \sigma)$   
Compute  $f(\theta_{\text{new}}) = L(\theta_{\text{new}}) \cdot p(\theta_{\text{new}})$
- If  $f(\theta_{\text{new}}) > f(\theta_{\text{old}})$   
→ accept  $\theta_{i+1} = \theta_{\text{new}}$
- If  $f(\theta_{\text{new}}) < f(\theta_{\text{old}})$   
→ accept  $\theta_{i+1} = \theta_{\text{new}}$  with probability  $\frac{f(\theta_{\text{new}})}{f(\theta_{\text{old}})}$   
(random draw)  
→ otherwise reject  $\theta_{\text{new}}$

Repeat e.g. 1000 times



# Markov Chain Monte Carlo

- $\theta_1, \theta_2, \theta_3, \dots, \theta_{1000}$  are called the „**chain**“
- They are samples from the underlying (but mathematically unknown) posterior distribution
- We can calculate empirical quantities
  - mean
  - standard deviation
  - quantiles
  - histogram (for visualization)
  - probability statements
- Without explicitly knowing the function  $p(\theta|y)$
- Even don't need to save computed values  $p(\theta_1|y), p(\theta_2|y), \dots$



# Markov Chain Monte Carlo

„**Markov**“ property:

each sample  $\theta_i$  **only** depends on the previous sample  $\theta_{i-1}$

„**Chain**“:

list of samples  $\theta_1, \theta_2, \theta_3, \dots, \theta_{1000}$

„**Monte Carlo**“:

each new sample involves a random draw

Very simple algorithm at its core (few lines of code)

Very sophisticated software to make it efficient  
(lots of maths go into good sample proposals)



# Example

Survival rate: 1 datapoint (7/10), prior  $\text{beta}(2,2)$

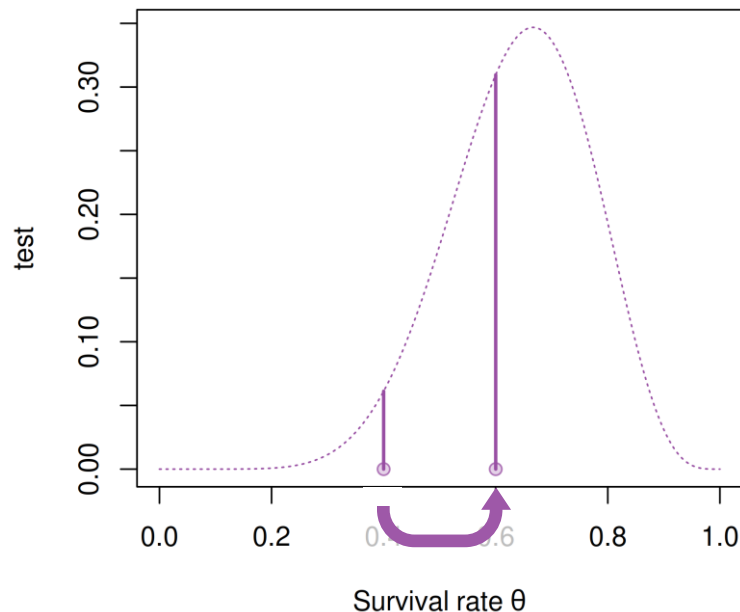
First sample  $\theta_1 = 0.4$

$$\begin{aligned} f(0.4) &= \text{dbinom}(7,10,0.4) \cdot \text{dbeta}(0.4,2,2) \\ &= 0.061 \end{aligned}$$

Propose  $\theta_{\text{new}} = 0.6$

$$\begin{aligned} f(0.6) &= \text{dbinom}(7,10,0.6) \cdot \text{dbeta}(0.6,2,2) \\ &= 0.310 \end{aligned}$$

$f(\theta_{\text{new}}) > f(\theta_{\text{old}}) \rightarrow$  accept  $\theta_2 = 0.6$  as next sample



# Example

Survival rate: 1 datapoint (7/10), prior beta(2,2)

Current sample  $\theta_2 = 0.6$

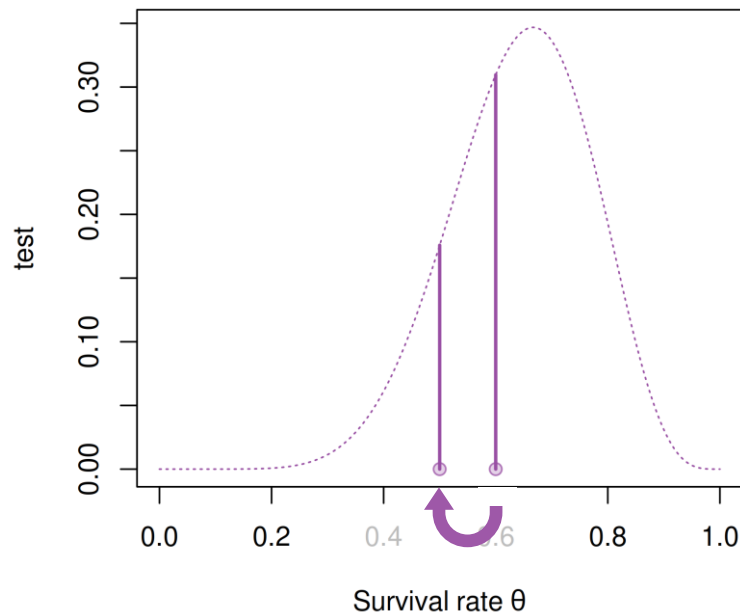
$$f(0.6) = 0.310$$

Propose  $\theta_{\text{new}} = 0.5$

$$\begin{aligned} f(0.5) &= \text{dbinom}(7,10,0.5) \cdot \text{dbeta}(0.5,2,2) \\ &= 0.175 \end{aligned}$$

$$f(\theta_{\text{new}}) < f(\theta_{\text{old}}) \rightarrow \text{accept with probability } \frac{f(\theta_{\text{new}})}{f(\theta_{\text{old}})} = \frac{0.175}{0.310} = 0.564$$

→ random draw with a 56.4% chance  
to accept  $\theta_3 = 0.5$  as next sample



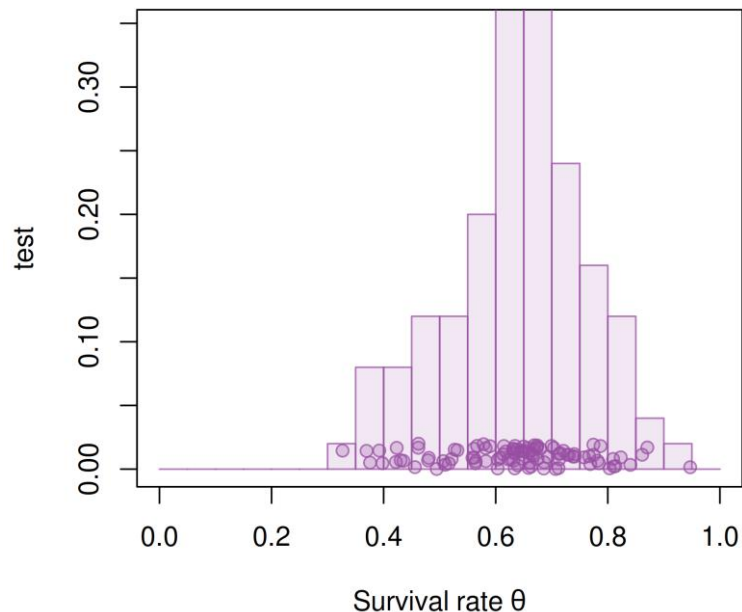
# Example

Survival rate: 1 datapoint (7/10), prior  $\text{beta}(2,2)$

Samples from posterior distribution  $\theta_1, \theta_2, \theta_3, \dots$   
describe probability for estimated survival rate.

Don't need to know the full curve  $p(\theta|y)$  !

- Empirical **histogram** for visualization





# Example

Survival rate: 1 datapoint (7/10), prior  $\text{beta}(2,2)$

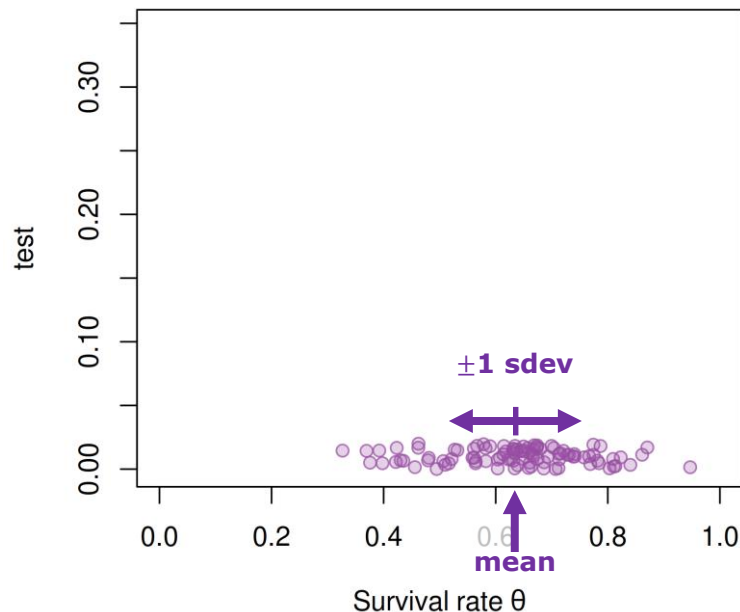
Samples from posterior distribution  $\theta_1, \theta_2, \theta_3, \dots$   
describe probability for estimated survival rate.

Don't need to know the full curve  $p(\theta|y)$  !

- Empirical **mean** and **standard deviation**

$$\text{mean} = \frac{1}{K} \sum_{i=1}^K \theta_i$$

$$\text{sdev} = \sqrt{\frac{1}{K} \sum_{i=1}^K (\theta_i - \text{mean})^2}$$



# Example

Survival rate: 1 datapoint (7/10), prior  $\text{beta}(2,2)$

Samples from posterior distribution  $\theta_1, \theta_2, \theta_3, \dots$   
describe probability for estimated survival rate.

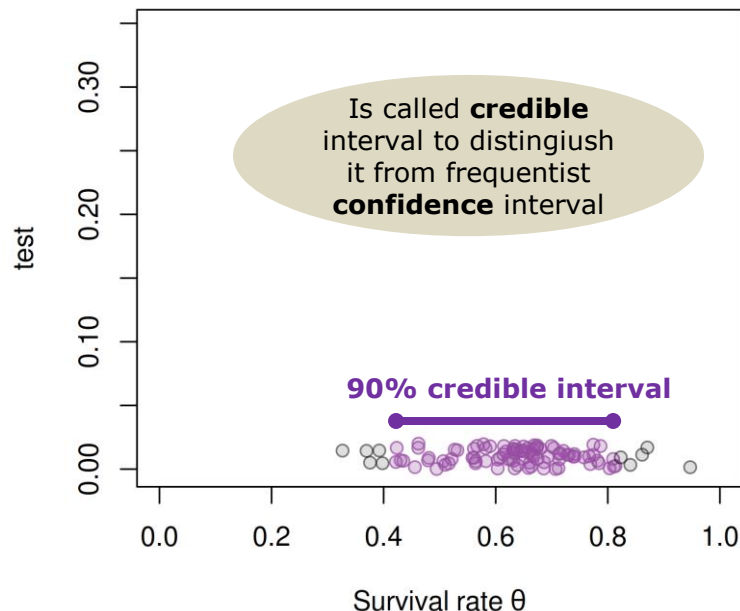
Don't need to know the full curve  $p(\theta|y)$  !

- „Credible intervals“

90% of samples between the 5% and the 95% quantiles.

$$P(\theta \in [0.42, 0.81]) = 0.9$$

„I am 90% sure the survival probability is between 0.42 and 0.81“



# Example

Survival rate: 1 datapoint (7/10), prior  $\text{beta}(2,2)$

Samples from posterior distribution  $\theta_1, \theta_2, \theta_3, \dots$   
describe probability for estimated survival rate.

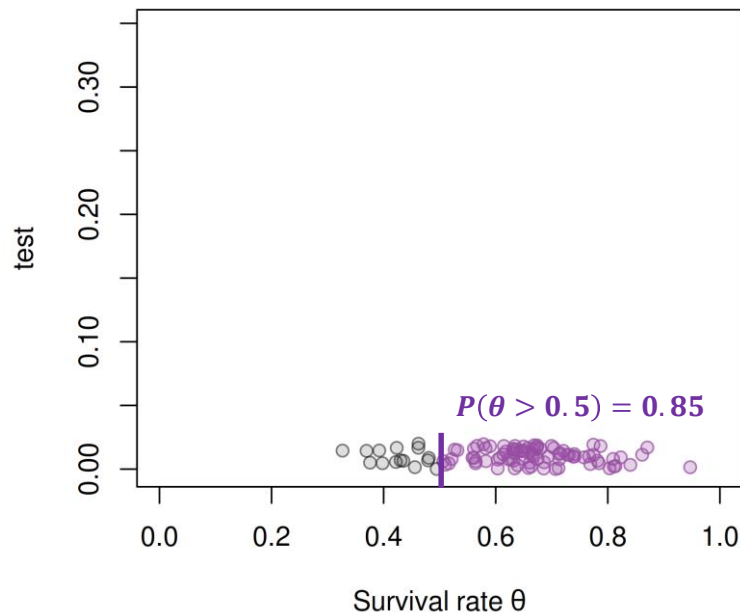
Don't need to know the full curve  $p(\theta|y)$  !

- **Probability statements** (about hypotheses)

85% of samples larger than a survival rate of 0.5

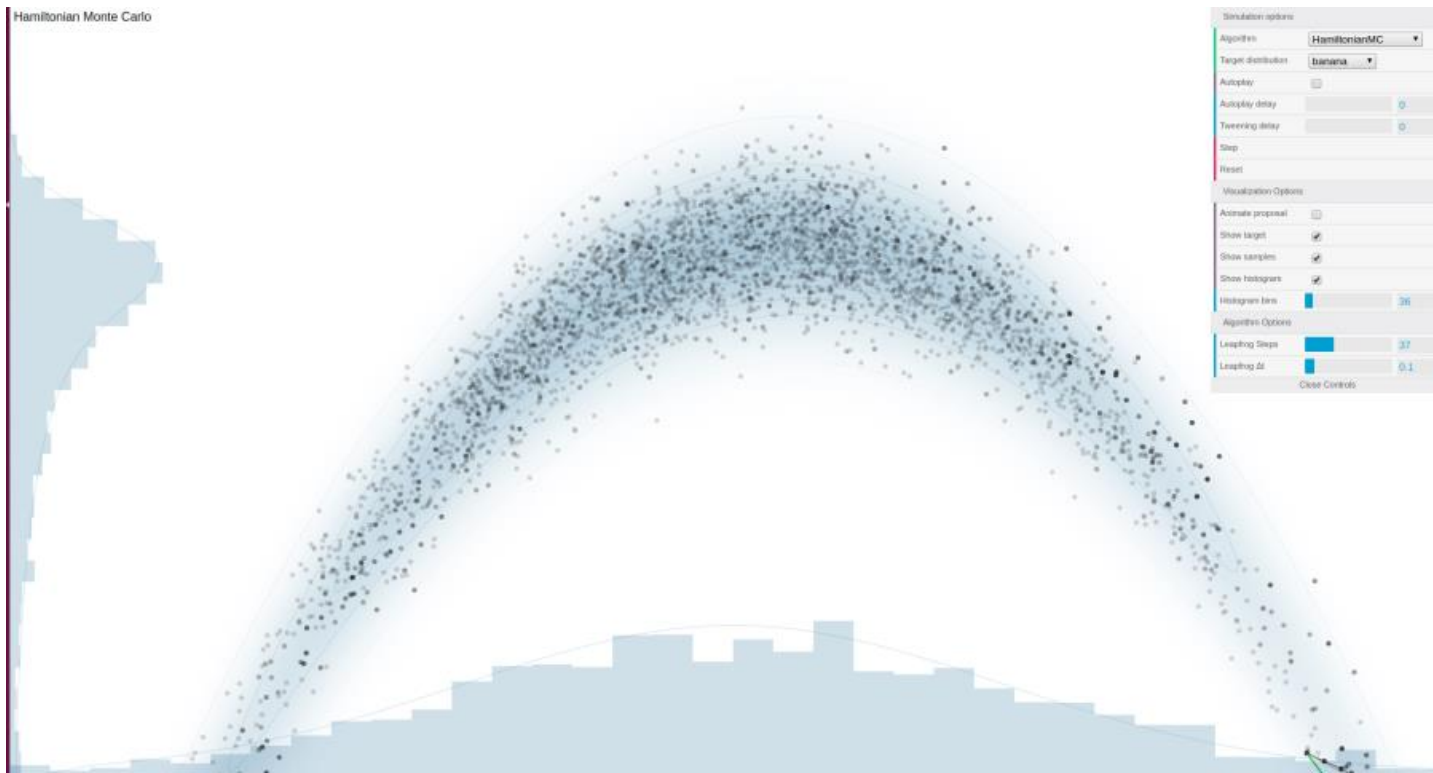
$$P(\theta > 0.5) = 0.85$$

„I am 85% sure the survival probability is larger than 0.5“



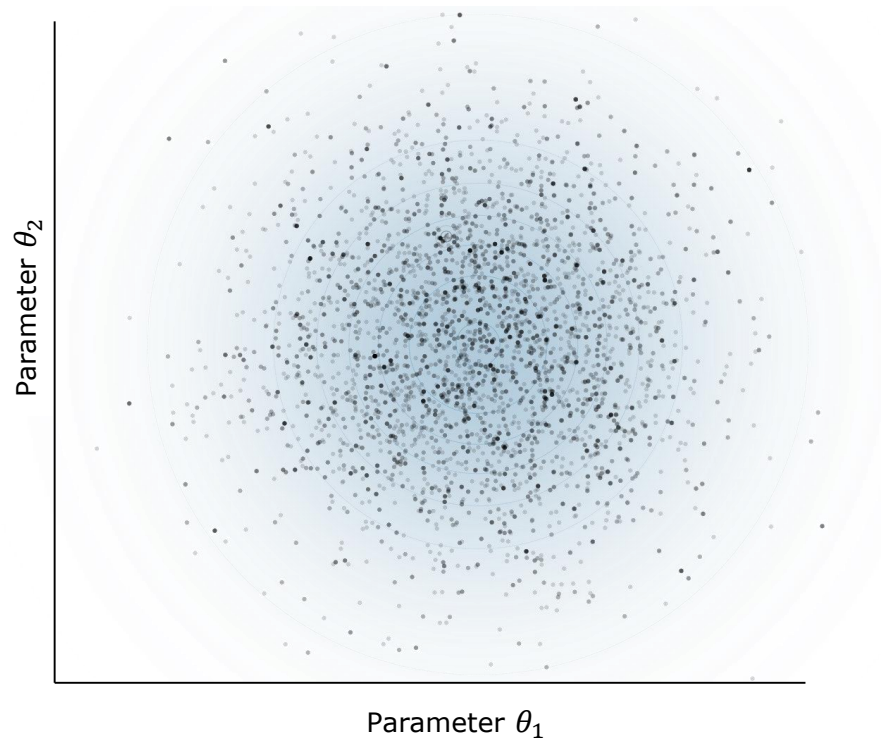
# MCMC Demo

<https://chi-feng.github.io/mcmc-demo/app.html>



# Convergence

- Mathematical theory says that MCMC will *eventually* be a good approximation of the posterior distribution
- How many samples are enough?
- Start with 1000-2000 samples
- Run multiple chains (3-4)
- Visual inspection
- Quantitative measures

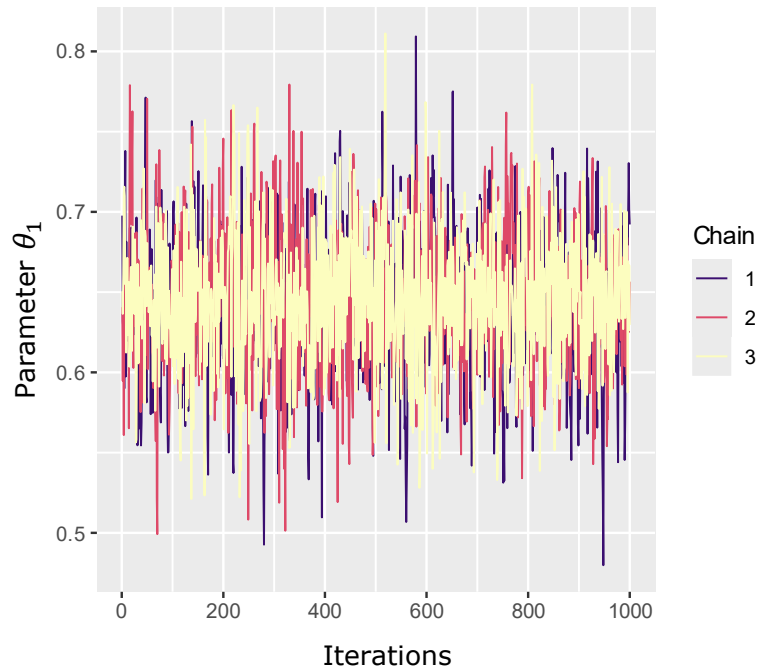


# Convergence

## Visual inspection

- Traceplots for each parameter
- Should look like random noise
- Centered around a constant mean
- Chains should look similar
- Like a fuzzy caterpillar!

→ MCMC has converged

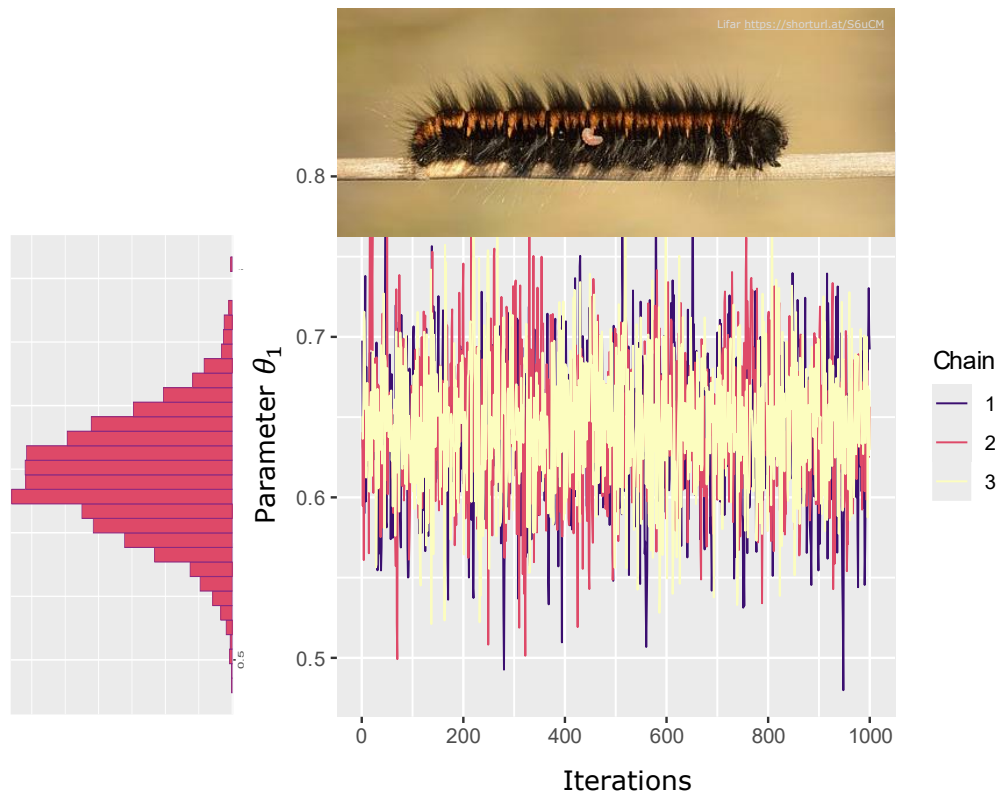


# Convergence

## Visual inspection

- Traceplots for each parameter
- Should look like random noise
- Centered around a constant mean
- Chains should look similar
- Like a fuzzy caterpillar!

→ **MCMC has converged**



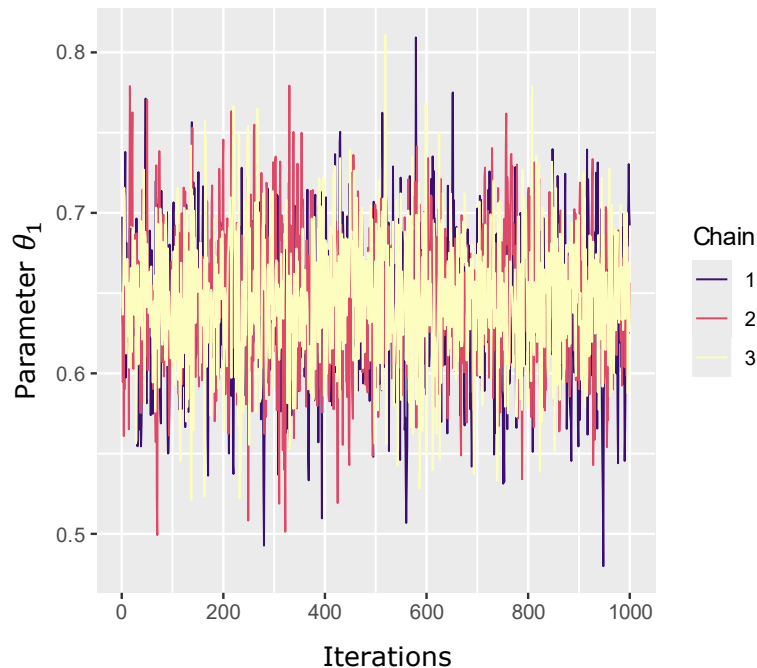
# Convergence

## Quantitative measures

- **Rhat** value („Gelman-Rubin statistic")
  - Compares the variation within and across chains
  - Value should be less than **1.1**
- **n\_eff** (Number of effective samples)
  - Chains usually have a bit of autocorrelation but it shouldn't be too strong
  - Small n\_eff values indicate a problem

## → MCMC has converged


```
3 chains with 1000 samples each  
3000 samples (post-warmup)  
Rhat = 1.001  
n_eff= 1770
```





*Software*

# Some history



**1700s** Bayes' theorem, Laplace formalized it

Bayes impractical  
Restricted to simple cases

**Early 1800s** Gauß: least squares, regression

**Late 1800s to early 1900s**

Birth of modern statistics. Pearson, Fisher, Neyman ... :  
max. likelihood, hypothesis testing, design of experiments

Frequentism superseded Bayes  
More practical in most cases

**Mid to late 1900s** MCMC algorithms

Still a niche topic in statistics

**2000s** Computational tools for MCMC  
BUGS, JAGS, Stan ...

Becoming more popular in sciences

**Today** Convenient R interfaces  
brms, rstanarm ...

Taught in gradschools

**Future**

Becoming the default  
instead of frequentism ??

# Software

All Bayesian software contains:

## **1) Modeling language**

User must define statistical model:

parameters, likelihood & priors

## **2) MCMC sampler**

Automated algorithm that takes care of sampling

# Bayesian programming languages

- + Maximum flexibility in statistical modeling
- + Total control over every part of the model
- Steep learning curve
- Coding can be time-consuming

## JAGS

Was the most popular once, now less and less used

## Nimble



Extends JAGS, more flexible

## Stan



Very efficient, runs in C++

Can all  
be called  
from **R**

```
a ~ dnorm(0, sd=1)
b ~ dnorm(0, sd=1)
sigma ~ dexp(1.0)
for(i in 1:n) {
  y[i] ~ dnorm(a+b*x[i], sd=sigma)
}
```

Nimble

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}
parameters {
  real a;
  real b;
  real<lower=0> sigma;
}
model {
  a ~ normal(0,1);
  b ~ normal(0,1);
  sigma ~ exponential(1.0);
  y ~ normal(a+b*x, sigma);
}
```

Stan

# Formula-based R packages

- + Model formulation similar to `lm` or `lme4`
- + Easy to learn
- + Less coding necessary
- + Handy functions for model analysis (after fitting)
- Limited to pre-defined model types
- „lm“-formulas deceive you into forgetting about model definition

## **rstanarm**

GLMMs only

## **brms**

Much more flexible, becoming quite popular

Both  
automatically  
translate model  
into **Stan**

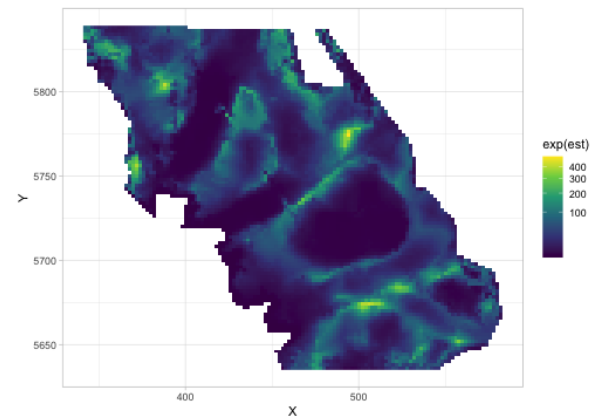


```
> stan_glm(y ~ x,  
           prior = normal(0,1),  
           data = dataset)
```

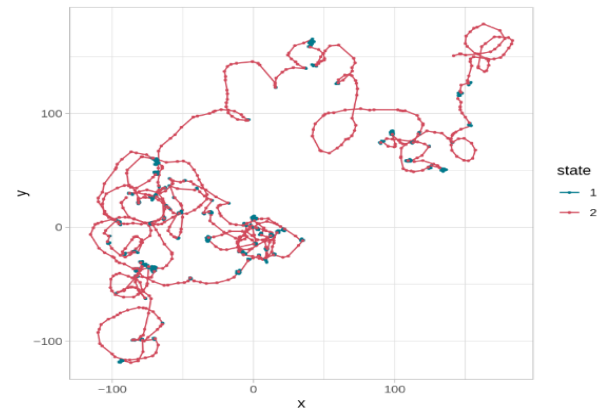
```
> brm(y ~ x,  
      prior = prior(normal(0,1),coef=x),  
      data = dataset)
```

# Specialized software (R-packages)

- R-INLA, sdmTMB, → Spatial models  
spBayes
- bsam, hmmTMB, → Animal movement  
bayesmove
- spOccupancy → Occupancy models
- spAbundance → Abundance models
- blavaan → Structural equation models
- ...



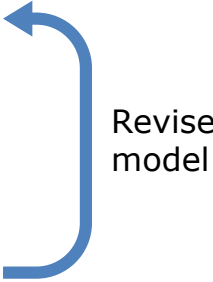
<https://pbs-assess.github.io/sdmTMB/>



<https://github.com/TheoMichelot/hmmTMB>

*Why Bayesian ?*

# Bayesian workflow

- 1) Research question (hypotheses)
  - 2) Data collection
  - 3) Statistical model
  - 4) Prior distribution choice
  - 5) Model fitting (MCMC)
  - 6) Evaluate model output
  - 7) Quantitative statements on hypotheses
- 
- ```
graph TD; 1[1) Research question (hypotheses)] --> 2[2) Data collection]; 2 --> 3[3) Statistical model]; 3 --> 4[4) Prior distribution choice]; 4 --> 5[5) Model fitting (MCMC)]; 5 --> 6[6) Evaluate model output]; 6 -- "Revise model" --> 3;
```

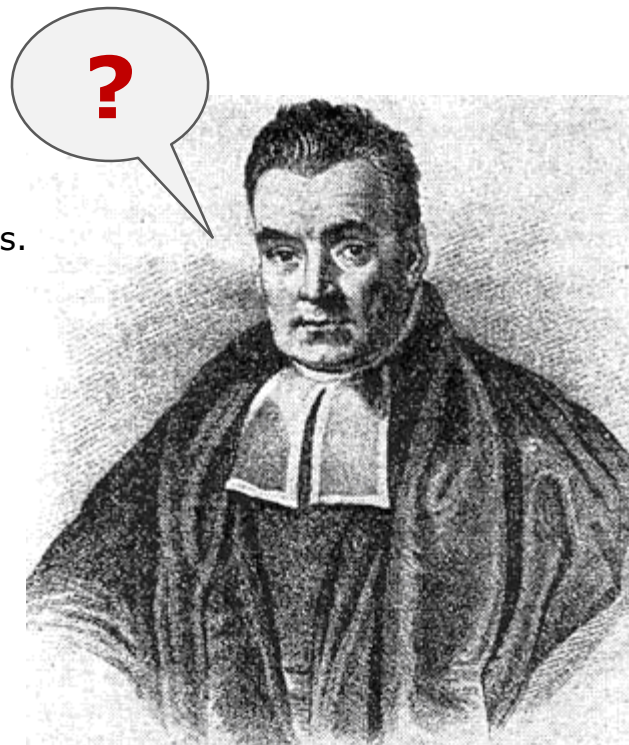
→ Workflow not that different from frequentist statistics.



# Why Bayesian?

## Philosophical answer:

- Frequentism assumes true and fixed underlying parameter values.
- Data are just a sample of the „true“ statistical model.
- Bayesian statistics embraces uncertainty and wants to quantify it correctly.
- Observed data are given, model parameters uncertain.



# Why Bayesian?

## Practical answer:

- Output is more intuitive:  
Direct inference on parameters / hypotheses instead of NHST  
What does the data tell me about my model?
- Full transparency and control over model and output
- Include prior belief / information
- Parameter regularization may be necessary
- Not limited to a specific toolbox, but full flexibility in modeling  
(especially with Stan or Nimble, but brms also very versatile)
- Fit complex models with lots of parameters

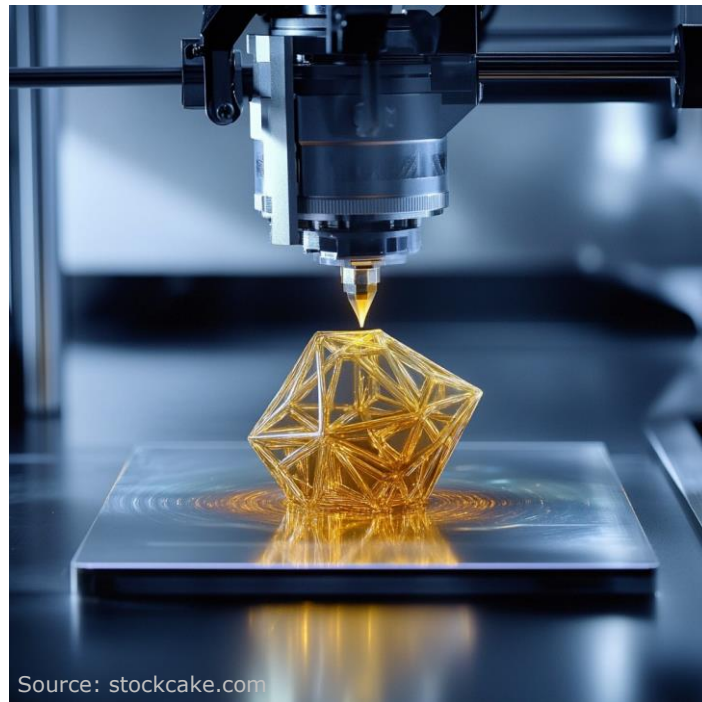


# Why Bayesian?

## What are complex models?

- Nonlinear models
- Hierarchical structure, mixed effects
- Combination of multiple, heterogeneous data sources and/or models
- Constraints on parameters
- Latent variables  
(occupancy models, animal movement models, SEM, HMM, ...)

## The Bayesian 3D printer



Source: stockcake.com

# Summary

- There is no such thing as a „Bayesian model“!
- Frequentist and Bayesian stats are different methodologies for estimating parameters of a statistical model
- **Frequentist** statistics cannot (mathematically) do direct inference  $P(\theta|y)$ , and requires a (methodological) detour via **NHST**  $P(y|\theta = 0)$
- **Bayesian** statistics can (conceptually) do direct inference  $P(\theta|y)$ , but requires a (computational) detour via **MCMC**

## Further reading

Bürkner, P. (2024). The brms Book [in progress]. <https://paulbuerkner.com/software/brms-book/>

Fieberg, J. (2024). Statistics 4 Ecologists. <https://statistics4ecologists-v2.netlify.app/> [Chapters 11-13]

Johnson, A. A., Ott, M. Q., Dogucu, M. (2021). Bayes Rules! *CRC Press*. <https://www.bayesrulesbook.com/> [Chapters 1-2]

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2nd ed.). *Chapman and Hall/CRC*. <https://doi.org/10.1201/9780429029608>

van de Schoot, R., Depaoli, S., King, R., et al. (2021). Bayesian statistics and modelling. *Nature Reviews. Methods Primers*, 1(1), 1–26. <https://doi.org/10.1038/s43586-020-00001-2>